ADVANCES IN INTELLIGENT AND SOFT COMPUTING 77

Christian Borgelt
Gil González-Rodríguez
Wolfgang Trutschnig
María Asunción Lubiano
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz (Eds.)

Combining Soft Computing and Statistical Methods in Data Analysis

Springer

cajAstur

Advances in Intelligent and Soft Computing 77

Editor-in-Chief: J. Kacprzyk

Advances in Intelligent and Soft Computing

Further volumes of this series can be found on our homepage: springer.com

Vol. 63. Á. Herrero, P. Gastaldo,
R. Zunino, E. Corchado (Eds.)
Computational Intelligence in Security for Information Systems, 2009
ISBN 978-3-642-04090-0

Vol. 64. E. Tkacz, A. Kapczynski (Eds.)
Internet – Technical Development and Applications, 2009
ISBN 978-3-642-05018-3

Vol. 65. E. Kącki, M. Rudnicki,
J. Stempczyńska (Eds.)
Computers in Medical Activity, 2009
ISBN 978-3-642-04461-8

Vol. 66. G.Q. Huang,
K.L. Mak, P.G. Maropoulos (Eds.)
Proceedings of the 6th CIRP-Sponsored International Conference on Digital Enterprise Technology, 2009
ISBN 978-3-642-10429-9

Vol. 67. V. Snášel, P.S. Szczepaniak,
A. Abraham, J. Kacprzyk (Eds.)
Advances in Intelligent Web Mastering - 2, 2010
ISBN 978-3-642-10686-6

Vol. 68. V.-N. Huynh, Y. Nakamori,
J. Lawry, M. Inuiguchi (Eds.)
Integrated Uncertainty Management and Applications, 2010
ISBN 978-3-642-11959-0

Vol. 69. E. Piętka and J. Kawa (Eds.)
Information Technologies in Biomedicine, 2010
ISBN 978-3-642-13104-2

Vol. 70. Y. Demazeau, F. Dignum,
J.M. Corchado, J. Bajo Pérez (Eds.)
Advances in Practical Applications of Agents and Multiagent Systems, 2010
ISBN 978-3-642-12383-2

Vol. 71. Y. Demazeau, F. Dignum,
J.M. Corchado, J. Bajo, R. Corchuelo,
E. Corchado, F. Fernández-Riverola,
V.J. Julián, P. Pawlewski, A. Campbell (Eds.)
Trends in Practical Applications of Agents and Multiagent Systems, 2010
ISBN 978-3-642-12432-7

Vol. 72. J.C. Augusto, J.M. Corchado,
P. Novais, C. Analide (Eds.)
Ambient Intelligence and Future Trends, 2010
ISBN 978-3-642-13267-4

Vol. 73. J.M. Corchado, P. Novais,
C. Analide, J. Sedano (Eds.)
Soft Computing Models in Industrial and Environmental Applications, 5th International Workshop (SOCO 2010), 2010
ISBN 978-3-642-13160-8

Vol. 74. M.P. Rocha, F.F. Riverola,
H. Shatkay, J.M. Corchado (Eds.)
Advances in Bioinformatics, 2010
ISBN 978-3-642-13213-1

Vol. 75. X.Z. Gao, A. Gaspar-Cunha,
M. Köppen, G. Schaefer, and J. Wang (Eds.)
Soft Computing in Industrial Applications, 2010
ISBN 978-3-642-11281-2

Vol. 76. T. Bastiaens, U. Baumöl,
and B.J. Krämer (Eds.)
On Collective Intelligence, 2010
ISBN 978-3-642-14480-6

Vol. 77. C. Borgelt, G. González-Rodríguez,
W. Trutschnig, M.A. Lubiano, M.Á. Gil,
P. Grzegorzewski, and O. Hryniewicz (Eds.)
Combining Soft Computing and Statistical Methods in Data Analysis, 2010
ISBN 978-3-642-14745-6

Christian Borgelt, Gil González-Rodríguez,
Wolfgang Trutschnig, María Asunción Lubiano,
María Ángeles Gil, Przemysław Grzegorzewski,
and Olgierd Hryniewicz (Eds.)

Combining Soft Computing and Statistical Methods in Data Analysis

Springer cajAstur

Editors

Christian Borgelt
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3^a Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: christian.borgelt@softcomputing.es

Gil González-Rodríguez
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3^a Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: gil.gonzalez@softcomputing.es

Wolfgang Trutschnig
Research Unit on Intelligent
Data Analysis and Graphical Models
European Centre for Soft Computing
Edificio Científico-Tecnológico. 3^a Planta
C/ Gonzalo Gutiérrez Quirós s/n
33600 Mieres, Spain
E-mail: wolfgang.trutschnig@softcomputing.es

María Asunción Lubiano
Departamento de Estadística e I.O. y D.M.
Universidad de Oviedo, Facultad de Ciencias
C/ Calvo Sotelo s/n, 33007 Oviedo, Spain
E-mail: lubiano@uniovi.es

María Ángeles Gil
Departamento de Estadística
e I.O. y D.M.
Universidad de Oviedo
Facultad de Ciencias
C/ Calvo Sotelo s/n
33007 Oviedo, Spain
E-mail: magil@uniovi.es

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

and

Faculty of Mathematics and
Information Science
Warsaw University of Technology
Plac Politechniki 1, 00-661 Warsaw
Poland
E-mail: pgrzeg@ibspan.waw.pl

Prof. Dr. Olgierd Hryniewicz
Systems Research Institute
Polish Academy of Science
Newelska 6, 01-447 Warsaw
Poland
E-mail: hryniewi@ibspan.waw.pl

ISBN 978-3-642-14745-6 e-ISBN 978-3-642-14746-3

DOI 10.1007/978-3-642-14746-3

Advances in Intelligent and Soft Computing ISSN 1867-5662

Library of Congress Control Number: 2010931605

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed on acid-free paper
5 4 3 2 1 0
springer.com

Preface

> “The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.”
>
> *Ronald A. Fisher*

Over the last forty years there has been a growing interest to extend probability theory and statistics and to allow for more flexible modelling of imprecision, uncertainty, vagueness and ignorance. The fact that in many real-life situations data uncertainty is not only present in the form of randomness (stochastic uncertainty) but also in the form of imprecision/fuzziness is but one point underlining the need for a widening of statistical tools. Most such extensions originate in a “softening” of classical methods, allowing, in particular, to work with imprecise or vague data, considering imprecise or generalized probabilities and fuzzy events, etc. The developed techniques frequently lead to more robust and interpretable models that better capture all the information contained in the given data.

About ten years ago the idea of establishing a recurrent forum for discussing new trends in the before-mentioned context was born and resulted in the first International Conference on Soft Methods in Probability and Statistics (SMPS) that was held in Warsaw in 2002. In the following years the conference took place in Oviedo (2004), in Bristol (2006) and in Toulouse (2008). In the current edition the conference returns to Oviedo. Apart from the rich number of topics already covered by the previous editions, the SMPS 2010 succeeded in incorporating statistics with censored data and robust statistics, both perfectly fitting the scope of the conference.

The wide variety of sessions taking place at the SMPS conference is reflected by the SMPS 2010’ plenary talks: Peter Filzmoser from the Vienna University of Technology on “Soft Methods in Robust Statistics”, Manfred Gilli from the University of Geneva on “An Introduction to Heuristic

Optimization Methods", Mario Guarracino from the High Performance Computing and Networking Institute in Naples on "Supervised Classification of Biological Data", and Enrique Ruspini from the European Centre for Soft Computing on "Ideas and Issues in Conceptual Fuzzy Clustering".

No conference can be organised without a lot of support from various people: We would like to thank all organizers of Invited Sessions, all members of the Program Committee and all additional reviewers - without their help a book like this would have been impossible. Moreover we would like to express our gratitude to "Obra social y cultural" of the main Savings Bank in Asturias, CajAstur, who kindly paid the production cost of these proceedings. The SMPS 2010 also benefited from the COST Action IC0702, this support is gratefully acknowledged. Last but not least we would like to express our gratitude to the European Centre for Soft Computing and the University of Oviedo.

Oviedo, June 2010

Christian Borgelt
Gil González-Rodríguez
Wolfgang Trutschnig
M. Asunción Lubiano
María Ángeles Gil
Przemysław Grzegorzewski
Olgierd Hryniewicz

Members of Committees

General Chairs

Christian Borgelt (Mieres, Spain)
Gil González Rodríguez (Mieres, Spain)
Wolfgang Trutschnig (Mieres, Spain)

Advisory Committee (Core SMPS Group)

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

Program Committee

Ricardo Cao (A Coruña, Spain)
Giulianella Coletti (Perugia, Italy)
Ana Colubi (Oviedo, Spain)
Renato Coppi (Rome, Italy)
Inés Couso (Gijón, Spain)
Bernard de Baets (Gent, Belgium)
Gert de Cooman (Gent, Belgium)
Thierry Denœux (Compiégne, France)
Didier Dubois (Toulouse, France)
Fabrizio Durante (Bolzano, Italy)
Juan Luis Fernández Martínez (Oviedo, Spain)
Peter Filzmoser (Vienna, Austria)
José Gámez (Albacete, Spain)
Pedro Gil (Oviedo, Spain)
Manfred Gilli (Geneve, Switzerland)
Lluis Godó (Barcelona, Spain)

Michel Grabisch (Paris, France)
Mario Guarracino (Naples, Italy)
Eyke Hüllermeier (Marburg, Germany)
Janusz Kacprzyk (Warsaw, Poland)
Etienne Kerre (Gent, Belgium)
Rudolf Kruse (Magdeburg, Germany)
Jonathan Lawry (Bristol, United Kingdom)
Shoumei Li (Beijing, China)
Uwe Ligges (Dortmund, Germany)
Miguel López (Oviedo, Spain)
Marloes Maathuis (Zurich, Switzerland)
Serafín Moral (Granada, Spain)
Domingo Morales (Elche, Spain)
Wolfgang Näther (Freiberg, Germany)
Mirko Navara (Praha, Czech Republic)
Hung T. Nguyen (Las Cruces, USA)
Dan Ralescu (Cincinnati, USA)
Marko Rojas-Medar (Campiñas, Brazil)
Enrique Ruspini (Mieres, Spain)
Antonio Salmerón (Almería, Spain)
Pedro Terán (Oviedo, Spain)
Stefan Van Aelst (Gent, Belgium)
Reinhard Viertl (Vienna, Austria)
Peter Winker (Giessen, Germany)
Marco Zaffalon (Lugano, Switzerland)

Additional Referees

Jorge Gabriel Adrover (Córdoba, Argentina)
José Alonso (Mieres, Spain)
Nicole Bäuerle (Karlsruhe, Germany)
Angela Blanco-Fernández (Oviedo, Spain)
Ulrich Bodenhofer (Linz, Austria)
Enea Bongiorno (Milano, Italy)
Ignacio Cascos (Madrid, Spain)
Etienne Côme (Paris, France)
Pierpaolo D'Urso (Rome, Italy)
Jacobo de Uña-Álvarez (Vigo, Spain)
Sébastien Destercke (Montpellier, France)
Giancarlo Diana (Padova, Italy)
Maria Brigida Ferraro (Rome, Italy)
Daan Fierens (Heverlee, Belgium)
Luis Angel García Escudero (Valladolid, Spain)
Ali Gholami (Tehran, Iran)

Paolo Giordani (Rome, Italy)
Inés González Rodríguez (Santander, Spain)
Sergio Guadarrama (Mieres, Spain)
Marc Hofmann (Neuchâtel, Switzerland)
María Amalia Jácome (A Coruña, Spain)
Osmo Kaleva (Tampere, Finland)
M. Asunción Lubiano (Oviedo, Spain)
Marek Malinowski (Zielona Góra, Poland)
Mylène Masson (Compiègne, France)
Enrique Miranda (Oviedo, Spain)
Isabel Molina (Madrid, Spain)
Susana Montes (Gijón, Spain)
Erik Quaeghebeur (Gent, Belgium)
José Juan Quesada Molina (Granada, Spain)
Ana Belén Ramos-Guajardo (Mieres, Spain)
Luis José Rodríguez Muñíz (Oviedo, Spain)
Luciano Sánchez (Gijón, Spain)
José María Sarabia (Santander, Spain)
Enrico Schumann (Geneve, Switzerland)
Fabio L. Spizzichino (Rome, Italy)
Matthias Troffaes (Durham, United Kingdom)
Paolo Vicig (Trieste, Italy)

Local Organization

Andrea Appe (Vienna, Austria)
Angela Blanco-Fernández (Oviedo, Spain)
María Rosa Casals (Oviedo, Spain)
Ana Colubi (Oviedo, Spain)
Maria Brigida Ferraro (Rome, Italy)
Marta García-Bárzana (Oviedo, Spain)
María Teresa López (Oviedo, Spain)
M. Asunción Lubiano (Oviedo, Spain)
Takehiko Nakama (Oviedo, Spain)
Antonio Palacio (Oviedo, Spain)
Ana Belén Ramos-Guajardo (Mieres, Spain)
Marc Segond (Mieres, Spain)
Beatriz Sinova (Oviedo, Spain)
Nadine Zwickl (Oviedo, Spain)

Technical Edition

M. Asunción Lubiano (Oviedo, Spain)

Contents

Prior Knowledge in the Classification of Biomedical Data

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino

Abstract. Standard data analysis techniques for biomedical problems cannot take into account existing prior knowledge, and available literature results cannot be incorporated in further studies. In this work we review some techniques that incorporate prior knowledge in supervised classification algorithms as constraints to the underlying optimization and linear algebra problems. We analyze a case study, to show the advantage of such techniques in terms of prediction accuracy.

Keywords: Supervised classification, Neural Networks, Support Vector Machines, Generalized Eigenvalue Classifier.

1 Introduction

The widespread availability of biomedical data is posing new and challenging problems to standard analysis algorithms. These problems are related to the quality of data, that are often affected by errors and uncertainty. This is the case of high throughput genomic and proteomic technologies, where the signal to noise ratio is very low. Other questions raise when data produced by comparable experimental protocols are available, because there is no clear strategy to systematically take advantage of previous results and knowledge. In the case of supervised classification, where models are built from data for which the class membership is known, available labeled data is added to the training sets. This has two major drawbacks. First, enlarging the training set increases the computational time needed to elaborate the model. Then, if data are affected by errors or uncertainties, these are introduced in the new classification model, reducing its generalization capabilities.

Danilo Abbate, Roberta De Asmundis, and Mario Rosario Guarracino
High Performance Computing and Networking Institute,
National Research Council (ICAR-CNR), 80131 Naples, Italy
e-mail: mario.guarracino@cnr.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 1–8.
springerlink.com

In this paper we show how to introduce prior knowledge in Support Vector Machines (SVM) [12], Generalized Eigenvalue Proximal SVM (GPSVM) [8], and Radial Basis Functions (RBF) Neural Networks [1]. The idea is if knowledge can be expressed in terms of regions of the data space, in which all points belong to a given class, then the geometrical expression of such regions can be used to constrain the underlying mathematical programming problem. The advantage of such strategy is that, although no points are added to the training set, the model is constrained to take into account available knowledge. We provide a case study that highlights the advantages of such strategy, in terms of classification accuracy.

2 Classification Algorithms

Support Vector Machines

SVM are the state of the art supervised classification methods, widely accepted in many application areas. SVM find a plane $\mathbf{w}^T\mathbf{x}+b=0$ with the objective to separate the elements belonging to two different classes. To this extend, we determine two parallel planes $\mathbf{w}^T\mathbf{x}+b=\pm 1$, of maximum distance, leaving all points of the two classes on different sides. Elements with the minimum distance from both classes are called *support vectors* and are the only elements needed to train the classifier.

Let us consider a data set composed of n pairs $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i \in \mathbf{R}^m$ is the feature vector of a point, and $y_i \in \{-1,1\}$ is the class label. The optimal separating plane is the solution to a quadratic linearly constrained problem.

The advantage of this method is that a very small number of support vectors are sufficient to define the optimal separating plane. In some cases, the relationship between points and class labels can be nonlinear and it is impossible to find a separating plane. In such a case, data can be nonlinearly embedded to a higher dimensional space in which the linear separation can be found. This nonlinear mapping can be implicitly done by kernel functions, which represent the inner product of the elements in the nonlinear space.

The nonlinear classification model cannot describe the discriminating function in terms of inequalities involving linear relations among features. This can be perceived as a problem in case of medical diagnosis, in which doctors prefer to find simple correlations between the results of a clinic exams and the diagnosis or prognosis of an illness. On the other hand, it is generally accepted that results achieved by nonlinear models provide higher classification accuracy. Furthermore, the number of exams to consider for a diagnosis can be very high and cannot be correlated only with the experience. Finally, methods that provide explicit classification rules are not guaranteed to find a set of rules small enough to be easy readable.

Generalized Eigenvalue Classifier

GEPSVM is an efficient algorithm in which the binary classification problem can be formulated as a generalized eigenvalue problem.

Let us consider two matrices $A \in \mathbf{R}^{n\times m}$ and $B \in \mathbf{R}^{k\times m}$, with $m \ll n+k$, representing the two classes, each row being a point in the feature space. Mangasarian et al. [8] propose to classify these sets of points A and B using two planes in the feature space, each closest to one set of points, and furthest from the other.

Suppose that points in classes A and B are not linearly separable, then a nonlinear embedding of each point $\mathbf{x}$ can be obtained using a Radial Basis Function kernel. Each component of the transformed point is given by $K(\mathbf{x},C_i) = exp(\|\mathbf{x}-C_i\|^2/\sigma)$, where C_i is the i-th row of $C = \left[A^T, B^T\right]^T \in \mathbf{R}^{(n+k)\times m}$, and σ is a parameter.

The two planes $K(\mathbf{x},C)\mathbf{u}_1 - \gamma_1 = 0$ and $K(\mathbf{x},C)\mathbf{u}_2 - \gamma_2 = 0$ in the feature space, can be obtained solving the generalized eigenvalue problem [6]:

$$\min_{\mathbf{u},\gamma\neq 0} \frac{\|K(A,C)\mathbf{u}-\mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_B\mathbf{u}-\mathbf{e}\gamma\|^2}{\|K(B,C)\mathbf{u}-\mathbf{e}\gamma\|^2 + \delta\|\tilde{K}_A\mathbf{u}-\mathbf{e}\gamma\|^2}. \tag{1}$$

Here $\tilde{K_A}$ and $\tilde{K_B}$ are diagonal matrices with the diagonal entries from the matrices $K(A,C)$ and $K(B,C)$; $\mathbf{e}$ is a vector of 1s of proper dimension, $\mathbf{u}$ is the coefficient vector of the plane, γ is the plane intercept and δ is the regularization parameter. The eigenvectors related to the minimum and the maximum eigenvalues of (1), provide the coefficients of the proximal planes P_i, $i = 1,2$. The class of a new point $\mathbf{x}$ is determined as

$$class(\mathbf{x}) = argmin_{i=-1,1}\{dist(\mathbf{x},P_i)\}, \tag{2}$$

where $dist(\mathbf{x},P_i)$ is the distance of a point $\mathbf{x}$ from plane P_i.

RBF Neural Networks

A RBF neural network is divided into two operative blocks: an inner hidden layer, and the output layer. The hidden layer creates a response localized on the input vector $\mathbf{x}$; the binary output will then be calculated as a weighted sum of these localized responses. Training a RBF network is a procedure divided into two phases: in the first one the parameters of the radial bases function are calculated using an unsupervised learning algorithm. In this phase the data set is divided in $\bar{n}+\bar{k}$ clusters. We define as $\bar{\mathbf{x}}$ the $\bar{n}+\bar{k}$ points closest to each centroid. In the second part of the training, we search for values of the weights w_i which determine the binary output:

$$h(\mathbf{x}) = \sum_{i=1}^{\bar{n}+\bar{k}} w_i K(\mathbf{x},\bar{\mathbf{x}}_i), \quad \bar{n} \ll n, \quad \bar{k} \ll k. \tag{3}$$

Such weights are calculated by minimizing the following error function, with respect to w_i:

$$E = \frac{1}{2}\sum_{i=1}^{n+k}(h(\mathbf{x}_i) - y_i)^2 \tag{4}$$

where y_i is the label of the point $\mathbf{x}_i$.

3 Prior Knowledge

SVM

We are now showing how it is possible to obtain, with a linear program [9], a nonlinear separating surface using a kernel function $K(\mathbf{x},C) : \mathbf{R}^m \times \mathbf{R}^{(n+k)\times m} \rightarrow \mathbf{R}^{n+k}$, to embed the points in a higher dimensional space. We recall that the resulting plane, projected in the feature space [11], has equation:

$$K(\mathbf{x},C)\mathbf{u} - \gamma\ = 0. \tag{5}$$

In standard SVM, parameters $\mathbf{u} \in \mathbf{R}^{n+k}$ and $\gamma \in \mathbf{R}$ are determined solving the following quadratic optimization problem [7], for some $\nu > 0$:

$$\begin{array}{lr} \min\limits_{\mathbf{u},\gamma,\mathbf{y}\in\mathbf{R}^{(n+k)+1+(n+k)}} & \nu\mathbf{e}^T\mathbf{y} + \frac{1}{2}\mathbf{u}^T\mathbf{u} \\ s.t. & D(K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e},\ \ \mathbf{y} \geq 0. \end{array} \tag{6}$$

where D is a diagonal matrix, with the diagonal elements equal to the labels of the corresponding element of the training set C, $\mathbf{y}$ is a vector of slack variables. Such condition places the points belonging to the two classes $+1$ and -1 on two different sides of the nonlinear separation surface (5). Problem (6) corresponds to the following linear programming problem [9]:

$$\begin{array}{lr} \mathop{min}\limits_{\mathbf{u},\gamma,\mathbf{y},\mathbf{s}} & \nu\mathbf{e}^T\mathbf{y} + \mathbf{e}^T\mathbf{s} \\ s.t. & (K(C,C)\mathbf{u} - \mathbf{e}\gamma) + \mathbf{y} \geq \mathbf{e}, \\ & -\mathbf{s} \leq \mathbf{u} \leq \mathbf{s}, \\ & \mathbf{y} \geq 0, \end{array} \tag{7}$$

where $\mathbf{s} \in \mathbf{R}^{n+k}$ is a vector of non negative slack variables.

In order to improve the results obtained by a classifier solely from the training set, it is possible to impose the knowledge of an expert into the learning phase of the function (5) [10]. Such expertise is represented by the following implication, which represents a knowledge region $\Delta \subset \mathbf{R}^m$ in the input space in which all points $\mathbf{x}$ are known to belong to class $+1$:

$$g(\mathbf{x}) \leq 0 \Rightarrow K(\mathbf{x},C)\mathbf{u} - \gamma \geq \alpha, \forall \mathbf{x} \in \Delta, \alpha \in \mathbf{R}^+, \tag{8}$$

where $g(\mathbf{x}) : \Delta \subset \mathbf{R}^m \rightarrow \mathbf{R}$.

To add positive nonlinear knowledge (8) to problem (7) we solve:

$$\begin{aligned} \min_{\mathbf{u},\gamma,\mathbf{y},\mathbf{s}} \quad & v\mathbf{e}^T\mathbf{y}+\mathbf{e}^T\mathbf{s} \\ s.t. \quad & D(K(C,C)\mathbf{u}-\mathbf{e}\gamma)+\mathbf{y}\geq\mathbf{e}, \\ & -\mathbf{s}\leq\mathbf{u}\leq\mathbf{s},\ \ \mathbf{y}\geq 0, \\ & K(\mathbf{x}_i,C)\mathbf{u}-\gamma-\alpha+vg(\mathbf{x}_i)+z_i\geq 0, \\ & v\geq 0, z_i\geq 0, \quad i=1,\ldots,l. \end{aligned} \tag{9}$$

Here $z_1,\ldots,z_l$ are non negative slack variables used to allow small deviation in prior knowledge and v is a parameter.

To add negative nonlinear knowledge just consider the following implication:

$$f(\mathbf{x})\leq 0 \Rightarrow K(\mathbf{x},C)\mathbf{u}-\gamma\leq-\alpha, \forall\mathbf{x}\in\Lambda, \alpha\in\mathbf{R}^+, \tag{10}$$

where $f(\mathbf{x}):\Lambda\subset\mathbf{R}^m\rightarrow\mathbf{R}$ represents the region in the input space where implication (10) forces the classification function to be less than or equal to $-\alpha$, in order to classify the points $\mathbf{x}\in\{\mathbf{x}|h(\mathbf{x})\leq 0\}$ as -1.

Now we can finally formulate the linear program (7) with nonlinear knowledge included in the cost function:

$$\begin{aligned} \min_{\mathbf{u},\gamma,\mathbf{y},\mathbf{s},\mathbf{v},\mathbf{p},z_i,q_i} \quad & v\mathbf{e}^T\mathbf{y}+\mathbf{e}^T\mathbf{s}+\mu(\sum_{i=1}^{l}z_i+\sum_{j=1}^{t}q_j) \\ s.t. \quad & D(K(C,C)\mathbf{u}-\mathbf{e}\gamma)+\mathbf{y}\geq\mathbf{e}, \\ & -\mathbf{s}\leq\mathbf{u}\leq\mathbf{s},\ \ \mathbf{y}\geq 0, \\ & K(\mathbf{x}_i,C)\mathbf{u}-\gamma-\alpha+vg(\mathbf{x}_i)+z_i\geq 0, \\ & v\geq 0,\ z_i\geq 0,\ i=1,\ldots,l \\ & -K(\mathbf{x}_j,C)\mathbf{u}+\gamma-\alpha+pf(\mathbf{x}_j)+q_j\geq 0, \\ & p\geq 0,\ q_j\geq 0,\ j=1,\ldots,t \end{aligned} \tag{11}$$

where μ is a positive weight, and p is a parameter.

The LP problem (11) minimizes the margin between the two classes constraining the classification model to leave the two prior knowledge regions Δ and Λ in the corresponding half spaces.

GEPSVM

It is possible to add nonlinear prior knowledge to GEPSVM formulating the model in terms of a constrained generalized eigenvalue problem. The latter has been extensively studied and a procedure for its solution has been proposed by Golub in [4].

If G, H and $\mathbf{z}$ are defined as:

$$\begin{aligned} G &= [K(A,C),-\mathbf{e}]^T [K(A,C),-\mathbf{e}] + \delta \left[\tilde{K}_B,-\mathbf{e}\right]^T \left[\tilde{K}_B,-\mathbf{e}\right], \\ H &= [K(B,C),-\mathbf{e}]^T [K(B,C),-\mathbf{e}] + \delta \left[\tilde{K}_A,-\mathbf{e}\right]^T \left[\tilde{K}_A,-\mathbf{e}\right], \\ & \qquad \mathbf{z} = \left[\mathbf{u}^T,\gamma\right]^T \in \mathbf{R}^{n+k+1}, \end{aligned} \tag{12}$$

constraints can be expressed by the equation:

$$V^T \mathbf{z} = 0, \tag{13}$$

where $V \in \mathbf{R}^{(n+k+1)\times p}$ is a matrix of rank r, with $r < p < n+k+1$. The constrained formulation of the eigenvalue problem (1) with positive knowledge becomes:

$$\begin{array}{ll} \underset{\mathbf{z}\in\mathbf{R}^{n+k+1}}{min} & \dfrac{\mathbf{z}^T G\mathbf{z}}{\mathbf{z}^T H\mathbf{z}} \\ s.t. & V^T\mathbf{z} = 0. \end{array} \tag{14}$$

Let Δ be the set of class $+1$ points describing nonlinear positive knowledge, then the constraint matrix V represents knowledge imposed on class $+1$ points, hence it will be:

$$V = \left[K(\Delta,C),\ -\mathbf{e}\right]^T \tag{15}$$

Matrix V needs to be rank deficient in order to have a non-trivial solution. The set of constraints (13) requires all points in Δ to have null distance from the plane, and thus to belong to class $+1$. Similarly, we can add a negative knowledge.

RBF Neural Networks

As for GEPSVM, [5], a classification model calculated by the RBF network must pass through the prior knowledge points.

Prior knowledge is then added as a set of constraints to problem (4) to obtain the following minimization problem:

$$\begin{array}{l} \min\limits_{w_i} \dfrac{1}{2}\sum\limits_{i=1}^{n+k} (h(\mathbf{x}_i) - y_i)^2 \\ s.t.\ \ V^T\mathbf{x} \geq 0. \end{array} \tag{16}$$

The constraints of this problem force the solution of equation (4) to pass through the points represented by the matrix V. Algebraically, this means the solution of the least squares problem has to be searched in the subspace generated by prior knowledge points. As pointed out by Golub [3], the original problem is reduced with a QR decomposition, or with a *singular value decomposition* as shown by Bjorck [2].

4 A Case Study

The prior knowledge introduced in the classification methods discussed above, has been tested on the UCI data set Thyroid composed of data coming from 215 patients. For each patient 5 cytological and clinical features are provided, which are useful to divide patients in two classes: *sick* and *not sick*. The first class is composed of 65 patients, while the second of 150 healthy patients. The features are: the percentage of T3-resin, total serum thyroxin measured by the isotopic displacement method, total serum triiodothyronine measured by radioimmuno assay, TSH measured by radioimmuno assay, and the maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value.

The results reported in Table 1 are comparison among GEPSVM, SVM and RBF Neural Network methods with and without prior knowledge. The values of accuracy, sensitivity and specificity have been obtained with the *leave one out* cross validation.To simulate the prior knowledge, points were chosen as the misclassified support vectors, obtained training the SVM on the complete data set during the *leave one out* cross validation.

Table 1 Values of accuracy, sensitivity and specificity obtained using GEPSVM, SVM and RBF methods The second line of each block in the table shows the results obtained introducing prior knowledge.

Method	Accuracy	Sensitivity	Specificity
GEPSVM	93.02%	87.69%	95.33%
GEPSVM with knowledge	99.07%	96.62%	100.00%
SVM	93.95%	92.23%	96.00%
SVM with knowledge	98.90%	96.92%	99.33%
RBF	85.12%	55.38%	98.00%
RBF with knowledge	90.23%	72.31%	98.00%

We note that all methods have a better prediction accuracy and higher values of sensitivity and specificity.

5 Conclusion

In this work, we described some classification methods that can take advantage of prior knowledge. We provided a case study to show the gain in terms of accuracy, sensitivity and specificity. Results confirm that prior knowledge substantially increase the classification accuracy of the considered methods. Further work need to be devoted to the automatic knowledge discovery in databases, when data are affected by noise and uncertainty.

Acknowledgements. Danilo Abbate and Roberta De Asmundis spent a research period at ICAR CNR as graduate students. This work has been partially funded by COST action ICT-0702 and PRIN 20078MHYS4.

References

1. Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1995)
2. Bjork, A.: Numerical methods for least squares. SIAM, Philadelphia (1996)
3. Golub, G., van Loan, C.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
4. Golub, G.H., Underwood, R.: Stationary values of the ratio of quadratic forms subject to linear constraints. Z. Angew. Math. Phys (ZAMP) 21(3), 318–326 (1970)
5. Guarracino, M., Abbate, D., Prevete, R.: Nonlinear knowledge in learning models. In: Proceedings of Workshop on Prior Conceptual Knowledge in Machine Learning and Knowledge Discovery, European Conference on Machine Learning, pp. 29–40 (2007), http://www.ecmlpkdd2007.org/CD/workshops/PRICKLWM2/P_Gua/GuarracinoPriCKL/Guarracino.pdf
6. Guarracino, M.R., Cifarelli, C., Seref, O., Pardalos, P.: A classification method based on generalized eigenvalue problems. Optim. Methods Softw. 22, 73–81 (2007)
7. Lee, Y., Mangasarian, O.L.: Ssvm: A smooth support vector machine for classification (1999), http://citeseer.ist.psu.edu/lee99ssvm.html
8. Mangasarian, O.L., Wild, E.W.: Multisurface proximal support vector classification via generalized eigenvalues. Tech. Rep. 04-03, Data Mining Institute (2004)
9. Mangasarian, O.L., Wild, E.W.: Nonlinear knowledge-based classification. Tech. rep., Data Mining Institute Technical Report 06-04, Computer Science Department, University of Wisconsin, Madison, Wisconsin (2006)
10. Pardalos, P.M., Abbate, D., Guarracino, M.R., Chinchuluun, A.: Neural network classification with prior knowledge for analysis of biological data. In: Proceedings of the International Symposium on Mathematical and Computational Biology, Biomat 2008, Brazil, pp. 223–234. World Scientific, Singapore (2008)
11. Schölop, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)

Estimation of a Simple Genetic Algorithm Applied to a Laboratory Experiment

Simone Alfarano, Eva Camacho, and Josep Domenech

Abstract. The aim of our contribution relies on studying the possibility of implementing a genetic algorithm in order to reproduce some characteristics of a simple laboratory experiment with human subjects. The novelty of our paper regards the estimation of the key-parameters of the algorithm, and the analysis of the characteristics of the estimator.

Keywords: Experiments, Genetic algorithm, Bounded rationality, Estimation.

1 Introduction

Nowadays, a large part of economists expresses dissatisfaction (or sometimes rejection) to the wide-spread paradigm of full or strict rationality in theorizing the behavior of economic agents. Laboratory experiments showed that, even in simple settings, human subjects are not consistent with the assumptions implied by their supposed perfect rationality. An existing alternative paradigm in economic theory considers that economic agents have limited capabilities in processing the information and in taking their decisions. Contrary to the fully rational paradigm, it does not exists a unified theory of bounded rationality. Therefore, many different models of human behavior which account for bounded rationality have been proposed in the literature (See for example [3]).

Simone Alfarano and Eva Camacho
Departamento de Economia, Universitat Jaume I, 12071 Castellón, Spain
e-mail: alfarano@eco.uji.es,camacho@eco.uji.es

Josep Domenech
Dpto. de Economía y Ciencias Sociales, Universidad Politecnica de Valencia,
46022 Valencia, Spain
e-mail: jdomenech@upvnet.upv.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 9–16.
springerlink.com

The adaptation of genetic algorithms (GA) from the realm of optimization literature to the description of human learning is an example of the creative ability of researchers to introduce bounded rational models.[1] A number of papers are now available in the literature which apply different versions of GAs in order to reproduce the behavior of economic agents in different contexts (See, for example, [1], [2], [7], [9]). GAs have also been applied in the context of laboratory experiments in order to reproduce the human subjects' behavior in different experimental settings (See [4], [10]).

However, up to now the different contributions are almost entirely based on a rough calibration of the underlying crucial parameters. To the best of our knowledge, our paper constitutes the first attempt to estimate the underlying parameters of a genetic algorithm. In this paper we provide a method to estimate the key parameters of the GA by means of an extensive simulation-based approach, using an extremely simple experimental setting of a common-pool resources problem.The experiment exhibits, in fact, a single dominant and symmetric Nash equilibrium as illustrated in the next section. The paper is organized as follows: in Section 2 we illustrate briefly the theoretical and empirical results of the experimental setting. In Section 3 we detail the characteristics of the implementation of our GA agents. In Section 4 we present the estimation procedure. Finally, in Section 5 we conclude.

2 Experiment: Setting and Results

In this section we will summarize the experimental setting and main results used as benchmark in order to build the GA and the corresponding parameters estimation (See [5] for the details on the experiment).

Consider an industry consisting of n symmetric firms where each firm $i = \{1,...,n\}$ is characterized by both its default profit Π^0, incurred without engaging in any abatement activity, and by its abatement technology represented by an abatement cost function $C(a_i)$, where we use a_i to denote the firm's abatement level.[2] Zero abatement leads to a maximal emission level $e^{\max}$. Accordingly, the profit function of each firm can be written as $\Pi_i = \Pi^0 - C(a_i)$. Total emissions by industry are then given by $E = \sum_{i=1}^{n}(e^{\max} - a_i)$ and are evaluated by using a social damage function $D(E) = d\,[\sum_{i=1}^{n}(e^{\max} - a_i)]$, where $d > 0$ denotes the marginal social damage.

In this industry the regulator decides to implement the tax-subsidy mechanism, proposed by [12]. This mechanism works as follows: Whenever the aggregate abatement level falls short of (exceeds) the socially optimal aggregate abatement level A^*, the regulator charges all the firms with a tax (or pays a subsidy to all the firms) proportional to the difference between optimal and actual abatement. Note that the total tax bill (subsidy payment) is the

[1] For more details on GA and their application to Economics see [6].

[2] The abatement cost function satisfies the following properties: $C(0) = 0$, $C' > 0$, and $C'' > 0$.

Table 1 Abatement cost schedule

Abated units	Marginal cost	Total cost
0	0	0
1	20	20
2	40	60
3	60	120
4	80	200

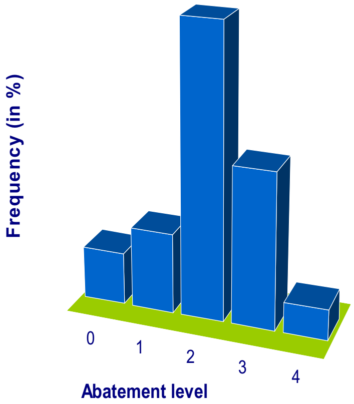


Fig. 1 Histogram of experimental subjects decisions.

same for each firm. Thus with this mechanism a typical firm's profit can be written as:

$$\Pi_i(a_i, a_{-i}) = \Pi^0 - C(a_i) - s\left[A^* - \sum_{i=1}^{n} a_i\right] , \tag{1}$$

where s denotes the tax or subsidy rate and a_{-i} the vector of the decisions by the other firms except from i. When implemented as a one-shot or finitely repeated game, the unique Nash equilibrium is characterized by the the following condition: $C'(a_i) = s$, i.e. the firms choose an abatement level with a marginal cost equaling the tax or subsidy rate. The Nash strategy is also a dominant strategy that leads to the first-best allocation, i.e. $a_i = a^*$, if s equals the marginal social damage d.[3]

In [5] they consider an industry consisting of 5 firms ($n = 5$) with a default profit $\Pi^0 = 200$ ECU (Experimental Currency Unit, which is then converted into Euros at a given exchange rate, known to the subjects at the beginning

[3] Note that the mechanism is not collusion-proof in a repeated setting as stressed by [8]. Therefore, if firms succeed in coordinating on a higher abatement level than is socially optimal, they can earn a higher profit than in the one–shot Nash equilibrium.

of the experiment), an optimal subsidy of $s = 50$ and a discrete abatement cost schedule presented in table 1. Abatement schedule and marginal damage imply a socially optimal abatement level of $a^* = 2$ for any $i = 1,...,5$, leading to an optimal aggregate abatement level of $A^* = 10$.

The mechanism was administered as a non-cooperative game and was repeated over 20 periods. In total 8 sessions with 5 subjects each were conducted. Figure 1 illustrates the aggregate results obtained in the experiments regarding the frequency of each possible abatement decision.

3 Genetic Algorithm

The basic philosophy in implementing our version of the GA is to be "as close as possible" to the laboratory setting described in the previous section. Therefore, the parameters of the algorithms and the implementations of its internal procedures are chosen, when possible, directly from the experimental design. Additionally, we do not intend to describe a general implementation of GA, neither mention all possible alternative implementations of its operators that can be found in the literature (See [6]). Instead, we directly illustrate what we have used to implement the experimental setting.

Our genetic algorithm is characterized by the following elements:

- **Strategy:** Each chromosome in the genetic algorithm represents a possible strategy that a subject can follow, that is, the abatement level decided by the subject. It is encoded as a single gene which takes integer values between 0 and 4. This is the basic element of the GA in the evolution of the algorithm. This choice follows directly the experimental setting.
- **Fitness Function:** It is associated to each strategy and accounts for the actual or potential payoff that derives from the use of a given strategy. In our setting, the GA player uses as measure of fitness the profit function that the experimental subjects face in the laboratory (as shown in Equation 1).
- **Time window:** In order to associate a fitness measure to each strategy, we compute the cumulative potential profit that a given strategy would have had if played in the past w time periods. This time window represents the time memory that the GA subjects use to evaluate each single strategy from its population.
- **Population:** Each subject is endowed with a set of P strategies. The limited size of this set bounds the sophistication of the GA subject when deciding which strategy to apply.
- **Mutation:** It implies that with a probability m one of the strategies included in the population will be randomly changed into any other strategy included in the entire set of potential strategies.
- **Choice rule:** Given the fitness measure and the population, for each single period, the GA subject chooses to play the fittest strategy available in its population.

- **Learning:** Typically there exists two different learning mechanisms: single population vs. multi population. Under the first one, each GA agent has a set of strategies that evolve independently of the strategies of the other agents. In a multi population approach, part of the genetic material is exchanged among the GA agents. This creates some sort of interaction or imitation among agents. Given that in the laboratory setting, total abatement was the only information provided to the subjects and that no communication among subjects was allowed, we decided to implement the leaning mechanism based on a single population approach.

The number of GA agents is $N = 5$, following the experimental setting. Moreover, given our limited number of possible strategies, and in order to simplify the estimation procedure, we decide not to implement the crossover operator, which is typically present in the GA (See [4]). The GA parameters we aim at estimating are population (P), time window (w) and mutation rate (m).

4 Estimation: Procedure and Results

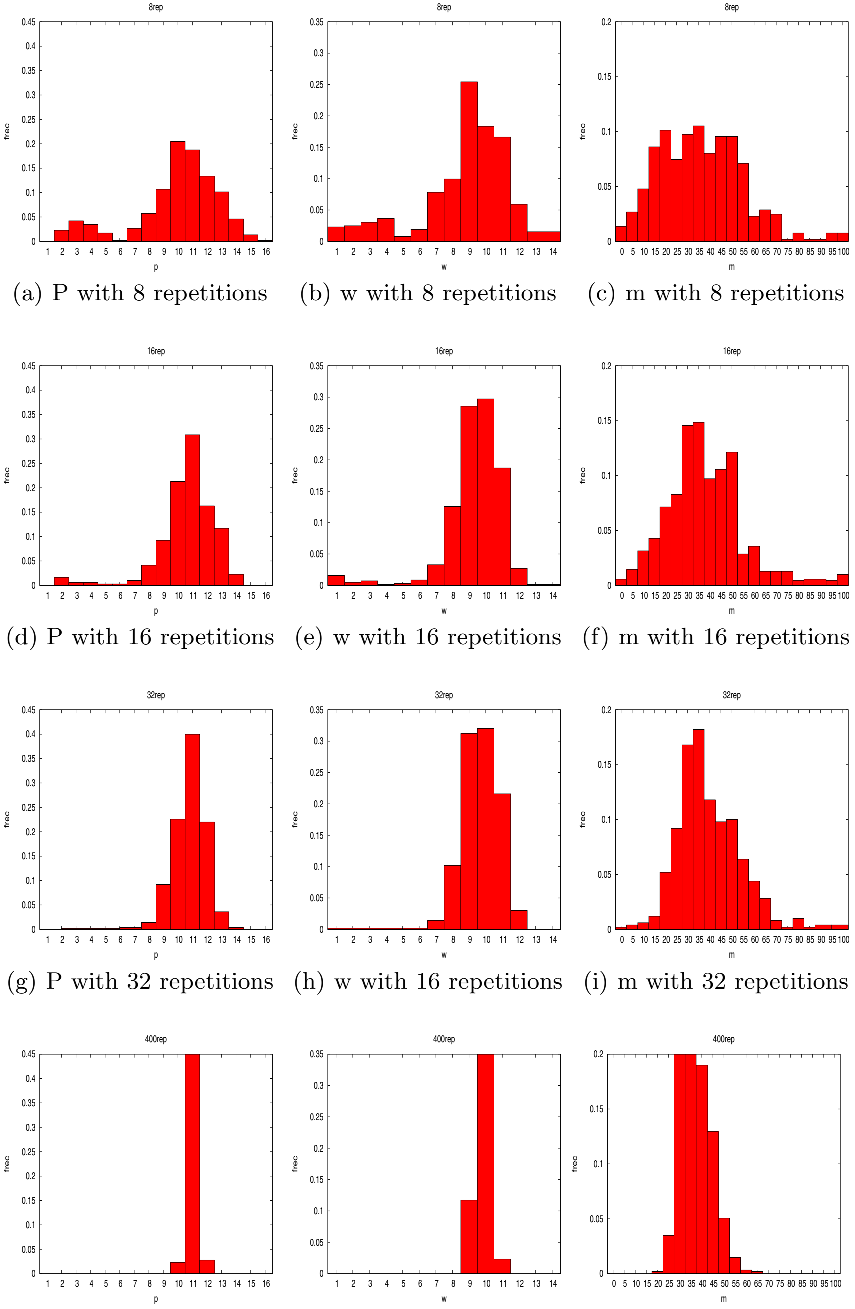
In order to estimate the key parameters of the GA described in the previous section, we fit the distribution of strategies observed in the 8 experimental sessions (See Figure 1).

Let us define as $\boldsymbol{\theta} = (P, w, m)$ the vector of the parameters to be estimated. The optimal value of $\boldsymbol{\theta}$ is calculated by minimizing the distance between the empirical histogram of the strategies from the experimental data (See Figure 1) and the histogram of the GA strategies computed using 5000 Monte Carlo simulations. The optimal value is then given by the following expression:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \sum_{i=0}^{4} \left[f_{exp}(i) - f_{sim}(i|\boldsymbol{\theta})\right]^2, \qquad (2)$$

where $f_{exp}(i)$ is the empirical frequency of the strategy i computed from the histogram of experimental data, and $f_{sim}(i|\boldsymbol{\theta})$ is the frequency of strategy i computed from 5000 Monte Carlo simulations of the GA with parameters $\boldsymbol{\theta}$. More precisely, given a vector of parameters $\boldsymbol{\theta}$, the GA runs 5000 times for a 20 periods[4] for each realization; then the distance between the resulting simulated histogram of strategies and the empirical one is evaluated and minimized using a Nelder-Mead optimization algorithm. The Nelder-Mead method was proposed by [11] as an unconstrained optimization algorithm. It is commonly used when the derivatives of the objective function are not available. The number of Monte Carlo repetitions has been decided taking into account the computational effort and the precision of evaluation of the simulated histogram. The optimization procedure takes around one hour, which is a reasonable time. The optimal value is $\boldsymbol{\theta}^* = (11, 10, 0.36)$.

[4] The number of periods is equal to the periods conducted in the experimental sessions.



(a) P with 8 repetitions (b) w with 8 repetitions (c) m with 8 repetitions

(d) P with 16 repetitions (e) w with 16 repetitions (f) m with 16 repetitions

(g) P with 32 repetitions (h) w with 16 repetitions (i) m with 32 repetitions

(j) P with 400 repetitions (k) w with 400 repetitions(l) m with 400 repetitions

Fig. 2 Distribution of parameter values: P, w and m for different number of repetitions.

In order to evaluate the performance of the entire estimation procedure, we run a series of Monte Carlo simulations using the previously described minimizing procedure with artificially generated histograms as benchmark instead of the experimental data. The vector of parameters of the GA is $\boldsymbol{\theta}^*$. Essentially, we re-estimate the known parameters of the GA, valuating then the *ex-post* resulting distribution of the estimated $\hat{\boldsymbol{\theta}}$. The benchmark histogram is computed averaging over an increasing number of single simulations of 20 periods (See details in Figure 2). We have computed 500 Monte Carlo replications of the re-estimation procedure for each benchmark histogram. The entire process required about 60 days of computing time, although it was parallelized in a 20-node cluster to cut the simulation time to three days.

5 Conclusions

The first important result of our computational exercise is to demonstrate that it is possible to estimate the parameters of a GA using experimental data. As it turns out, the estimation of the key-parameters of GA applied to this set of experiments gives satisfactory results, considering the small data sample available and the highly complex nature of the GA algorithm. The different parameters can be, in fact, estimated with reasonable errors, as the Monte Carlo numerical re-estimation exercise shows. We have performed the re-estimation procedure with a benchmark histogram averaged over 8, 16, 32 and 400 replications of the genetic algorithm. The case using 400 repetitions was conducted as a computational exercise to see the asymptotic properties of the estimator. From an experimental point of view, our Monte Carlo exercise shows that are enough few experimental sessions to generate a sufficiently large data set in order to reliably estimate the parameters.

As final remarks, we would like to stress that our computational exercise, although promising, it is just a first step in developing a general computational approach to complement the laboratory experiments in analyzing economic phenomena. The robustness of the GAs with respect to changes in the experimental setting, the flexibility of GA under changes in its internal operators, the importance to obtain *reasonable* and *consistent* values of the parameters in describing human behavior are just few examples of open problems that we have in our research agenda.

Acknowledgements. Financial support from the Spanish Ministry of Science and Innovation under research project ECO2008-00510 and from Universitat Jaume I-Bancaixa under the research project P11A2009-09 is gratefully acknowledged.

References

1. Arifovic, J.: Genetic algorithm learning and the cobweb model. J. Econom. Dynam. Control 18(1), 3–28 (1994)
2. Arifovic, J.: The behavior of the exchange rate in the genetic algorithm and experimental economies. J. Political Economy 104(3), 510–541 (1996)

3. Aumann, R.J.: Rationality and Bounded Rationality. Games Econ. Behav. 21, 2–14 (1997)
4. Casari, M.: Can Genetic Algorithms Explain Experimental Anomalies? Comput. Econ. 24(3), 257–275 (2004)
5. Camacho, E., Requate, T.: The Regulation of Non-Point Source Pollution and Risk Preferences: An Experimental Approach. Mimeo (2010)
6. Dawid, H.: On the convergence of genetic learning in a double auction market. J. Econom. Dynam. Control 23(9-10), 1545–1569 (1999)
7. Duffy, J.: Agent-Based Models and Human Subject Experiments. In: Tesfatsion, L., Judd, K.L. (eds.) Handbooks of Computational Economics. Handbooks in Economic Series, vol. 2. Elsevier, Amsterdam (2006)
8. Hansen, L.G.: A Damage Based Tax Mechanism for Regulation of Non-Point Emissions. Environ. Resource Econom. 12(1), 99–112 (1998)
9. Lux, T., Schornstein, S.: Genetic learning as an explanation of stylized facts of foreign exchange markets. J. Math. Econ. 41(1-2), 169–196 (2005)
10. Lux, T., Hommes, C.: Individual Expectations and Aggregate Behavior in Learning to Forecast Experiments. Kiel Working Papers, 1466. Kiel Institute for the World Economy (2008)
11. Nelder, J.A., Mead, R.: A Simplex Method for Function Minimization. Comput. J. 7(4), 308–313 (1965)
12. Segerson, K.: Uncertainty and incentives for nonpoint pollution control. J. Environ. Econ. Manage. 15(1), 87–98 (1988)

A Comparison of Robust Methods for Pareto Tail Modeling in the Case of Laeken Indicators

Andreas Alfons, Matthias Templ, Peter Filzmoser, and Josef Holzer

Abstract. The Laeken indicators are a set of indicators for measuring poverty and social cohesion in Europe. However, some of these indicators are highly influenced by outliers in the upper tail of the income distribution. This paper investigates the use of robust Pareto tail modeling to reduce the influence of outlying observations. In a simulation study, different methods are evaluated with respect to their effect on the quintile share ratio and the Gini coefficient.

1 Introduction

As a monitoring system for policy analysis purposes, the European Union introduced a set of indicators, called the *Laeken indicators*, to measure risk-of-poverty and social cohesion in Europe. The basis for most of these indicators is the EU-SILC *(European Union Statistics on Income and Living Conditions)* survey, which is an annual panel survey conducted in EU member states and other European countries. Most notably for this paper, EU-SILC data contain information on the income of the sampled households. Each person of a household is thereby assigned the same *equivalized disposable income* [9].

Andreas Alfons, Matthias Templ, Peter Filzmoser, and Josef Holzer
Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstraße 8-10, 1040 Vienna, Austria
e-mail: alfons@statistik.tuwien.ac.at,templ@tuwien.ac.at,
p.filzmoser@tuwien.ac.at

Matthias Templ
Methods Unit, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria

Josef Holzer
Landesstatistik Steiermark, Hofgasse 13, 8010 Graz, Austria
e-mail: josef.holzer@stmk.gv.at

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 17–24.
springerlink.com

The subset of Laeken indicators based on EU-SILC is computed from this equivalized income, taking into account the sample weights.

In general the upper tail of an income distribution behaves differently than the rest of the data and may be modeled with a *Pareto* distribution. Moreover, EU-SILC data typically contain some extreme outliers that not only have a strong influence on some of the Laeken indicators, but also on fitting the Pareto distribution to the tail. Modeling the tail in a robust manner should therefore improve the estimates of the affected indicators.

The rest of the paper is organized as follows. Section 2 gives a brief description of selected Laeken indicators, while Section 3 discusses Pareto tail modeling. A simulation study is performed in Section 4 and Section 5 concludes.

2 Selected Laeken Indicators

This paper investigates the influence of promising robust methods for Pareto tail modeling on the *quintile share ratio* and the *Gini coefficient*. Both indicators are measures of inequality and are highly influenced by outliers in the upper tail. Strictly following the Eurostat definitions [9], the indicators are implemented in the R package `laeken` [2].

For the following definitions, let $\boldsymbol{x} := (x_1, \dots, x_n)'$ be the equivalized disposable income with $x_1 \leq \dots \leq x_n$ and let $\boldsymbol{w} := (w_i, \dots, w_n)'$ be the corresponding personal sample weights, where n denotes the number of observations.

2.1 Quintile Share Ratio

The income quintile share ratio is defined as the ratio of the sum of equivalized disposable income received by the 20% of the population with the highest equivalized disposable income to that received by the 20% of the population with the lowest equivalized disposable income [9]. Let $q_{0.2}$ and $q_{0.8}$ denote the weighted 20% and 80% quantiles of $\boldsymbol{x}$ with weights $\boldsymbol{w}$, respectively. With $I_{\leq q_{0.2}} := \{i \subset \{1, \dots, n\} : x_i \leq q_{0.2}\}$ and $I_{> q_{0.8}} := \{i \subset \{1, \dots, n\} : x_i > q_{0.8}\}$, the quintile share ratio is estimated by

$$QSR := \frac{\sum_{i \in I_{> q_{0.8}}} w_i x_i}{\sum_{i \in I_{\leq q_{0.2}}} w_i x_i}. \tag{1}$$

2.2 Gini Coefficient

The Gini coefficient is defined as the relationship of cumulative shares of the population arranged according to the level of equivalized disposable income, to the cumulative share of the equivalized total disposable income received by them [9]. In mathematical terms, the Gini coefficient is estimated by

$$Gini := 100 \left[\frac{2\sum_{i=1}^{n}\left(w_i x_i \sum_{j=1}^{i} w_j\right) - \sum_{i=1}^{n} w_i^2 x_i}{\left(\sum_{i=1}^{n} w_i\right)\sum_{i=1}^{n}(w_i x_i)} - 1 \right]. \tag{2}$$

3 Pareto Tail Modeling

The *Pareto* distribution is defined in terms of its cumulative distribution function

$$F_\theta(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}, \qquad x \geq x_0, \tag{3}$$

where $x_0 > 0$ is the scale parameter and $\theta > 0$ is the shape parameter [12]. Furthermore, the density is given by

$$f_\theta(x) = \frac{\theta x_0^\theta}{x^{\theta+1}}, \qquad x \geq x_0. \tag{4}$$

In Pareto tail modeling, the cumulative distribution function on the whole range of x is modeled as

$$F(x) = \begin{cases} G(x), & \text{if } x \leq x_0, \\ G(x_0) + (1 - G(x_0))F_\theta(x), & \text{if } x > x_0, \end{cases} \tag{5}$$

where G is an unknown distribution function [8].

Let n be the number of observations and let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ denote the observed values with $x_1 \leq \ldots \leq x_n$. In addition, let k be the number of observations to be used for tail modeling. In this scenario, the threshold x_0 is estimated by

$$\hat{x}_0 := x_{n-k}. \tag{6}$$

On the other hand, if an estimate $\hat{x}_0$ for the scale parameter of the Pareto distribution has been obtained, k is given by the number of observations larger than $\hat{x}_0$. Thus estimating x_0 and k directly corresponds with each other. Various methods for the estimation of x_0 or k have been proposed [5, 6, 8, 17]. However, this paper is focused on evaluating robust methods for estimating the shape parameter θ (with respect to their influence on the selected Laeken indicators) once the threshold is fixed.

3.1 Hill Estimator

The maximum likelihood estimator for the shape parameter of the Pareto distribution was introduced by [10] and is referred to as the *Hill* estimator. It is given by

$$\hat{\theta} = \frac{k}{\sum_{i=1}^{k} \log x_{n-i+1} - k \log x_{n-k}}. \tag{7}$$

Note that the Hill estimator is non-robust, therefore it is included for benchmarking purposes.

3.2 Weighted Maximum Likelihood (WML) Estimator

The weighted maximum likelihood (WML) estimator [7, 8] falls into the class of M-estimators and is given by the solution $\hat{\theta}$ of

$$\sum_{i=1}^{k} \Psi(x_{n-i+1}, \theta) = 0 \tag{8}$$

with

$$\Psi(x,\theta) := w(x,\theta) \frac{\partial}{\partial \theta} \log f(x,\theta) = w(x,\theta) \left(\frac{1}{\theta} - \log \frac{x}{x_0} \right), \tag{9}$$

where $w(x,\theta)$ is a weight function with values in $[0,1]$. In this paper, a Huber type weight function is used, as proposed in [8]. Let the logarithms of the relative excesses be denoted by

$$y_i := \log \left(\frac{x_{n-i+1}}{x_{n-k}} \right), \qquad i = 1, \ldots, k. \tag{10}$$

In the Pareto model, these can be predicted by

$$\hat{y}_i := -\frac{1}{\theta} \log \left(\frac{k+1-i}{k+1} \right), \qquad i = 1, \ldots, k. \tag{11}$$

The variance of y_i is given by

$$\sigma_i{}^2 := \sum_{j=1}^{i} \frac{1}{\theta^2 (k-i+j)^2}, \qquad i = 1, \ldots, k. \tag{12}$$

Using the standardized residuals

$$r_i := \frac{y_i - \hat{y}_i}{\sigma_i}, \tag{13}$$

the Huber type weight function with tuning constant c is defined as

$$w(x_{n-i+1}, \theta) := \begin{cases} 1, & \text{if } |r_i| \leq c, \\ \frac{c}{|r_i|}, & \text{if } |r_i| > c. \end{cases} \tag{14}$$

For this choice of weight function, the bias of $\hat{\theta}$ is approximated by

$$B(\hat{\theta}) = -\frac{\sum_{i=1}^{k} \left(w_i \frac{\partial}{\partial \theta} f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-i+1}) - F_{\hat{\theta}}(x_{n-i}) \right)}{\sum_{i=1}^{k} \left(\frac{\partial}{\partial \theta} w_i \frac{\partial}{\partial \theta} f_i + w_i \frac{\partial^2}{\partial \theta^2} f_i \right) |_{\hat{\theta}} \left(F_{\hat{\theta}}(x_{n-i+1}) - F_{\hat{\theta}}(x_{n-i}) \right)}, \tag{15}$$

where $w_i := w(x_{n-i+1}, \theta)$ and $f_i := f(x_{n-i+1}, \theta)$. This term is used to obtain a bias-corrected estimator

$$\tilde{\theta} := \hat{\theta} - B(\hat{\theta}). \tag{16}$$

For details and proofs of the above statements, the reader is referred to [7, 8].

3.3 Partial Density Component (PDC) Estimator

For the partial density component (PDC) estimator [16], the Pareto distribution is modeled in terms of the relative excesses

$$y_i := \frac{x_{n-i+1}}{x_{n-k}}, \qquad i = 1, \ldots, k. \tag{17}$$

The density function of the Pareto distribution for the relative excesses is approximated by

$$f_\theta(y) = \theta y^{-(1+\theta)}. \tag{18}$$

The PDC estimator is then given by

$$\hat{\theta} = \arg\min_\theta \left[w^2 \int f_\theta^2(y)dy - \frac{2w}{k} \sum_{i=1}^{k} f_\theta(y_i) \right], \tag{19}$$

i.e., by minimizing the integrated squared error criterion [15] using an incomplete density mixture model wf_θ. The parameter w can be interpreted as a measure of the uncontaminated part of the sample and is estimated by

$$\hat{w} = \frac{\frac{1}{k} \sum_{i=1}^{k} f_{\hat{\theta}}(y_i)}{\int f_{\hat{\theta}}^2(y)dy}. \tag{20}$$

See [16] and references therein for more information on the PDC estimator.

4 Simulation Study

Various robust methods for the estimation of poverty and inequality indicators, mostly non-parametric, have been investigated in [17], but neither the WML nor the PDC estimator for Pareto tail modeling are considered there. Preliminary results with income generated from theoretical distributions [11] are an indication that both estimators are promising in the context of Laeken indicators. This is further investigated in this section. However, variance estimation is not yet considered in this paper.

The simulations are carried out in R [14] using the package `simFrame` [1, 4], which is a general framework for statistical simulation studies. A synthetic data set consisting of 35 041 households and 81 814 individuals is used as population data in the simulation study. This data set has been generated with the R package `simPopulation` [3, 13] based on Austrian EU-SILC survey data from 2006 and is about 1% of the size of the real Austrian population. A thorough investigation in a close-to-reality environment using real-life sized synthetic Austrian population data is future work.

From the synthetic data, 500 samples are drawn using simple random sampling. Each sample consists of 6000 households, which is roughly the sample size used in the real-life survey. With these samples, two scenarios are investigated. In the first scenario, no contamination is added. In the second, the equivalized disposable income of 0.25% of the households is contaminated. The contamination is thereby drawn from a normal distribution with mean $\mu = 1\,000\,000$ and standard deviation $\sigma = 10\,000$. Note that the *cluster effect* is considered, i.e., all persons in a contaminated household receive the same income. The threshold for Pareto tail modeling is in both cases set to $k = 275$ based on graphical exploration of the original EU-SILC sample with a Pareto quantile plot [5]. Furthermore, the tuning constant $c = 2.5$ is used for the bias-corrected WML estimator due to favorable robustness properties [11].

Figure 1 shows the results of the simulations without contamination for the quintile share ratio *(left)* and the Gini coefficient *(right)*. The three methods for tail modeling as well as the standard estimation method without tail modeling behave very similarly and are very close to the true values, which are represented by the grey lines. This is also an indication that the choice of k is suitable.

Figure 2 shows the results of the simulations with 0.25% contamination for the quintile share ratio *(left)* and the Gini coefficient *(right)*. Even such a small amount of contamination completely corrupts the standard estimation of these inequality indicators. Fitting the Pareto distribution with the Hill estimator is still highly influenced by the outliers. The best results are obtained with the PDC estimator, while the WML estimator shows a small negative bias.

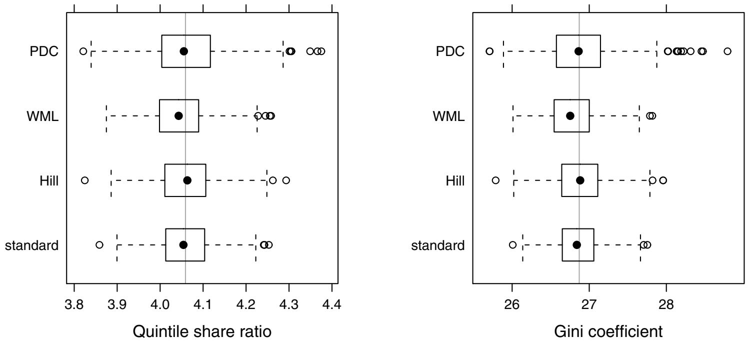


Fig. 1 Simulation results for the quintile share ratio *(left)* and the Gini coefficient *(right)* without contamination.

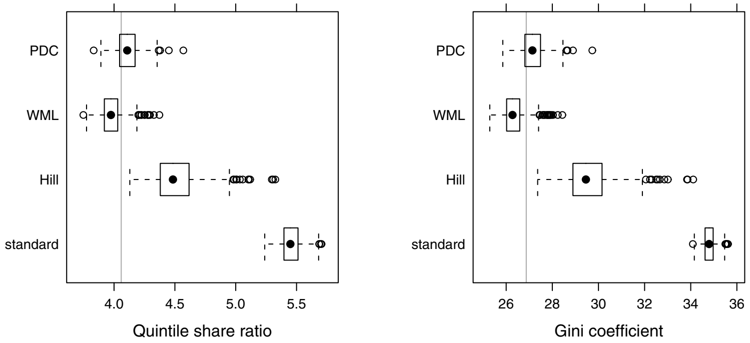


Fig. 2 Simulation results for the quintile share ratio *(left)* and the Gini coefficient *(right)* with 0.25% contamination.

5 Conclusions and Outlook

The quintile share ratio and the Gini coefficient, which are inequality indicators belonging to the set of Laeken indicators, are highly influenced by outliers. A simulation study for the case of simple random sampling showed that robust Pareto tail modeling can be used to reduce the influence of the outlying observations. The partial density component (PDC) estimator thereby performed best.

The simulation study in this paper is limited to simple random sampling because the estimators for Pareto tail modeling do not account for sample weights. Future work is to modify the estimators such that sample weights are taken into account, to investigate variance estimation, and to perform simulations using real-life sized synthetic population data.

Acknowledgements. This work was partly funded by the European Union (represented by the European Commission) within the 7th framework programme for research (Theme 8, Socio-Economic Sciences and Humanities, Project AMELI (Advanced Methodology for European Laeken Indicators), Grant Agreement No. 217322). Visit `http://ameli.surveystatistics.net` for more information on the project.

References

1. Alfons, A.: `simFrame`: Simulation Framework. R package version 0.1.2 (2009), `http://CRAN.R-project.org/package=simFrame`
2. Alfons, A., Holzer, J., Templ, M.: `laeken`: Laeken indicators for measuring social cohesion. R package version 0.1 (2010), `http://CRAN.R-project.org/package=laeken`

3. Alfons, A., Kraft, S., Templ, M., Filzmoser, P.: Simulation of synthetic population data for household surveys with application to EU-SILC. Research Report CS-2010-1, Department of Statistics and Probability Theory, Vienna University of Technology (2010), `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2010-1complete.pdf`
4. Alfons, A., Templ, M., Filzmoser, P.: `simFrame`: An object-oriented framework for statistical simulation. Research Report CS-2009-1, Department of Statistics and Probability Theory, Vienna University of Technology (2009), `http://www.statistik.tuwien.ac.at/forschung/CS/CS-2009-1complete.pdf`
5. Beirlant, J., Vynckier, P., Teugels, J.L.: Tail index estimation, Pareto quantile plots, and regression diagnostics. J. Amer. Statist. Assoc. 31(436), 1659–1667 (1996)
6. Beirlant, J., Vynckier, P., Teugels, J.L.: Excess functions and estimation of the extreme-value index. Bernoulli 2(4), 293–318 (1996)
7. Dupuis, D.J., Morgenthaler, S.: Robust weighted likelihood estimators with an application to bivariate extreme value problems. Canad. J. Statist. 30(1), 17–36 (2002)
8. Dupuis, D.J., Victoria-Feser, M.-P.: A robust prediction error criterion for Pareto modelling of upper tails. Canad. J. Statist. 34(4), 639–658 (2006)
9. EU-SILC: Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Eurostat, Luxembourg (2004)
10. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Statist. 3(5), 1163–1174 (1975)
11. Holzer, J.: Robust methods for the estimation of selected Laeken indicators. Master's Thesis, Vienna University of Technology (2009)
12. Kleiber, C., Kotz, S.: Statistical Size Distributions in Economics and Actuarial Sciences. Wiley, Hoboken (2003)
13. Kraft, S., Alfons, A.: `simPopulation`: Simulation of synthetic populations for surveys based on sample data. R package version 0.1.1 (2010), `http://CRAN.R-project.org/package=simPopulation`
14. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna (2010) ISBN 3-900051-07-0, `http://www.R-project.org`
15. Terrel, G.: Linear density estimates. In: Proceedings of the Statistical Computing Section of the American Statistical Association, pp. 297–302 (1990)
16. Vandewalle, B., Beirlant, J., Christmann, A., Hubert, M.: A robust estimator for the tail index of Pareto-type distributions. Comput. Statist. Data Anal. 51(12), 6252–6268 (2007)
17. Van Kerm, P.: Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC. IRISS Working Paper Series 2007-01, CEPS/INSTEAD (2007)

R Code for Hausdorff and Simplex Dispersion Orderings in the 2D Case

Guillermo Ayala

Abstract. This paper proposes a software implementation using R of the Hausdorff and simplex dispersion orderings. A copy can be downloaded from `http://www.uv.es/~ayala/software/fun-disp.R`. The paper provides some examples using the functions *exactHausdorff* for the Hausdorff dispersion ordering and the function *simplex* for the simplex dispersion orderings. Some auxiliary functions are commented too.

1 The Introduction

The Hausdorff and simplex dispersion orderings have proposed in [6] and [1] respectively. Although the definitions are considered for d-dimensional random vectors we will assume in this paper 2-dimensional random vectors.

First, let us give basic notation used later. If $x \in \mathbb{R}^d$ and $r \in [0,\infty)$ then $B(x,r)$ is the ball centered at x with radius r. If $A \subset \mathbb{R}^d$ then $co(A)$ will denote the convex hull of the set A. For $A,B \subset \mathbb{R}^d$ then the Hausdorff distance between them will be denoted as $d_H(A,B)$ and $A+B$ the Minkowski addition. The usual stochastic ordering will be denoted as $\preceq_{st}$.

Let us begin by remembering those definitions. If X and Y are two random vectors and $r \in [0,\infty)$ then X is less dispersive than Y in the *Hausdorff dispersion ordering* for the index r, denoted as $X \preceq_H^r Y$, if

$$\begin{aligned} & d_H\big(co\,(\{X\} \cup B_r(EX)), co\,(\{X'\} \cup B_r(EX))\big) \qquad (1) \\ & \preceq_{st} d_H\big(co\,(\{Y\} \cup B_r(EY)), co\,(\{Y'\} \cup B_r(EY))\big), \end{aligned}$$

with X and X' i.i.d. (respectively Y and Y' i.i.d.).

Guillermo Ayala
Dpto. de Estadística e Investigación Operativa, Universidad de Valencia,
46100-Burjasot, Spain
e-mail: `Guillermo.Ayala@uv.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 25–31.
springerlink.com

Let X and Y be random vectors then X is less dispersive than Y in the *simplex dispersion ordering*, denoted as $X \preceq_{sx} Y$, if

$$d_H\big(\mathscr{S}_{\boldsymbol{X}},\mathscr{S}_{\boldsymbol{X}'}\big) \preceq_{st} d_H\big(\mathscr{S}_{\boldsymbol{Y}},\mathscr{S}_{\boldsymbol{Y}'}\big) \tag{2}$$

where $\boldsymbol{X} = (X_1,X_2,X_3)$ and $\boldsymbol{Y} = (Y_1,Y_2,Y_3)$ are two random samples of X and Y and $\mathscr{S}_{\boldsymbol{X}}$ is the convex hull of $\boldsymbol{X}$.

I have developed a collection of R functions in order to evaluate both dispersion orderings. This collection of R functions can be download from `http://www.uv.es/~ayala/software/fun-disp.R`. In this paper, we explain how they can be used. Note that the analyses included in this paper can be reproduced by a simple copy-paste of the code.

2 Data

First, we declare our functions and load the packages needed later. In particular, we will need the R packages *geometry, mvtnorm* and *Hmics* [4],[3], [5]. We will give later details about their use.

```
> source("fun-disp.R"); library(geometry); library(mvtnorm)
```

We will use multivariate normal distribution data generated using the package *mvtnorm* [3]. Let us consider the $\mathbb{R}^d$-valued random vectors X and Y with normal distributions, $X \sim_{st} N(\mu_X,\Sigma_X)$ and $Y \sim_{st} N(\mu_Y,\Sigma_Y)$, where $\Sigma_X = AA^t$, $A \in M_{d\times d}$ being a matrix whose values are randomly chosen with uniform distribution in the interval $(0,1)$, the super index ^{t} denoting the transpose matrix, and $\Sigma_Y = \Sigma_X + \lambda I_d$, with $\lambda \geq 0$. It is well-known that the eigenvalues of Σ_Y are those of Σ_X plus the value λ. It holds that $X \preceq_{sx} Y$ [1]. Roughly speaking, larger values of λ will produce larger dispersion for the random vector Y.

Let us generate two point sets from the model just considered.

```
> n = 100; mu1 = rep(0,2); mu2 = mu1; lambda = 0.5; n1 = n2 = n = 100
> sigma1 = matrix(runif(4), nrow = 2, ncol = 2)
> sigma1 = t(sigma1) %*% sigma1
> sigma2 = sigma1 + lambda * diag(1, 2)
> A = rmvnorm(n1, mean = mu1, sigma = sigma1)
> B = rmvnorm(n2, mean = mu2, sigma = sigma2)
```

Figure 1 shows the original point set A and the corresponding convex hull obtained by using the package *geometry* [4] that contains an interface to *qhull* (`http://www.qhull.org/`).

```
> plot(A); chA = convhulln(A); lines(A[chA[1, ], ])
> for (i in 2:nrow(chA)) lines(A[chA[i, ], ])
```

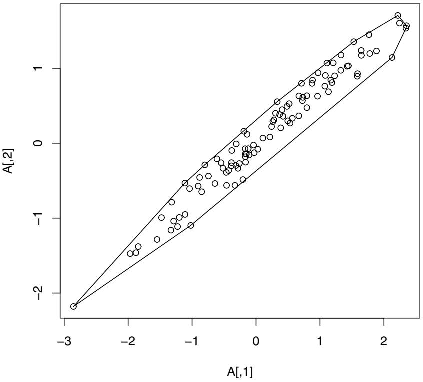


Fig. 1 A data set and the corresponding convex hull

3 Hausdorff Dispersion Ordering

The basic reference is [6]. We need to calculate the Hausdorff distance between $x_1 + B(y,r)$ and $x_2 + B(y,r)$ where $x_1, x_2, y \in \mathbb{R}^2$ and $B(m,r)$ is the disc centered at m with radius r. Our first approach will be to discretize $co(x + \partial B(y,r))$. The number of points used in the discretization is *NOP*. Then the continuous set $x + B(y,r)$ is replaced by the corresponding discrete set. Finally we calculate the Hausdorff distance between the corresponding convex hulls of the discrete sets composed by 100 points. The function *cone2* calculates this distance. Let us see how to calculate this Hausdorff distance.

```
> x1 = c(3, 5); x2 = c(7, 9); center = c(4, 4); radius = 1;
> NOP = 100
> cone2(x1, x2, center, radius, NOP)
```

```
[1] 5.656854
```

Given a random vector X and a random sample $\{x_1, \ldots, x_n\}$, The function *bootcone* provides us a bootstrap sample of $d_H(co(X_1^* + B(y,r)), co(X_2^* + B(y,r)))$ where X_1^*, X_2^* is a random sample without replacement from $\{x_1, \ldots, x_n\}$. Let us generate a sample and see the empirical distribution function in Figure 2.

```
> A.bc = bootcone(A, radius = 0.2, NOP = 100, nresamples = 10)
> Ecdf(A.bc)
```

If we have two samples x and y from X and Y then it can be tested if X is less dispersive than Y in the Hausdorff dispersion ordering using the following code. Note that the function *uso.test* tests the usual stochastic ordering using the Wilcoxon and Kolmogorov-Smirnov tests.

```
> AB.bc = bootcone2(A, B, radius, NOP = 100, nresamples = 10)
> uso.test(AB.bc$dhx, AB.bc$dhy)
```

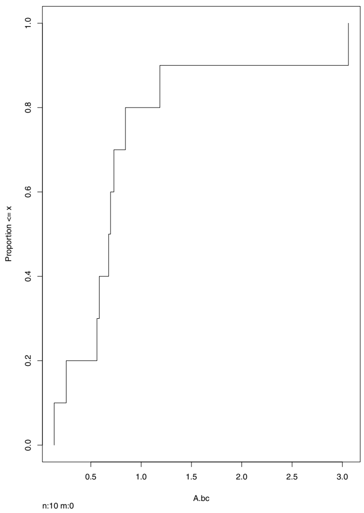


Fig. 2 Empirical distribution function of the Hausdorff distances $d_H(co(X_1^* + B(y,r)), co(X_2^* + B(y,r)))$.

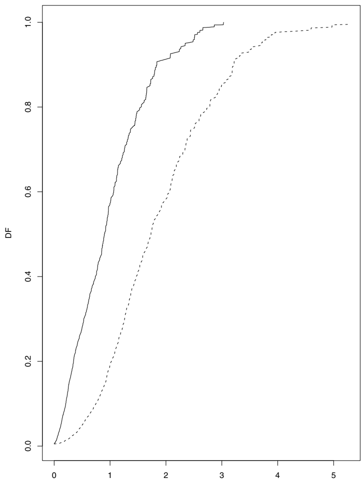


Fig. 3 Empirical distribution functions of the Hausdorff distances $d_H(co(X_1^* + B(y,r)), co(X_2^* + B(y,r)))$.

In the 2-dimensional case, an exact algorithm to calculate the distance $d_H(co(X_1^* + B(y,r)), co(X_2^* + B(y,r)))$ has been proposed [2]. It has been implemented in the function *exactHausdorff*. The following code calculates these distance from the two data sets and displays the empirical distribution functions. See Figure 3.

```
> r = 0.1; n= 100; prob = rep(1/n, n)
> HA = exactHausdorff(A, prob, r); HB = exactHausdorff(B, prob, r)
> plot(HA$distance, cumsum(HA$probability), type = "l", xlab = "",
+     ylab = "DF", xlim = range(c(HA,HB)))
> lines(HB$distance, cumsum(HB$probability), lty = 2)
```

Finally, the test if performed using the following code. The Kolmogorov-Smirnov test is used to test the usual stochastic ordering.

```
> y = rbind(A, B); x = c(rep(1, nrow(A)), rep(2, nrow(B)))
> z = testHDO(y, x, r = 0.2)
```

4 Simplex Dispersion Ordering

A detailed explanation of this algorithm can be found in [2].

If $A=\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n_1}\}$ and $B=\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_{n_2}\}$, let $\{i_1,\ldots,i_{d+1},i_{d+2},\ldots,i_{2(d+1)}\}$ be a sample without replacement from $\{1,\ldots,n_1\}$, and $U=d_H(\mathrm{co}(\boldsymbol{x}_{i_1},\ldots,\boldsymbol{x}_{i_{d+1}}),\mathrm{co}(\boldsymbol{x}_{i_{d+2}},\ldots,\boldsymbol{x}_{i_{2(d+1)}}))$. Therefore, s_1 independent extractions from the set $\{1,\ldots,n_1\}$ will produce a random sample of the corresponding bootstrap distribution $u_1,\ldots,u_{s_1}$. Replacing $\boldsymbol{x}$ by $\boldsymbol{y}$, we obtain $v_1,\ldots,v_{s_2}$, a random sample of the bootstrap distribution associated to the vector $\boldsymbol{y}$. Now, these values can be used for the proposed tests.

First, we describe some auxiliary functions. The function *rotatePHA* provides the angle to rotate a point w to the positive x-axis.

```
> w = c(-1, 5)
```

The angle is given by

```
> (tau = rotatePHA(w))

[1] 1.768192
```

and the rotated point can be found using

```
> AA = rbind(c(cos(tau), sin(tau)), c(-sin(tau), cos(tau)))
> w.rotated = t(AA %*% t(t(w)))
```

Given three points (corresponding with the rows of pp), we need to know if the convex hull of these points is a triangle, a segment or just they are the same point. This is given by the function *whichShape*.

```
> pp = matrix(data = c(0, -1, 0, -3, -3, 0), ncol = 2, byrow = T)
> whichShape(pp)
 [1] "triangle"
```

In order to calculate the Hausdorff distance between the point z and the convex hull of the points corresponding to the rows of pp, we have to move all points.

```
> z = c(-9, 1)
> pp = matrix(data = c(0, -1, 0, -3, -3, 0), ncol = 2, byrow = T)
> zpp = moveShapeAndPoint(z, pp, dib = T)
```

The different steps are illustrated in figure 4.

The function *distShape* calculates the Hausdorff distance by taking into account if the convex hull is a triangle, a segment or just a point. A detailed explanation of the calculations can be found in [2].

```
> z = c(-9, 1)
> pp = matrix(data = c(0, -1, 0, -3, -3, 0), ncol = 2, byrow = T)
> distShape(z, pp)
         1
2 6.082763
```

The function *simplex* provides us a sample of *u*'s.

```
> d1 = simplex(A, withBootstrap = TRUE, nresamples = 100)
```

If we consider two different samples we can test the simplex dispersion ordering using

```
> d1 = simplex2(A, B, withBootstrap = TRUE, nresamples = 10)
> uso.test(d1$dhx, d1$dhy)
```

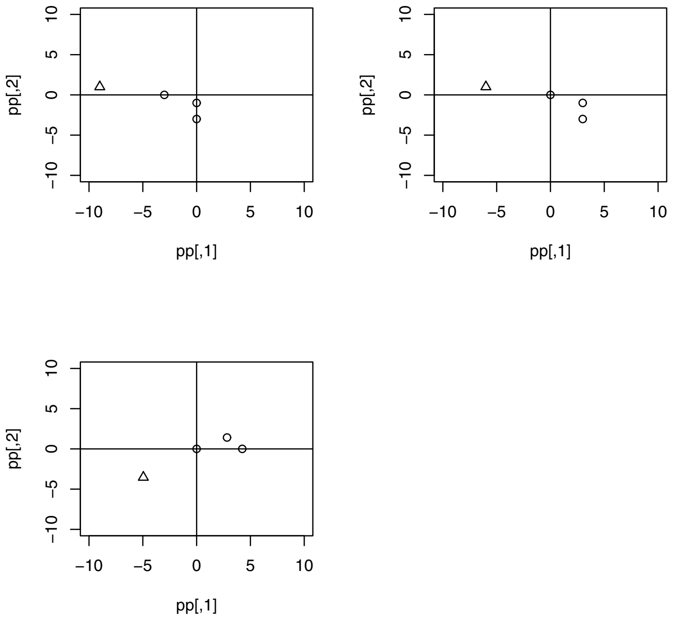


Fig. 4 Moving a triangle.

Acknowledgements. This paper has been funded by grant TIN2009-14392-C02-01. The software is a result of the collaboration with Miguel López-Díaz from the University of Oviedo.

References

1. Ayala, G., López-Díaz, M.: The simplex dispersion ordering and its application to the evaluation of human corneal endothelia. J. Multivariate Anal. 100, 1447–1464 (2009)
2. Ayala, G., López-Díaz, M.C., López-Díaz, M., Martínez-Costa, L.: Methods and algorithms to test the simplex and hausdorff dispersion orders with a simulation study and an ophthalmological application. Technical report, Universidad de Oviedo (2010)
3. Genz, A., Bretz, F., Hothorn, T.: mvtnorm: Multivariate Normal and t Distributions. R package version 0.9-2 (2008)
4. Grasman, R., Gramacy, R.B.: geometry: Mesh generation and surface tesselation. R package version 0.1-3 (2008)
5. Harrell, F.E.: Hmisc: Harrell Miscellaneous. R package version 3.7-0 (2009)
6. López-Díaz, M.: An indexed multivariate dispersion ordering based on the Hausdorff distance. J. Multivariate Anal. 97(7), 1623–1637 (2006)

On Some Confidence Regions to Estimate a Linear Regression Model for Interval Data

Angela Blanco-Fernández, Norberto Corral,
Gil González-Rodríguez, and Antonio Palacio

Abstract. Least-squares estimation of various linear models for interval data has already been considered in the literature. One of these models allows different slopes for mid-points and spreads (or radii) integrated in a unique equation based on interval arithmetic. A preliminary study about the construction of confidence regions for the parameters of that model on the basis of the least-squares estimators is presented. Due to the lack of realistic parametric models for random intervals, bootstrap approaches are proposed. The empirical suitability of the bootstrap confidence sets will be shown by means of some simulation studies.

Keywords: Confidence region, Simple linear regression model, Interval random set, Bootstrap approach.

1 Introduction

The study of the linear relationship between two random intervals has been addressed in the literature on the basis of several set arithmetic-based regression models (see, for instance, [2, 3, 4, 5, 6, 7, 8]). In order to analyze those models the mid-spread representation of the involved intervals is employed. The utility of this representation is twofold. On one hand, it captures the location and imprecision of the intervals, and on the other hand, it is technically

Angela Blanco-Fernández and Norberto Corral
Statistics and Operational Research Department, Oviedo University,
33007 Oviedo, Spain
e-mail: blancoangela@uniovi.es,norbert@uniovi.es

Gil González-Rodríguez and Antonio Palacio
European Centre for Soft Computing, 33600 Mieres, Spain
e-mail: gil.gonzalez@softcomputing.es,antoniopalacio1982@gmail.com

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 33–40.
springerlink.com

easier to handle than the minimum-maximum representation. The linear model presented in [3], denoted by Model M, generalizes those in [4] and [6].

Least squares estimation problems of Model M has been also considered in [3]. On the basis of least-squares estimators different approaches to determine confidence regions can be proposed. Contrary to what happens when the linear regression problem between real random variables is addressed, in the interval scenario no realistic parametric models to describe the distribution of the random sets have been defined up to now. Thus, exact methods are not feasible. Inferential studies about Model M can be developed by means of asymptotic techniques, based on the study of the limit distributions of the regression estimators. To improve the results for finite sample sizes, bootstrap methods are widely considered. In this work several bootstrap approaches are considered in order to build confidence sets for the parameters of the model.

2 Preliminaries

Let $(\mathscr{K}_c(\mathbb{R}), +, \cdot)$ be the space of nonempty compact intervals of $\mathbb{R}$ endowed with the semilinear structure induced by the Minkowski addition and the product by a scalar, that is, $A+B = \{a+b \,|\, a \in A, b \in B\}$ and $\lambda A = \{\lambda a \,|\, a \in A\}$ for all $A, B \in \mathscr{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. Moreover, given $A, B \in \mathscr{K}_c(\mathbb{R})$, if there exists $C \in \mathscr{K}_c(\mathbb{R})$ so that $A = B + C$, then C is defined as the Hukuhara difference between A and B, denoted by $A -_H B$. The interval A can be characterized by means of the real vector $(\inf A, \sup A) \in \mathbb{R}^2$ such that $\inf A \leq \sup A$, or equivalently, by its mid-point (or centre) and its spread (or radius), that is, $(\text{mid}A, \text{spr}A)$ with $\text{spr}A \geq 0$, where $\text{mid}A = (\sup A + \inf A)/2$ and $\text{spr}A = (\sup A - \inf A)/2$. The notation $A = [\inf A, \sup A]$ or $A = [\text{mid}A \pm \text{spr}A]$, respectively, will be considered in each case.

Several metrics can be defined on the space $\mathscr{K}_c(\mathbb{R})$. For least squares problems associated with regression studies, an L_2-type metric is suitable. Taking inspiration on the family of metrics for compact convex sets introduced in [10], a generalized L_2-type distance between two intervals A and B can be defined as

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

with $\theta > 0$.

Given a probability space $(\Omega, \mathscr{A}, P)$, a mapping $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ is said to be an *interval-valued random set* (or *random interval*), if it is $\mathscr{A}|\mathscr{B}_{d_\theta}$-measurable, $\mathscr{B}_{d_\theta}$ denoting the σ-field generated by the topology induced by the metric d_θ on $\mathscr{K}_c(\mathbb{R})$.

Let $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ be a random interval such that $E(|X|) < \infty$ (with $|X|(\omega) = \sup\{|x| \,\big|\, x \in X(\omega)\}$ for all $\omega \in \Omega$), then, the *expected value of X in Kudō-Aumann's sense* (see, e.g., [1]) is the interval $E(X) = [E(\inf X), E(\sup X)]$. The variance of X is defined in the classical statistical way, in terms of the d_θ metric, as $\sigma_X^2 = E\big(d_\theta(X, E(X))^2\big)$, whenever $E(|X|^2) < \infty$. However, it is not possible to define the covariance analogously to the usual concept, due to the lack of linearity on $\mathcal{K}_c(\mathbb{R})$. Through the (*mid-spr*) parametrization it is possible to define the covariance between X and Y by means of the natural concept of covariance in Hilbert spaces as $\sigma_{X,Y} = E\left(\langle t_X - E_{t_X}, t_Y - E_{t_Y}\rangle_\theta\right)$, where $t_X = (\text{mid}X, \text{spr}X) \in \mathbb{R}^2$ (analogously for t_Y), and $\langle\cdot,\cdot\rangle_\theta$ is an inner product on $\mathbb{R}^2$ defined in terms of the constant $\theta > 0$ (see [1]) as $\langle \mathbf{a}, \mathbf{b}\rangle_\theta = \mathbf{a}'\begin{pmatrix} 1 & 0 \\ 0 & \theta \end{pmatrix}\mathbf{b}$ for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$. The covariance can be expressed in terms of mids and spreads as $\sigma_{X,Y} = \text{Cov}(\text{mid}X, \text{mid}Y) + \theta\text{Cov}(\text{spr}X, \text{spr}Y)$. The variance of the interval X can be also expressed as $\sigma_X^2 = \text{Var}(\text{mid}X) + \theta\,\text{Var}(\text{spr}X)$.

The estimators for the moments of random intervals presented above are the usual ones. Given a simple random sample $\{(X_i, Y_i)\}_{i=1}^n$ from (X, Y), let us define by $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$, $\widehat{\sigma}_X^2 = \overline{d_\theta(X, \overline{X})^2}$ (analogously $\overline{Y}$ and $\widehat{\sigma}_Y^2$) and $\widehat{\sigma}_{X,Y} = \overline{\langle t_X - \overline{t_X}, t_Y - \overline{t_Y}\rangle_\theta}$ the sample mean, the sample variance and the sample covariance for random intervals, respectively.

3 Linear Regression Model M

A natural way to model the relationship between two random intervals has been previously proposed by the expression $Y = \alpha X + \varepsilon$, with $\alpha \in \mathbb{R}$ and ε a random interval such that $E(\varepsilon|X) = B \in \mathcal{K}_c(\mathbb{R})$ (see [6]). Nevertheless, this model is not flexible enough for many real-life applications. As an example, it can be easily checked that this interval linear model tranfers relationships between the *mid* and *spr* real variables by means of the expressions $\text{mid}Y = \alpha\text{mid}Y + \text{mid}\varepsilon$ and $\text{spr}Y = |\alpha|\text{spr}Y + \text{spr}\varepsilon$. Since both equations involve the same regression coefficient (in absolute value), the model is somehow restrictive.

With the aim of considering the *mid* and *spr* components of the intervals separately, but keeping the good properties of the interval arithmetic, a new representation has been introduced in [3]. Each interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed as $A = \text{mid}A[1 \pm 0] + \text{spr}A[0 \pm 1]$. This notation gives the inspiration to formalize the called **Model M** between X and Y in [3] as

$$Y = \alpha\text{mid}X[1 \pm 0] + \beta\text{spr}X[0 \pm 1] + \varepsilon \tag{2}$$

with $\alpha, \beta \in \mathbb{R}$ and $E(\varepsilon|X) = B \in \mathcal{K}_c(\mathbb{R})$. For simpler notation, the linear model (2) will be denoted by $Y = \alpha X^M + \beta X^S + \varepsilon$. Moreover, it is easy to check that $X^S = -X^S$ (since $X^S(\omega) = [-\text{spr}X(\omega), \text{spr}X(\omega)]$, for all $\omega \in \Omega$), so it is possible to consider $\beta \geq 0$ without loss of generality.

From (2) the linear relationships for *mid* and *spr* variables of X and Y are $\text{mid}Y = \alpha\text{mid}Y + \text{mid}\varepsilon$ and $\text{spr}Y = |\beta|\text{spr}Y + \text{spr}\varepsilon$, which clearly entails more flexibility. The flexibility is associated with the extra parameter of Model M, which depends on two scalars and one interval value.

3.1 Least Squares Estimation of Model M

The least-squares (LS) estimation of the regression parameters of the model (2) has been developed in [3]. The LS approach leads to a contrained minimization problem, namely,

$$\left.\begin{array}{l} (\widehat{\alpha},\widehat{\beta},\widehat{B}) = \operatorname{argmin}_{\left\{a\in\mathbb{R},b\geq 0,C\in\mathcal{K}_c(\mathbb{R})\right\}} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} d_\theta^2(Y_i, aX_i^M + bX_i^S + C) \\ \\ \text{subject to } \; b\in S \end{array}\right\} \tag{3}$$

where $S = \{b \in [0,\infty) : Y_i -_H bX_i^S \text{exists, for all } i = 1,\ldots,n\}$. It is easy to check that $b \in S$ implies that $Y_i -_H (aX_i^M + bX_i^S)$ exists for all $i = 1,\ldots,n$ and for all $a \in \mathbb{R}$. The existence of these Hukuhara differences assures the existence of the residuals of the sample model, and thus, the coherence of the solutions as suitable estimators of the regression parameters.

It should be underlined that, as it was shown in [6] for the simpler model, if the restriction is overseen, the obtained estimates of the parameters could not work as estimates for the model (because the residuals could not exist).

The resolution of problem (3) provides the following expressions:

$$\widehat{\alpha} = \frac{\widehat{\sigma}_{X^M,Y}}{\widehat{\sigma}^2_{X^M}} \;, \quad \widehat{\beta} = \min\left\{\widehat{s}_0, \max\left\{0, \frac{\widehat{\sigma}_{X^S,Y}}{\widehat{\sigma}^2_{X^S}}\right\}\right\} \text{ and} \tag{4}$$

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

where $\widehat{s}_0 = \min\{\text{spr}Y_i/\text{spr}X_i : \text{spr}X_i \neq 0\}$ ($\widehat{s}_0 = \infty$ if $\text{spr}X_i = 0$ for all $i = 1,\ldots,n$).

4 Bootstrap Confidence Regions for the Regression Parameters

Since it is not feasible to look for the exact distribution of the LS estimators and since the asymptotic results usually provide good results only for very large sample sizes, in this section some alternatives based on bootstrapping are explored.

Different schemes to generate bootstrap samples from Model (2) can be followed. When a fixed design is considered (that is, the independent variable is not random but deterministic), the most usual procedure is the *residual bootstrap*. On other hand, when both variables in the linear model are considered as random elements, the natural resampling is made from a simple random sample of the pair of variables by means of the *paired bootstrap* (see [9] for a complete description of both procedures). The linear model (2) is formalized for two random intervals, so the paired bootstrap approach will be used for the development of inferential studies about Model M.

Several bootstrap confidence sets can be constructed for the regression parameters of linear models involving real-valued random variables (see [9]).

The best known ones are the *percentile*, *hybrid* and t- bootstrap confidence set. Each of them is based on the sample distribution of a different bootstrap expression obtained from the bootstrap estimator of the parameter.

Let X and Y be random intervals verifying Model (2). The separate expressions for the least-squares estimators of the parameters α and β presented in (4) allow us to build confidence sets for each parameter separately. Let $\widehat{\alpha}$ be the least-squares estimator of α obtained from a simple random sample $\{X_i,Y_i\}_{i=1}^n$ from (X,Y). We denote by $\{X_i^*,Y_i^*\}_{i=1}^n$ a bootstrap sample, generated by means of the election of n elements uniformly and with replacement from $\{X_i,Y_i\}_{i=1}^n$. Let $\widehat{\alpha}^*$ be the least-squares estimator of α with respect to the bootstrap sample. From the bootstrap estimator $\widehat{\alpha}^*$ the procedure to build the three confidence intervals (CI) for parameter α follows.

- *Bootstrap percentile CI*: If we denote by K_{BOOT} the distribution function of the bootstrap estimator $\widehat{\alpha}^*$, the bootstrap percentile CI for α at a confidence level $1-\rho$ is defined by means of the corresponding percentiles of K_{BOOT}, that is,

$$IC_P(\alpha)_{1-\rho} = \left[K_{BOOT}^{-1}(\rho/2)\ ,\ K_{BOOT}^{-1}(1-\rho/2) \right], \tag{5}$$

  where K_{BOOT}^{-1} denotes the pseudoinverse of K_{BOOT}.
- *Bootstrap hybrid CI*: Let H_{BOOT} be the distribution function of the term $n^l(\widehat{\alpha}^* - \widehat{\alpha})$, where l is an arbitrary constant. $H_{BOOT}(x) = P[n^l(\widehat{\alpha}^* - \widehat{\alpha}) \leq x]$, for $x \in \mathbb{R}$. The most usual election for l is $1/2$. Thus, the bootstrap hybrid CI for α at significance level ρ has got the expression

$$IC_H(\alpha)_{1-\rho} = \left[\widehat{\alpha} - \frac{1}{\sqrt{n}} H_{BOOT}^{-1}(1-\rho/2)\ ,\ \widehat{\alpha} - \frac{1}{\sqrt{n}} H_{BOOT}^{-1}(\rho/2) \right] \tag{6}$$

- *t-bootstrap CI*: We consider the standarized pivot $R = \dfrac{\widehat{\alpha}-\alpha}{\widehat{\sigma}_{\widehat{\alpha}}}$, where $\widehat{\sigma}^2_{\widehat{\alpha}}$ is an estimator of the variance of $\widehat{\alpha}$, and the bootstrap replica of R, $R^* = \dfrac{\widehat{\alpha}^* - \widehat{\alpha}}{\widehat{\sigma}^*_{\widehat{\alpha}^*}}$, with $\widehat{\sigma}^*_{\widehat{\alpha}^*}$ the analogous estimator for the variance of $\widehat{\alpha}^*$. If we denote by G_{BOOT} the distribution function of R^*, the t-bootstrap CI for α at confidence level $1-\rho$ is given by

$$IC_T(\alpha)_{1-\rho} = \left[\widehat{\alpha} - \widehat{\sigma}_{\widehat{\alpha}} G_{BOOT}^{-1}(1-\rho/2)\ ,\ \widehat{\alpha} - \widehat{\sigma}_{\widehat{\alpha}} G_{BOOT}^{-1}(\rho/2) \right] \tag{7}$$

Remark 1. The percentiles of the functions K_{BOOT}, H_{BOOT} and G_{BOOT} (in each case) can be approximated from the empirical distribution of $\widehat{\alpha}^*$ by means of MonteCarlo Method.

Remark 2. It can be shown that the estimator of the variance of $\widehat{\alpha}$ can be expressed as $\widehat{\sigma}^2_{\widehat{\alpha}} = \dfrac{\widehat{\sigma}^2_{\mathrm{mid}\widehat{\varepsilon}}}{n\widehat{\sigma}^2_{\mathrm{mid}X}}$. However, it is difficult to obtain an analytic expression for $\widehat{\sigma}_{\widehat{\beta}}$. In this case, a bootstrap estimator of the variance of $\widehat{\beta}$ can be approximated by means of MonteCarlo Method based on B_2 bootstrap replications (see [9]).

Taking into account the definitions and remarks presented above, an algorithm for the construction of the *percentile*, *hybrid* and *t*- bootstrap confidence set for parameter α of Model M has the following form.

Algorithm: bootstrap confidence sets for α

Let $\{X_i, Y_i\}_{i=1}^n$ be a random sample obtained from (X,Y). Let ρ be a fixed significance level and $B \in \mathbb{N}$ large enough.

P1. Compute the estimates $\widehat{\alpha}$ and $\widehat{\sigma}^2_{\widehat{\alpha}}$.

P2. Generate B bootstrap samples $\{X_i^*, Y_i^*\}_{i=1}^n$ of size n, resampling with replacement from the original sample $\{X_i, Y_i\}_{i=1}^n$.

P3. For each iteration $b = 1, \ldots, B$, compute the estimate for α from the corresponding bootstrap sample, $\widehat{\alpha}^*(b) = \dfrac{\widehat{\sigma}_{X^{M*},Y^*}}{\widehat{\sigma}^2_{X^{M*}}}$, and the bootstrap estimator of its variance, $\widehat{\sigma}^{*2}_{\widehat{\alpha}^*} = \dfrac{\widehat{\sigma}^2_{\mathrm{mid}\varepsilon^*}}{n\widehat{\sigma}^2_{\mathrm{mid}X^*}}$.

P4. Aproximate the lower and upper limits of the intervals (5), (6) and (7) substituting the quantiles of the distributions with the corresponding quantiles from the empirical distribution of $\widehat{\alpha}^*$. That is, the values $\{\widehat{\alpha}^*(b)\}_{b=1}^B$ are increasing ordered, and the ones in position $[(\rho/2)B] + 1$ and $[(1-\rho/2)B]$ are selected (where $[\cdot]$ denotes the integer function). Let $\widehat{\alpha}^*_{C1}$ and $\widehat{\alpha}^*_{C2}$ be that values. Thus, the *percentile*, *hybrid* and *t*- confidence sets for α at a confidence level $1-\rho$ are given by

$$IC_P(\alpha)_{1-\rho} = \left[\ \widehat{\alpha}^*_{C1}\ ,\ \widehat{\alpha}^*_{C2}\ \right],$$

$$IC_H(\alpha)_{1-\rho} = \left[\ 2\widehat{\alpha} - \widehat{\alpha}^*_{C2}\ ,\ 2\widehat{\alpha} - \widehat{\alpha}^*_{C1}\ \right],\ \text{and}$$

$$IC_T(\alpha)_{1-\rho} = \left[\ \widehat{\alpha} - \widehat{\sigma}_{\widehat{\alpha}} \frac{\widehat{\alpha}^*_{C2} - \widehat{\alpha}}{\widehat{\sigma}^*_{\widehat{\alpha}^*_{C2}}}\ ,\ \widehat{\alpha} - \widehat{\sigma}_{\widehat{\alpha}} \frac{\widehat{\alpha}^*_{C1} - \widehat{\alpha}}{\widehat{\sigma}^*_{\widehat{\alpha}^*_{C1}}}\ \right]$$

respectively.

An analogous algorithm can be developed for the construction of the bootstrap confidence sets for the regression parameter β in Model M, taking into account the details explained in Remark 2.

4.1 Simulation Studies

The empirical behaviour of the bootstrap procedure can be shown by means of some simulation studies. Let us define a theoretical situation for two random intervals X and Y associated by means of the expression

$$Y = X^M + X^S + \varepsilon \tag{8}$$

where the independent interval X is characterized through the real random vector $(\text{mid}X, \text{spr}X)$ such that $\text{mid}X \sim N(0,1)$ and $\text{spr}X \sim \chi^2_1$, and the error interval term is also defined by $\text{mid}\varepsilon \sim N(0,1)$ and $\text{spr}\varepsilon \sim \chi^2_1 + 1$ independent from X.

For different samples sizes n, a random sample from (X,Y) is simulated. Let $\{X_i, Y_i\}_{i=1}^n$ be one of them. For $k = 10000$ iterations of the suggested bootstrap algorithms, the $0.95-$ bootstrap confidence sets for α (and analogously for β) based on $B = 1000$ bootstrap replications are computed, checking for each of them if the theoretical parameter $\alpha = 1$ (and $\beta = 1$) belongs to the corresponding confidence interval. Finally, the coverage rates are gathered in Table 1.

Table 1 Empirical confidence level of the bootstrap CIs for α and β

		α			β	
n	$IC_P(\alpha)$	$IC_H(\alpha)$	$IC_t(\alpha)$	$IC_P(\beta)$	$IC_H(\beta)$	$IC_t(\beta)$
30	0.9301	0.9318	0.9374	0.8852	0.8911	0.8969
50	0.9360	0.9458	0.9466	0.8985	0.9061	0.9067
100	0.9460	0.9465	0.9476	0.9012	0.9082	0.9124
200	0.9475	0.9487	0.9494	0.9111	0.9123	0.9152

Since the success rates are close to the nominal confidence level 0.95 (the larger sample size, the closer they are), the empirical correctness of the bootstrap procedure is justified. Indeed, for parameter α, the rate of convergence of the empirical significance level can be found in [9]. $IC_t(\alpha)$ is the most accurate, $IC_H(\alpha)$ the second one, and $IC_P(\alpha)$ is the less accurate of the three approaches. In the case of parameter β, the approximation to the nominal level is slower. A preliminary analysis of this result has shown that the expression of the estimator $\widehat{\beta}$ depending on the sample term $\widehat{s_0}$ entails that the bootstrap estimator $\widehat{\beta}^*$ does not always perform well. Let us recall that $\widehat{s_0}$ is an order statistic (it is defined as the minimum of several real random variables), for which classic bootstrap methods are inconsistent in some situations (see [9]).

5 Concluding Remarks

Different procedures to construct bootstrap confidence sets for the parameters α and β of Model M have been proposed. Their empirical correctness has been shown by means of some simulation studies. With respect to the parameter β, a wider study and a possible improvement of the bootstrap procedure for the construction of confidence sets will be addressed in future research. The statistical study of Model M will be extended by means of the development of other inferential studies, like hypothesis testing, the study of linear independence, among others.

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, the Principality of Asturias Grants IB09-042C1 and IB09-042C2, the COST Action IC0702 and a Research Grant from *Fundacion Banco Herrero*. Their financial support is gratefully acknowledged.

References

1. Aumann, R.J.: Integrals of set-valued functions. J. Math. Anal. Appl. 12, 1–12 (1965)
2. Blanco-Fernández, A., Colubi, A., Corral, N., González-Rodríguez, G.: On a linear independence test for interval-valued random sets. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Handling Variability and Imprecision. Advances in Soft Computing, vol. 48, pp. 331–337. Springer, Heidelberg (2008)
3. Blanco-Fernández, A., Corral, N., González-Rodríguez, G.: Estimation of a flexible simple linear model for interval data based on the set arithmetic (submitted for publication, 2010)
4. Gil, M.A., Lubiano, M.A., Montenegro, M., López-García, M.T.: Least squares fitting of an affine function and strength of association for interval-valued data. Metrika 56, 97–111 (2002)
5. Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval-valued data. Comput. Statist. Data Anal. 51(6), 3002–3015 (2007)
6. González-Rodríguez, G., Blanco-Fernández, A., Corral, N., Colubi, A.: Least squares estimation of linear regression models for convex compact random sets. Adv. Data Anal. Class. 1, 67–81 (2007)
7. Montenegro, M., Casals, M.R., Lubiano, M.A., Gil, M.A.: Two-sample hypothesis tests of means of a fuzzy random variable. Inf. Sci. 133, 89–100 (2001)
8. Montenegro, M., Colubi, A., Casals, M.R., Gil, M.A.: Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59(1), 31–49 (2004)
9. Shao, J., Tu, D.: The Jackknife and Bootstrap. Springer, New York (1995)
10. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact convex (fuzzy) sets based on a generalized concept of mid and spread. Inf. Sci. 179(23), 3964–3972 (2009)

Possibilistic Coding: Error Detection vs. Error Correction

Luca Bortolussi and Andrea Sgarro

Abstract. Possibilistic information theory is a flexible approach to old and new forms of coding; it is based on possibilities and patterns, rather than pointwise probabilities and traditional statistics. Here we fill up a gap of the possibilistic approach, and extend it to the case of error detection, while so far only error correction had been considered.

1 Introduction

The possibilistic approach to source and channel coding (to compression codes and error-correcting codes) arose as a formal game in which pointwise probabilities, as currently used in Shannon's information theory, were replaced by possibilities, so as to find a possibilistic equivalent for probabilistic notions as are error probability, source entropy and channel capacity, cf. [14, 15]. The formal game has proved to be more stimulating than was expected: the possibilistic approach could be applied to the design of error-correcting phone keyboards [11], to more theoretic questions like defining the nature of channel noise in biological computation [1, 2], or introducing adequate "operational" information measures [8, 13, 15], and also, more recently, to the construction of codes which correct *twiddles*, i.e. transpositions between consecutive letters, inadvertently made [3]. It turns out that possibilistic information theory includes as sub-cases both Shannon's zero-error information theory [9] and the standard approach to error correcting codes [10] based on checking Hamming distances between codewords. As for the first inclusion, suffice it to say that the possibilistic approach may be seen as a multi-step generalisation of

Luca Bortolussi and Andrea Sgarro
Dept. of Mathematics and Informatics, University of Trieste, Italy

Luca Bortolussi and Andrea Sgarro
Centre for Biomolecular Medicine, Area Science Park, Trieste, Italy
e-mail: `luca@dmi.units.it,sgarro@units.it`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 41–48.
springerlink.com

Shannon's approach, which is two-step only, without intermediate degrees of possibility between possible and impossible. As for the second inclusion, the probabilistic layer (or at least the traditional pointwise probabilistic layer) has always been felt to be rather thin, so that resorting to probabilistic symmetric channels, as done in introductory textbooks, e.g. [10], might appear to be a homage[1] paid to Shannon, rather than an intrinsic need of coding theory (as opposed to information theory proper, or Shannon theory).

The possibilistic approach to coding is rigorously Shannon-theoretic: based as it is on patterns rather than traditional statistics, it responds to a general need of new approaches to information theory felt in the computer science community [4]. A basic Shannon-theoretic notion like *channel noise* can be safely exported to the new setting, as we did in [1, 2] in the case of DNA word design (codewords are DNA strings, cf. [5]). One might want to mimic also in the DNA case what one successfully does for standard codes (the noise of symmetric channels is probabilistic), but we have proven in [2] that nothing like this holds in the DNA case, where *no* probabilistic description of channel noise is feasible. By this we have been able to give a remarkable example where channel noise is *intrinsically possibilistic and non-probabilistic.* Clearly, the probability which is ruled out here is *pointwise* probability: the interpretation of possibilities as *upper probabilities* suggests instead the feasibility of a probabilistic approach to information theory and coding which might prove to be quite comprehensive, even if its impact on coding practice remains to be assessed.

In this paper we emend a fault of the possibilistic approach, which appears to be able to deal only with *error correction* and not with *error detection*: so, at least seemingly, it has a weak point with respect not only to standard codes but also to DNA word design as covered in the literature. Below we give a solution to the problem of possibilistic error detection which is fully general, and which is based on the notion of *even codeword couples*, to be defined in Section 4; Sections 2 and 3 introduce our problem and ensure self-readability, while the problem itself is tackled in the final Section 4.

2 Distinguishability and Confusability

Let $\mathscr{X}$ be a finite metric space and let $d(x,y)$ be the corresponding metric distance. The idea is that an element $x \in \mathscr{X}$ is fed to a *transmission channel* and at the other end of the channel an element $z \in \mathscr{X}$ is observed, which might be different from x due to *channel noise*. The aim is to recover the correct input x from the observed output z. A *codebook* $\mathscr{C}$, or for short a

[1] Shannon's original approach is probabilistic even in the zero-error case: for him *possible* means that the probability is positive, however small it may be. Note that the notions of codeword distinguishability and codeword confusability are already present in the zero-error theory, even if not in the general form as below, Section 2, and so are due to Shannon.

code, is simply a non-void subset of elements called *codewords*, to be used as possible inputs to the channel.

Once we have a distance on $\mathscr{X}$ with maximal value N, a corresponding transition possibility from x to z can be obtained in a "canonical" way:

$$\text{Poss}(z|x) \;=\; 1 - N^{-1} d(x,z) \tag{1}$$

These transition possibilities can be arranged into an $|\mathscr{X}| \times |\mathscr{X}|$ *possibility matrix*: the entries in each row of a possibility matrix, rather than summing up to 1 as in a stochastic matrix, have a 1 as their maximum, as typical of possibility theory, which is maxitive rather than additive (for possibility theory cf. e.g. [7]). In channel coding, a transition possibility as (1) can be interpreted as follows: the possibility of receiving z when x is sent over the noisy channel is high or low according whether the "pattern similarity" between input x and output z is high or low. We stress that a *possibilistic noisy channel* is completely described by a possibility matrix. The distance-based (geometric) approach will be more palatable to coding theorists, but an explicit use of possibilities[2] has the advantage of better emphasising the links with information theory, on the base of the opposition probability vs. possibility. Even if overlooked in the literature on standard coding and DNA word design, and this for reasons explained below, basic coding-theoretic notions are codeword distinguishability or, equivalently, codeword confusability.

Definition 1. *The distinguishability of a couple* (x,y) *is defined to be*

$$\delta(x,y) \;=\; \min_{z \in \mathscr{X}} \; \max\{d(x,z), d(y,z)\}$$

From now on, unless otherwise specified, distances are assumed to be consecutive integers, $d(x,y) \in \{0,1,\ldots,N\}$. One soon proves the following[3] bounds:

$$\left\lceil \frac{d(x,y)}{2} \right\rceil \;\le\; \delta(x,y) \;\le\; d(x,y) \tag{2}$$

In a possibilistic setting one might prefer to deal with *confusabilities*

$$\gamma(x,y) = \max_{z \in \mathscr{X}} \; \min\{\text{Poss}(z|x), \text{Poss}(z|y)\} = 1 - N^{-1}\,\delta(x,y) \tag{3}$$

The rightmost equality assumes (1). Nothing much changes, and so in the following we shall stick to distinguishabilities. A situation when the lower bound in (2) is achieved is the following:

[2] The possibilistic framework can be readily and naturally enlarged to more general situations, e.g. when the input alphabet and the output alphabet are distinct, cf. [12, 14]: then dissimilarities (which take up the role of distances) are between input and output, while distinguishabilities involve two inputs.

[3] The lower bound follows from the triangle inequality; if the distances are not constrained to be integers the integer ceiling must be understood as the smallest available distance which is $\ge d(x,y)/2$, c.f. [12].

Definition 2. *An integer metric space is* dense *when, whatever the couple (x,y), for any integer $m \in [0, d(x,y)]$ one can find an element z at distance m from x and at distance $d(x,y) - m$ from y.*

An obvious example is given by Hamming distances for strings of the same length, a less obvious example is DNA word design, cf. [1, 2]. Instead, the upper bound in (2) is always achieved if and only if the (not necessarily integer) metric is an *ultrametric*, i.e. if and only if the fuzzy triangle inequality $\max\{d(x,z), d(z,y)\} \geq d(x,y)$ is always verified, cf. [12]. Cf. [3] for the case of significant string distances, e.g. variants of the edit distance or Spearman footrule [6], which might be used to correct twiddles, as hinted at in Section 1; the distinguishabilities corresponding to these distances are sometimes equal to the lower bound, sometimes to the upper bound, and sometimes have intermediate values, depending on (x,y), cf. [3].

3 Two Equivalent Approaches to Coding

For the moment being we deal only with error-correcting codes and so ignore error detection; the definition below might be equivalently given in term of confusabilities (3).

Definition 3. Optimal codes*: once the integer threshold Δ is chosen, construct maximum-size codes with* guaranteed minimum distinguishability *Δ, i.e. with $\delta(x,y) \geq \Delta$ for all couples of distinct codewords in $\mathscr{C}$ ($0 < \Delta \leq \max_{x,y} \delta(x,y) \leq N$).*

Whatever the code size, when threshold Δ is guaranteed one proves the following reliability criterion, given in two equivalent phrasings, cf. [12, 14]. To enhance self-readability, a quick proof is given; for more details cf. [12, 14].

Reliability criterion 1. *Decode to a codeword x which maximises the transition possibility* $\text{Poss}(z|x)$ *to the output z: the error possibility*[4] *is at most equal to $1 - \Delta/N$.*

Reliability criterion 2. *Once the output string z is received, decode to a codeword x which minimises the distance $d(x,z)$ between input and output: if the input string x was such that $d(x,z) < \Delta$, decoding is successful.*

Proof. If x is sent, z is received, and $y \neq x$ is decoded to, then $d(y,z) \leq d(x,z)$ and so, by definition 1, $\delta(x,y) \leq d(x,z)$: by comparison, $d(x,z) \geq \Delta$. □

If two codewords x and y have distinguishability $\delta(x,y) = \Delta$, one can provide an output z at distance $d(x,z) = \Delta$ from x and at distance $d(y,z) \leq \Delta$ from y,

[4] For each codeword y sent over the channel, its error possibility is the possibility of the set of the outputs z which lead to a decoding error, and so, according to the maxitive rules of possibility theory, it is the maximum possibility $\text{Poss}(z|y)$ of such z's.

or the other way round, which will bring about a decoding error of "weight" Δ and of possibility $1-\Delta/N$: in this sense, the Reliability criterion cannot be improved.

Actually, optimal codes of both standard coding and DNA word design are constructed by choosing a threshold T and checking directly distances rather than distinguishabilities. In general, ignoring distinguishabilities can lead to inconsistent results, in the sense that the resulting codebooks are nice combinatorial constructions devoid of error-correcting capabilities, cf. [1]. This is not so in the standard case or in the DNA case, because standard and DNA distinguishabilities are a *monotone*[5] function of the corresponding distances (recall that the lower bound (2) is always achieved in the case of dense spaces, Definition 2). When monotonicity is strict, everything is fine: here, however, monotonicity is only weak, and so the reader will object that one ends up "losing" all optimal codebooks which had been obtained by constraining the minimum distance $d(x,y)$ against an *even* integer threshold T. As a matter of fact, the Reliability criteria soon imply that even bounds on distances are completely useless if one insists on *hard* decoding (the decoder decides to a single codeword, however fishy the situation might be). Instead, even bounds on distances are quite relevant in *error detection*, when a *soft* decoder is used; in the next section we show how the possibilistic approach can deal with error detection quite in general.

4 Even Couples in Error Detection

Definition 4. *The couple (x,y) is an* even couple *when any z achieving $\delta(x,y)$ as in Definition 1 is at the same distance from both x and y, it is an* odd couple *if for any such z the two distances from x and y are distinct, else it is a* mixed couple*.*

Snags with error detection occur with mixed couples, i.e. when one can provide a *skew quadruple* (x,y,u,w) where u and w both achieve $\delta(x,y)$, but $\delta(x,y)=d(x,w)=d(w,y)$ while $\delta(x,y)=d(u,y)>d(u,x)$. Set $d(x,y)=d$, $\delta(x,y)=\rho$, $d(x,u)=\mu$, $d(u,w)=\xi$.

Lemma 1. *Four positive real numbers d,ρ,μ,ξ as above are the lengths of a skew quadruple in a metric space of size 4 if and only if they verify the constraints $\lceil d/2\rceil \le \rho \le d$, $\rho \ne d/2$, $d-\rho \le \mu < \rho$, $\rho-\mu \le \xi \le \rho+\mu$.*

Proof. Choose d; as for ρ the bounds (2) must hold. Forget ξ for the moment being: one is left with two triangles, and a check of the corresponding triangle inequalities gives $d-\rho \le \mu \le d+\rho$, but since μ should be strictly smaller

[5] The reader will have appreciated that, thinking of optimal codes and reliability, the possibilistic approach is basically invariant with respect to strictly monotone transformations of the transition possibilities involved, or of the distances involved.

than ρ one ends up imposing $d-\rho \le \mu < \rho$. To avoid that this interval be void one must rule out the value $d/2$ for ρ. Adding ξ gives two more triangles, and a check of the corresponding triangle inequalities completes the proof. □

The proof does not assume that the distances should be integers; if it is so, the constraints on ρ can be subsumed to $\lceil (d+1)/2 \rceil < \rho \le d$. The lemma allows one, after choosing d, to find ρ, μ, ξ in this order; it can be used in spaces of any size to spot mixed couples. E.g. take $d=2$ to find the three integer solutions $\rho = 2$, $\mu = 1$, $\xi \in \{1,2,3\}$. Of these, the third gives the distance matrix below; transition possibilities are soon obtained from (1) setting $N=3$; distinguishabilities are equal to distances, as soon checked, save $\delta(u,w) = 3 = \lceil d(u,w)/2 \rceil$:

	x	y	u	w
x	0	2	1	2
y	2	0	2	2
u	1	2	0	3
w	2	2	3	0

Theorem 1. *If x and y achieve the lower bound* (2) *and $d(x,y)$ is an integer, the couple (x,y) is even or odd according whether their distance $d(x,y)$ is even or odd, respectively. If x and y achieve the upper bound* (2)*, the couple (x,y) cannot be even.*

Proof. The first claim in the theorem below is a straightforward by-product of the lemma, which rules out mixed couples for $d = \delta/2$; the rest soon follows from the triangle inequality. As for the second claim just think that x and y are both minimizing z's as in definition 1. □

So, the set of even couples (x,y) is made up of *all* the couples at an even distance in the dense case, but in general it is not so. E.g., in spite of even distances, there are no even couples with $x \ne y$ for the distance matrix above, as soon checked (out of the six couples two are odd, four are mixed).
We modify optimality and reliability as follows; we assume $\Delta < N$ to no loss of generality (use theorem 1: if $\delta(x,y) = N$ then the upper bound in (2) is achieved and so (x,y) cannot be even).

Definition 5. Optimal codes for error detection*: once the integer threshold Δ is chosen, construct maximum-size codes as in definition 3, but adding the constraint: if $d(x,y) = \Delta$ then (x,y) is an even couple.*

Reliability criterion 3 for error detection. *Decode to the single codeword x which maximises the transition possibility* $\mathrm{Poss}(z|x)$ *to the output z, and in case of ties declare a detected error: the undetected error possibility is strictly less than $1-\Delta/N$.*

Reliability criterion 4 for error detection. *Once the output string z is received, decode to the single codeword x which minimises the distance $d(x,z)$*

between input and output; in case of ties declare instead a detected error. No undetected error occurs if the input string x was such that $d(x,z) \leq \Delta$.

Proof. Re-take the proof of the criterion for error correction. The equality $d(y,z) \leq d(x,z)$ must be modified to a strict inequality $d(y,z) < d(x,z)$, else a detected error would have been declared. Once more one gets $\Delta \leq \delta(x,y) \leq d(x,z)$; however, if $\Delta = d(x,z)$ and so $\Delta = \delta(x,y) = d(x,z)$, z would achieve $\delta(x,y)$ and (x,y) would be an odd or mixed couple. □

Criteria 3 and 4 do not require that the codes are maximum-size (optimal). In practice, whenever the lower bound (2) holds, the Reliability criterion 4 can be re-stated in a way which is quite familiar to coding-theorists, just use Theorem 1 (requiring that a couple (x,y) is even amounts to requiring that its distance should be an even integer). At the other end of the spectrum, we have ultrametric spaces, where error detection does not offer any advantage, and so can be safely ignored, use again Theorem 1. As for the intermediate and stimulating case of the string distances for twiddles mentioned in Section 2, to construct error-detecting codes one will have to carefully understand the structure of even couples in the corresponding string "geometry".

The gap of error detection having been filled, possibilistic information theory stands out as a full-fledged approach to information theory and coding, able to deal with situations, as is channel noise in DNA word design or the correction of twiddles, where the traditional probabilistic and distance-based approach falls short of the mark.

Acknowledgements. Work partially supported by FIRB-RBLA039M7M_005 LIBI Project on Bioinformatics.

References

1. Bortolussi, L., Sgarro, A.: Possibilistic channels for DNA word design. In: Lawry, J., Miranda, E., Bugarin, A., Li, S., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Integrated Uncertainty Modelling. Advances in Soft Computing, vol. 37, pp. 327–335. Springer, Heidelberg (2006)
2. Bortolussi, L., Sgarro, A.: Noise of DNA word design is not stochastic (submitted for publication, 2010), `www.dmi.units.it/~sgarro/nostochasticDNA.pdf`
3. Bortolussi, L., Dinu, L.P., Sgarro, A.: Twiddle correction and codeword distinguishability (in preparation, 2010), Preliminary version at: `www.dmi.units.it/~sgarro/rankCODES.pdf`
4. Brooks Jr., F.P.: Three great challenges for half-century-old computer science. J. ACM. 50(1), 25–26 (2003)
5. Condon, A., Corn, R.M., Marathe, A.: On combinatorial dna word design. J. Comput. Biol. 8(3), 201–220 (2001)
6. Deza, E., Deza, M.M.: Dictionary of Distances. Elsevier, Amsterdam (2006)
7. Dubois, D., Prade, H.: Fundamentals of Fuzzy Sets. The Handbooks of Fuzzy Sets Series. Kluwer Academic Publishers, Dordrecht (2000)

8. Guiaşu, S.: Comments on: "On possibilistic entropies" by Sgarro, A., Dinu, L.P. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10(6), 655–657 (2002)
9. Körner, J., Orlitsky, A.: Zero-error information theory. IEEE Trans. Inform. Theory 44(6), 2207–2229 (1998)
10. van Lint, J.: Introduction to Coding Theory. Springer, Berlin (1999)
11. Luccio, F., Sgarro, A.: Fuzzy graphs and error-proof keyboards. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge Based Systems, IPMU 2002, Annecy, France, pp. 1503–1508 (2002)
12. Sgarro, A., Bortolussi, L.: Codeword distinguishability in minimum diversity decoding. J. Discrete Math. Sci. Cryptogr. 9(3), 487–502 (2006)
13. Sgarro, A., Dinu, L.P.: Possibilistic entropies and the compression of possibilistic data. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 10(6), 635–653 (2002)
14. Sgarro, A.: Possibilistic information theory: a coding-theoretic approach. Fuzzy Sets Syst. 132(1), 11–32 (2002)
15. Sgarro, A.: An axiomatic derivation of the coding-theoretic possibilistic entropy. Fuzzy Sets Syst. 143, 335–353 (2003)

Coherent Correction for Conditional Probability Assessments with R

A. Brozzi, A. Morelli, and F. Vattari

Abstract. We present an R implementation for a procedure proposed to correct incoherent conditional probability assessments. We obtain a coherent correction $\mathbf{q}$ for the initial assessment $\mathbf{p}$ given on a family $\mathscr{E}$ of conditional events by minimizing a discrepancy measure between $\mathbf{p}$ and the set of all probability distributions over the sample space spanned by $\mathscr{E}$.

Keywords: Coherent correction, Conditional probability assessments, Nonlinear optimization, R.

1 Introduction

In this paper we describe an R implementation for a procedure proposed in [1] and analysed in [3] to correct incoherent probability assessments. We start with an initial assessment $\mathbf{p} = (p_1, \dots, p_n)$ given by an expert on a finite domain of conditional events $\mathscr{E} = \{E_1|H_1, \dots, E_n|H_n\}$ with logical contraints representing their particular configurations. We consider the sample space $\boldsymbol{\Omega} = \{\omega_1, \dots, \omega_k\}$ spanned by $\mathscr{E}$. By minimizing a discrepancy measure between $\mathbf{p}$ and the set of all probability distributions over $\boldsymbol{\Omega}$, we generate a coherent correction for the initial opinion $\mathbf{p}$. In cases where the initial assessment is coherent the correction coincides with $\mathbf{p}$, so the procedure might also be used to verify the coherence of the initial assessment.

In the following subsection we give basic notions about the coherence for conditional probabilities and we analyse the discrepancy measure used to

A. Brozzi
Istituto Europeo di Oncologia (IEO), 20139 Milan, Italy
e-mail: alessandro.brozzi@ifom-ieo-campus.it

A. Morelli and F. Vattari
Dipartimento di Matematica e Informatica, Università degli Studi di Perugia, Italy
e-mail: morelliangelo84@libero.it,francesca.vattari@dmi.unipg.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 49–56.
springerlink.com

correct incoherent assessments. In Section 2 we detail the R code used and in Section 3 we provide some explanatory examples.

1.1 Conditional Probabilities and Discrepancy Measure

A conditional assessment $\mathbf{p}$ on a finite family of conditional events $\mathscr{E}$ is said to be coherent if there exists a conditional probability defined on $\mathscr{E}' = \mathscr{G} \times \mathscr{B}^o$ ($\mathscr{B}^o = \mathscr{B} \setminus \emptyset$) which extends $\mathbf{p}$, where $\mathscr{G}$ is a boolean algebra and $\mathscr{B}$ is an additive set contained in $\mathscr{G}$. We will use the following characterization given by Coletti and Scozzafava in [5]:

Theorem 1. *Let $\mathscr{E} = \{E_1|H_1, \dots, E_n|H_n\}$ be an arbitrary family of conditional events, Ω be the set of atoms ω_j generated by the events $E_1, H_1, \dots, E_n, H_n$ and $\mathscr{G}$ be the algebra spanned by them. For an assessment on $\mathscr{E}$ given by a real function $\mathbf{p}$, the following statements are equivalent:*
(i) $\mathbf{p} = (p_1, \dots, p_n)$ is a coherent conditional probability on $\mathscr{E}$;
(ii) there exists a sequence of compatible systems, indexed with l (layer), with unknowns $\alpha_j^l \geq 0$

$$\begin{cases} \sum\limits_{\omega_j \subseteq E_i H_i} \alpha_j^l = p_i \sum\limits_{\omega_j \subseteq H_i} \alpha_j^l, \quad for\ i = 1, \dots, n\ s.t. \sum\limits_{\omega_j \subseteq H_i} \alpha_j^{l-1} = 0, \quad l \geq 1 \\ \sum\limits_{\omega_j \subseteq H_0^l} \alpha_j^l = 1 \end{cases}$$

with $l = 0, 1, \dots, t \leq n$ where $E_i H_i = E_i \wedge H_i$, $H_0^0 = H_0 = H_1 \vee \ldots \vee H_n$ and H_0^l denotes, for $l \geq 1$, the union of the H_i's such that $\sum_{\omega_j \subseteq H_i} \alpha_j^{l-1} = 0$.

Therefore an assessment $\mathbf{p}$ is incoherent if this sequence of systems has no solution; in [1] the authors present a procedure to correct incoherent assessment using a discrepancy measure between the conditional assessment $\mathbf{p}$ and the set of all probability distributions over the sample space Ω generated by $\mathscr{E}$. By minimizing this discrepancy measure they generate a coherent assessment $\mathbf{q}$ which is the closest to the initial opinion with respect to this measure. In the next steps we describe this discrepancy and its theoretical properties which will be used to correct the incoherence with a procedure written in R.

Let us consider the following sets of probability distributions over Ω:

$\mathscr{A} := \left\{\alpha = [\alpha_1, \dots, \alpha_k], \sum \alpha_j = 1, \alpha_j \geq 0, j = 1, \dots, k\right\}$;
$\mathscr{A}_0 := \{\alpha \in \mathscr{A} | \alpha(H_0) = 1\}$;
$\mathscr{A}_1 := \{\alpha \in \mathscr{A}_0 | \alpha(H_i) > 0, i = 1, \dots, n\}$;
$\mathscr{A}_2 := \{\alpha \in \mathscr{A}_1 | 0 < \alpha(E_i H_i) < \alpha(H_i), i = 1, \dots, n\}$.

For each $\alpha \in \mathscr{A}_1$ we define a coherent assessment $\mathbf{q}_\alpha$ that, for sake of simplicity, we call $\mathbf{q}$:

$$q_i = \frac{\sum\limits_{\omega_j \subseteq E_i H_i} \alpha_j}{\sum\limits_{\omega_j \subseteq H_i} \alpha_j}, \quad \forall i = 1, \ldots, n. \tag{1}$$

Let us now consider the following scoring rule defined for each $\mathbf{p} \in (0,1)^n$:

$$S(\mathbf{p}) := \sum_{i=1}^{n} |E_i H_i| \ln p_i + \sum_{i=1}^{n} |E_i^c H_i| \ln(1 - p_i) \tag{2}$$

where $|\cdot|$ is the indicator function of unconditional events. $S(\mathbf{p})$ is an adaptation to partial conditional probability assessments of the *proper scoring rule* for probability distributions proposed by Lad in [7]. Using this scoring rule, in the aforementioned work [1] the authors introduce a discrepancy between a partial conditional assessment $\mathbf{p}$ over $\mathscr{E}$ and a distribution $\alpha \in \mathscr{A}_2$ as

$$\Delta(\mathbf{p}, \alpha) := E_\alpha(S(\mathbf{q}_\alpha) - S(\mathbf{p})) = \sum_{j=1}^{k} \alpha_j [S_j(\mathbf{q}_\alpha) - S_j(\mathbf{p})]. \tag{3}$$

This definition can be extended by continuity in $\mathscr{A}_0$ as

$$\Delta(\mathbf{p}, \alpha) = \sum_{i | \alpha(H_i) > 0} \alpha(H_i) \left(q_i \ln \frac{q_i}{p_i} + (1 - q_i) \ln \frac{(1 - q_i)}{(1 - p_i)} \right)$$

with the usual convention $0 \ln 0 = 0$. In [3] it is formally proved that $\Delta(\mathbf{p}, \alpha)$ is a non negative function on $\mathscr{A}_0$ and that $\Delta(\mathbf{p}, \alpha) = 0$ if and only if $\mathbf{p} = \mathbf{q}$; moreover $\Delta(\mathbf{p}, \cdot)$ is a convex function on $\mathscr{A}_2$ and it admits a minimum on $\mathscr{A}_0$. This measure has been introduced in [1] to correct incoherent assessments and it has been also developed for the imprecise probabilities [2] and to aggregate expert opinions [4]. We generate a coherent correction for $\mathbf{p}$ solving the following nonlinear optimization problem

$$\min_{\alpha \in \mathscr{A}_0} \Delta(\mathbf{p}, \alpha). \tag{4}$$

Since we seek a solution in $\mathscr{A}_0$ we restrict our attention only to the atoms contained in H^0 considering $\Omega \equiv H^0$. Every q_i is properly defined as in (1) only if $\alpha(H_i) > 0$ so we minimize $\Delta(\mathbf{p}, \alpha)$ in $\mathscr{A}_0$ getting $\underline{\alpha}$ solution of (4) and we generate

$$q_i = \frac{\sum\limits_{\omega_j \subseteq E_i H_i} \underline{\alpha}_j}{\sum\limits_{\omega_j \subseteq H_i} \underline{\alpha}_j}, \ \forall \ i \mid \underline{\alpha}(H_i) > 0 \tag{5}$$

for every i such that $\underline{\alpha}(H_i) > 0$. Hence we start a new layer: we restrict our attention to indexes i such that $\underline{\alpha}(H_i) = 0$ and we solve a new optimization problem of the same kind. The procedure is iterated at most n times[1].

[1] Notice that $S(\mathbf{p})$ is defined only for $\mathbf{p} \in (0,1)^n$ but to correct every $\mathbf{p} \in [0,1]^n$ is reasonable just taking $q_i = 0$ for all i s.t. $p_i = 0$ and $q_i = 1$ for all i s.t. $p_i = 1$.

2 Coherent Correction with R

We wrote in R a main user-level function called `coherentCorrection` to correct conditional probability assessments. The function takes in input the following arguments:

```
> args(coherentCorrection)
```

```
function (Events = Events, condEvents = condEvents, relations =
relations,  p = p)
```

where

`Events` is the vector of all (unconditional) events
`condEvents` is the list of conditional events
`relations` is a vector of logical relations between the events
`p` is the vector of initial probability assessment.

`coherentCorrection` is an interface to a function written in C^{++}[2] which takes the logical constraints (stored in `relations`) and conditional events (stored in `condEvents`) to generate a matrix $M = (m_{ij}) \in \mathscr{M}_{2n \times k}$ such that

$$m_{ij} = \begin{cases} 1, & \omega_j \subseteq A_i \\ 0, & \omega_j \subseteq A_i^c \end{cases}$$

where A_i is the i-*th* (unconditional) event.

In the procedure we use <1> and <0> to indicate Ω and $\emptyset$ respectively and the relations between the events coded as in Table 1.

Table 1 Table of operators

operators	logical connectives
=>	implication
<=>	equivalence
*	conjunction
+	disjunction
-	negation

When the `coherentCorrection` function is called the matrix M is automatically generated and imported into R. The matrix is passed together with the vector **p** inside the objective function `DeltaPalpha`. This function has to be minimized in order to find out the closest coherent probability assessment with respect to the discrepancy measure Δ above introduced. To issue the minimization task we took advantage of the functions contained in the package `Rdonlp2`. This package is an R implementation of a software written by

[2] The function generates the atoms solving a SAT (Boolean satisfiability) problem.

Peter Spellucci for solving nonlinear programming problems. The output of the `coherentCorrection` function is a vector $\mathbf{q}$ which represents the coherent correction of the initial assessment $\mathbf{p}$.

A sketch of the procedure behind the R function is described below. The index l has been intentionally highlighted in the algorithm to give a straightforward view of the coherence layers. We start with the initializations: $l = 0$, n^l = number of initial conditional events, $\mathbf{p}^l$ = initial assessment and M^l = initial matrix of atoms.

```
while $n^l > 0$ do
   $\underline{\alpha}^l = min(\Delta(\mathbf{p}^l, \alpha^l))$
   for $i = 1$ to $n^l$ do
      if $\underline{\alpha}^l(H_i) > 0$ then
         $q_i = \underline{\alpha}^l(E_i H_i)/\underline{\alpha}^l(H_i)$
         delete the $i$-th row of $M^l$
         delete the $(n^l + i)$-th row of $M^l$
   for $j = 1$ to $k^l$ do
      if $M^l[i, j] = 0 \;\; \forall i = \text{nrow}(M^l)/2 + 1, \ldots, \text{nrow}(M^l)$ then
         delete j-th column of $M^l$
   $n^{l+1} = nrow(M^l)/2$
   $\mathbf{p}^{l+1} = (p_i^l)_{i|\underline{\alpha}^l(H_i)=0}$
   $M^{l+1} = M^l$
   $l = l + 1$
```

3 Examples

We illustrate the functionality of `coherentCorrection` using several inputs.

Example 1. We start with the Example 3 reported in [1]. Five basic events A,B,C,D,E are given with the following interpretations: A = "cardiac insufficiency", B ="asthma attack", C ="asthma attack and cardiac lesion", D ="taking drug for asthma does not reduce choking symptoms", E ="taking the drug M for asthma increases tachycardia". Events are characterized by the following logical constraints: $C \subseteq AB$, $DA^cB \equiv \emptyset$, $EA^cB \equiv \emptyset$.

The set $\mathscr{E}$ of conditional events and corresponding initial assessment $\mathbf{p}$ are reported in Table 2.

Table 2 Initial assessment of Example 1

$\mathscr{E}$	A	B	C	$A \vee B$	$D \mid A$	$D \mid A^c$	$D \mid B$	$D \mid B^c$	$D \mid C$	$D \mid C^c$	$E \mid DA$	$E \mid DB$
$\mathbf{p}$	0.57	0.333	0.2	0.6	0.9	0.7	0.45	0.75	0.75	0.6	0.875	0.6

The output of our R function is a vector $\mathbf{q}$ which represents the coherent correction of the initial assessment $\mathbf{p}$; in this example the procedure gives the correction after one layer:

```
> condEvents1 = list("A", "B", "C", "A + B", "D | A", "D | - A",
"D | B", "D | - B", "D | C", "D | - C", "E | D * A", "E | D * B")
> relations1 = c("C => A * B", "D * - A * B <=> <0>",
"E * - A * B <=> <0>")
> Events1 = c("A", "B", "C", "D", "E")
> p1 = c(0.5, 0.333, 0.2, 0.6, 0.9, 0.7, 0.45, 0.75, 0.75, 0.6,
     0.875, 0.6)
> coherentCorrection(Events = Events1, condEvents = condEvents1,
     relations = relations1, p = p1)
Processing layer... 0 Correction q =
 A     B     C    A+B   D|A   D|-A  D|B  D|-B   D|C   D|-C  E|D*A E|D*B
0.477 0.319 0.200 0.595 0.840 0.571 0.502 0.792 0.800 0.674 0.856 0.639
```

Example 2. In the following example two layers have been processed. The data are taken from Example 14, pag. 94 [5]. We start with six events $F_1, F_2, F_3, G_1, G_2, G_3$ under the following constraints $G_1 \equiv G_2 \vee G_3$, $G_2 G_3 \equiv \emptyset$, $G_1 \equiv \Omega$, $F_1 \subseteq G_2$, $F_2 \subseteq G_2$, $F_3 \subseteq G_3$, $F_1 F_2 \equiv \emptyset$, $F_1 \vee F_2 \equiv G_2$. The initial probability assessment $\mathbf{p}$ is given on $\mathscr{E}$ as in Table 3:

Table 3 Initial assessment of Example 2

$\mathscr{E}$	$F_1 \mid G_1$	$F_2 \mid G_2$	$F_3 \mid G_3$
$\mathbf{p}$	0.75	0.25	0.5

Being the initial assessement coherent, our R function returns a correction vector $\mathbf{q}$ coincident with the initial vector $\mathbf{p}$. Such result shows how the procedure can also be used to check the coherence for conditional probability assessment.

```
> p2 = c(3/4, 1/4, 1/2)
> Events2 = c("F1", "G1", "F2", "G2", "F3", "G3")
> condEvents2 = list("F1 | G1", "F2 | G2", "F3 | G3")
> relations2 = c("G1 <=> G2 + G3", "G2 * G3 <=> <0>", "G1 <=> <1>",
"F1 => G2","F2 => G2","F3 => G3","F1 * F2 <=> <0>","F1 + F2 <=>
G2")
> coherentCorrection(Events = Events2, condEvents = condEvents2,
+     relations = relations2, p = p2)

Processing layer... 0 Processing layer... 1 Correction q =
F1|G1 F2|G2 F3|G3
 0.75  0.25  0.50
```

Example 3. In the same framework of the previous example, adding another conditional event $F_4|G_4$ with the constraints $F_4 \subseteq G_3, F_4 F_3 \equiv \emptyset, G_4 \equiv G_3, F_3 \vee F_4 \equiv G_3$ and taking $p_4 = 1/3$, we get an incoherent initial assessment. Our procedure corrects it in two layers.

```
> p3 = c(3/4, 1/4, 1/2, 1/3)
> Events3 = c("F1", "G1", "F2", "G2", "F3", "G3", "F4", "G4")
> condEvents3 = list("F1 | G1", "F2 | G2", "F3 | G3", "F4 | G4")
> relations3 =c("G1 <=> G2 + G3","G2 * G3 <=> <0>","G1 <=> <1>",
"F1 => G2", "F2 => G2", "F3 => G3", "F1 * F2 <=> <0>", "F4 => G4",
"F4 * F3 <=> <0>", "G4 <=> G3", "F1 + F2 <=> G2", "F3 + F4 <=>
G3")
> coherentCorrection(Events = Events3, condEvents = condEvents3,
+     relations = relations3, p = p3)

Processing layer... 0 Processing layer... 1 Correction q =
F1|G1 F2|G2 F3|G3 F4|G4
0.750 0.250 0.586 0.414
```

4 Conclusions

In several research fields it might happen that probability assessments are given only on events of particular interest and not generally over all the possible states of the world. Moreover in many cases information gathered on phenomenons being evaluated might be partial, conditional or even incoherent. In this scenario becomes important to draw the best representative coherent assessement from this partial information. We have presented a new procedure using R to issue this task; we have chosen R because it has become the most widely used enviroment in many fields of research (e.g. economics, sociology, life sciences) where such kind of problems usually rise. By means of the flexibility of R, we provided an integrative framework of different computational solutions usually accomplished by different approches (satisfiability problem, nonlinear optimization under constraints). The usage of the proposed procedure might also be extended to check the coherence of the initial assessment.

Availability: The function is implemented in an R package. It is currently available for download at
`http://sites.google.com/site/alessandrobrozzi/`

Acknowledgements. We thank Andrea Capotorti and Giuliana Regoli for their supervision and critical comments on the manuscript; Davide Petturiti for his technical help; Peter Spellucci and Ryuichi Tamura for making their softwares (donlp2 and Rdonlp2) available as well as the contributors to R (`http://www.R-project.org` [8]).

References

1. Capotorti, A., Regoli, G.: Coherent correction of inconsistent conditional probability assessments. In: Proceedings of the 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2008, Malaga, Spain (2008)

2. Capotorti, A., Regoli, G., Vattari, F.: On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities. In: Augustin, T., Coolen, F.P.A., Moral, S., Troffaes, M.C.M. (eds.) Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK (2009)
3. Capotorti, A., Regoli G., Vattari F.: Correction of incoherent conditional probability assessments. Internat. J. Approx. Reason. (in press, 2010)
4. Capotorti, A., Regoli, G., Vattari, F.: Merging different probabilistic information sources through a new discrepancy measure. In: Kroupa, T., Vejnarova, J. (eds.) Proceedings of the 8th Workshop on Uncertainty Processing, WUPES 2009, Liblice, Czech Republic (2009)
5. Coletti, G., Scozzafava, R.: Probabilistic Logic in a Coherent Setting. Series Trends in Logic, vol. 15. Kluwer Academic Publishers, Dordrecht (2002)
6. De Finetti, B.: Sull'impostazione assiomatica del calcolo delle probabiliá. Ann. Triestini. Sez. 2(4) 3(19) (1949), 29–81 (1950); Engl. transl. of Probability, Induction, Statistics. The art of guessing. ch. 5. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York (1972)
7. Lad, F.: Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction. John Wiley, New York (1996)
8. R Development Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, http://www.R-project.org

Inferential Rules for Weak Graphoid

Giuseppe Busanello and Barbara Vantaggi

Abstract. We deal with the problem of computing efficiently the closure of a set of independencies, compatible with a coherent conditional probability, under cs–independence. For this aim we provide two inferential rules, which allow to build a basis for the closure.

Keywords: Conditional independence, Closure, Graphoid, Inference rules.

1 Introduction

In probability and multivariate statistics graphical models have a fundamental role, in particular for representing independence models, whose properties depend on the independence notion taken into account. Actually, independence models, arising from the classic independence, are closed in general under semi–graphoid properties and if the probability is strictly positive under graphoid properties [6].

However, it is well known that the classical definition of independence leads to some counter-intuitive situations when some events with probability 0 or 1 are involved and when logical constraints among the variables are present. Then, different definitions of independence have been introduced to encompass such situations as, for example, cs-independence [2], given within the framework of coherent conditional probability [3], which allows also to deal with partial assessments with conditioning events of zero probability. This represents a very crucial feature not only from a merely theoretical point of view, in fact they can be met in many problems as, e.g., in medical diagnosis, statistical mechanics, physics.

The graphoid properties under cs-independence have been tested in [9], in particular, symmetry can fail. Actually, the relevant independence models are

Giuseppe Busanello and Barbara Vantaggi
Dip. Metodi e Modelli Matematici, Università "La Sapienza", Roma, Italy
e-mail: busanello,vantaggi@dmmm.uniroma1.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 57–64.
springerlink.com

closed under weak graphoid properties (decomposition and its reverse, weak union, contraction and its reverse, intersection).

The aim of this paper is to consider a set J of conditional independence statements, compatible with a coherent conditional probability assessment, and to build efficiently the closure under weak graphoid properties.

This topic has already been faced successfully for semi-graphoids by Studený [7, 8] and for graphoids in [1], now the aim is to extend these results to the more general structure of weak graphoid.

Since the computation of the closure is infeasible due to its size, which is exponentially larger than the size of J, our aim is to find a suitable subset of the closure, which represents the same independence structure. This set should be as small as possible and from it all the relations in the closure should be easily deducible, so it can be considered a basis for the closure.

In order to introduce such a set we define a suitable notion of inclusion between independence statements and we provide two inferential rules, which generalize classic weak graphoid properties. The obtained reduced set is relevant for the model selection complexity on the basis of data for building, for example, the relevant directed acyclic graph. This is one of the reasons of our effort.

In this paper we follow a strategy similar to that introduced in [1, 7, 8] and this shows that it can be carried out also for other independence models (see e.g. [4, 5, 10]) satisfying some other properties among that of graphoids.

2 Weak Graphoid

Let $\tilde{S} = \{Y_1,...,Y_n\}$ be a finite non-empty set of variables and $S = \{1,...,n\}$ the related set of indices. Given a subset $I \subseteq S$, Y_I is the vector $(Y_i : i \in I)$ and $Y_A \perp\!\!\!\perp Y_B | Y_C$ is an independence statement, which is simply denoted as an ordered triple (A,B,C). A conditional independence model, related to a coherent (conditional) probability P, is a subset of the set $S^{(3)}$ formed by all triples (A,B,C) of disjoint subsets of S, with A and B non-empty.

In this paper, as already claimed in the Introduction, we deal with models arising under cs–independence [2]. These models have a weak graphoid structure [9], that can be seen as a couple $(S,\mathscr{I})$, where $\mathscr{I}$ is a ternary relation on the set S, which satisfies the following properties:

De (Decomposition) $(A,B\cup C,D) \in \mathscr{I} \Rightarrow (A,B,C) \in \mathscr{I}$;
DeR (Reverse Decomposition) $(A\cup B,C,D) \in \mathscr{I} \Rightarrow (A,C,D) \in \mathscr{I}$;
WU (Weak Union) $(A,B\cup C,D) \in \mathscr{I} \Rightarrow (A,B,C\cup D) \in \mathscr{I}$;
Co (Contraction) $(A,B,C\cup D),(A,C,D) \in \mathscr{I} \Rightarrow (A,B\cup C,D) \in \mathscr{I}$;
CoR (Reverse Contraction) $(A,B,C\cup D),(C,B,D) \in \mathscr{I} \Rightarrow (A\cup C,B,D) \in \mathscr{I}$;
In (Intersection) $(A,B,C\cup D),(A,C,B\cup D) \in \mathscr{I} \Rightarrow (A,B\cup C,D) \in \mathscr{I}$;

where A, B, C and D are pairwise disjoint and nonempty subsets of S.

The symbol $\theta \vdash_R \theta'$ denotes that θ' is obtained from θ by applying once the property R, which can be De, DeR, WU. Similarly, $\theta_1, \theta_2 \vdash_R \theta$ means that θ is obtained from θ_1, θ_2 by applying once R, where R can be Co, CoR or In.

Now, given a set $J \subseteq S^{(3)}$ of triples, compatible with P, we are interested to establish whether a triple θ can be derived from J (in symbol $J \vdash^* \theta$), that is θ can be obtained by applying a finite number of times the weak graphoid properties by starting from J. A strictly related problem is to compute the closure of a set J, defined as

$$\bar{J} = \{\theta \in S^{(3)} : J \vdash^* \theta\}.$$

The computation of the closure is infeasible since its size is exponentially larger than the size of J. Then, in the following we introduce a suitable subset of $\bar{J}$ having the same information as $\bar{J}$.

3 Generalized Inference Rules

In order to compute efficiently the closure of a set of conditional independence statements under weak graphoid properties, we need to introduce a notion of weak inclusion (briefly w–inclusion): a triple θ_1 is said to be *w-included* in θ_2 (in symbol $\theta_1 \sqsubseteq_w \theta_2$), if θ_1 can be obtained from θ_2 by applying a finite number of times De, DeR and WU.

We provide a characterization of w–inclusion.

Proposition 1. *Let $\theta_1 = (A_1,B_1,C_1)$ and $\theta_2 = (A_2,B_2,C_2)$, then $\theta_1 \sqsubseteq_w \theta_2$ if and only if the following conditions hold*

1. $C_2 \subseteq C_1 \subseteq (B_2 \cup C_2)$;
2. $A_1 \subseteq A_2$ *and* $B_1 \subseteq B_2$.

Proof. If *1.* and *2.* hold, the triple θ_1 is obtained from θ_2 by the following steps: take $B_2' = B_2 \setminus C_1$ and $C_1 = C_2 \cup (C_1 \cap B_2)$, so $B_1 \subseteq B_2'$ and $(A_2,B_2,C_2) \vdash_{WU} (A_2,B_2',C_1) \vdash_{De} (A_2,B_1,C_1)$ and by $A_1 \subseteq A_2$, one has $(A_2,B_1,C_1) \vdash_{DeR} (A_1,B_1,C_1)$.

Now, we prove the reverse implication. If $\theta_1 \sqsubseteq_w \theta_2$, then there exist θ_i' and $R_i \in \{\text{De, WU, DeR}\}$, $i = 1,\ldots,n$, such that $\theta_1' = \theta_2$, $\theta_{n+1}' = \theta_1$, $\theta_i' \vdash_{R_i} \theta_{i+1}'$. By induction on i we show that $\theta_i' \sqsubseteq_w \theta_2$. For $i = 1$, it is trivial, if it is true for i (i.e. $\theta_i' \sqsubseteq_w \theta_2$) we have the following three cases

1. $\theta_i' \vdash_{De} \theta_{i+1}'$ with $A_{i+1}' = A_i' \subseteq A_2$, $B_{i+1}' \subseteq B_i' \subseteq B_2$, $C_{i+1}' = C_i'$;
2. $\theta_i' \vdash_{WU} \theta_{i+1}'$ with $A_{i+1}' = A_i' \subseteq A_2$, $B_{i+1}' \subseteq B_i' \subseteq B_2$, $C_{i+1}' = C_i' \cup (B_i' \setminus B_{i+1}')$, $C_2 \subseteq C_i' \subseteq C_{i+1}'$. Furthermore , $B_i' \setminus B_{i+1}' \subseteq B_2$ and $C_i' \subseteq (B_2 \cup C_2)$ imply $C_{i+1}' \subseteq (B_2 \cup C_2)$;
3. $\theta_i' \vdash_{DeR} \theta_{i+1}'$ with $A_{i+1}' \subseteq A_i' \subseteq A_2$, $B_{i+1}' = B_i' \subseteq B_2$, $C_{i+1}' = C_i'$.

□

This definition of w–inclusion can be extended to sets of triples: J is a *covering* of H with respect to w–inclusion (in symbol $H \sqsubseteq_w J$) if and only if for any triple $\theta \in H$ there exists a triple $\theta' \in J$ such that $\theta \sqsubseteq_w \theta'$.

Let

$$J_{\big/\sqsubseteq_w} = \{\tau \in J : \nexists \bar{\tau} \in J \text{ with } \bar{\tau} \neq \tau \text{ such that } \tau \sqsubseteq_w \bar{\tau}\} \qquad (1)$$

a set J is said *"maximal"* if $J = J_{\big/\sqsubseteq_w}$.

We show that the w–inclusion on maximal sets is a partial order:

Proposition 2. *The w–inclusion on maximal sets is reflexive, transitive and anti–symmetric.*

Proof. Reflexivity is trivial. To prove transitivity firstly refer to triples: let $\theta_1 \sqsubseteq_w \theta_2$ and $\theta_2 \sqsubseteq_w \theta_3$, then, by Proposition 1, $A_1 \subseteq A_2 \subseteq A_3$, $B_1 \subseteq B_2 \subseteq B_3$ and $C_3 \subseteq C_2 \subseteq C_1$, moreover, $C_2 \subseteq (B_3 \cup C_3)$ and $C_1 \subseteq (B_2 \cup C_2)$, so $C_1 \subseteq (B_3 \cup C_3)$.

Suppose $H \sqsubseteq_w K$ and $K \sqsubseteq_w J$, with H, K, J maximal sets of $S^{(3)}$, then, for any $\theta \in H$ there exists $\theta' \in K$ with $\theta \sqsubseteq_w \theta'$. For $\theta' \in K$, since $K \sqsubseteq_w J$, there exists $\theta'' \in J$ with $\theta' \sqsubseteq_w \theta''$. From transitivity on triples $\theta \sqsubseteq_w \theta''$.

To prove anti–symmetry, we refer again firstly to triples. If $\theta_1 \sqsubseteq_w \theta_2$ and $\theta_2 \sqsubseteq_w \theta_1$, one has again, by Proposition 1, $A_1 \subseteq A_2 \subseteq A_1$, $B_1 \subseteq B_2 \subseteq B_1$, $C_2 \subseteq C_1 \subseteq (B_2 \cup C_2)$ and $C_1 \subseteq C_2 \subseteq (B_1 \cup C_1)$. Therefore $A_1 = A_2$, $B_1 = B_2$ and $C_1 = C_2$.

Let $H \sqsubseteq_w J$ and $J \sqsubseteq_w H$. From $H \sqsubseteq_w J$, for any $\theta \in H$, there exists $\tau \in J$ with $\theta \sqsubseteq_w \tau$. Analogously, for any τ since $J \sqsubseteq_w H$ there exists $\sigma \in H$ with $\tau \sqsubseteq_w \sigma$. By transitivity $\theta \sqsubseteq_w \sigma$, but, since H is a maximal set, $\theta = \sigma$, so $\tau \sqsubseteq_w \theta$ and $\theta \sqsubseteq_w \tau$ imply that $\theta = \tau$, by anti–symmetry between triples. □

On the other hand it is easy to show that anti–symmetry on general sets (not maximal) can fail.

3.1 Closure through a Generalization of Co, CoR and In

In order to provide general inferential rules, we prove a sort of monotonicity property for the binary operations Co, CoR and In.

Proposition 3. *Let θ_1, θ_2, θ_3, θ_4 be triples such that $\theta_1 \sqsubseteq_w \theta_3$, $\theta_2 \sqsubseteq_w \theta_4$.*

1. *If $\theta_1, \theta_2 \vdash_{Co} \theta$ and $\theta_3, \theta_4 \vdash_{Co} \theta'$, then $\theta \sqsubseteq_w \theta'$.*
2. *If $\theta_1, \theta_2 \vdash_{CoR} \theta$ and $\theta_3, \theta_4 \vdash_{CoR} \theta'$, then $\theta \sqsubseteq_w \theta'$.*
3. *If $\theta_1, \theta_2 \vdash_{In} \theta$ and $\theta_3, \theta_4 \vdash_{In} \theta'$, then $\theta \sqsubseteq_w \theta'$.*

Proof. From $\theta_1 \sqsubseteq_w \theta_3$ and $\theta_2 \sqsubseteq_w \theta_4$, then $A_1 \subseteq A_3$, $B_1 \subseteq B_3$, $C_3 \subseteq C_1 \subseteq (B_3 \cup C_3)$, $A_2 \subseteq A_4$, $B_2 \subseteq B_4$, $C_4 \subseteq C_2 \subseteq (B_4 \cup C_4)$.

If there exists (as in *1.*) a triple $\theta = (A, B, C)$ such that $\theta_1, \theta_2 \vdash_{Co} \theta$, $A = A_1 = A_2$, $C_1 = (B_2 \cup C_2)$, $(B_1 \cap B_2) = \emptyset$, $B = (B_1 \cup B_2)$ and $C = C_2$.

Analogously, if there is $\theta' = (A', B', C')$ such that $\theta_3, \theta_4 \vdash_{Co} \theta'$, then $A' = A_3 = A_4$, $C_3 = (B_4 \cup C_4)$, $(B_3 \cap B_4) = \emptyset$, $B' = (B_3 \cup B_4)$ and $C' = C_4$.

Then, $\theta \sqsubseteq_w \theta'$ being $A = A_1 \subseteq A_3 = A'$, $B = B_1 \cup B_2 \subseteq B_3 \cup B_4 = B'$, $C' = C_4 \subseteq C_2 = C = C_2 \subseteq (B_4 \cup C_4) \subseteq (B_4 \cup C_4) \cup B_3 = (B' \cup C')$.

Since $\theta_1 \sqsubseteq_w \theta_3$ and $\theta_2 \sqsubseteq_w \theta_4$, then $A_1 \subseteq A_3$, $B_1 \subseteq B_3$, $C_3 \subseteq C_1 \subseteq (B_3 \cup C_3)$, $A_2 \subseteq A_4$, $B_2 \subseteq B_4$, $C_4 \subseteq C_2 \subseteq (B_4 \cup C_4)$.

The proof of points *2.* and *3.* goes along the same lines. □

Now, our target is to find an efficient method to compute a reduced (with respect to w–inclusion) set J^* included in $\bar{J}$ and having the same information of $\bar{J}$; this means that, for any $\theta \in \bar{J}$, there is $\theta' \in J^*$ with $\theta \sqsubseteq_w \theta'$.

Firstly, we characterize the binary weak graphoid properties.

Proposition 4. *Let $\theta_1 = (A_1,B_1,C_1)$, $\theta_2 = (A_2,B_2,C_2)$ be in $S^{(3)}$, then*

1. $W_C(\theta_1,\theta_2) = \{\tau : \theta_1', \theta_2' \vdash_{Co} \tau,$ *with* $\theta_1' \sqsubseteq_w \theta_1, \theta_2' \sqsubseteq_w \theta_2\}$ *is non-empty if and only if all the following conditions hold:*

a. $(A_1 \cap A_2) \neq \emptyset$;
b. $C_1 \subseteq (B_2 \cup C_2)$ *and* $C_2 \subseteq (B_1 \cup C_1)$;
c. $(B_1 \setminus C_2) \neq \emptyset$, $B_2 \cap (B_1 \cup C_1) \neq \emptyset$ *and* $|(B_1 \setminus C_2) \cup (B_2 \cap (B_1 \cup C_1))| \geq 2$.

Moreover, if $W_C(\theta_1,\theta_2)$ *is non-empty, then*

$$gc(\theta_1,\theta_2) = (A_1 \cap A_2, (B_1 \setminus C_2) \cup (B_2 \cap (B_1 \cup C_1)), C_2)$$

is in $W_C(\theta_1,\theta_2)$ *and* $\theta \sqsubseteq_w gc(\theta_1,\theta_2)$ *for any* $\theta \in W_C(\theta_1,\theta_2)$.

2. $W_{Cr}(\theta_1,\theta_2) = \{\tau : \theta_1', \theta_2' \vdash_{CoR} \tau,$ *with* $\theta_1' \sqsubseteq_w \theta_1, \theta_2' \sqsubseteq_w \theta_2\}$ *is non-empty if and only if all the following conditions hold:*

a. $(B_1 \cap B_2) \neq \emptyset$;
b. $C_1 \subseteq (A_2 \cup B_2 \cup C_2)$ *and* $C_2 \subseteq (B_1 \cup C_1)$;
c. $(A_1 \setminus C_2) \neq \emptyset$, $A_2 \cap (B_1 \cup C_1) \neq \emptyset$ *and* $|(A_1 \setminus C_2) \cup (A_2 \cap (B_1 \cup C_1))| \geq 2$.

Moreover, if $W_{Cr}(\theta_1,\theta_2)$ *is non-empty, then*

$$gcr(\theta_1,\theta_2) = ((A_1 \setminus C_2) \cup (A_2 \cap (B_1 \cup C_1)), B_1 \cap B_2, C_2 \cup (B_2 \cap C_1))$$

is in $W_{Cr}(\theta_1,\theta_2)$ *and* $\theta \sqsubseteq_w gcr(\theta_1,\theta_2)$ *for any* $\theta \in W_{Cr}(\theta_1,\theta_2)$.

3. $W_I(\theta_1,\theta_2) = \{\tau : \theta_1', \theta_2' \vdash_{In} \tau,$ *with* $\theta_1' \sqsubseteq_w \theta_1, \theta_2' \sqsubseteq_w \theta_2\}$ *is non-empty if and only if all the following conditions hold:*

a. $A_1 \cap A_2 \neq \emptyset$;
b. $C_1 \subseteq (B_2 \cup C_2)$ *and* $C_2 \subseteq (B_1 \cup C_1)$;
c. $B_1 \cap (B_2 \cup C_2) \neq \emptyset$, $B_2 \cap (B_1 \cup C_1) \neq \emptyset$ *and* $|(B_1 \cap (B_2 \cup C_2)) \cup (B_2 \cap (B_1 \cup C_1))| \geq 2$.

Moreover, if $W_I(\theta_1,\theta_2)$ *is non-empty, then*

$$gi(\theta_1,\theta_2) = (A_1 \cap A_2, (B_1 \cap B_2) \cup (B_1 \cap C_2) \cup (B_2 \cap C_1), (C_1 \cap C_2))$$

is in $W_I(\theta_1,\theta_2)$ *and* $\theta \sqsubseteq_w gi(\theta_1,\theta_2)$ *for any* $\theta \in W_I(\theta_1,\theta_2)$.

Proof. Concerning point *1.*, if $W_C(\theta_1,\theta_2)$ is non-empty, then for any $\tau = (A,B,C)$ in $W_C(\theta_1,\theta_2)$ there are $\theta_1' = (A_1',B_1',C_1') \sqsubseteq_w \theta_1$ and $\theta_2' = (A_2',B_2',C_2') \sqsubseteq_w \theta_2$ such that $\theta_1', \theta_2' \vdash_{Co} \tau$. Then, these conditions hold:

- $A'_1 \subseteq A_1$, $A'_2 \subseteq A_2$, $A'_1 = A'_2$, then $A_1 \cap A_2 \neq \emptyset$.
- $C'_1 = B'_2 \cup C'_2$, $C_1 \subseteq C'_1 \subseteq (B_1 \cup C_1)$, $C_2 \subseteq C'_2 \subseteq (B_2 \cup C_2)$ and $\emptyset \neq B'_2 \subseteq B_2$. This implies $C_1 \subseteq C'_1 = B'_2 \cup C'_2 \subseteq (B_2 \cup C_2)$, so $C_1 \subseteq (B_2 \cup C_2)$.
  From $C'_2 \subseteq C'_1$ it follows $C_2 \subseteq (B_1 \cup C_1)$.
- $B'_2 \subseteq C'_1 \subseteq (B_1 \cup C_1)$, $B'_2 \subseteq B_2$, so $B'_2 \subseteq B_2 \cap (B_1 \cup C_1)$ and then $B_2 \cap (B_1 \cup C_1) \neq \emptyset$.
- $B'_1 \cap C'_1 = \emptyset$, $C'_1 = B'_2 \cup C'_2$, $B'_1 \cap C'_2 = \emptyset$, $\emptyset \neq B'_1 \subseteq B_1$ and $C_2 \subseteq C'_2$, then it follows $B'_1 \subseteq B_1 \setminus C_2$ and hence $B_1 \setminus C_2 \neq \emptyset$.
- Moreover, from $B'_1 \cap B'_2 = \emptyset$, $B'_1 \subseteq B_1 \setminus C_2$ and $B'_2 \subseteq (B_1 \cup C_1)$ it follows $|(B_1 \setminus C_2) \cup (B_2 \cap (B_1 \cup C_1))| \geq 2$. In fact, $B'_1 \neq \emptyset$ and $B'_2 \neq \emptyset$ so $(B_1 \setminus C_2) \cup (B_2 \cap (B_1 \cup C_1))$ contains at least two elements (otherwise there are no two disjoint subsets).

Suppose that the conditions *a.–c.* hold, it is possible to find two disjoint nonempty sets B^1 and B^2 such that $B^1 \subseteq B_1 \setminus C_2$, $B^2 \subseteq B_2 \cap (B_1 \cup C_1)$ and $B^1 \cup B^2 = (B_1 \setminus C_2) \cup (B_2 \cap (B_1 \cup C_1))$. Let $C^2 = C_2$, the triples $\theta_a = (A_1 \cap A_2, B^1, B^2 \cup C^2)$ and $\theta_b = (A_1 \cap A_2, B^2, C^2)$ are such $\theta_a \sqsubseteq_w \theta_1$, $\theta_b \sqsubseteq_w \theta_2$ and $\theta_a, \theta_b \vdash_{Co} gc(\theta_1, \theta_2) = (A_{gc}, B_{gc}, C_{gc})$. This implies that $W_C(\theta_1, \theta_2)$ is non-empty and $gc(\theta_1, \theta_2) \in W_C(\theta_1, \theta_2)$.

Now, it is simple to show $\tau \sqsubseteq_w gc(\theta_1, \theta_2)$: in fact it is straightforward to show that $A \subseteq A_{gc}$ and $B \subseteq B_{gc}$. Since $C_2 \subseteq C'_2 = C$, then $C_{gc} = C_2 \subseteq C$. On the other hand, since $C'_2 \subseteq C'_1 \subseteq (B_1 \cup C_1)$ and $C'_2 \subseteq (B_2 \cup C_2)$ then $C \subseteq ((B_1 \cup C_1) \cap (B_2 \cup C_2))$ which is a subset of $B_{gc} \cup C_{gc}$.

The proofs of points *2.* and *3.* go along the same lines. □

Let $GC(\theta_1, \theta_2)$ be the set of the possible (i.e. belonging to $S^{(3)}$) triples among $gc(\theta_1, \theta_2)$ and $gcr(\theta_1, \theta_2)$. Note that generally $GC(\theta_1, \theta_2)$ is different from $GC(\theta_2, \theta_1)$.

We introduce two new inference rules:

Co* (Generalized Contraction) from θ_1, θ_2 deduce any triple $\tau \in GC(\theta_1, \theta_2)$;
In* (Generalized Intersection) from θ_1, θ_2 deduce the triple $\tau = gi(\theta_1, \theta_2)$;

which, as explained above, generalize the three classical inference rules.

Given a subset J of $S^{(3)}$, let

$$J^* = \{\tau : J \vdash^*_G \tau\}$$

be the set closed with respect to generalized contraction Co* and generalized intersection In*, where $J \vdash^*_G \tau$ means that τ is obtained by applying a finite number of times the rules Co* and In*.

In order to show the relationship between the two sets J^* and $\bar{J}$, we show that a triple can be deduced through Co* or In* if and only if it can be deduced by means of weak graphoid properties.

Proposition 4 implies the following result:

Proposition 5. *Let J be a subset of $S^{(3)}$, denote by J^* and $\bar{J}$ the closure, respectively, with respect to Co*–In* and the weak graphoid properties. Then $J^* \subseteq \bar{J}$.*

Now, we show that any triple obtained through weak graphoid is w–included in a triple deduced from Co* and In*.

Proposition 6. *Let J be a subset of $S^{(3)}$, denote by J^* and $\bar{J}$ the closure, respectively, with respect to Co*–In* and the weak graphoid properties. Then $\bar{J} \sqsubseteq_w J^*$.*

Proof. The proof is done by induction: starting from $J_0 = J$, let J_i be the union of J_{i-1} and the triples obtained by applying some graphoid properties to J_{i-1}, and put $\bar{J} = \bigcup_{i=0}^{\infty} J_i$.

Since J is finite this iterative process ends when $J_k = J_{k+1}$, $k \in \mathbb{N}$ and $J_k = \bar{J}$. We need to show $J_i \sqsubseteq_w J^*$. For $i = 0$ it is trivial, suppose $J_i \sqsubseteq_w J^*$ and let $\tau \in J_{i+1} \setminus J_i$.
If τ is obtained by means of De, DeR, WU from $\theta \in J_i$, then $\tau \sqsubseteq_w \theta$ and, since $\theta \in J_i$, by hypothesis $\exists \bar{\theta} \in J^*$ such that $\theta \sqsubseteq_w \bar{\theta}$, so by transitivity $\tau \sqsubseteq_w \bar{\theta}$. If $\theta_1, \theta_2 \vdash_{Co} \tau$ with $\theta_1, \theta_2 \in J_i$, then there exist $\bar{\theta}_1, \bar{\theta}_2 \in J^*$ such that $\theta_1 \sqsubseteq_w \bar{\theta}_1$ and $\theta_2 \sqsubseteq_w \bar{\theta}_2$, $\tau \in W_C(\bar{\theta}_1, \bar{\theta}_2)$ and, from Proposition 4, $\tau \sqsubseteq_w gc(\bar{\theta}_1, \bar{\theta}_2) \in J^*$.

The proof of the cases $\theta_1, \theta_2 \vdash_{In} \tau$ and $\theta_1, \theta_2 \vdash_{CoR} \tau$ goes along the same lines of the previous one by Proposition 4. □

Then, J^* is a subset of $\bar{J}$, and it has the same information of $\bar{J}$. Actually, J^* contains some "redundant" triples, which are w–included in other ones.

In order to eliminate redundant triples, for any J consider the set $J_{\big/\sqsubseteq_w}$ defined in equation (1). The set $\bar{J}_{\big/\sqsubseteq_w}$ has the same information of $\bar{J}$ as the following result shows:

Lemma 1. *Let $J \subseteq S^{(3)}$. Then, $J \sqsubseteq_w J_{\big/\sqsubseteq_w}$.*

Proof. Let $\theta \in J$, if $\nexists \bar{\theta} \in J$ such that $\theta \sqsubseteq_w \bar{\theta}$, $\theta \neq \bar{\theta}$, then $\theta \in J_{\big/\sqsubseteq_w}$. Otherwise, i.e. $\theta \in J \setminus J_{\big/\sqsubseteq_w}$, since J is finite, any chain $\theta_1 \sqsubseteq_w \theta_2 \sqsubseteq_w \cdots \sqsubseteq_w \theta_n \sqsubseteq_w \ldots$, with $\theta_i \in J$ and $i \geq 1$, must have a maximal element θ_n, which necessarily belongs to $J_{\big/\sqsubseteq_w}$. □

Then, given a set J, we compute J^* and then we cut redundant triples by taking its *"maximal"* triples, i.e. $J^*_{\big/\sqsubseteq_w}$. We call the set $J^*_{\big/\sqsubseteq_w}$ "fast closure", with respect to w–inclusion, and we denote it, for simplicity, with J_*.

The proof of the following relationships is trivial.

Proposition 7. *Given a subset J of $S^{(3)}$, then $J_* \subseteq \bar{J}$ and $\bar{J} \sqsubseteq_w J_*$.*

Now, it is interesting to show that $\bar{J}_{\big/\sqsubseteq_w}$ and J_* coincide.

Proposition 8. *Given a subset J of $S^{(3)}$, then $J_* = \bar{J}_{\big/\sqsubseteq_w}$.*

Proof. By Proposition 7 it follows that $\bar{J}\big/_{\sqsubseteq_w} \subseteq \bar{J} \sqsubseteq_w J_*$; $J_* = J^*\big/_{\sqsubseteq_w} \subseteq J^* \subseteq \bar{J} \sqsubseteq_w \bar{J}\big/_{\sqsubseteq_w}$. Now, being both J_* and $\bar{J}\big/_{\sqsubseteq_w}$ maximal sets, by Proposition 2, it follows that $J_* = \bar{J}\big/_{\sqsubseteq_w}$. □

The set J_* is computed by eliminating redundant triples at the end; while to improve the computational performance by saving space and time we could eliminate redundant triples at any step (by Proposition 3).

The set J_* allows to test whether a given triple is implied by J. Actually a linear search has to be performed by looking for whether this triple is w–include in some triples of J_*.

References

1. Baioletti, M., Busanello, G., Vantaggi, B.: Conditional independence structure and its closure: Inferential rules and algorithms. Internat. J. Approx. Reason. 50, 1097–1114 (2009)
2. Coletti, G., Scozzafava, R.: Zero probabilities in stochastical independence. In: Bouchon-Meunier, B., Yager, R.R., Zadeh, L.A. (eds.) Information, Uncertainty, Fusion, pp. 185–196. Kluwer Academic Publishers, Dordrecht (2000)
3. Coletti, G., Scozzafava, R.: Probabilistic logic in a coherent setting. Trends in logic, vol. 15. Kluwer Academic Publishers, Dordrecht (2002)
4. Coletti, G., Vantaggi, B.: Possibility theory: conditional independence. Fuzzy Sets Syst. 157, 1491–1513 (2006)
5. Cozman, F.G., Seidenfeld, T.: Independence for full conditional measures, graphoids and Bayesian networks. Bol. BT/PMR/0711 Escola Pol. da Universidade de Sao Paulo, Sao Paulo, Brazil (2007)
6. Dawid, A.P.: Conditional independence in statistical theory. J. Roy. Stat. Soc. Ser. B 41, 15–31 (1979)
7. Studený, M.: Semigraphoids and structures of probabilistic conditional independence. Ann. Math. Artif. Intell. 21, 71–98 (1997)
8. Studený, M.: Complexity of structural models. In: Proceedings of Prague Stochastics 1998, pp. 521–528. Union of Czech Mathematicians and Physicists, Prague (1998)
9. Vantaggi, B.: Conditional independence in a coherent setting. Ann. Math. Artif. Intell. 32, 287–313 (2001)
10. Vantaggi, B.: Conditional independence structures and graphical models. Int. J. Uncertain. Fuzziness Knowledge-Based Systems 11(5), 545–571 (2003)

Fast Factorization of Probability Trees and Its Application to Recursive Trees Learning

Andrés Cano, Manuel Gómez-Olmedo,
Cora B. Pérez-Ariza, and Antonio Salmerón

Abstract. We present a fast potential decomposition algorithm that seeks for proportionality in a probability tree. We give a measure that determines the accuracy of a decomposition in case that exact factorization is not possible. This measure can be used to decide the variable with respect to which a tree should be factorized in order to obtain the most accurate decomposed model.

Keywords: Fast factorization, Probability trees, Recursive probability trees.

1 Introduction

The outperformance of trees over other structures in the field of Bayesian networks inference has been analyzed [1, 3]. Trees can be decomposed in order to improve the overall inference process efficiency, by locating the parts of the tree where a concrete operation must be performed. Many algorithms take this into account, being able to work with lists of potentials, as Lazy propagation [6] or Lazy-penniless [4]. Recursive Probability Trees (RPTs) [2] are a generalization of probability trees, and are able to store any possible decomposition of a potential.

There have been previous works on tree-based potential decomposition [7, 8], but the limitations of probability trees make those attempts quite time demanding during the inference process. In this work we present a potential

Andrés Cano, Manuel Gómez-Olmedo, and Cora B. Pérez-Ariza
Dept. of Computer Science and Artificial Intelligence, University of Granada,
18071 Granada, Spain
e-mail: `acu,mgomez,cora@decsai.ugr.es`

Antonio Salmerón
Dept. of Statistics and Applied Mathematics, University of Almería,
04120 Almería, Spain
e-mail: `antonio.salmeron@ual.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 65–72.
springerlink.com

decomposition algorithm that quickly divides a probability tree into (approximately) proportional factors that can be stored in a RPT.

2 Recursive Probability Trees

A Recursive Probability Tree [2] is a directed tree with two different kinds of inner nodes: Split and List nodes, and two types of leaf nodes: Value and Potential nodes. A Split node represents a discrete variable. A List node represents a multiplicative factorization by listing all the factors in which a potential is decomposed. It has as many outgoing arcs as factors in the decomposition. A Value node represents a non-negative real number, and finally a Potential node stores a full potential internally represented in whatever representation.

With this structure it is possible to represent context-specific independencies within a probability distribution as well as factorizations (involving the whole potential or parts of it). Sometimes potentials present proportionality relations between several parts of the tree. In the simplest case, a part of the tree can be derived from another just by multiplying by a certain factor. This is the case of a probability tree encoding a joint distribution for X_1 and X_2 (see left part in Fig. 1): the potential for $X_1 = 0$ is proportional to the one corresponding to $X_1 = 1$. The second one can be obtained from the first one multiplying by 4. The right part of the figure shows the recursive probability tree for this potential. This recursive tree needs to store only 6 numbers

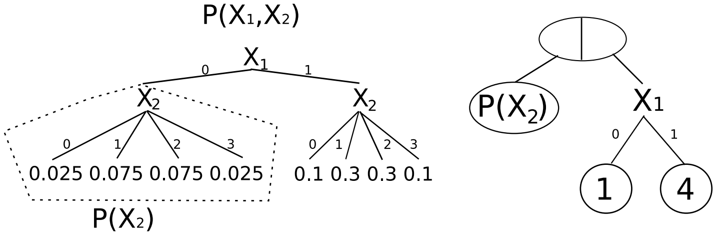


Fig. 1 Potential with proportional values

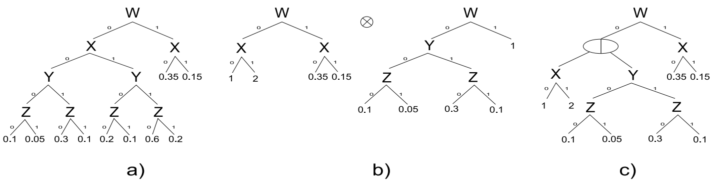


Fig. 2 Decomposition of tree using b) classical decomposition and c) RPTs

instead of 8. For example, an operation involving X_1 need not to work with the left part of the tree.

A RPT can handle this situation and can represent the factorization in a single structure (see Fig. 1). Probability Trees have been used for this kind of decomposition, but usually the factorization must be stored in two independent structures, as in the following example. In the left part of Fig. 2 it can be seen a probability tree encoding a joint distribution for X,Y,Z and W. The result of the factorization taking into account proportional values is shown in the middle part, represented as a multiplication of two probability trees. Finally, this decomposition can be held in just one RPT, as shown in the right part of the figure.

3 Algorithm for Quick Probability Trees Decomposition

When probability distributions are represented with probability trees the efficiency of inference algorithms can be improved by applying tree decomposition, reducing its size and locating the parts that will be affected by a concrete operation. We propose an approximate method for probability trees decomposition, looking for proportional values within itself. The result of the algorithm is a list of factors representing the factorized potential. All these factors can be compactly stored in a RPT.

3.1 Exact Decomposition

The use of probability trees allows to decompose a given tree encoding proportionality into a set of subtrees representing such factorization. However, previous factorization methods search for proportional subtrees located below the variable to delete. Therefore, the performance is highly dependent of the order of the variables in the tree. Let's see an example of this. Consider the probability distribution for variables X, Y and Z represented as the probability tree shown in the left part of Fig. 3. Imagine that X is the next variable to delete during the inference process. Then using classical factorization the most compact decomposition would be the one shown in Fig. 3.

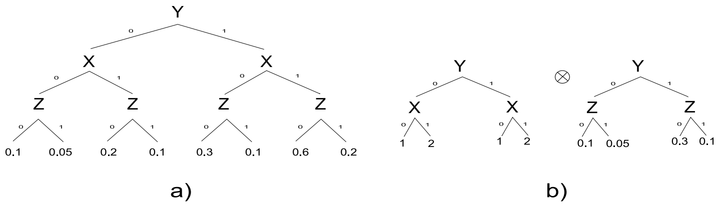


Fig. 3 a) Probability Tree b) Decomposition using classical factorization

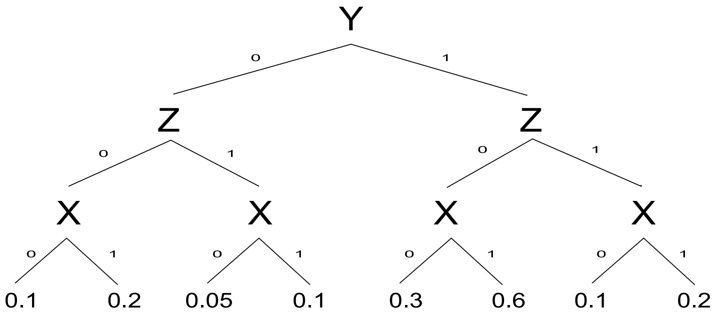


Fig. 4 Probability Tree that can not be factorized by variable X with classical factorization

If the variable to delete is not positioned in the root of the tree, the algorithms that look for factorizations may not be able to find proportionality at all. This happens in the example presented in Fig.4.

We propose an algorithm that finds the most compact factorization of a probability tree for a given variable, independently of the order of the variables in the tree. It returns two probability trees, and we aim to store them as children of a list node within a RPT to benefit from this structure.

```
Input: Probability Tree T and variable to factorize X
Output: two probability trees $\{T_1, T_2\}$
begin
    Let $x_0,\ldots,x_{r-1}$ be the possible values for variable $X$
    Let $(W = w)$ be any configuration for all variables of tree $T$, but $X$
    Let $T^{R(W=w)}$ denote the tree obtained from $T$ by keeping only the
    branches compatible with configuration $(W = w)$
    Let $\alpha_0$, ..., $\alpha_{r-1}$ be the leaves of tree $T^{R(W=w)}$
    for $i \leftarrow 0$ to $r-1$ do
        Let $\beta_i = \frac{\alpha_i}{\alpha_0}$
    end
    Let $T_1$ be a tree with $X$ as only inner node and $\beta_0$, ..., $\beta_{r-1}$ as leaves
    Set $T_2 = T^{R(X=x_0)}$
end
```

Algorithm 1. Factorize(T,X), quick tree factorization algorithm

3.2 Approximate Decomposition

Given a tree, Algorithm 1 finds its most compact decomposition for any variable. However, the exact decomposition for a variable may be impossible just because the tree is not proportional with respect to it. In such case it would be helpful to have a measure about the degree of exact decomposability of a tree for a given variable. Such measure of degree of decomposability could be used, for instance, to establish a threshold to control the accuracy of the decompositions, so that a tree would be allowed to be factorized with respect to a given variable whenever the degree of decomposability surpasses a previously established limit. In this way, it would be possible to obtain an approximate factorization of a tree with a fixed accuracy. It would be achieved by decomposing the tree with respect to any variable that surpass the established limit, and then repeating the same process recursively with the resulting factors, while the limit is surpassed.

If a tree is not exactly decomposable with respect to a given variable, we propose to use the Kullback-Leibler (KL) divergence [5] as a basis to determine how far from exact factorization a given decomposition is. The key result is given in the next theorem, where we give an upper bound of the KL divergence for a given decomposition.

Theorem 1. *Let T be a probability tree to be decomposed with respect to variable X. Let $\mathbf{Y}$ be the set of variables for which T is defined. Let $\mathbf{Z} = \mathbf{Y} \setminus \{X\}$. Let T_1 and T_2 be the output of Algorithm 1 applied to T and X. Then if $D(\cdot,\cdot)$ denotes the KL divergence, it holds that*

$$D(T, T_1 \times T_2) \leq -H(T) - \left(\sum_{\mathbf{y}} t(\mathbf{y})\right)\left(\sum_x \log t_1(x) + \sum_{\mathbf{z}} \log t_2(\mathbf{z})\right), \qquad (1)$$

where H denotes Shannon's entropy, and t, t_1 and t_2 are the real functions represented by trees T, T_1 and T_2 respectively.

Proof

$$\begin{aligned} D(T, T_1 \times T_2) &= \sum_{\mathbf{y}} t(\mathbf{y}) \log \frac{t(\mathbf{y})}{t_1(x) t_2(\mathbf{z})} = \sum_{x,\mathbf{z}} t(x,\mathbf{z}) \log \frac{t(x,\mathbf{z})}{t_1(x) t_2(\mathbf{z})} \\ &= \sum_{x,\mathbf{z}} t(x,\mathbf{z}) \left(\log t(x,\mathbf{z}) - \log t_1(x) - \log t_2(\mathbf{z})\right) \\ &= \sum_{x,\mathbf{z}} t(x,\mathbf{z}) \log t(x,\mathbf{z}) - \sum_{x,\mathbf{z}} t(x,\mathbf{z}) \log t_1(x) - \sum_{x,\mathbf{z}} t(x,\mathbf{z}) \log t_2(\mathbf{z}) \\ &= -H(T) - \sum_x \left(\log t_1(x) \sum_{\mathbf{z}} t(x,\mathbf{z})\right) - \sum_{\mathbf{z}} \left(\log t_2(\mathbf{z}) \sum_x t(x,\mathbf{z})\right). \end{aligned}$$

Now, since all the values in T are non-negative real numbers, we find that

$$\begin{aligned} D(T, T_1 \times T_2) &= -H(T) - \sum_x \left(\log t_1(x) \sum_{\mathbf{z}} t(x,\mathbf{z})\right) - \sum_{\mathbf{z}} \left(\log t_2(\mathbf{z}) \sum_x t(x,\mathbf{z})\right) \\ &\leq -H(T) - \left(\sum_{x,\mathbf{z}} t(x,\mathbf{z})\right)\left(\sum_x \log t_1(x) + \sum_{\mathbf{z}} \log t_2(\mathbf{z})\right) \\ &= -H(T) - \left(\sum_{\mathbf{y}} t(\mathbf{y})\right)\left(\sum_x \log t_1(x) + \sum_{\mathbf{z}} \log t_2(\mathbf{z})\right). \end{aligned}$$

□

Note that, if a decomposition is exact, then $D(T, T_1 \times T_2)$ is equal to 0, and the further a decomposition is to the exact one, the higher the value of the divergence reaches. Observe also that the upper bound given in Equation (1), actually depends on the specific decomposition through the term $S = \sum_x \log t_1(x) + \sum_{\mathbf{z}} \log t_2(\mathbf{z})$, which suggests that S could be used as a measure of the degree of decomposability of a tree T with respect to variable X. The computation of S is rather fast, as it requires a time linear on the size of T_1 and T_2. But if computing time is a critical issue, S can be again bounded using Jensen's inequality as follows:

$$S = \sum_x \log t_1(x) + \sum_{\mathbf{z}} \log t_2(\mathbf{z}) \leq \log \sum_x t_1(x) + \log \sum_{\mathbf{z}} t_2(\mathbf{z}). \tag{2}$$

Note that the right side of the inequality is faster to compute because the logarithm is applied to the result of the sum, instead of being calculated for all the terms. Hence, we use the reasoning above to formally define the degree of decomposability of a tree with respect to a given variable, called the factorization degree, as follows.

Definition 1 (Factorization degree). *Let T be a probability tree. Let $\mathbf{Y}$ be the set of variables for which T is defined, and $X \subset \mathbf{Y}$. Let $\mathbf{Z} = \mathbf{Y} \setminus \{X\}$. Let T_1 and T_2 be the output of Algorithm 1 applied to T and X. We define the* factorization degree *of T with respect to X as*

$$\text{fd}(T, X) = \log \sum_x t_1(x) + \log \sum_{\mathbf{z}} t_2(\mathbf{z}), \tag{3}$$

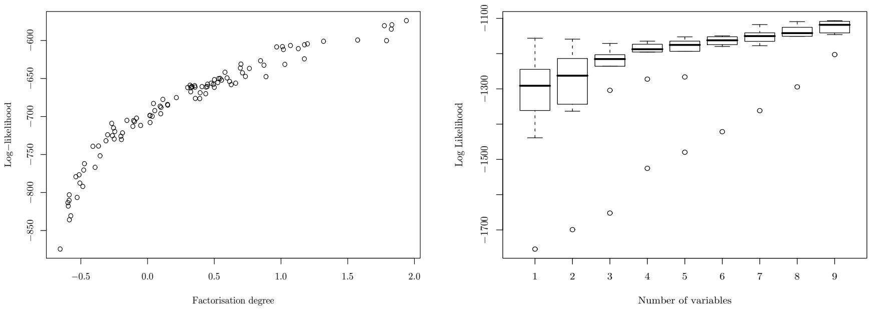
where t_1 and t_2 are the real functions represented by T_1 and T_2 respectively.

We have restricted this study to the simplest case, in which the first term of the decomposition contains only one variable. However, the bound given by Theorem 1 can be extended to the case in which T_1 has more than one variable. The complexity of computing the factorization would be higher, but the size of the resulting decomposition would be smaller as well. We leave for a future work the analysis of this case.

4 Experimental Evaluation

In order to illustrate the behavior of the approximate decomposition method introduced in Section 3.2, we have carried out two experiments. The first one is aimed at checking the capability of the factorization degree in Def. 1 for ranking the variables in a probability tree according to the accuracy of the decompositions produced when factorizing with respect to them. The goal of the second experiment is to show that the factorization degree is also able to guide the decomposition of a single probability tree into several factors.

Experiment 1 consists of generating a random tree with 10 binary variables and then computing the factorization degree for each one of the variables and decomposing the tree for the given variable. Then, we measure, using the obtained factorization, the log-likelihood of a test data set sampled from



(a) (b)

Fig. 5 Results of the experiments carried out.

the original tree. This process is repeated 10 times for different random trees. The obtained results can be seen in Fig. 5.(a).

In experiment 2, we decompose a random tree with 10 binary variables into as many factors as variables it has. Initially, we get the variable with highest factorization degree for the original tree, and decompose it with respect to that variable using Algorithm 1. Then we repeat the same process with the resulting factors, as long as they contain more than one variable, until no more factors can be decomposed. We annotate the log-likelihood of a test data set sampled from the original tree together with the number of variables in the largest remaining factor. Again, the process is repeated 10 times for different random trees. The results are displayed in the box plot in Fig. 5.(b).

The results of experiment 1 support the idea that the factorization degree of the variable with respect to which the decomposition is carried out actually influences the quality of the decomposed model, in the sense that higher factorization degrees result in models with higher likelihood (see Fig. 5.(a)). This fact suggests that the factorization degree could be used as a means for deciding which is the best variable with respect to which a potential should be split. This can be taken into account when designing specific approximate inference algorithms that deal with factorized representations of potentials as, for instance, RPTs. Experiment 2 shows how the factorization degree can also be used to decompose a potential into several factors, by dividing the potential with respect to the variable with higher factorization degree in each step. The decomposition process could be continued until a significant variation in the likelihood of the resulting model happens.

5 Conclusions

In this paper we have introduced a new and fast procedure for factorizing probability trees. An important feature of the proposed algorithm is related to its capability for obtaining optimal decompositions for trees hiding proportionality. Therefore, the resulting decompositions are the most compact ones. We have also shown that the decomposition can be carried out even if the tree does not really contain proportional subtrees, in which case the obtained factorization will be approximate. In order to deal with the degree of approximation of the possible factorizations of a potential, we have introduced a measure called the *factorization degree*, that ranks the variables in the domain of a potential according to the accuracy of the decompositions that they induce. The computation of such measure is fast enough as to be included in any inference algorithm, where computing time is a crucial issue.

The experimental evaluation shows that the methodology developed in this paper can be useful in the design of new inference algorithms that can handle decomposed representations of potentials. Also, it can be used as a tool for learning RPTs or, more precisely, learning parts of a RPT that corresponds to list nodes, and therefore could be a part of a more general learning algorithm for RPTs able to capture all the regularities that a RPT can represent, as for instance, context specific independencies.

Acknowledgements. This work has been supported by the Spanish Ministry of Science and Innovation, under projects TIN2007-67418-C03-02,03, by EFRD (FEDER) funds and the FPI scholarship programme (BES-2008-002049).

References

1. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Horvitz, E., Jensen, F. (eds.) Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, UAI 1996, Portland, Oregon, pp. 115–123. Morgan Kaufmann, San Francisco (1996)
2. Cano, A., Gómez-Olmedo, M., Moral, S., Pérez-Ariza, C.: Recursive probability trees for Bayesian networks. In: Proceedings of the XIII Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2009, Sevilla, Spain (2009)
3. Cano, A., Moral, S.: Propagación exacta y aproximada con árboles de probabilidad. In: Actas de la VII Conferencia de la Asociación Española para la Inteligencia Artificial, CAEPIA 1997, Málaga, Spain, pp. 635–644 (1997)
4. Cano, A., Moral, S., Salmerón, A.: Lazy evaluation in Penniless propagation over join trees. Networks 39, 175–185 (2002)
5. Kullback, S., Leibler, R.: On information and sufficiency. Ann. Math. Statist. 22, 76–86 (1951)
6. Madsen, A., Jensen, F.: Lazy propagation: a junction tree inference algorithm based on lazy evaluation. Artificial Intelligence 113, 203–245 (1999)
7. Martínez, I., Moral, S., Rodríguez, C., Salmerón, A.: Factorisation of probability trees and its application to inference in Bayesian networks. In: Gámez, J., Salmerón, A. (eds.) Proceedings of the First European Workshop on Probabilistic Graphical Models, PGM 2002, Cuenca, Spain, pp. 127–134 (2002)
8. Martínez, I., Moral, S., Rodríguez, C., Salmerón, A.: Approximate factorisation of probability trees. In: Godo, L. (ed.) ECSQARU 2005. LNCS (LNAI), vol. 3571, pp. 51–62. Springer, Heidelberg (2005)

Option Pricing in Incomplete Markets Based on Partial Information

Andrea Capotorti, Giuliana Regoli, and Francesca Vattari

Abstract. In this paper we describe a new approach for the valuation problem in incomplete markets with $m \geq 1$ stocks which can be used when the available information about the uncertainty model is only a partial conditional probability assessment $\mathbf{p}$. We select a risk neutral probability minimizing a discrepancy measure between $\mathbf{p}$ and the convex set of all possible risk neutral probabilities.

Keywords: Risk neutral valuation, Partial conditional probability assessments, Incomplete markets.

1 Introduction

In a viable single-period model with $m \geq 1$ stocks and $k \geq 2$ scenarios the completeness of the market is equivalent to the uniqueness of the risk neutral probability; this equivalence allows to price every derivative security with a unique fair price. In literature, different methods have been proposed in order to select a risk neutral probability when the market is incomplete (see for example [8], [9] and [10]). Contrary to the complete case, where the so called *"real world probability"* $\mathbf{p}$ expressing the agent behaviors is not involved in the valuation problem, in the methods for incomplete markets $\mathbf{p}$ is really used and its elicitation is a crucial point for the option pricing. In particular, $\mathbf{p}$ is supposed to be given over all the possible scenarios while usually the available information about the possible states of the world is partial, conditional or even incoherent. In this paper our purpose is to present a method for the

Andrea Capotorti, Giuliana Regoli, and Francesca Vattari
Dipartimento di Matematica e Informatica, Università degli Studi di Perugia, Rome, Italy
e-mail: capot@dmi.unipg.it,regoli@dmi.unipg.it,
francesca.vattari@dmi.unipg.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 73–80.
springerlink.com

risk neutral valuation in incomplete markets which can be used when $\mathbf{p}$ is a partial conditional probability assessment as well as when there are more partial conditional probability assessments given by different expert opinions. In the next Subsections 1.1 and 1.2 we introduce the valuation problem and a discrepancy measure which will be minimized in order to select a risk neutral probability among those characterizing the incomplete market. In Section 2 we describe such selection procedure and we give some illustrative examples.

1.1 Single-Period Models with m Stocks

In this section we describe the risk neutral valuation problem for a single-period financial model with m risky assets and a risk-free interest rate r (see [1], [7] and [12] for more details).

Let $\mathbf{S}_t = (S_t^1, \dots, S_t^m)$ be the vector of the stock prices at time t with $t = 0, 1$ and let us suppose that the initial prices $S_0^1, \dots, S_0^m$ are known at time 0 while the prices of each stock S_1^l, $l = 1, \dots, m$, are finite random variables

$$S_1^l : \Omega \to \mathbb{R},$$

where $\Omega = \{\omega_1, \dots, \omega_k\}$ is the set of possible scenarios (states of the world). If we denote with $\overline{\mathbf{S}}_t := \mathbf{S}_t/(1+r)^t$, $t = 0, 1$ the discounted stock price process, we can define a risk neutral probability as a probability distribution α over Ω under which the discounted stock price process is a martingale, that is

$$\overline{\mathbf{S}}_0 = \mathbb{E}_{\alpha}(\overline{\mathbf{S}}_1). \tag{1}$$

Notice that this means that $\mathbf{S}_0 = \mathbb{E}_{\alpha}(\mathbf{S}_1)/(1+r)$ and this expression explains why a probability measure α which verifies (1) is called risk neutral: α is a probability distribution such that the price S_0^l of each stock can be computed as the expected value with respect to α of S_1^l discounted with the risk free interest rate r. A market model is said to be *viable* if there are no arbitrage opportunities and it is said to be *complete* if every derivative security [1] admits a replicating portfolio (i.e. a portfolio with the same payoff). A single period model with a finite number of scenarios and an arbitrary number of stocks admits a risk neutral probability if and only if it is viable.

Let us consider the problem of completeness for a viable single-period model with m stocks and k scenarios. Since a market model is complete if and only if every derivative security D is attainable, the completeness is equivalent to have, for every derivative D, a portfolio (x, y) such that

$$\begin{cases} xB_1 + y\mathbf{S}_1(\omega_1) = D(\omega_1) \\ \dots \\ xB_1 + y\mathbf{S}_1(\omega_k) = D(\omega_k) \end{cases}$$

where $B_1 = (1+r)B_0$ is the price at $t = 1$ of the risk-free asset.

[1] We recall that a derivative security is a security whose value at time 1 depends on the values of risky assets $S_1^1, \dots, S_1^m$.

Thus a single-period model with k scenarios is complete if and only if

$$A := \begin{pmatrix} B_1 \; S_1^1(\omega_1) \ldots S_1^m(\omega_1) \\ \vdots \qquad\qquad \vdots \\ B_1 \; S_1^1(\omega_k) \ldots S_1^m(\omega_k) \end{pmatrix} \tag{2}$$

has rank k, that is it contains at least k independent assets. Notice that if a model is complete then we must have $m+1 \geq k$. It is easy to see that α is a risk neutral probability if and only if

$$A^T \alpha = (1+r) \begin{pmatrix} B_0 \\ \mathbf{S}_0 \end{pmatrix}$$

and $\alpha_j \geq 0, \;\; j = 1, \ldots, k$. Therefore it follows that in a viable single period model, with $k \geq 2$ scenarios, a unique risk free asset and m risky assets, the completeness is equivalent to the uniqueness of the risk neutral probability.

When the market is viable and complete, the fair price π of any derivative security D is given by

$$\pi = \mathbb{E}_{\alpha}(\overline{D}) \tag{3}$$

where $\overline{D}$ is the discounted price of D and α is the risk neutral probability.

When the market is incomplete the set F of all possible fair prices for a derivative D is

$$F = [l, u]$$

where

$$l := \inf\{\mathbb{E}_{\alpha}(\overline{D}) \mid \alpha \text{ is a risk neutral probability}\},$$

$$u := \sup\{\mathbb{E}_{\alpha}(\overline{D}) \mid \alpha \text{ is a risk neutral probability}\}.$$

Obviously a derivative security is attainable if and only if $l = u$; otherwise $l < u$ and we have to consider the interval $[l, u]$. If we denote by $\mathscr{Q}$ the convex set of possible risk neutral probabilities, taken $\alpha^* \in \mathscr{Q}$, for every derivative security there will be a corresponding fair price π given by

$$\pi = \mathbb{E}_{\alpha^*}(\overline{D}) \in F.$$

In Section 2 we will describe how to select such a α^* that should be as close as possible to the agent behaviors expressed by partial conditional probability assessment $\mathbf{p}$. To express the closeness between α^* and $\mathbf{p}$ we will profit from a recently introduced ([2, 4]) discrepancy measure that, for the sake of completeness, we briefly describe in the following subsection.

1.2 Discrepancy Measure

Let $\mathbf{p} = (p_1, \ldots, p_n) \in (0,1)^n$ be a conditional probability assessment given by an agent over a set of conditional events $\mathscr{E} = [E_1|H_1, \ldots, E_n|H_n]$. $\mathscr{E}$ expresses events, the E_i's, considered under specific situations or hypothesis, the H_i's, over which the agent possess, or is able to express, probabilistic behaviors. In the following E_iH_i will denote the logical conjunction "$E_i \wedge H_i$", while E_i^c will denote the negation "$\neg E_i$". To be meaningful, the E_i's and the H_i's must be expressible through possible values of the assets present in the market, hence without loss of generality we consider the E_i's and the H_i's as subsets of the set of all possible states of the world $\Omega = \{\omega_1, \dots, \omega_k\}$, with each ω_j representing a specific assets evaluation situation.

To properly define a pseudo-distance between a probabilistic evaluation over $\mathscr{E}$ and an other over Ω, we need to introduce the following hierarchy of probability mass function on Ω:
$\mathscr{A} := \left\{\alpha = [\alpha_1, \dots, \alpha_k], \sum \alpha_j = 1, \alpha_j \geq 0, j = 1, \dots, k\right\}$;
$\mathscr{A}_0 := \{\alpha \in \mathscr{A} | \alpha(\bigcup_{i=1}^n H_i) = 1\}$;
$\mathscr{A}_1 := \{\alpha \in \mathscr{A}_0 | \alpha(H_i) > 0, i = 1, \dots, n\}$;
$\mathscr{A}_2 := \{\alpha \in \mathscr{A}_1 | 0 < \alpha(E_iH_i) < \alpha(H_i), i = 1, \dots, n\}$.
Any $\alpha \in \mathscr{A}_1$ induces a coherent conditional assessment on $\mathscr{E}$ given by

$$\mathbf{q}_\alpha := [q_i = \frac{\sum\limits_{j:\,\omega_j \subset E_iH_i} \alpha_j}{\sum\limits_{j:\,\omega_j \subset H_i} \alpha_j}, i = 1, \dots, n]. \tag{4}$$

Associated to any assessment $\mathbf{p} \in (0,1)^n$ over $\mathscr{E}$ we can define a scoring rule

$$S(\mathbf{p}) := \sum_{i=1}^n |E_iH_i| \ln p_i + \sum_{i=1}^n |E_i^cH_i| \ln(1 - p_i) \tag{5}$$

with $|\cdot|$ indicator function of unconditional events. This score $S(\mathbf{p})$ is an "adaptation" to partial and conditional probability assessments of the "proper scoring rule" for probability distributions proposed by Lad in [11]. By adopting the difference between the expected penalties suffered by the two evaluations $\mathbf{p}$ and $\mathbf{q}_\alpha$ as distance criterion, it is possible to define the "discrepancy" between a partial conditional assessment $\mathbf{p}$ over $\mathscr{E}$ and a distribution $\alpha \in \mathscr{A}_2$ through the expression

$$\Delta(\mathbf{p}, \alpha) := E_\alpha(S(\mathbf{q}_\alpha) - S(\mathbf{p})) = \sum_{j=1}^k \alpha_j [S_j(\mathbf{q}_\alpha) - S_j(\mathbf{p})]. \tag{6}$$

It is possible to extend by continuity the definition of $\Delta(\mathbf{p}, \alpha)$ in $\mathscr{A}_0$ as

$$\Delta(\mathbf{p}, \alpha) = \sum_{i|\alpha(H_i)>0} \alpha(H_i) \left(q_i \ln \frac{q_i}{p_i} + (1 - q_i) \ln \frac{(1 - q_i)}{(1 - p_i)} \right)$$

adopting the usual convention $0 \ln 0 = 0$.

In [4] is formally proved that $\Delta(\mathbf{p}, \alpha)$ is a non negative function on $\mathscr{A}_0$ and that $\Delta(\mathbf{p}, \alpha) = 0$ if and only if $\mathbf{p} = \mathbf{q}_\alpha$; moreover $\Delta(\mathbf{p}, \cdot)$ is a convex function on $\mathscr{A}_2$ and it admits a minimum on $\mathscr{A}_0$. Finally if $\alpha, \alpha^0 \in \mathscr{A}_0$ are distributions that minimize $\Delta(\mathbf{p}, \cdot)$, then for all $i \in \{1, \dots, n\}$ such that $\alpha(H_i) > 0$ and $\alpha^0(H_i) > 0$

we have $(\mathbf{q}_\alpha)_i = (\mathbf{q}_{\alpha^0})_i$; in particular if $\Delta(\mathbf{p},\cdot)$ attains its minimum value on $\mathscr{A}_1$ then there is a unique coherent assessment $\mathbf{q}_{\underline{\alpha}}$ such that $\Delta(\mathbf{p},\underline{\alpha})$ is minimum. The discrepancy measure $\Delta(\mathbf{p},\alpha)$ can be used to correct incoherent assessments [2], to aggregate expert opinions [5] and it can be even applied with imprecise probabilities [3]. Here we propose a particular optimization problem involving $\Delta(\mathbf{p},\alpha)$ which will be used to select the risk neutral probability in the set of all possible martingale measures which better represents the agent's behaviors.

2 Selection of a Risk-Neutral Probability

In order to keep the market tractable, we start with a viable single period model like in Subsection 1.1, without transaction costs and with the following stock prices structure:

	ω_1	ω_2	$\dots$	ω_k
S_1^l	$a_1^l\ S_0^l$	$a_2^l\ S_0^l$	$\dots$	$a_k^l\ S_0^l$

l=1,...,m.

In this model a probability distribution α on Ω is risk neutral if and only if

$$\mathbf{S}_0 = \frac{1}{1+r}\left[\alpha_1\mathbf{a}_1\mathbf{S}_0 + \ldots + \alpha_k\mathbf{a}_k\mathbf{S}_0\right] \Leftrightarrow \mathbf{1} = \frac{1}{1+r}\left[\alpha_1\mathbf{a}_1 + \ldots + \alpha_k\mathbf{a}_k\right]$$

where $\mathbf{a}_j = (a_j^1, \ldots, a_j^m)$ for $j = 1, \ldots, k$. Thus we can define the set of all possible martingale measures as:

$$\mathscr{Q} := \{\alpha \in \mathbb{R}^k : \alpha \cdot \mathbf{1} = 1,\ \alpha \geq 0,\ \sum_{j=1}^{k} \alpha_j a_j^l = 1 + r,\ l = 1, \ldots, m\}.$$

Finally we assume that $\mathbf{p} = (p_1, \ldots, p_n)$ is a partial conditional probability assessment given over the set of conditional events $\mathscr{E} = [E_1|H_1, \ldots, E_n|H_n]$. Notice that it is not required that the assessment $\mathbf{p}$ is coherent[2]; we can have an assessment which is inconsistent with all the distributions in $\mathscr{A}$ or we can have a coherent assessment that is anyhow inconsistent with the set of all martingale measures. Our purpose is to find the risk neutral probability which is the *closest* to the initial opinion with respect to the discrepancy measure Δ. Let $\mathscr{Q}_0$ be the convex set $\mathscr{Q}_0 := \mathscr{Q} \cap \mathscr{A}_0$; we propose to select a martingale measure in $\mathscr{Q}_0$ starting from the assessment $\mathbf{p}$. In fact, we suggest a selection procedure which is based on the following result:

Theorem 1. *Let $\mathscr{M} := \arg\min\{\Delta(\mathbf{p},\alpha), \alpha \in \mathscr{Q}_0\}$ be the set of all martingale measures minimizing $\Delta(\mathbf{p},\alpha)$; then $\mathscr{M}$ is a non-empty convex set.*

Proof. $\Delta(\mathbf{p},\alpha)$ is a convex function on $\mathscr{Q}_0$ and then there is at least one $\underline{\alpha}$ in $\mathscr{Q}_0$ such that

[2] For coherence notion of partial conditional assessments refer e.g. to [6]

$$\Delta(\mathbf{p},\underline{\alpha}) = \min_{\alpha\in\mathscr{Q}_0} \Delta(\mathbf{p},\alpha) \tag{7}$$

and then $\mathscr{M}$ is non empty. Notice that the convexity of $\Delta(\mathbf{p},\alpha)$ guarantees the existence of this minimum but it is possible that more than one distribution minimize $\Delta(\mathbf{p},\alpha)$ in $\mathscr{Q}_0$ and in this case $\mathscr{M}$ is not a singleton. However, since $\Delta(\mathbf{p},\alpha)$ is a convex function and $\mathscr{M}$ is the set of minimal points of $\Delta(\mathbf{p},\alpha)$ in $\mathscr{Q}_0$, $\mathscr{M}$ is a convex set. □

Thanks to this result we can select exactly one of such minimizer distributions $\underline{\alpha}$. Let us see how it works with first example.

Example 1. Let us consider a model with two risky assets S^1, S^2 with initial prices $S_0^1 = 200$, $S_0^2 = 150$ and final values

	ω_1	ω_2	ω_3	ω_4
S_1^l	220	210	200	180
S_1^2	180	150	150	120

Since we have $m = 2$ assets and $k = 4$ states of the world, the rank of matrix A in (2) is surely less than k, so that the market is incomplete. For $r = 0$ the set of all possible martingale measures is $\mathscr{Q}_\lambda = \{(\lambda, 0, 1-2\lambda, \lambda), \lambda \in [0,1/2]\}$. Let us suppose that the agent assesses the probabilities $p_1 = P(\omega_4) = 1/3$ and $p_2 = P(\omega_1|\omega_1 \vee \omega_2) = 1/4$. Then

$$\Delta(\mathbf{p},\alpha) = \alpha_4 \ln 3\alpha_4 + (1-\alpha_4)\ln\frac{3}{2}(1-\alpha_4) + \alpha_1 \ln\frac{4\alpha_1}{(\alpha_1+\alpha_2)} + \alpha_2 \ln\frac{4\alpha_2}{3(\alpha_1+\alpha_2)}$$

that is $\Delta(\mathbf{p},\lambda) = \lambda\ln 3\lambda + (1-\lambda)\ln\frac{3}{2}(1-\lambda) + \lambda\ln 4$.

Since $\Delta'(\mathbf{p},\lambda) = \ln\lambda - ln(1-\lambda) + \ln 8 = 0 \Leftrightarrow \frac{\lambda}{1-\lambda} = \frac{1}{8} \Leftrightarrow \lambda = \frac{1}{9}$ we get $\underline{\alpha} = \left(\frac{1}{9}, 0, \frac{7}{9}, \frac{1}{9}\right)$.

Theorem 1 guarantees the existence of a solution $\underline{\alpha}$ for the optimization problem (7) but it does not assure its uniqueness; when the information that we have is not sufficient to give us a unique solution for (7) we need another criterion to choose, between the martingale measure minimizing $\Delta(\mathbf{p},\alpha)$, a unique α^* as risk-neutral probability. The idea is to select one distribution in $\mathscr{M}$ which in some sense minimizes the exogenous information. In fact, we will define α^* as

$$\alpha^* := arg \min_{\alpha\in\mathscr{M}} \sum_{j=1}^{k} \alpha_j \ln\alpha_j \tag{8}$$

that is the distribution which minimizes the relative entropy with respect to the uniform distribution (i.e. the distribution with maximum entropy).

Let us see how it can be operationally done with an exemplifying situation

Example 2. Let us consider again a model with two risky assets S^1, S^2 but now with initial prices $S_0^1 = 19$, $S_0^2 = 21$ and final values

	ω_1	ω_2	ω_3	ω_4	ω_5
S_1^l	22	21	20	19	18
S_1^2	25	24	21	21	20

so that the set of martingale measures is

$$\mathscr{Q}_{\lambda,\mu} = \left\{ (\lambda, \mu - \lambda, \mu, 1 - \lambda - 5\mu, \lambda + 3\mu), \lambda \leq \frac{1}{6}, \mu \in [\lambda, \frac{1-\lambda}{5}] \right\}.$$

Let us suppose that the agent assesses the probabilities $p_1 = P(\omega_3) = 1/3$ and $p_2 = P(\omega_1 \vee \omega_2 | \omega_1 \vee \omega_2 \vee \omega_3) = 9/10$. Then

$$\begin{aligned} \Delta(\mathbf{p}, \alpha) &= \alpha_3 \ln 3\alpha_3 + (1-\alpha_3) \ln \frac{3}{2}(1-\alpha_3) + \\ &+ (\alpha_1 + \alpha_2) \ln \frac{10(\alpha_1+\alpha_2)}{9\,(\alpha_1+\alpha_2+\alpha_3)} + \alpha_3 \ln \frac{10\alpha_3}{(\alpha_1+\alpha_2+\alpha_3)} \end{aligned}$$

that is $\Delta(\mathbf{p}, \mu) = \mu \ln 3\mu + (1-\mu) \ln \frac{3}{2}(1-\mu) + \mu \ln \frac{5}{9} + \mu \ln 5$.

Since

$$\Delta'(\mathbf{p}, \mu) = \ln \mu - ln(1-\mu) - \ln 4 - \ln(1/3) + \ln(2/3) - \ln(9/100)$$

we have

$$\Delta'(\mathbf{p}, \mu) = 0 \Leftrightarrow \frac{\mu}{1-\mu} = \frac{18}{100} \Leftrightarrow \mu = \frac{9}{59}$$

so that the set of martingale measures which minimize $\Delta(\mathbf{p}, \cdot)$ is

$$\mathscr{M} = \{(\lambda, 9/59 - \lambda, 9/59, 14/59 - \lambda, \lambda + 27/29) : \lambda \leq 9/59\}.$$

Among all such distributions we can select $\alpha^* \in \mathscr{M}$ maximizing the entropy

$$\begin{aligned} H(\lambda) &= -\lambda \ln \lambda - (\frac{9}{59} - \lambda) \ln(\frac{9}{59} - \lambda) - \frac{9}{59} \ln \frac{9}{59} + \\ &- (\frac{14}{59} - \lambda) \ln(\frac{14}{59} - \lambda) - (\lambda + \frac{27}{59}) \ln(\lambda + \frac{27}{59}) \end{aligned}$$

obtaining $\alpha^* = (.043, .110, .153, .195, .5)$.

3 Conclusion

Thanks to the minimization of a pseudo-distance among a partial conditional probability assessments $\mathbf{p}$ and probability distributions α we have shown that it is possible to aggregate disparate fonts of information like those induced by market prices structures, which are usually extremely rich even in the context of incomplete markets, and those induced by human agents, which are typically non structured and partial.

References

1. Bingham, N.H., Kiesel, R.: Risk-Neutral Valuation: pricing and hedging of financial derivatives. Springer, London (2004)
2. Capotorti, A., Regoli, G.: Coherent correction of inconsistent conditional probability assessments. In: Proceedings of the 12th International Conference on Information Processing and management of Uncertainty in Knowledge-based Systems, IPMU 2008, Malaga, Spain (2008)
3. Capotorti, A., Regoli, G., Vattari, F.: On the use of a new discrepancy measure to correct incoherent assessments and to aggregate conflicting opinions based on imprecise conditional probabilities. In: Augustin, T., Coolen, F.P.A., Moral, S., Troffaes, M.C.M. (eds.) Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK (2009)
4. Capotorti, A., Regoli, G., Vattari, F.: Correction of incoherent conditional probability assessments. Intern. J. Approx. Reason. (to appear, 2010)
5. Capotorti, A., Regoli, G., Vattari, F.: Merging different probabilistic information sources through a new discrepancy measure. In: Kroupa, T., Vejnarova, J. (eds.) Proceedings of the 8th Workshop on Uncertainty Processing, WUPES 2009, Liblice, Czech Republic (2009)
6. Coletti, G., Scozzafava, R.: Probabilistic Logic in a Coherent Setting. Trends in Logic Series, vol. 15. Kluwer, Dordrecht (2002)
7. Elliot, R.J., Kopp, P.E.: Mathematics of Financial Markets. Springer Finance, New York (2005)
8. Follmer, H., Schied, A.: Stochastic Finance: an introduction in discrete time. Walter de Gruyter, Berlin (2004)
9. Follmer, H., Sondermann, D.: Hedging of Non-Redundant Contingent Claim. In: Hildenbrand, W., Mas-Colell, A. (eds.) Contributions to Mathematical Economics, pp. 205–223. North-Holland, Amsterdam (1986)
10. Frittelli, M.: The minimal entropy martingale measure and the valuation problem in incomplete markets. Math. Finance 10, 39–52 (2000)
11. Lad, F.: Operational Subjective Statistical Methods: a mathematical, philosophical, and historical introduction. John Wiley, New York (1996)
12. Musiela, M., Rutkowski, M.: Martingale Methods in Financial Modelling. Springer, New York (2005)

Lorenz Curves of *extrema*

Ignacio Cascos and Miguel Mendes

Abstract. We study the Generalized Lorenz curves of the minima of samples from a random variable. These Lorenz curves can be used to compare distributions in terms of their variability and define coherent risk measures. The dual Lorenz curves of the maxima of samples from the random variable are also considered.

1 Introduction

Risk measures are quite a fashionable tool in financial mathematics. They quantify the risk of an investment and are widely used in the financial markets. Banks are obliged to maintain their risks below the level fixed by the regulatory agency (capital requirement) given in terms of risk measures and further employ risk measures in order to optimize their capital allocation.

An investor would benefit from having investments ordered after their risks. We will define stochastic orders in terms of Lorenz curves and *coherent* risk measures (see Artzner et al. [3]) that are consistent with them.

Sections 2 and 3 are devoted to Lorenz curves and stochastic orders, Sections 4 and 5 to risks and deviations. Conclusions are presented in Section 6.

2 Lorenz Curves

Throughout the present manuscript, we assume that all random variables have finite first moment.

Ignacio Cascos
Department of Statistics, Universidad Carlos III de Madrid, Spain
e-mail: ignacio.cascos@uc3m.es

Miguel Mendes
Faculdade de Engenharia e Centro de Matemática, Universidade do Porto, Portugal
e-mail: migmendx@fe.up.pt

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 81–88.
springerlink.com

Definition 1. *The* Generalized (nonnormalized) Lorenz curve *(GL) of X is defined for any* $0 \le t \le 1$ *as*

$$\underline{GL}(X;t) := \int_0^t F_X^{-1}(s)\mathrm{d}s$$

and its dual *(dual GL) as*

$$\overline{GL}(X;t) := \int_{1-t}^1 F_X^{-1}(s)\mathrm{d}s = \mathbb{E}X - \underline{GL}(X;1-t)\,,$$

where $F_X^{-1}(t) := \inf\{x : F_X(x) > t\}$ *is the quantile function of X, whose cdf we denote by F_X.*

The duality between $\underline{GL}$ and $\overline{GL}$ can be expressed by means of the symmetry of their respective graphs with respect to the point $(1/2, \mathbb{E}X/2)$. Alternatively, it is also true that $\overline{GL}(-X;t) = -\underline{GL}(X;t)$.

The *Lorenz curve* is defined for positive random variables as $L_X(t) = \underline{GL}(X;t)/\mathbb{E}X$, where $\mathbb{E}X$ is the expectation of X. It has an interesting interpretation if X determines the wealth distribution in a population. In such a case $L_X(t)$ is the fraction of total wealth possed by the fraction t of poorest individuals in the population. The interpretation of the dual curve is analogous to this one but with a viewpoint on the richest individuals.

In the name Generalized (nonnormalized) Lorenz curve, GL, the generalization reflects that it is defined for random variables that might take negative values and the nonnormalization that it is not divided by $\mathbb{E}X$.

Lorenz Curves of *extrema*

We will consider the GL of the *minima* of sequences of independent random variables and the dual GL of the *maxima* of the same sequences. We refer to this minima and maxima generically as *extrema.*

Given $X_1, \dots, X_n$ a sample of n independent copies of X, we denote its minimum by $X_{1:n}$ and its maximum by $X_{n:n}$.

Definition 2. *For any* $n \ge 1$ *and* $0 \le t \le 1$*, the* n-GL curve of X *is given by*

$$\underline{GL}_n(X;t) := \underline{GL}(X_{1:n};t) = \int_0^t F_X^{-1}(1-(1-s)^{1/n})\mathrm{d}s\,,$$

and its dual n-GL curve *by*

$$\overline{GL}_n(X;t) := \overline{GL}(X_{n:n};t) = \int_{1-t}^1 F_X^{-1}(s^{1/n})\mathrm{d}s\,.$$

Since the endpoint ($t = 1$) of the GL curve (and its dual) is the expectation, we obtain $\underline{GL}_n(X;1) = \mathbb{E}X_{1:n}$ and $\overline{GL}_n(X;1) = \mathbb{E}X_{n:n}$.

Remark 1. Since $(-X)_{n:n} = -X_{1:n}$, we have $\overline{GL}_n(-X;t) = -\underline{GL}_n(X;t)$.

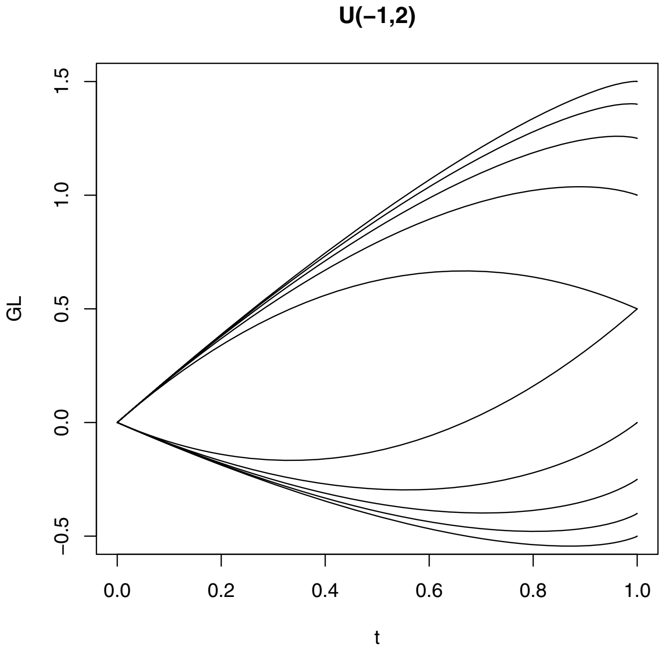


Fig. 1 $\underline{GL}_n(X)$ -convex curves- and $\overline{GL}_n(X)$ -concave curves- for X uniform $(-1,2)$ and $n = 1,2,3,4,5$.

Proposition 1. *For any $n \geq 1$, each of the two curves $\underline{GL}_n(X;\cdot)$ and $\overline{GL}_n(X;\cdot)$ characterizes the distribution of X.*

Proposition 2. *The n-GL curve and its dual satisfy the following properties for $x \in \mathbb{R}$ and $0 \leq t \leq 1$:*

0. If $X = x$ a.s., then $\underline{GL}_n(X;t) = tx$ and $\overline{GL}_n(X;t) = tx$;
1. $\underline{GL}_n(X;t) \leq t\mathbb{E}X_{1:n} \leq t\mathbb{E}X \leq t\mathbb{E}X_{n:n} \leq \overline{GL}_n(X;t)$;
2. If $X \leq Y$ a.s., then $\underline{GL}_n(X;\cdot) \leq \underline{GL}_n(Y;\cdot)$ and $\overline{GL}_n(X;\cdot) \leq \overline{GL}_n(Y;\cdot)$;
3. $\underline{GL}_n(X+x;t) = \underline{GL}_n(X;t) + tx$ and $\overline{GL}_n(X+x;t) = \overline{GL}_n(X;t) + tx$;
4. $\underline{GL}_n(\lambda X;t) = \lambda \underline{GL}_n(X;t)$ and $\overline{GL}_n(\lambda X;t) = \lambda \overline{GL}_n(X;t)$ for $\lambda > 0$;
5. $\underline{GL}_n(X+Y;\cdot) \geq \underline{GL}_n(X;\cdot) + \underline{GL}_n(Y;\cdot)$ and $\overline{GL}_n(X+Y;\cdot) \leq \overline{GL}_n(X;\cdot)$ $+\overline{GL}_n(Y;\cdot)$.

3 Stochastic Orderings

Stochastic orders are partial order relations between probability distributions, see Müller and Stoyan [7]. Here we will make use of the *increasing convex*, *increasing concave* and *convex* stochastic orders, defined as follows:

- *Increasing convex*: $X \leq_{\mathrm{icx}} Y$ if $\overline{GL}(X;\cdot) \leq \overline{GL}(Y;\cdot)$;
- *Increasing concave*: $X \leq_{\mathrm{icv}} Y$ if $\underline{GL}(X;\cdot) \leq \underline{GL}(Y;\cdot)$ (equiv. $-Y \leq_{\mathrm{icx}} -X$) ;
- *Convex*: $X \leq_{\mathrm{cx}} Y$ if $\overline{GL}(X;\cdot) \leq \overline{GL}(Y;\cdot)$ and $\underline{GL}(X;\cdot) \geq \underline{GL}(Y;\cdot)$ (equiv. $X \leq_{\mathrm{icx}} Y$ and $Y \leq_{\mathrm{icv}} X$) .

The former three relations are *integral stochastic orders*. This means that they can be characterized by $\mathbb{E}f(X) \leq \mathbb{E}f(Y)$ for all functions f from a given family, whenever both expectations exist. The respective families of functions is clear by the name of the orderings.

Definition 3. *We define the following stochastic orderings:*

- $X \leq_{\mathrm{icx}_n} Y$ *if* $\overline{GL}_n(X;\cdot) \leq \overline{GL}_n(Y;\cdot)$ *(equiv.* $X_{n:n} \leq_{\mathrm{icx}} Y_{n:n}$*);*
- $X \leq_{\mathrm{icv}_n} Y$ *if* $\underline{GL}_n(X;\cdot) \leq \underline{GL}_n(Y;\cdot)$ *(equiv.* $X_{1:n} \leq_{\mathrm{icv}} Y_{1:n}$*);*
- $X \leq_{\mathrm{cx}_n} Y$ *if* $\overline{GL}_n(X;\cdot) \leq \overline{GL}_n(Y;\cdot)$ *and* $\underline{GL}_n(X;\cdot) \geq \underline{GL}_n(Y;\cdot)$ *(equiv.* $X_{n:n} \leq_{\mathrm{icx}} Y_{n:n}$ *and* $Y_{1:n} \leq_{\mathrm{icv}} X_{1:n}$*).*

Since the 1-GL curve is the classical GL curve, for $n = 1$ we obtain the classical increasing convex, increasing concave and convex stochastic orders.

Lemma 1. *For any* $n \geq 1$*,* $X \leq_{\mathrm{icv}_n} Y$ *is equivalent to* $-Y \leq_{\mathrm{icx}_n} -X$*.*

The relations given in Definition 3 are partial order relations for probability distributions of random variables.

Proposition 3. *If* $\preceq$ *stands for* $\leq_{\mathrm{icx}_n}$*,* $\leq_{\mathrm{icv}_n}$ *or* $\leq_{\mathrm{cx}_n}$*, then for r.v.* X, Y, Z*:*

1. Reflexivity*:* $X \preceq X$
2. Transitivity*: if* $X \preceq Y$ *and* $Y \preceq Z$*, then* $X \preceq Z$*;*
3. Antisymmetry*: if* $X \preceq Y$ *and* $Y \preceq X$*, then* X *and* Y *are identically distributed.*

Proposition 4. *If* $X \leq_{\mathrm{icx}_n} Y$*, then* $X \leq_{\mathrm{icx}_m} Y$ *for all* $m \geq n$*.*

Proof. Let $X \leq_{\mathrm{icx}_n} Y$, $m > n$ and $0 < t < 1$. Applying several changes of variables and Fubini's Theorem, we have

$$\begin{aligned}
\overline{GL}_m(X,t) &= \int_{1-t}^{1} F_X^{-1}(s^{1/m})\mathrm{d}s = \int_{(1-t)^{n/m}}^{1} F_X^{-1}(r^{1/n})\frac{m}{n} r^{m/n-1}\mathrm{d}r \\
&= \int_{(1-t)^{n/m}}^{1} F_X^{-1}(r^{1/n}) \int_0^r \frac{m(m-n)}{n^2} u^{m/n-2}\mathrm{d}u\mathrm{d}r \\
&= \frac{m(m-n)}{n^2} \int_0^1 \int_{\max\{(1-t)^{n/m},u\}}^{1} F_X^{-1}(r^{1/n}) u^{m/n-2}\mathrm{d}r\mathrm{d}u \\
&= \frac{m(m-n)}{n^2} \int_0^1 \overline{GL}_n(X, 1-\max\{(1-t)^{n/m}, u\}) u^{m/n-2}\mathrm{d}u\,.
\end{aligned}$$

As a consequence, if $\overline{GL}_n(X,\cdot) \leq \overline{GL}_n(Y,\cdot)$, the same holds for m, and we have $X \leq_{\mathrm{icx}_m} Y$. □

Corollary 1. *1. If $X \leq_{\mathrm{icv}_n} Y$, then $X \leq_{\mathrm{icv}_m} Y$ for all $m \geq n$;*
2. if $X \leq_{\mathrm{cx}_n} Y$, then $X \leq_{\mathrm{cx}_m} Y$ for all $m \geq n$.

Proof. We will only prove statement *1.* If $X \leq_{\mathrm{icv}_n} Y$, then by Lemma 1 $-Y \leq_{\mathrm{icx}_n} -X$ and after Proposition 4 $-Y \leq_{\mathrm{icx}_m} -X$ for all $m \geq n$, which is equivalent to $X \leq_{\mathrm{icv}_m} Y$ for all $m \geq n$. □

After Proposition 4 and Corollary 1, we have three chains of stochastic orders, the strongest (more restrictive) in each of them being the increasing convex, increasing concave and convex stochastic orders.

Example 1. Let X take values -1 and $1/4$ each with probability $1/2$. Let Y be equal to -1 with probability $2/5$ and to 0 with probability $3/5$. It holds that $X \leq_{\mathrm{icv}_n} Y$ for all $n \geq 2$, but it is not true that $X \leq_{\mathrm{icv}} Y$.

4 Risk Measures

Let us assume that a portfolio is modeled as a random variable X representing its net worth after discounting. A *coherent risk measure* in the sense of Artzner *et al.* [3] is a functional on the set of random variables $X : \Omega \to \mathbb{R}$, Ω being the set of possible scenarios, satisfying the following four axioms,

R1 If $X \geq 0$ *a.s.* then $\rho(X) \leq 0$;
R2 (Cash-invariance) For all $c \in \mathbb{R}$, $\rho(X + c) = \rho(X) - c$;
R3 (Positive homogeneity) For all $\lambda \geq 0$ we have $\rho(\lambda X) = \lambda \rho(X)$;
R4 (Subadditivity) For all X and Y, $\rho(X + Y) \leq \rho(X) + \rho(Y)$;

One example of such a measure is given by the $\alpha-$*expected shortfall* (denoted by ES_α) which is defined for $0 < \alpha < 1$ by

$$ES_\alpha(X) := -\frac{1}{\alpha}\int_0^\alpha F_X^{-1}(s)\,\mathrm{d}s = -\frac{1}{\alpha}\underline{GL}(X;\alpha) \ .$$

For the properties of ES_α, see Acerbi and Tasche [2] or Proposition 2 .

Here we define a family of risk measures which verifies all axioms of coherence and generalizes the expected shortfall. For any $n \geq 1$ and $0 \leq t \leq 1$, let us write

$$R_{n;t}(X) := -\frac{1}{t}\underline{GL}_n(X;t) = -\frac{1}{t}\int_0^t F_X^{-1}(1-(1-s)^{1/n})\mathrm{d}s\,. \tag{1}$$

By Proposition 2 this is a coherent risk measure and as we shall see in the following paragraphs it admits a very special representation.

Spectral Risk Measures

Acerbi [1] showed that given a function $\phi \in \mathscr{L}^1([0,1])$ which is positive, decreasing and with unit norm, i.e., $\|\phi\| = \int_0^1 |\phi(p)|\,\mathrm{d}p = 1$, then the functional

$$M_\phi(X) = -\int_0^1 F_X^{-1}(p)\,\phi(p)\,\mathrm{d}p$$

is a coherent risk measure. These coherent risk measures are termed *spectral risk measures* and the function ϕ is called their *risk aversion function.* Spectral risk measures are *law-invariant*, that is, they only depend on the distribution of X, not on its particular realization.

Note that the $\alpha-$expected shortfall can be retrieved in the context of spectral measures by noting that $ES_\alpha \equiv M_{\phi_\alpha}$ for $\phi_\alpha(p) = \frac{1}{\alpha}\mathbf{1}_{\{0\le p\le\alpha\}}$ where $\mathbf{1}_A$ is the indicator function of a set A.

Let us consider the risk measure $R_{n;t}$ defined in (1). Clearly, writing $t_n^* = 1-(1-t)^{1/n}$, we have that

$$R_{n;t}(X) = -\frac{1}{t}\int_0^{t_n^*} F_X^{-1}(p)n(1-p)^{n-1}\,\mathrm{d}p = -\int_0^1 F_X^{-1}(p)\phi_{n;t}(p)\,\mathrm{d}p$$

by defining $\phi_{n;t}(p) = t^{-1}n(1-p)^{n-1}\mathbf{1}_{\{0\le p\le t_n^*\}}$ which can be easily seen to be positive, decreasing and satisfying the normalization condition $\|\phi\| = 1$. Hence $R_{n;t}$ is a spectral risk measure for every $n \ge 1$ and $0 \le t \le 1$ with risk aversion function $\phi_{n;t}$. Note that $ES_\alpha \equiv R_{1;\alpha}$.

Remark 2. We cannot obtain a coherent risk measure by applying a similar construction to the dual GL. Nevertheless, we can make use of its subadditivity (*5.* in Proposition 2) in order to define a general deviation in terms of it, see Section 5.

Stochastic Orderings

Bäurle and Müller [4] proved that given two random portfolios X and Y that are ordered in terms of the increasing concave stochastic order as $X \le_{\mathrm{icv}} Y$, then for any law-invariant coherent risk measure ρ it holds that $\rho(Y) \le \rho(X)$.

Here we have a family of stochastic orderings, all of them weaker than the increasing concave one and for each of them, a family of coherent risk measures such that if $X \le_{\mathrm{icv}_n} Y$, then $R_{m;t}(Y) \le R_{m;t}(X)$ for all $m \ge n$.

5 General Deviation Measures

A generalization of the standard deviation was introduced in Rockafellar *et al.* [8] by means of the following definition. Consider the space $\mathscr{L}^2(\Omega)$ of all functions which have finite $\mathscr{L}^2$-norm. A *general deviation* is a functional $D : \mathscr{L}^2(\Omega) \to [0,\infty]$ satisfying:

D1 $D(X) \ge 0$ for all X, with $D(X) > 0$ for nonconstant X ;
D2 $D(X+c) = D(X)$ for all X and $c \in \mathbb{R}$;
D3 $D(0) = 0$, and $D(\lambda X) = \lambda D(X)$ for all X and all $\lambda > 0$;
D4 $D(X+Y) \le D(X) + D(Y)$ for all X and Y .

The class of general deviations contains the standard deviation $\sigma(X)$ and standard upper and lower semideviations.

Rockafellar *et al.* [8] proved that one can build general deviations from coherent risk measures as long as they are *strictly expectation bounded*, that is, $\rho(X) > \mathbb{E}(-X)$ for all nonconstant X by taking $D(X) = \rho(X) + \mathbb{E}X$. In this setting, for $0 < t < 1$, we define

$$D_{n;t}(X) := R_{n;t}(X) + \mathbb{E}X = -\frac{1}{t}\underline{GL}_n(X;t) + \mathbb{E}X\,. \tag{2}$$

Essentially, we have added a coherent risk measure on X, $R_{n;t}(X)$, and a coherent risk measure on $-X$, minus its expectation. As long as the positivity (**D1**) is guaranteed, given ρ_1 and ρ_2 coherent risk measures, then $\rho_1(X) + \rho_2(-X)$ is a general deviation. In our current setting, we will add a coherent risk measure on $X_{1:n_1}$ plus a coherent risk measure on $-(X_{1:n_2}) = -X_{n_2:n_2}$ in order to obtain the following general deviation

$$D_{n_1,n_2;t_1,t_2}(X) := \frac{1}{t_2}\overline{GL}_{n_2}(X;t_2) - \frac{1}{t_1}\underline{GL}_{n_1}(X;t_1)\,. \tag{3}$$

Finally, we define a general deviation inspired by the *Gini mean difference of* X which amounts to the area between the GL curve of X and its dual

$$\begin{aligned} D_{n_1,n_2}(X) :&= \mathrm{vol}\Big(\big\{(t,x)\,:\,0 \le t \le 1,\,\underline{GL}_{n_1}(X;t) \le x \le \overline{GL}_{n_2}(X;t)\big\}\Big) \\ &= \int_0^1 \Big(\overline{GL}_{n_2}(X;t) - \underline{GL}_{n_1}(X;t)\Big)\mathrm{d}t\,. \end{aligned} \tag{4}$$

The functionals defined in (2), (3) and (4) are deviation measures as can easily be seen from Proposition 2.

Stochastic Orderings

If a general deviation can be decomposed as $D(X) = \rho_1(X) + \rho_2(-X)$ for ρ_1, ρ_2 law-invariant coherent risk measures, then whenever $X + c \le_{\mathrm{cx}} Y$ for some $c \in \mathbb{R}$ (classically referred to as *dilation order*), we have $D(X) \le D(Y)$.

If $X \le_{\mathrm{cx}_n} Y$ and $m, n_1, n_2 \ge n$, then:

- $D_{m;t}(X) \le D_{m;t}(Y)$ for any $0 \le t \le 1$;
- $D_{n_1,n_2;t_1,t_2}(X) \le D_{n_1,n_2;t_1,t_2}(Y)$ for any $0 \le t_1, t_2 \le 1$;
- $D_{n_1,n_2}(X) \le D_{n_1,n_2}(Y)$.

6 Conclusions and Future Research

The new stochastic order relations that have been defined here are weaker than the classical variability ones. This means that they allow us to compare more distributions than the classical orderings. As the index n grows larger in the relation $\le_{\mathrm{icv}_n}$, a stronger focus is put on the lower tail of the distribution, which is the crucial one when measuring risks. We are studying under what conditions a risk measure is consistent with $\le_{\mathrm{icv}_n}$.

We plan to work in a multivariate generalization of the current constructions and results. The n-GL curve studied here (as well as its dual) can be obtained as part of the boundary of the (set-valued) expectation of a certain random set built using a similar construction to the one that leads to the *lift zonoid*, see Mosler [6]. The multivariate structures that we obtain this way can be applied to measure risks of vector portfolios in the framework described by Cascos and Molchanov [5].

Acknowledgements. This research was initiated when Ignacio Cascos was working with Ilya Molchanov at the University of Berne, thanks are due to him.

The first author acknowledges the support of the Spanish Ministry of Science and Innovation under grants MTM2008-01519 and ECO2008-05080. The second author acknowledges the partial support of the Centro de Matemática da Universidade do Porto (CMUP), financed by FCT (Portugal) through the programmes POCTI and POSI, with national and European structural funds.

References

1. Acerbi, C.: Spectral measures of risk: A coherent representation of subjective risk aversion. J. Bank Financ. 26, 1505–1518 (2002)
2. Acerbi, C., Tasche, D.: On the coherence of expected shortfall. J. Bank Financ. 26, 1487–1503 (2002)
3. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Financ. 9, 203–228 (1999)
4. Bäuerle, N., Müller, A.: Stochastic orders and risk measures: consistency and bounds. Insur. Math. Econ. 38, 132–148 (2006)
5. Cascos, I., Molchanov, I.: Multivariate risks and depth-trimmed regions. Financ. Stoch. 11, 373–397 (2007)
6. Mosler, K.: Multivariate Dispersion, Central Regions and Depth. In: The Lift Zonoid Approach. Lecture Notes in Statistics, vol. 165. Springer, Berlin (2002)
7. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley Series in Probability and Statistics. Wiley, New York (2002)
8. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Generalized Deviations in Risk Analysis. Financ. Stoch. 10, 51–74 (2006)

Likelihood in a Possibilistic and Probabilistic Context: A Comparison

Giulianella Coletti, Davide Petturiti, and Barbara Vantaggi

Abstract. We provide a comparison between a probabilistic and a possibilistic likelihood both as point and set functions.

1 Introduction

We consider conditional probabilities and T-conditional possibilities (where T stands either for min or any strict t-norm). We focus on likelihood functions, regarded as coherent probability or possibility assessment on a class of conditional events $\{E|H_i\}$, with $\{H_i\}$ a finite partition of the sure event Ω. Then, we characterize their coherent extensions on the conditional events $\{E|H\}$, with H belonging to the algebra $\mathscr{H}$ spanned by the H_i's.

Our aim is to give, from a syntactic point of view, a thorough comparison of probabilistic and possibilistic likelihoods both as point and set functions. The interest arises from "bayesian-like" inferential situations where the available information is expressed by different uncertainty measures; for instance, when the prior is a possibility assessment (possibly obtained as supremum of a class of probabilities, see e.g. [1, 5, 8, 9]) and the likelihood comes from a probabilistc data base.

As the point function likelihood is concerned, we find that, from a syntactic point of view, any possibilistic likelihood is also a probabilistic likelihood, and vice versa. Moreover, both conditional possibility and conditional probability, regarded as set functions, are capacities if and only if they are not

Giulianella Coletti and Davide Petturiti
Dept. of Matematica e Informatica, University of Perugia, 06100 Perugia, Italy
e-mail: coletti,davide.petturiti@dmi.unipg.it

Barbara Vantaggi
Dept. of Metodi e Modelli Matematici per le Scienze Applicate,
The Sapienza University of Rome, 00185 Rome, Italy
e-mail: vantaggi@dmmm.uniroma1.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 89–96.
springerlink.com

necessarily normalized possibilities (they are normalized if and only if at least in a point the likelihood is equal to one). An interesting difference is instead the following: in probabilistic setting no kind of monotonicity is required, while in the possibilistic one there is a local form of monotonicity (i.e. it is monotone on the elements of a suitable partition of the algebra).

2 Coherent Conditional Possibility Assessments

The concept of coherence, introduced by de Finetti in probability theory (see [7]), has a fundamental role in managing partial assessments of an uncertainty measure. In fact, coherence is a tool to check consistency, with respect to a specific measure, of a function defined on an arbitrary set of events, and to extend it to new (conditional) events, maintaining consistency.

In this paper we refer to coherent conditional probabilities (see, for instance, [3]) and coherent T-conditional possibility (with T triangular norm) starting from the following definition of T-conditional possibility (see [2]):

Definition 1. *Let $\mathscr{F} = \mathscr{B} \times \mathscr{H}$ be a set of conditional events with $\mathscr{B}$ a Boolean algebra and $\mathscr{H}$ an additive set (i.e. closed with respect to finite logical sums), such that $\mathscr{H} \subseteq \mathscr{B} \setminus \{\emptyset\}$. Let T be a t-norm, a function $\Pi : \mathscr{F} \to [0,1]$ is a T-conditional possibility if it satisfies the following properties:*

1. *$\Pi(E|H) = \Pi(E \wedge H|H)$, for every $E \in \mathscr{B}$ and $H \in \mathscr{H}$;*
2. *$\Pi(\cdot|H)$ is a possibility, for any $H \in \mathscr{H}$;*
3. *$\Pi(E \wedge F|H) = T(\Pi(E|H), \Pi(F|E \wedge H))$, for any $H, E \wedge H \in \mathscr{H}$ and $E, F \in \mathscr{B}$.*

An assessment Π on an arbitrary set $\mathscr{E}$ of conditional events is a coherent T-conditional possibility if (and only if) Π is a restriction of a T-conditional possibility (in the sense of Definition 1) defined on $\mathscr{F} = \mathscr{B} \times \mathscr{H} \supseteq \mathscr{E}$.

We recall a characterization of a coherent T-conditional possibility assessment given in [6]:

Theorem 1. *Let $\mathscr{E} = \{E_1|H_1, ..., E_n|H_n\}$ be an arbitrary set of conditional events, and $\mathscr{C}_0$ denotes the set of atoms spanned by $\{E_1, H_1, ..., E_n, H_n\}$.*
For a real function $\Pi : \mathscr{E} \to [0,1]$, the following statements are equivalent:

a) Π is a coherent T-conditional possibility assessment on $\mathscr{E}$;
b) there exists a sequence of compatible systems S_α ($\alpha = 0, ..., k$), with unknowns $x_r^\alpha = \Pi_\alpha(C_r) \geq 0$ for $C_r \in \mathscr{C}_\alpha$,

$$S_\alpha \begin{cases} \max\limits_{C_r \subseteq E_i \wedge H_i} x_r^\alpha = T(\Pi(E_i|H_i), \max\limits_{C_r \subseteq H_i} x_r^\alpha) & \textit{if } \max\limits_{C_r \subseteq H_i} \mathbf{x}_r^{\alpha-1} < 1 \\ x_r^\alpha \geq \mathbf{x}_r^{\alpha-1} & \textit{if } C_r \in \mathscr{C}_\alpha \\ \mathbf{x}_r^{\alpha-1} = T(x_r^\alpha, \max\limits_{C_j \in \mathscr{C}_\alpha} \mathbf{x}_j^{\alpha-1}) & \textit{if } C_r \in \mathscr{C}_\alpha \\ \max\limits_{C_r \in \mathscr{C}_\alpha} x_r = 1 & \end{cases} \tag{1}$$

with $\alpha = 0,...,k$, *where* $\mathbf{x}^{\alpha}$ *(with r-th component* $\mathbf{x}_r^{\alpha}$*) is the solution of the system* S_{α} *and* $\mathscr{C}_{\alpha} = \{C_r \in \mathscr{C}_{\alpha-1} : \mathbf{x}_r^{\alpha-1} < 1\}$, *moreover* $\mathbf{x}_r^{-1} = 0$ *for any* C_r *in* $\mathscr{C}_o$.

We recall that, in possibility theory as well as in probability theory, coherence assures the extension of any (coherent) assessment to new events, by preserving coherence (see [6]). In particular, the coherent extension on any conditional event lays on a closed interval.

3 Likelihood as Point and Set Function

This session is devoted to a comparative analysis of likelihood in probabilistic and possibilistic framework. By Theorem 1 and by an analogous characterization (see e.g. [3]) of coherent conditional probability assessments the following result easily follows:

Theorem 2. *Let* $\mathscr{C} = \{E|H_i\}_{i=1,...,n}$, *be a set of conditional events, with* $\mathscr{I} = \{H_i\}_{i=1,...,n}$ *a partition of* Ω. *For any function* $f : \mathscr{C} \rightarrow [0,1]$ *such that*

$$f(E|H_i) = 0 \text{ if } E \wedge H_i = \emptyset \text{ and } f(E|H_i) = 1 \text{ if } H_i \subseteq E \tag{2}$$

the following statements hold:
i) f is a coherent conditional probability,
ii) f is a coherent T-conditional possibility.

Proof. Condition i) has been proved in [3]. To prove ii) let us consider the characterization of coherent T-conditional possibility given in Theorem 1. By the incompatibility of the events H_i, the equations of the system S_0 have different unknowns (each of them is linked only with the last equation), and so the system S_0 admits a solution assigning possibility 1 to each conditioning event H_i. Then, the assessment f is a coherent T-conditional possibility. □

The above theorem shows a syntactical coincidence between probabilistic and possibilistic (point) likelihood, so this allows to regard a probabilistic likelihood as a possibilistic one and vice versa without introducing inconsistency. Moreover, the above result puts in evidence that (in both contexts) no significant property characterizes likelihood as point function.

We are now interested on studying properties of aggregated likelihoods, that is all the coherent extensions of the assessment $f(E|H_i)$, $(H_i \in \mathscr{I})$ to the events $E|K$, with K any logical sum of the events H_i.

Theorem 3. *Let* $\mathscr{C}, \mathscr{I}$ *and* f *be as in Theorem 2 and let* $\mathscr{A}$ *be the algebra spanned by by* $\mathscr{I}$ *and* $\mathscr{H} = \mathscr{A} \setminus \emptyset$. *For any extension* g *of* f *on* $\{E\} \times \mathscr{H}$, *which is either a coherent conditional probability or a coherent T-conditional possibility assessment, the following condition holds for every* $H \in \mathscr{H}$:

$$\min_{H_i \subseteq H} f(E|H_i) \leq g(E|H) \leq \max_{H_i \subseteq H} f(E|H_i). \tag{3}$$

Proof. For coherent conditional probability the condition (3) is proved in [3]. Let f be a coherent T-conditional possibility assessment, then there is an extension $g = \Pi$ on $\mathscr{B} \times \mathscr{H}$, where $\mathscr{B}$ is the algebra generated by E and $\mathscr{H}$, and $\Pi(E|H) = \max_{H_i \subseteq H} T(\Pi(E|H_i), \Pi(H_i|H)) \leq \max_{H_i \subseteq H} \Pi(E|H_i)$. Moreover, by distributivity of maximum with respect to any t-norm T we have:

$$\Pi(E|H) = \max_{H_i \subseteq H} T(\Pi(E|H_i), \Pi(H_i|H)) \geq \max_{H_i \subseteq H} T(\beta, \Pi(H_i|H)) =$$

$$T\big(\beta, \max_{H_i \subseteq H} \Pi(H_i|H)\big) = T\big(\beta, 1\big) = \beta$$

where $\beta = \min_{H_i \subseteq H} \Pi(E|H_i)$. □

Remark 1. By condition (3) it follows immediately that both probability and possibility aggregated likelihood can be monotone, with respect to $\subseteq$, only if the extension can be obtained, for every H, as $\max_{H_i \subseteq H} f(E|H_i)$.

The following Theorem 4 assures that this (particular) extension is coherent. So we can conclude that both probability and possibility conditional probabilities, when they are regarded as function of the conditioning events, are capacities if and only if they are obtained through the maximum.

Theorem 4. *Let $\mathscr{C}$, $\mathscr{I}$, f and $\mathscr{H}$ be as in Theorem 3. For any extension g of f on $\{E\} \times \mathscr{H}$, and such that*

$$g(E|H \vee K) = \max\{g(E|H), g(E|K)\} \tag{4}$$

for every $H, K \in \mathscr{H}$ the following conditions hold:
i) g is a coherent conditional probability,
ii) g is a coherent T-conditional possibility.

Proof. The proof of *i)* is in [3]. To prove *ii)* consider as solution of system S_0 in Theorem 1 the possibility $\Pi_0(H_i) = 1$, for any $H_i \in \mathscr{C}$. □

Remark 2. Note that in Theorem 4 we state that g is a coherent T-conditional possibility, but $g(E|.)$ is not necessarily a possibility even if condition (4) holds for every $H, K \in \mathscr{H}$. In fact $g(E|\Omega)$ can be strictly less than 1.

Actually, $g(E|\Omega)$ is 1 if and only if there is H_i with $f(E|H_i) = 1$: this requirement could seem natural since it claims the existence of an event H_i supporting the evidence E.

In order to deepen the comparison of possibilistic and probabilistic aggregated likelihoods we need to introduce the notion of *scale* and then a relevant local form of monotonicity.

Definition 2. *Let $\mathscr{I} = \{H_1, ..., H_n\}$ be a partition of the sure event and $\mathscr{H} = \mathscr{A} \setminus \emptyset$ with $\mathscr{A}$ the algebra spanned by by $\mathscr{I}$. A scale of $\mathscr{H}$ is any partition $\{\mathscr{H}^0, ..., \mathscr{H}^k\}$ of $\mathscr{H}$, such that every $\mathscr{H}^i$ (with $i = 0, ..., k$) is an additive set and it contains at least one element $H_j \in \mathscr{I}$ and any $K \supseteq H_j$, with $K \in \mathscr{H} \setminus \bigcup_{k<i} \mathscr{H}^k$.*

Definition 3. *Let $\mathscr{H}$ as in Definition 2, a function $\varphi : \mathscr{H} \to [0,1]$ is scale monotone, with respect to a scale $\mathscr{H}^0,...,\mathscr{H}^k$ of $\mathscr{H}$, if φ, restricted to any $\mathscr{H}^i$ with $i = 0,...,k$, is monotone, with respect to the inclusion $\subseteq$.*

We give an example of a scale and a scale monotone function:

Example 1. Let $\mathscr{I} = \{H_1,H_2,H_3\}$ be a partition of the sure event and denote by $\mathscr{H}$ the algebra spanned by $\mathscr{I}$ less the impossible event. Consider the set $\mathbb{K} = \{\mathscr{H}^0,\mathscr{H}^1\}$ with $\mathscr{H}^0 = \{H_2,H_1 \vee H_2,H_2 \vee H_3,\Omega\}$ and $\mathscr{H}^1 = \{H_1,H_3,H_1 \vee H_3\}$. It is easy to check that $\mathbb{K}$ is a scale.

Consider now the assessment $\varphi(H_1) = 0.3$, $\varphi(H_2) = 0.5$, $\varphi(H_3) = 0.8$, $\varphi(H_1 \vee H_2) = 0.5$, $\varphi(H_1 \vee H_3) = 0.8$, $\varphi(H_2 \vee H_3) = 0.7$ and $\varphi(\Omega) = 0.75$.

For every $K \in \mathscr{H}$, condition (3) holds, so $\varphi(\cdot)$ is scale monotone with respect to $\mathbb{K}$.

Theorem 5. *Let $\mathscr{A}$ be a finite algebra, $\mathscr{H} = \mathscr{A} \setminus \emptyset$ and $\varphi(\cdot) = \Pi(E|\cdot)$ be a coherent T-conditional possibility on $\{E\} \times \mathscr{H}$. Then, there exists a scale $\{\mathscr{H}^0,...,\mathscr{H}^k\}$ of $\mathscr{H}$ such that φ is scale monotone with respect to it.*

Proof. If $\Pi(E|\cdot)$ is a coherent T-conditional possibility, then there is an extension (that we continue to denote by Π) on $\mathscr{B} \times \mathscr{H}$, where $\mathscr{B}$ is the algebra generated by E and $\mathscr{H}$. Put $\mathscr{H}^0 = \{H \in \mathscr{H} : \Pi(H|\Omega) = 1\}$ and define, for any $j = 1,...,k$, the set $\mathscr{H}^j = \{H \in \mathscr{H} : \Pi(H|H_0^{j-1}) = 1\}$ where $\mathscr{C}^{j+1} = \mathscr{H} \setminus \bigcup_{i=0}^{j}(\mathscr{H}^i)$ and $H_0^j = \bigvee_{H \in \mathscr{C}^j} H$. The class $\{\mathscr{H}^0,...,\mathscr{H}^k\}$ is a partition of $\mathscr{H}$ and each element is an additive set. Let $\mathscr{I} = \{H_1,...,H_n\}$ be the set of atoms of $\mathscr{H}$, since $\Pi(\cdot|H_0^{j-1})$ is a possibility, there is at least an atom H_i of $\mathscr{I}$ such that $\Pi(H_i|H_0^{j-1}) = 1 = \Pi(K|H_0^{j-1})$ for any $K \supseteq H_i$. Then, $\{\mathscr{H}^0,...,\mathscr{H}^k\}$ is a scale and, by construction, Π is monotone with respect to it: in fact, for any $H, H \vee K \in \mathscr{H}^i$ it follows $\Pi(E|H) = T(\Pi(E|H),\Pi(H|H \vee K)) = \Pi(E|H \vee K)$ since $\Pi(H|H \vee K) = 1$. □

We notice that for coherent conditional probabilities a similar result does not hold as the following example shows:

Example 2. Let $\mathscr{I} = \{H_1,H_2,H_3\}$ be a partition of the sure event, and consider the likelihood assessment: $P(E|H_1) = \frac{1}{2}, P(E|H_2) = \frac{1}{4}, P(E|H_3) = \frac{1}{8}$.

It is easy to prove that the following assessment
$P(E|H_1 \vee H_2) = \frac{3}{8}, P(E|H_1 \vee H_3) = \frac{5}{16}, P(E|H_2 \vee H_3) = \frac{3}{16}, P(E|\Omega) = \frac{7}{24}$
is a coherent conditional probability extending the above likelihood (the extension is obtained by giving $P(H_i) = 1/3$, for i=1,...,3).

Nevertheless, this probabilistic aggregated likelihood is not scale monotone. In fact, by definition, the first element $\mathscr{H}_0$ of any scale must contain an atom and all its supersets, so the thesis follows immediately by the following inequalities: $P(E|H_1) > P(E|H_1 \vee H_2)$, $P(E|H_2) > P(E|H_2 \vee H_3)$ and $P(E|H_2 \vee H_3) > P(E|H_1 \vee H_2 \vee H_3)$.

We notice that the necessary conditions given in Theorem 3 and Theorem 5 are not sufficient to characterize possibilistic aggregated likelihood as coherent extensions of a point likelihood, as shown in the following example:

Example 3. Let us consider again the scale $\mathbb{K}$ and the function φ of Example 1. We show that $\varphi(\cdot) = \Pi(E|\cdot)$ cannot be seen as a coherent conditional possibility. Let $C_i = E \wedge H_i$ and $C_{i+3} = E^c \wedge H_i$ $(i = 1, \ldots, 3)$ be the atoms spanned by $\{E, H_1, H_{2,3}\}$ and consider the following system with unknowns $x_r^0 = \Pi(C_r) \geq 0$ for $r = 1, \ldots, 6$

$$S_0^m = \begin{cases} x_1^0 = \min\{0.3, \max\{x_1^0, x_2^0\}\} \\ x_3^0 = \min\{0.5, \max\{x_3^0, x_4^0\}\} \\ x_5^0 = \min\{0.8, \max\{x_5^0, x_6^0\}\} \\ \max\{x_1^0, x_3^0\} = \min\{0.5, \max\{x_1^0, x_2^0, x_3^0, x_4^0\}\} \\ \max\{x_1^0, x_5^0\} = \min\{0.8, \max\{x_1^0, x_2^0, x_5^0, x_6^0\}\} \\ \max\{x_3^0, x_5^0\} = \min\{0.7, \max\{x_3^0, x_4^0, x_5^0, x_6^0\}\} \\ \max\{x_1^0, x_3^0, x_5^0\} = \min\{0.75, \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\}\} \\ \max\{x_1^0, x_2^0, x_3^0, x_4^0, x_5^0, x_6^0\} = 1 \end{cases}$$

The systems S_0^m admits no solution: in fact, only x_2^0, x_4^0 and x_6^0 can assume value 1, but the seventh equation forces to be $x_2^0 < 1$ and $x_6^0 < 1$ in the fifth one, while the sixth equation implies $x_4^0 < 1$ in the seventh one.

Similarly, it is possible to prove that φ is not a coherent T-conditional possibility, for any strict t-norm T.

The following result characterizes all the coherent extensions of a likelihood $f(E|\cdot)$ as coherent T-conditional possibility (with T either strict t-norm or $\min$). This allows an easy comparison.

Theorem 6. *Let $\mathscr{C}$, $\mathscr{I}$, f and $\mathscr{H}$ be as in Theorem 3. For any extension Π of f on $\{E\} \times \mathscr{H}$, and for every strict t-norm T the following two statements are equivalent:*

i) Π is a coherent T-conditional possibility extending f to $\mathscr{K} = \{E\} \times \mathscr{H}$;

ii) there exist a class of subfamilies $\mathscr{H}_\alpha$ $(\alpha = 0, \ldots, k \leq n-1)$, with $\mathscr{H}_\alpha \supset \mathscr{H}_{\alpha+1}$, and sets of coefficients $\lambda_i^\alpha \geq 0$ with $\max_i \lambda_i^\alpha = 1$, $\lambda_i^{-1} = 0$ for any i, and $H_i \in \mathscr{H}_\alpha$ if and only if $\lambda_i^{\alpha-1} = 0$, such that for every $H \in \mathscr{H}$ the value $x = \Pi(E|H)$ is a solution of

$$T(x, \max_{H_i \subseteq H} \lambda_i^\alpha) = \max_{H_i \subseteq H} T(\lambda_i^\alpha, \Pi(E|H_i)) \tag{5}$$

for every α such that $H_i \in \mathscr{H}_\alpha$ when $H_i \subseteq H$.

Proof. Since f is a coherent T-conditional possibility then there is at least a T-conditional possibility Π on $\mathscr{B} \times \mathscr{H}$ extending it. So, by Theorem 1 there exists a sequence of compatible systems and for any $H \in \mathscr{H}$ there is a j such that $\max_{C_r \subseteq H} \mathbf{x}_r^j = 1$ and $\Pi(E|H)$ is solution of the following equation for any $i = 0, \ldots, j$

$$T(\Pi(E|H), \max_{C_r \subseteq H} \mathbf{x}_r^i) = \max_{C_r \subseteq E \wedge H} \mathbf{x}_r^i.$$

That implies the existence of λ_r^i with $H_r \in \mathscr{H}_i$ such that

$$T(\Pi(E|H), \max_{H_r \subseteq H} \lambda_r^i(H)) = \max_{H_r \subseteq H} T(\pi(E|H_r), \lambda_r^i). \qquad \square$$

Note that in the case of product t-norm the aggregate likelihood is the maximum of a **weighted combination** of the $\Pi(E|H_i)$'s (with weights λ_i^j) over the maximum of the weights.

A similar characterization can be given for the t-norm minimum:

Theorem 7. *Let $\mathscr{C}$, $\mathscr{I}$, f and $\mathscr{H}$ be as in Theorem 3. For any extension Π of f on $\{E\} \times \mathscr{H}$, the following two statements are equivalent:*

i) Π is a coherent conditional possibility extending f to $\mathscr{K} = \{E\} \times \mathscr{H}$;

ii) there exist a class of subfamilies $\mathscr{H}_\alpha$ $(\alpha = 0, \dots, k \leq n-1)$, with $\mathscr{H}_\alpha \supset \mathscr{H}_{\alpha+1}$, and sets of coefficients $\lambda_i^\alpha \geq 0$ with $\max_i \lambda_i^\alpha = 1$, $\lambda_i^{-1} = 0$ for any i, and $H_i \in \mathscr{H}_\alpha$ if and only if $\lambda_i^{\alpha-1} \leq f(E|H_i)$, such that for every $H \in \mathscr{H}$ the value $x = \Pi(E|H)$ is a solution of

$$\min\{x, \max_{H_i \subseteq H} \lambda_i^\alpha\} = \max_{H_i \subseteq H} \min\{\lambda_i^\alpha, \Pi(E|H_i)\} \tag{6}$$

for every α such that $H_i \in \mathscr{H}_\alpha$ when $H_i \subseteq H$.

We recall the quoted characterization for the probabilistic aggregated likelihood.

Theorem 8. *Let $\mathscr{C}$, $\mathscr{I}$, f and $\mathscr{H}$ be as in Theorem 3. For any extension P of f on $\{E\} \times \mathscr{H}$, the following two statements are equivalent:*

i) P is a coherent conditional probability extending f to $\mathscr{K} = \{E\} \times \mathscr{H}$;

ii) there exist a class of subfamilies $\mathscr{H}_\alpha$ $(\alpha = 0, \dots, k \leq n-1)$, with $\mathscr{H}_\alpha \supset \mathscr{H}_{\alpha+1}$, and sets of coefficients $\lambda_i^\alpha \geq 0$ with $\max_i \lambda_i^\alpha = 1$, $\lambda_i^{-1} = 0$ for any i, and $H_i \in \mathscr{H}_\alpha$ if and only if $\lambda_i^{\alpha-1} = 0$, such that for every $H \in \mathscr{H}$ the value $x = P(E|H)$ is a solution of

$$x \sum_{H_i \subseteq H} \lambda_i^\alpha = \sum_{H_i \subseteq H} \lambda_i^\alpha P(E|H_i) \tag{7}$$

for every α such that $H_i \in \mathscr{H}_\alpha$ when $H_i \subseteq H$.

Note that while in the probabilistic context the aggregated likelihood is obtained as **weighted mean** of the $P(E|H_i)$'s, in the possibilistic one the aggregated likelihood is obtained as solution of equation (5) for strict t-norms and (6) for the minimum. However in the case of product t-norm the aggregate likelihood, as recalled above, is the maximum of a weighted combination of the $\Pi(E|H_i)$'s. In both probabilistic and possibilistic cases weights equal to zero or one are allowed and in the case where the likelihood is scale monotone with respect to a scale with the maximum number n of elements (n equal to the number of elementary events), then the extensions coincides.

Remark 3. Notice that if we can assign $\lambda_i^0 = 1$ for every i, then we obtain only one class of λ_i^α, and we get a coherent T-conditional possibility taking only the values assumed by the likelihood. Moreover, the above aggregated likelihood is obtained (for every α) by giving value 1 to only one λ_{i*}^α, and value 0 to all others. So in this case the aggregated likelihood is also a coherent conditional probability.

A particular case of this situation is when at any step α the value 1 corresponds, for $E|H_i \in \mathscr{H}_\alpha$, to the *maximum* (or *minimum*) value of $\Pi(E|H_i)$.

4 Conclusions

We compare the likelihood function in probabilistic and possibilistic contexts both as point and set functions. For this aim we provide also a characterization of coherent extensions of a possibilistic likelihood. The results can be extended to any conditional decomposable measure (for the definition see e.g. [3]): this extension is immediate for Theorems 2, 3, 4.

References

1. Baudrit, C., Couso, I., Dubois, D.: Joint propagation of probability and possibility in risk analysis: Towards a formal framework. Intern. J. Approx. Reason. 45, 82–105 (2007)
2. Bouchon-Meunier, B., Coletti, G., Marsala, C.: Independence and possibilistic conditioning. Ann. Math. Artif. Intell. 35, 107–124 (2002)
3. Coletti, G., Scozzafava, R.: Probabilistic logic in a coherent setting. Kluwer Academic Publisher, Dordrecht (2002)
4. Coletti, G., Scozzafava, R.: Conditional Probability and Fuzzy Information. Comput. Statist. Data Anal. 51, 115–132 (2006)
5. Coletti, G., Scozzafava, R., Vantaggi, B.: Possibility measures in probabilistic inference. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Handling Variability and Imprecision. Advances in Soft Computing., vol. 48, pp. 51–58. Springer, Heidelberg (2008)
6. Coletti, G., Vantaggi, B.: T-conditional possibilities: coherence and inference. Fuzzy Sets Syst. 160, 306–324 (2009)
7. de Finetti, B.: Sul significato soggettivo della probabilità. Fund.Math. 17, 298–329 (1931)
8. Dubois, D., Prade, H.: Qualitative possibility theory and its probabilistic connections. In: Grzegorzewski, P., Hryniewicz, O., Gil, M.A. (eds.) Soft Methods in Probability, Statistics and Data Analysis. Advances in Soft Computing, vol. 16, pp. 3–26. Physica Verlag, Heidelberg (2002)
9. Moral, S.: Constructing a possibility distribution from a probability distribution. In: Jones, A., Kauffmann, A., Zimmermann, H.J. (eds.) Fuzzy sets theory and Applications, pp. 51–60. D. Reidel, Dordrecht (1986)

Nonparametric Predictive Inference for Order Statistics of Future Observations

Frank P.A. Coolen and Tahani A. Maturi

Abstract. Nonparametric predictive inference (NPI) is a powerful frequentist statistical framework which uses only few assumptions. Based on a post-data exchangeability assumption, precise probabilities for some events involving one or more future observations are defined, based on which lower and upper probabilities can be derived for all other events of interest. We present NPI for the *r*-th order statistic of *m* future real-valued observations and its use for comparison of two groups of data.

1 Introduction

Nonparametric predictive inference (NPI) [3, 5] is a statistical framework which uses few modelling assumptions, with inferences explicitly in terms of future observations. For real-valued random quantities attention has thus far been mostly restricted to a single future observation, although multiple future observations have been considered for some NPI methods for statistical process control [1, 2]. For Bernoulli quantities, NPI has also been presented for $m \geq 1$ future observations [4], with explicit study of the influence of the choice of m for comparison of groups of proportions data [6].

In this paper, we consider m future real-valued observations, given n observations, and as main contribution we focus on the r-th ordered observation of these m future observations, including comparison of two groups of data via comparison of their corresponding r-th ordered future observations. Without making further assumptions, these inferences require the use of lower and upper probabilities for several events of interest, as such this work fits in the theory of imprecise probability [12] and interval probability [13].

Frank P.A. Coolen and Tahani A. Maturi
Dept. of Mathematical Sciences, Durham University, Durham, United Kingdom
e-mail: frank.coolen@durham.ac.uk,tahani.maturi@durham.ac.uk

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 97–104.
springerlink.com

Assume that we have real-valued ordered data $x_1 < x_2 < \ldots < x_n$, with $n \geq 1$. We assume that ties do not occur, in Example 2 in Section 3 we explain how to deal with ties. For ease of notation, define $x_0 = -\infty$ and $x_{n+1} = \infty$. The n observations partition the real-line into $n+1$ intervals $I_j = (x_{j-1}, x_j)$ for $j = 1, \ldots, n+1$. If we wish to allow ties between past and future observations explicitly, we could use closed intervals $[x_{j-1}, x_j]$ instead of these open intervals I_j, the difference is rather minimal and to keep presentation easy we have opted not to do this here. We are interested in $m \geq 1$ future observations, X_{n+i} for $i = 1, \ldots, m$. We link the data and future observations via Hill's assumption $A_{(n)}$ [10], or, more precisely, via $A_{(n+m-1)}$ (which implies $A_{(n+k)}$ for all $k = 0, 1, \ldots, m-2$; we will refer to this generically as 'the $A_{(n)}$ assumptions'), which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. $A_{(n+m-1)}$ implies that all possible orderings of the n data observations and the m future observations are equally likely, where the n data observations are not distinguished among each other, and neither are the m future observations. Let $S_j = \#\{X_{n+i} \in I_j,\ i = 1, \ldots, m\}$, then assuming $A_{(n+m-1)}$ we have

$$P\left(\bigcap_{j=1}^{n+1} \{S_j = s_j\}\right) = \binom{n+m}{n}^{-1} \tag{1}$$

where s_j are non-negative integers with $\sum_{j=1}^{n+1} s_j = m$. Another convenient way to interpret the $A_{(n+m-1)}$ assumption with n data observations and m future observations is to think that n randomly chosen observations out of all $n+m$ real-valued observations are revealed, following which you wish to make inferences about the m unrevealed observations. The $A_{(n+m-1)}$ assumption then implies that one has no information about whether specific values of neighbouring revealed observations make it less or more likely that a future observation falls in between them. For any event involving the m future observations, (1) implies that we can count the number of such orderings for which this event holds. Generally in NPI a lower probability for the event of interest is derived by counting all orderings for which this event has to hold, while the corresponding upper probability is derived by counting all orderings for which this event can hold [3, 5].

NPI is close in nature to predictive inference for the low structure stochastic case as briefly outlined by Geisser [9], which is in line with many earlier nonparametric test methods where the interpretation of the inferences is in terms of confidence intervals. In NPI the $A_{(n)}$ assumptions justify the use of these inferences directly as probabilities. Using only precise probabilities or confidence statements, such inferences cannot be used for many events of interest, but in NPI we use the fact, in line with De Finetti's Fundamental Theorem of Probability [7], that corresponding optimal bounds can be derived for all events of interest [3]. NPI provides exactly calibrated frequentist inferences [11], and it has strong consistency properties in theory of interval probability [3]. In NPI the n observations are explicitly used through the $A_{(n)}$ assumptions, yet as there is no use of conditioning as in the Bayesian framework, we do not use an explicit notation to indicate this use of the data. It is important to emphasize that there is no assumed population from which the n observations were randomly drawn, and hence also no assumptions on the sampling process. NPI is totally based on the $A_{(n)}$ assumptions, which however should be considered with care as they imply e.g. that the specific ordering in which the data appeared is irrelevant, so accepting $A_{(n)}$ implies an exchangeability judgement for the n observations. It is attractive that the appropriateness of this approach can be decided upon after the n observations have become available. NPI is always in line with inferences based on empirical distributions, which is an attractive property when aiming at objectivity [5].

2 NPI for Order Statistics

Let $X_{(r)}$, for $r=1,\ldots,m$, be the r-th ordered future observation, so $X_{(r)}=X_{n+i}$ for one $i=1,\ldots,m$ and $X_{(1)}<X_{(2)}<\ldots<X_{(m)}$. The following probabilities are derived by counting the relevant orderings, and hold for $j=1,\ldots,n+1$, and $r=1,\ldots,m$,

$$P(X_{(r)}\in I_j)=\binom{j+r-2}{j-1}\binom{n-j+1+m-r}{n-j+1}\binom{n+m}{n}^{-1} \tag{2}$$

For this event NPI provides a precise probability, as each of the $\binom{n+m}{n}$ equally likely orderings of n past and m future observations has the r-th ordered future observation in precisely one interval I_j.

As an example, suppose that one is interested in the minimum $X_{(1)}$ of the m future observations. Formula (2) gives $P(X_{(1)}\in I_j)=\binom{n-j+m}{n-j+1}\binom{n+m}{n}^{-1}$, with for example $P(X_{(1)}\in I_1)=\frac{m}{n+m}$. Clearly, the event $X_{(1)}\in I_1$ occurs if the smallest of all $n+m$ observations, so the n data observations and m future observations, is not in the data set, which would occur with probability $\frac{n}{n+m}$. A further special case of interest is $P(X_{(1)}\in I_{n+1})=\binom{n+m}{n}^{-1}$, following from the fact that there is only one ordering for which all n data observations occur before all m future observations.

In theory of mathematical statistics and probability, much attention is paid to limit results. Many popular statistical methods are justified through limit properties, with limits taken with regard to the number n of data observations, leading to 'large-sample' methods that are often applied in cases with relatively small samples without due consideration of the quality of the approximations involved and lacking clear foundational justification. Considering limits for n going to infinity is not very exciting in NPI as one just ends up with the empirical distribution function and corresponding inferences. However, in NPI it might be of some interest to consider the limiting behaviour of the predictive probabilities (2) if m goes to infinity, hence if we consider an ever increasing future. Defining $\theta \in [0,1]$ through the relationship $r = \theta m$ (of course, this only makes sense when θm, and therefore also $(1-\theta)m$, is integer, we only sketch the argument here without giving the detailed mathematical presentation), the following limiting result is easily proven, for $j = 1,\ldots,n+1$,

$$\lim_{m\to\infty} P(X_{(\theta m)} \in I_j) = \binom{n}{j-1} \theta^{j-1}(1-\theta)^{n-j+1} \tag{3}$$

It is important to emphasize the difference with established statistical methods. The θ in (3) is not a characteristic of an assumed population from which the data are sampled, indeed no population assumption is made. Furthermore, (3) is not a probability distribution nor a likelihood function for θ. Instead, θ only serves for notation of this event of interest, and indicates the specific relative (with regard to m) future order statistic of interest. Of course, the actual $A_{(n)}$ assumptions required for this limit imply infinite exchangeability of the future observations, hence De Finetti's Representation Theorem [7] indicates that a parametric representation can be assumed, yet this is different from the explicitly predictive use in NPI, most noticeably through the absence of a probability distribution for θ. The limiting probability (3) can be understood from the consideration that for the event $X_{(\theta m)} \in I_j$ to hold, precisely $j-1$ of the n data observations must be smaller than $X_{(\theta m)}$, but it must be emphasized again that (3) specifies probabilities for $X_{(\theta m)}$, not for any aspect of the observed data for which no concept of randomness, e.g. as following from assumed sampling from a population, is used in NPI. In NPI, the data are given, all randomness is explicitly with regard to the future observations, which nicely reflects where the uncertainty really is in applications.

Analysis of the probability (2) leads to some interesting results, including the obvious symmetry $P(X_{(r)} \in I_j) = P(X_{(m+1-r)} \in I_{n+2-j})$. For all r, the probability for $X_{(r)} \in I_j$ is unimodal in j, with the maximum probability assigned to interval I_{j^*} with $\left(\frac{r-1}{m-1}\right)(n+1) \leq j^* \leq \left(\frac{r-1}{m-1}\right)(n+1)+1$. This carries through to the limiting situation (3), where for given θ the maximum probability is assigned to interval I_{j^*} with $(n+1)\theta \leq j^* \leq (n+1)\theta + 1$. It is worth commenting on extreme values, in particular inference involving $X_{(1)}$ or $X_{(m)}$ for m large compared to the value of n. In these cases, NPI assigns large probabilities to the intervals I_1 or I_{n+1}, respectively, which are outside the range of the observed data and unbounded unless the random quantities of interest are logically bounded (e.g. zero as lower bound for lifetime data). This indicates that, for such inferences, very little can be concluded without further assumptions on the probability masses within these end intervals beyond the observed data. This will be illustrated in the examples in Section 3. There are several inferential problems where one is explicitly interested in such a future order statistic $X_{(r)}$. It may be of explicit interest to compare different groups or treatments by comparing particular future order statistics, this is presented in Section 3.

3 Comparing Two Groups

Suppose we have two independent groups of real-valued observations, X and Y, their ordered observed values are $x_1 < x_2 < \ldots < x_{n_x}$ and $y_1 < y_2 < \ldots < y_{n_y}$. For ease of notation, let $x_0 = y_0 = -\infty$ and $x_{n_x+1} = y_{n_y+1} = \infty$. And let $I^x_{j_x} = (x_{j_x-1}, x_{j_x})$ and $I^y_{j_y} = (y_{j_y-1}, y_{j_y})$. We are interested in $m \geq 1$ future observations from each group (i.e. $m_x = m_y = m$), so in X_{n_x+i} and Y_{n_y+i} for $i = 1, \ldots, m$. We wish to compare the r-th future order statistics from these two groups by considering the event $X_{(r)} < Y_{(r)}$, for which the NPI lower and upper probabilities, based on the $A_{(n_x)}$ and $A_{(n_y)}$ assumptions per group, are derived by

$$\underline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x} < y_{j_y-1}\} P(X_{(r)} \in I^x_{j_x}) P(Y_{(r)} \in I^y_{j_y}) \tag{4}$$

$$\overline{P}(X_{(r)} < Y_{(r)}) = \sum_{j_x=1}^{n_x+1} \sum_{j_y=1}^{n_y+1} \mathbf{1}\{x_{j_x-1} < y_{j_y}\} P(X_{(r)} \in I^x_{j_x}) P(Y_{(r)} \in I^y_{j_y}) \tag{5}$$

where $\mathbf{1}\{E\}$ is an indicator function which is equal to 1 if event E occurs and 0 else. This NPI lower (upper) probability follows by putting all probability masses for $Y_{(r)}$ corresponding to the intervals $I^y_{j_y} = (y_{j_y-1}, y_{j_y})$, $j_y = 1, \ldots, n_y + 1$, to the left (right) end points of these intervals, and by putting all probability masses for $X_{(r)}$ corresponding to the intervals $I^x_{j_x} = (x_{j_x-1}, x_{j_x})$, $j_x = 1, \ldots, n_x + 1$, to the right (left) end points of these intervals. We illustrate this NPI method for comparison of two groups based on the r-th future order statistic in two examples, first a small artificial example followed by one considering a real data set.

Example 1. To get a basic feeling for these inferences, we consider three small artificial data sets (cases A,B,C) as given in Table 1. For $m = 5, 25, 200$, the NPI lower and upper probabilities for the events $X_{(r)} < Y_{(r)}$ for all $r = 1, \ldots, m$ are displayed in Fig. 1, with row 1,2,3 corresponding to cases A,B,C. Actually, the plotted lines per value of r represent the intervals bounded by the corresponding lower and upper probabilities, so the length of each line is the imprecision for that event.

Table 1 Data sets, Example 1

A	X:	1	4							Y:	2	3						
B	X:	1	2	7	8					Y:	3	4	5	6				
C	X:	1	2	3	4	13	14	15	16	Y:	5	6	7	8	9	10	11	12

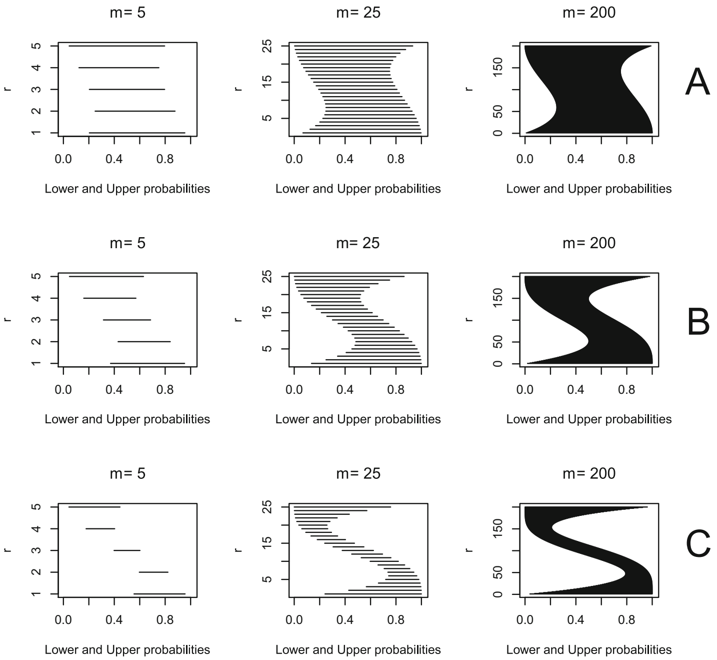


Fig. 1 NPI lower and upper probabilities, Example 1

These results illustrate clearly the effect of increased sample sizes, leading to decreasing imprecision for future order statistics that are most likely to fall within the observed data range. For extreme future order statistics, imprecision remains high as no assumptions are made about the spread of probability mass within any interval $I^x_{j_x}$ or $I^y_{j_y}$, so also not in the end intervals. This makes clear that, without additional assumptions, no strong inferences can be achieved for events involving extreme future order statistics if m is substantially larger than n.

Example 2. We consider the data set of a study of the effect of ozone environment on rats growth [8, p.170]. One group of 22 rats were kept in an ozone environment and the second group of 23 similar rats were kept in an ozone-free environment. Both groups were kept for 7 days and their weight gains are given in Table 2.

The NPI lower and upper probabilities (4) and (5) for the events $X_{(r)} < Y_{(r)}$, $r = 1,\ldots,m$, are displayed in Fig. 2, where the first row gives figures corresponding to the full data for the cases with $m = 5, 25, 200$, while the second row gives the corresponding figures but with the observation -16.9 removed from group Y. This is done as this value could perhaps be considered to be an outlier, hence it might be interesting to see its influence on these inferences. Note that the data for group X and for group Y both contain two tied observations, at -9.0 and 26.0, respectively. As tied observations are within the same group, we just

Table 2 Rats' weight gains data, Example 2

Ozone group (X)						Ozone-free group (Y)					
-15.9	-14.7	-12.9	-9.9	-9.0	-9.0	-16.9	13.1	15.4	17.4	17.7	18.3
6.1	6.6	6.8	7.3	10.1	12.1	19.2	21.4	21.8	21.9	22.4	22.7
14.0	14.3	15.5	15.7	17.9	20.4	24.4	25.9	26.0	26.0	26.6	27.3
28.2	39.9	44.1	54.6			27.4	28.5	29.4	38.4	41.0	

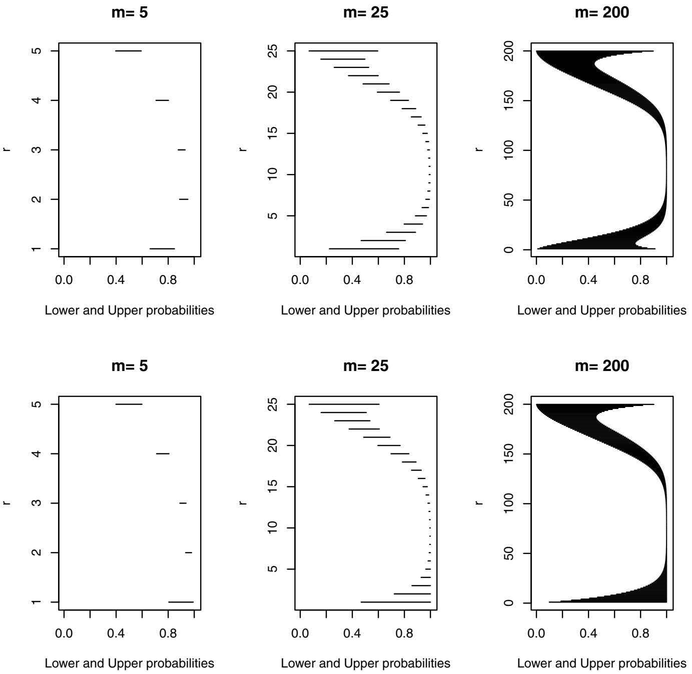


Fig. 2 NPI lower and upper probabilities, Example 2

add a very small amount to one of them, not affecting their rankings within the group nor with the data for both groups combined and therefore not affecting the inferences. This can be interpreted as assuming that these values actually differ in a further decimal, not reported due to rounding. If observations where tied among the two groups, the same breaking of ties could be performed, with the NPI method presented in this paper applied to all possible ways to do so, and the smallest (largest) of the corresponding lower (upper) probabilities for the event of interest would be used as the NPI lower (upper) probability. The possibility to break ties in this manner is an attractive feature of statistical methods using lower and upper probabilities, as it does not require further assumptions for such tied values.

This example shows that these data strongly support the event $X_{(r)} < Y_{(r)}$ for future order statistics that are likely to be in the middle area of the data ranges, with the values of the NPI lower and upper probabilities reflecting the amount of overlap in the observed data for groups X and Y. For extreme future order statistics the imprecision is again very large, and the effect of deleting the smallest Y value from the data has caused quite a difference between the inferences for small values of r, as the lower ends of the plots in rows 1 and 2 in Fig. 2 clearly illustrate.

References

1. Arts, G.R.J., Coolen, F.P.A., van der Laan, P.: Nonparametric predictive inference in statistical process control. Qual. Technol. Quant. Manag. 1, 201–216 (2004)
2. Arts, G.R.J., Coolen, F.P.A.: Two nonparametric predictive control charts. J. Stat. Theory Pract. 2, 499–512 (2008)
3. Augustin, T., Coolen, F.P.A.: Nonparametric predictive inference and interval probability. J. Statist. Plann. Inference 124, 251–272 (2004)
4. Coolen, F.P.A.: Low structure imprecise predictive inference for Bayes' problem. Statist. Probab. Lett. 36, 349–357 (1998)
5. Coolen, F.P.A.: On nonparametric predictive inference and objective Bayesianism. J. Logic Lang. Inform. 15, 21–47 (2006)
6. Coolen, F.P.A., Coolen-Schrijner, P.: Nonparametric predictive comparison of proportions. J. Statist. Plann. Inference 137, 23–33 (2007)
7. De Finetti, B.: Theory of probability, vol. 2. Wiley, London (1974)
8. Desu, M.M., Raghavarao, D.: Nonparametric statistical methods for complete and censored data. Chapman and Hall, Boca Raton (2004)
9. Geisser, S.: Predictive inference: an introduction. Chapman and Hall, London (1993)
10. Hill, B.M.: Posterior distribution of percentiles: Bayes' theorem for sampling from a population. J. Amer. Statist. Assoc. 63, 677–691 (1968)
11. Lawless, J.F., Fredette, M.: Frequentist prediction intervals and predictive distributions. Biometrika 92, 529–542 (2005)
12. Walley, P.: Statistical reasoning with imprecise probabilities. Chapman and Hall, London (1991)
13. Weichselberger, K.: Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als umfassendes Konzept. Physika, Heidelberg (2001)

Expected Pair-Wise Comparison of the Outcomes of a Fuzzy Random Variable

Inés Couso, Laura Garrido, Susana Montes, and Luciano Sánchez

Abstract. We introduce the notion of expected pair-wise comparison of a fuzzy random variable. It includes some well-known parameters such as the quadratic entropy of a random variable, the upper probability induced by a random set or the scalar variance of a fuzzy random variable as particular cases. The special case of expected dissimilitude is highlighted and shown as a useful alternative to the scalar variance when the images of the fuzzy random variable are not necessarily convex, nor in a numerical scale.

Keywords: Comparison measure, Divergence measure, Similarity measure, Semantics of fuzzy sets, Fuzzy random variable.

1 Introduction

Fuzzy random variables (frv for short) were first introduced by Féron in 1976, as functions that assign a fuzzy subset to each possible output of a random experiment, extending the notions of random variable and random set. Later on, several variants were proposed. The different definitions in the literature vary on the measurability conditions imposed to this mapping, and in the properties of the output space, but all of them intend to model situations that combine fuzziness and randomness. Apart from the differences among the formal definitions, fuzzy random variables have been also given different interpretations. Thus, a frv can be viewed ([3]) as a random object, an ill-known random variable or as a conditional upper probability. Each of those

Inés Couso, Laura Garrido, and Susana Montes
Dep. of Statistics and O.R, University of Oviedo, Spain
e-mail: couso,garridolaura,montes@uniovi.es

Luciano Sánchez
Dep. of Computer Sciences, University of Oviedo, Spain
e-mail: luciano@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 105–113.
springerlink.com

interpretations leads to a different way of extending parameters as the expectation, the variance, etc. The case of the variance is discussed in detail in [3]. In this paper, we will treat fuzzy random variables as random objects. We will introduce the notion of *expected pair-wise comparison* and we will discuss in some detail the specific notion of *expected dissimilitude measure*. The expected dissimilitude will average the "degrees of difference" or "divergence" between pairs of outcomes of the frv. In order to find a suitable quantification the differences between two outcomes of the frv, we will provide a brief discussion about some previous notions in the literature such as divergence measures ([8]) or distance measures ([6]), and we will introduce the notion of *dissimilitude*. Once being able to quantify such differences, the expected divergence will average them into a single quantity. We will show how the expected dissimilitude encompasses some different kinds of existing measures: on the one side, it extend the notion of scalar variance of a frv ([3, 7]). On the other hand, it also extends the quadratic entropy of a random variable. So, depending on the specific dissimilitude measure we use, we can extend the notion of variance, entropy, or a mixture of them. Furthermore, it allows us to quantify the expected difference between the different outcomes of the frv when the universe is not a numerical scale. The existing definitions of scalar variance are not easily adaptable to this kind of universes, because they involve the notion of expectation. In this paper, we do not average the dissimilitude degree between each possible outcome and the expectation, but between pairs of outcomes, by taking into account a pair of independent copies of the frv.

The rest of the paper is organized as follows: in Section 2, we briefly discuss the state of art about comparison measures, we provide some new results relating the notions of dissimilarity [2], divergence [8], distance [6] and metric, and we introduce the new notion of *dissimilitude measure*. In Section 3, we propose the concepts of expected pair-wise comparison and expected pair-wise dissimilitude of a frv, studying some interesting properties, and illustrating them with examples. We end the paper with some concluding remarks.

2 Dissimilitude Measures for Pairs of Fuzzy Sets

As we pointed out in the last section, an initial step in the construction of an expected dissimilitude measure will be the study of different options to compare pairs of outcomes of a fuzzy random variable. Let us denote by $\mathscr{F}(U)$ the family of fuzzy subsets of a universe U. A *comparison measure* [2] is a mapping $S : \mathscr{F}(U) \times \mathscr{F}(U) \to [0,1]$ expressed as:

$$S(A,B) = G_S(A \cap B, A-B, B-A),\ \forall A,B \in \mathscr{F}(U),$$

for some $G_S : \mathscr{F}(U) \times \mathscr{F}(U) \times \mathscr{F}(U) \to [0,1]$. This notion includes the idea of similarity and dissimilarity, and other kinds of comparison of pairs of fuzzy sets in a common framework. In this paper, we will slightly relax the assumptions for a comparison measure, and we will not force them to take values in the unit interval. We will pay attention to the quantification of the the degree of "difference" between two fuzzy outcomes. To this purpose, we will survey some previous proposals in the literature and we will check some properties and relations between them.

Montes el al. introduced in [8] an axiomatic definition for the *divergence* between pairs of fuzzy subsets based on the following natural properties:

- It is nonnegative and symmetric.
- It becomes zero when the two fuzzy sets coincide.
- It decreases when two fuzzy sets become "more similar".

Different formalizations of the third idea lead to different axiomatic definitions. Two of them are:

Definition 1. *(Bouchon-Meunier et al. [2]) Consider a universe U and let $\mathscr{F}(U)$ the family of fuzzy subsets of U. A comparison measure $S:\mathscr{F}(U)\times\mathscr{F}(U)\to\mathbb{R}$ is a* dissimilarity measure *when, for any pair $A,B\in\mathscr{F}(U)$ the following conditions hold:*

Diss1.- G_S *does not depend on its first argument (intersection) and it is increasing in the other two (differences) w.r.t. the fuzzy inclusion.*

Diss2.- $S(A,A)=0$

Definition 2. *(Montes et al. [8]) Consider a universe U and let $\mathscr{F}(U)$ the family of fuzzy subsets of U. A mapping $D:\mathscr{F}(U)\times\mathscr{F}(U)\to\mathbb{R}$ is a* divergence measure *when, for any pair $A,B\in\mathscr{F}(U)$ the following conditions hold:*

Div1.- $D(B,A)=D(A,B)$.

Diss2.- $D(A,A)=0$.

Div3.- $D(A\cup C,B\cup C)\leq D(A,B)$.

Div4.- $D(A\cap C,B\cap C)\leq D(A,B)$.

There is a strong relationship between dissimilarities and divergences: according to the following result, the divergence between two crisp sets A and B does not depend on their intersection, and it increases with their difference. (We omit the proof.)

Proposition 1. *Consider the function $D:\mathscr{F}(U)\times\mathscr{F}(U)\to\mathbb{R}$*

- *If D satisfies axiom Div3, then $D(A,B)\leq D(A-B,B-A)$, $\forall A,B\in\wp(U)$ (the class of crisp subsets of U).*
- *If D satisfies axiom Div4, then $D(A,B)\geq D(A-B,B-A)$, $\forall A,B\in\mathscr{F}(U)$.*
- *If D satisfies axiom Div4, then $D(A,B)\leq D(C,B)$, for all $A,B,C\in\wp(U)$ such that $C\cap B=\emptyset$ and $A\subseteq C$.*

Thus, the above measures (divergence and dissimilarity measures) focus on the differences between two fuzzy sets, but they do not care about their similarities. Sometimes, we need to take into account the similarities between sets, other times, we do not. Let us illustrate this with an easy example.

Example 1. Consider the set of languages:

$U = \{$English (e), Spanish (s), French (f), Italian (i), Dutch (d), Russian (r)$\}$

and let the crisp subsets $E = \{e,s,f,i,d\}$, $G = \{e,s,f,d,r\}$, $A = \{i\}$ and $V = \{r\}$ denote the respective communication skills of four persons called Enrique, Gert, Angelo and Vladimir. Enrique and Gert share much more language skills than Angelo and Vladimir, but those commonalities cannot be detected by means of the above measures (divergences and dissimilarities). If we just wanted to focus on the differences, those measures would be useful. But, if also take into account their common skills, we should use different comparison measures.

We can find in the literature some other measures that detect the differences between two fuzzy subsets, but they are not necessarily independent on the commonalities. Let us show some of them:

Definition 3. *(Fan, J. and Xie, W. [6]) Consider a universe U and let $\mathscr{F}(U)$ the family of fuzzy subsets of U. A mapping $d : \mathscr{F}(U) \times \mathscr{F}(U) \to \mathbb{R}$ is a* distance measure *when:*
Div1.- $d(B,A) = d(A,B),\ \forall A,B \in \mathscr{F}(U)$.
Diss2.- $d(A,A) = 0,\ \forall A \in \mathscr{F}(U)$.
DM3.- $d(D,D^c) = \max_{A,B\in\mathscr{F}(U)} d(A,B)$, *for any crisp set* $D \in \wp(U)$.
DM4.- *If* $A \subseteq B \subseteq C$, *then* $\max\{d(A,B), d(B,C)\} \leq d(A,C)$.

There are some relationships between divergence and distance measures. In fact, it is checked in [8] that any function satisfying Div3 and Div4 fulfills DM4. Furthermore, any *local*[1] divergence satisfies DM3. Thus, any local divergence measure satisfies Definition 3. Let us mention that the term *distance measure* is used in [6] without referring to the mathematical notion of *metric*. Nevertheless, both notions are somehow related, as they quantify the degree of difference between fuzzy subsets. We can find in the recent literature some metrics and pseudo-metrics defined on classes of fuzzy sets, such as the well known Hamming distance, the Puri-Ralescu [9] pseudo-metric[2] and other families of metrics proposed in [1, 7, 10] on some specific classes of convex fuzzy sets, for instance.

We can find some relationships between the above metrics and the notions of divergence, distance and comparison measure. In this short paper, we will only list them, without referring to formal details about the domain of definition of each measure, and without detailing the proofs:

[1] A divergence measure is called *local* ([8]) when there exists a function $h : [0,1] \times [0,1] \to \mathbb{R}$ such that: $D(A,B) - D(A \cup \{x\}, B \cup \{x\}) = h(A(x), B(x)),\ \forall x \in U$.

[2] Puri and Ralescu introduce a metric in the class of fuzzy subsets of $\mathbb{R}^n$ with compact and non-empty level cuts. It can be easily extended to more general families of fuzzy subsets as a pseudo-metric.

Proposition 2

- *All the metrics and pseudo-metrics cited above can be expressed as comparison measures on their respective domains of definition and they satisfy axioms Div1, Diss2, Div3 and DM4.*
- *Only the Hamming distance satisfies axioms Diss1, Div4 and DM3.*

According to the above proposition, axioms Diss1, Div4 and DM3 exclude most of the mentioned (pseudo)-metrics. In the rest of the paper, we will use the term *dissimilitude measure* for those comparison measures safisfying Div1, Diss2, Div3 and DM4.

3 Expected Pair-Wise Comparison of a Fuzzy Random Variable

Consider a probability space $(\Omega,\mathscr{A},P)$ and a frv defined on it, i.e., an $\mathscr{A}-\sigma$ measurable mapping $\tilde{X}:\Omega\to\mathscr{F}$, where σ is a σ-field defined on a certain class of fuzzy subsets $\mathscr{F}\subseteq\mathscr{F}(U)$. (This general definition encompasses several specific proposals in the literature). Any fuzzy random variable induces a probability measure on σ by means of the formula:

$$P_{\tilde{X}}(\mathscr{C})=P(\{\omega\in\Omega\,:\,\tilde{X}(\omega)\in\mathscr{C}\}),\ \forall\mathscr{C}\in\sigma.$$

Now consider the product probability $P\otimes P:\mathscr{A}\otimes\mathscr{A}\to[0,1]$ as the only probability measure satisfying the restriction

$$(P\otimes P)(A\times B)=P(A)\cdot P(B)\ \forall A,B\in\mathscr{A}.$$

Let $\tilde{X}_1$ and $\tilde{X}_2$ two (identically distributed) copies of $\tilde{X}$, and consider a comparison measure on $\mathscr{F}$, $S:\mathscr{F}\times\mathscr{F}\to[0,1]$.

Definition 4. *We define the* expected pair-wise comparison *of $\tilde{X}$ as the quantity*

$$E_S(\tilde{X})=\int_{\Omega\times\Omega}S(\tilde{X}_1(\omega),\tilde{X}_2(\omega'))\,d(P\otimes P)(\omega,\omega'),$$

provided that the mapping $g(\omega,\omega')=S(\tilde{X}_1(\omega),\tilde{X}_2(\omega'))$, $\forall(\omega,\omega')\in\Omega\times\Omega$ is $\mathscr{A}\otimes\mathscr{A}-\beta_{\mathbb{R}}$ measurable.

The above definition generalizes some well-known quantities, as we show in the following examples.

Example 2. Consider a finite population Ω and the set of languages $U=\{e,s,f,i,d,\,r\}$ of Example 1. Consider the multi-valued mapping $\Gamma:\Omega\to\wp(U)$ that assigns to each person $\omega\in\Omega$ the subset of languages in U (s)he can speak[3]. The following expected pair-wise comparison measures provide interesting information about such attribute:

- Let us fix an arbitrary subset $D \subseteq U$. Consider the comparison measure S_1 such that G_{S_1} is defined as $G_{S_1}(A,B,C) = \max\{M(A \cap B), M(A \cap B^c)\}$, where
$$M(E) = \begin{cases} 1 & \text{if } E \cap D \neq \emptyset, \\ 0 & otherwise. \end{cases}$$
The expected pair-wise comparison of Γ, $E_{S_1}(\Gamma)$ coincides with the *upper probability* ([4]) of D and it represents the proportion of persons in the population that speak some of the languages included in D. If, for instance, D is equal to $\{d\}$, then $E_{S_1}(\Gamma)$ represents the proportion of persons that can speak Dutch, at least.
- Let us fix again an arbitrary subset $D \subseteq U$. Consider the comparison measure S_2 such that G_{S_2} is defined as $G_{S_2}(A,B,C) = \min\{M(A \cap B), M(A \cap B^c)\}$, where
$$M(E) = \begin{cases} 1 & \text{if } E \subseteq D, \\ 0 & otherwise. \end{cases}$$
The expected pair-wise comparison of Γ, $E_{S_2}(\Gamma)$, coincides with the *lower probability* ([4]) of D and it represents the proportion of persons in the population that do not speak any language outside D.
- Consider the comparison measure $S_3(A,B) = \#(A \cap B)$. The expected pair-wise comparison $E_{S_3}(\Gamma)$ averages the capacity of communication between pairs of people in the population.
- Consider the Hamming distance $S_4(A,B) = d_H(A,B) = \#(A \triangle B)$. The expected pair-wise comparison $E_{S_4}(\Gamma)$ represents a degree of divergence about the language skills of the people in the population.

Example 3. The above example can be modified if we have more refined information about the communication skills of the people. We can use a frv $\tilde{X} : \Omega \to \mathscr{F}(U)$ to represent those abilities. The membership value $\tilde{X}(\omega)(u)$ will represent a degree of preference ([5]) in a $[0,1]$ scale for the language $u \in U$. Thus $\tilde{X}(\omega)(u) > \tilde{X}(\omega)(u')$ will mean that the person ω prefers to speak u than u', because (s)he is more familiar with it. Those degrees of preference can be determined as a function of the CEFR levels, for instance. For a specific dissimilitude measure, the expected dissimilitude of $\tilde{X}$ reflects an expected degree of difference in the language skills between pairs of persons in the population.

Example 4. Consider a set of days, Ω, and consider the multi-valued mapping $\Gamma : \Omega \to \wp(\mathbb{R})$, where $\Gamma(\omega) = [L(\omega), U(\omega)]$ represents the interval of minimum

[3] Let us assume that $\Gamma(\omega) \neq \emptyset$, $\forall \omega \in \Omega$, so everybody is assumed to be able to speak some language in the set U. Multi-valued mappings represent special cases of fuzzy-valued mappings. Furthermore, if we consider the power set as the initial σ-field, they are measurable with respect to any σ-field on the final space.

and maximum temperatures attained in Mieres on a date ω. Several expected pair-wise comparison measures return different informative quantities such as: the variance of the min temperatures, the variance of the max temperatures, a mixture (linear combination) of both variances, the variance of the amplitudes of the min-max intervals, the proportion of days where the min temperature exceeds a certain threshold, the variance of the middle points of the intervals, etc.

The particular case where the comparison measure S is a dissimilitude is remarkable. It extends some key notions in the literature, as we show in the following remarks.

Remark 1. On the one hand, it extends the notion of quadratic entropy of a random variable: If $\tilde{X}$ represents a random variable X on a finite universe $U = \{u_1, \ldots, u_n\}$ in the sense that $\tilde{X}(\omega) = \{X(\omega)\}$, $\forall \omega \in \Omega$, and $S(A,B) = d_H(A,B) = \#A \triangle B$ is the Hamming distance, then the expected pair-wise comparison of $\tilde{X}$ is the quadratic entropy of X:

$$E_S(\tilde{X}) = \sum_{i=1}^{n}\sum_{j=1}^{n} d_H(\{x_i\},\{x_j\}) p_i \cdot p_j = \sum_{i=1}^{n}\sum_{j=1}^{n}(1-\delta_{ij}) p_i \cdot p_j = 1 - \sum_{i=1}^{n} p_i^2,$$

where p_i denotes the probability $P(X = u_i)$, $i = 1, \ldots n$.

Remark 2. On the other hand, it extends some notions of scalar variance of a fuzzy random variable [3] in the literature: all the (pseudo-)metrics considered at the end of Section 2 satisfy the properties of dissimilitude measures. Furthermore, any non-decreasing function of a similitude satisfying the boundary condition $g(0) = 0$ is also a dissimilitude. If we construct the similitude measure $S = \frac{d^2}{2}$ on the basis of any of those distances d, and we take into account the specific arithmetic used in each context, in order to avoid the explicit use of the expectation, we can extend the existing notions of scalar variances [3] in the literature. Furthermore, expected dissimilitude measures even apply when the images of the frv are not necessarily convex, and/or they do not lay in a numerical scale, as we have illustrated in Example 2.

Some general properties of expected dissimilitude measures are given in the following proposition.

Proposition 3. *Let $S : \mathscr{F} \times \mathscr{F} \to \mathbb{R}$ be a dissimilitude measure. Then:*

- *$E_S(A) = 0$, $\forall A \in \mathscr{F}$.*
- *$E_S(\tilde{X} \cup A) \leq E_S(\tilde{X})$, for all frv $\tilde{X}$ and all $A \in \mathscr{F}$.*
- *If $S(A,B) = \sum_{x \in U} g(A(x),B(x))$ then $E(\tilde{X} \cup \tilde{Y}) \leq E_S(\tilde{X}) + E_S(\tilde{Y})$, for all frv $\tilde{X}$ and $\tilde{Y}$.*

Example 5. We can illustrate the above properties by referring to the language skills of Example 2. The first property would mean that the expected dissimilitude is null when everybody in the population owns the same communication skills. For the second property, let us assume that all the people in that population that do not speak Spanish take a course on this language. Then, the expected dissimilitude measure should decrease. Finally, suppose that we consider two separate groups of languages, and we consider the communication skills of the people within each group ($\tilde{X}$ denotes the abilities within the first group of languages, and $\tilde{Y}$ denotes the abilities within the second group.) Then, the expected dissimilitude in the whole set of languages cannot be strictly greater than the sum of the expected dissimilitude values within each group.

4 Concluding Remarks

The notion of expected comparison of a fuzzy random variable encompasses several well known parameters associated to random variables, random sets and fuzzy random variables. In particular, the expected dissimilitude quantifies the dispersion of the outcomes of a fuzzy random variable. It generalizes some entropies for random variables and also some scalar variances of fuzzy random variables. The existing definitions of scalar variances that we can find in the literature [3, 7] are restricted to those situations where the outcomes of the frv are convex fuzzy subsets of $\mathbb{R}^n$. The new definition applies in a variety of situations, even for the cases where there is not a numerical scale. We have illustrated the utility of the new notion with several examples. In future works we plan to study some additional properties of the expected dissimilitude, for some specific dissimilitude measures, trying to lay bare the connection with the general notions of entropy and dispersion.

Acknowledgements. This work has been supported by the FEDER-MEC Grants TIN2007-67418-C03-03, TIN2008-06681-C06-04 and MTM2007-61193.

References

1. Bertoluzza, C., Corral, N., Salas, A.: On a new class of distances between fuzzy numbers. Mathware Soft Comput. 2, 71–84 (1995)
2. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy Sets Syst. 84, 143–153 (1996)
3. Couso, I., Dubois, D.: On the variability of the concept of variance of a fuzzy random variable. IEEE Trans. Fuzzy Syst. 17, 1070–1080 (2009)
4. Dempster, A.P.: Upper and Lower Probabilities Induced by a Multivalued Mapping. Ann. Math. Stat. 38, 325–339 (1967)
5. Dubois, D., Prade, H.: The three semantics of fuzzy sets. Fuzzy Sets Syst. 90, 141–150 (1997)
6. Fan, J., Xie, W.: Some notes on similarity measure and proximity measure. Fuzzy Sets Syst. 101, 403–412 (1999)
7. Körner, R.: On the variance of fuzzy random variables. Fuzzy Sets Syst. 92, 83–93 (1997)

8. Montes, S., Couso, I., Gil, P., Bertoluzza, C.: Divergence measure between fuzzy sets. Internat. J. Approx. Reason. 30, 91–105 (2002)
9. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
10. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inform. Sci. 179(23), 3964–3972 (2009)

The Behavioral Meaning of the Median

Inés Couso and Luciano Sánchez

Abstract. We generalize the notion of statistical preference to the theory of imprecise probabilities, by proposing an alternative notion of desirability of a gamble. As a natural consequence, we derive a general definition of median, providing it with a behavioral meaning. Furthermore, we show that, when we restrict to absolutely continuous probability distributions, a random variable is statistically preferred to another one if and only if the the median of their difference is positive.

Keywords: Stochastic orderings, Statistical preference, Imprecise Probabilities, Desirable Gambles, Median.

1 Introduction

Several preference relations between random variables have been proposed in the literature. One of them, called *statistical preference* [2, 3] is based on the probabilistic relation $Q(X,Y) = P(X > Y) + \frac{1}{2}P(X = Y)$ and it states that X is preferred to Y when $Q(X,Y) \geq 0.5$. Independently, a similar criterion has been proposed in [4, 5] in the framework of possibility theory. In this paper, we aim to extend the notion of statistical preference to the general theory of imprecise probabilities, relating it to the notions of *desirability* and *preference* between gambles. The problem of reconciling two different ways of treating preference relations will arise. In fact, a preference relation for pairs of variables (or *gambles*) can be understood in two different ways:

Inés Couso
Dep. of Statistics and O.R, University of Oviedo, Spain
e-mail: couso@uniovi.es

Luciano Sánchez
Dep. of Computer Sciences, University of Oviedo, Spain
e-mail: luciano@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 115–122.
springerlink.com

- The expert initial information is assessed by means of a preference criterion and, afterwards, a set of joint feasible *linear previsions*[1] is derived from it. This is the approach followed in the general theory of imprecise probabilities (see [1, 6]).
- A joint probability is assumed for any pair of gambles on the universe, and a preference relation is derived from it. This is the approach considered in [4, 5, 2, 3], for instance.

Taking into account the above duality, we will first start from an initial set of desirable gambles and we will say that a gamble X is signed-preferred to another gamble Y when the sign of their difference is a desirable gamble. Afterwards, we will show that signed-almost-preference becomes into statistical preference, when the initial set of desirable gambles induces a singleton as a credal set. In a second approach, we will state signed-preference and signed-desirability as primary concepts, appealing to a new idea of desirability: X will be said to be desirable when we have stronger beliefs about $X > 0$ than about $X < 0$. In words, we accept the gamble X, because we have stronger beliefs on making money than on loosing it –no matter how much money–. Based on this desirability definition, we can define a lower prevision as the supremum of the constants c satisfying that $X - c$ is desirable, according to the new definition. Such supremum makes sense as a threshold for buying prices: for any strictly lower price, you have stronger beliefs on earning money that on loosing it. Analogously, we will define an upper prevision as an infimum threshold for selling prices. Once introduced both approaches (the interpretation of sign-desirability as a secondary an as a primary concept), we will relate them, and we will derive an interesting conclusion: the pair of lower and upper previsions defined for the set of signed-desirable gambles generalizes the notion of median, providing it with a meaningful behavioral interpretation. As a consequence of that, we will be able to show that there exists a very strong connection between the relation of statistical preference of two random variables and the sign of the median of their difference. This result adds another piece to the puzzle about the relationships between different stochastic orderings proposed in the literature.

2 Sets of Desirable Gambles and Partial Preference Orderings

Let Ω denote the set of outcomes of an experiment. A *gamble*, X, on Ω is a bounded mapping from Ω to $\mathbb{R}$ (the real line). If you were to accept gamble X and ω turned to be true, then you would gain $X(\omega)$. (This reward can be negative, and then it will represent a loss.) Let $\mathscr{L}$ denote the set of all gambles (bounded mappings from Ω to $\mathbb{R}$). A subset $\mathscr{D}$ of $\mathscr{L}$ is said to be a *coherent set of desirable gambles* [6] when it satisfies the following four axioms:

[1] The notion of linear prevision generalizes the notion of probability.

D1. If $X \leq 0$ then $X \notin \mathscr{D}$, *(Avoiding partial loss)*
D2. If $X \in \mathscr{L}$, $X \geq 0$ and $X \neq 0$, then $X \in \mathscr{D}$. *(Accepting partial gain)*
D3. If $X \in \mathscr{D}$ and $c \in \mathbb{R}^+$, then $cX \in \mathscr{D}$. *(Positive homogeneity)*
D4. If $X \in \mathscr{D}$ and $Y \in \mathscr{D}$ then $X+Y \in \mathscr{D}$. *(Addition)*

For a detailed justification of each of the above axioms concerning coherence in assessments of a subject, we refer the reader to (cf.[6], Section 2.2.4).

The *lower prevision induced by a set of desirable gambles $\mathscr{D}$* is the set function $\underline{P} : \mathscr{L} \to \mathbb{R}$ defined as follows:

$$\underline{P}(X) = \sup\{c : X - c \in \mathscr{D}\}.$$

It is interpreted as your supremum acceptable buying price for X, so you are disposed to pay $\underline{P}(X) - \varepsilon$, for the reward determined by the gamble X, for any $\varepsilon > 0$. The *upper prevision induced by $\mathscr{D}$* is the set function $\overline{P} : \mathscr{L} \to \mathbb{R}$ defined as follows:

$$\overline{P}(X) = \inf\{c : c - X \in \mathscr{D}\}.$$

It can be regarded as an infimum selling price for the gamble X.

The set of linear previsions[2] induced by a coherent set of gambles $\mathscr{D}$ is defined as:

$$\mathscr{P}_{\mathscr{D}} = \{P : P(X) \geq 0 \text{ for all } X \in \mathscr{D}\}.$$

$\mathscr{P}_{\mathscr{D}}$ is always a *credal set* (a closed and convex set of finitely additive probability measures). $\underline{P}$ and $\overline{P}$ are dual and they respectively coincide with the infimum and the supremum of $\mathscr{P}_{\mathscr{D}}$. There is not a one-to correspondence between sets of desirable gambles and credal sets, as there can be two different sets of desirable gambles $\mathscr{D} \neq \mathscr{D}'$ inducing the same class of linear previsions $\mathscr{P}_{\mathscr{D}} = \mathscr{P}_{\mathscr{D}'}$. On the other hand, a subset $\mathscr{D}^- \subset \mathscr{L}$ satisfying axioms D2–D4 and

D1'. If $\sup X < 0$ then $X \notin \mathscr{D}^-$. *(Avoiding sure loss)*
D5. If $X + \delta \in \mathscr{D}^-$, for all $\delta > 0$ then $X \in \mathscr{D}^-$. *(Closure)*

is called a coherent set of *almost desirable gambles.* (Let the reader notice that axiom D1' is weaker than D1.) A set of almost desirable gambles $\mathscr{D}^-$ determines a pair of lower and upper previsions, and a credal set, by means of expressions analogous to the case of desirable gambles. Conversely, a credal set univocally determines a coherent set of almost desirable gambles via the formula:

$$\mathscr{D}^-_{\mathscr{P}} = \{X \in \mathscr{L} : P(X) \geq 0, \ \forall P \in \mathscr{P}\}.$$

Finally, a set $\mathscr{D}^+ \subset \mathscr{L}$ is said to be a coherent set of *strict desirable gambles* if it is a coherent set of desirable gambles, and it satisfies, in addition, the following axiom:

D6. If $X \in \mathscr{D}^+$, then either $X \geq 0$ or $X - \delta \in \mathscr{D}^+$, for some $\delta > 0$. (openness)

[2] A *linear prevision* is a linear functional $P : \mathscr{L} \to \mathbb{R}$ satisfying the constraint $P(1) = 1$. So it generalizes the notions of expectation and (finitely additive) probability measure at the same time.

A coherent set of strict desirable gambles can be derived from a credal set as follows:

$$\mathscr{D}^{+}_{\mathscr{P}} = \{X : X \geq 0 \text{ and } X \neq 0 \text{ or } P(X) > 0 \; \forall P \in \mathscr{P}\}.$$

The notion of *desirability* of gambles is closely related to *partial preference ordering* between gambles. A gamble X is said to be *preferred* to another gamble Y $(X \succ Y)$, Coherent preference orderings can be characterized through a set of axioms closely related to D1–D5.

3 Generalized Statistical Preference

Probabilistic relations are usual representation of several relational preference models. A *probabilistic relation* (see [3]) Q on a set of alternatives A is a mapping from $A \times A$ to $[0,1]$ satisfying the equality $Q(a,b) + Q(b,a) = 1$ for any pair of alternatives $(a,b) \in A^2$. On the other hand, De Schuymer et al. [2, 3] introduced the notions of *strict preference*, $P(X,Y) = P(X > Y)$, and *indifference*, $I(X,Y) = P(X = Y)$, for comparing pairs of random variables. A probabilistic relation can be naturally derived from P and I as follows:

$$Q(X,Y) = P(X,Y) + \frac{1}{2} I(X,Y).$$

Based on it, a total preorder can be defined on the class of random variables defined on a probability space:

Definition 1. *[3] A random variable X is* statistically preferred *to another random variable Y if $Q(X,Y) \geq 0.5$. We will denote it by $X \geq_{SP} Y$. Furthermore, we will use the notation $X >_{SD} Y$ when $X \geq_{SP} Y$, but not $Y \geq_{SP} X$.*

The following result follows from the fact that the probabilistic relation $D(X,Y)$ is greater than 0.5 if and only if it is greater than $D(Y,X)$.

Proposition 1. *Consider two random variables defined on the same probability space. Then, $X \geq_{SP} Y$ if and only if $P(X > Y) \geq P(X < Y)$. Consequently $X >_{SP} Y$ iff $P(X > Y) > P(X < Y)$.*

According to the last straightforward result, a random variable (from now on, a *gamble*) is statistically preferred to another gamble Y if and only if they satisfy the inequality $P(X - Y > 0) \geq P(Y - X > 0)$. According to the behavioral interpretation of previsions in the general theory of imprecise probabilities, the above inequality is related to the following preference assessment: you are disposed to give up $1_{Y-X>0}$ in return for $1_{X-Y>0}$, where 1_A denotes the indicator of A. So, statistical preference of X over Y is connected to your acceptance of a reward of one unit of *probability currency* [6] if X takes an strictly higher value than Y in exchange to the reward or one unit if Y takes a strictly higher valued than X. (Because you have stronger belief on the occurrence of $X > Y$ than on the occurrence of $Y > X$.)

As we pointed out in the last section, there is a strong connection between the notions of *desirability* and *preference* of gambles, as a gamble X is preferred to another one Y when $X-Y$ is desirable, and, conversely, X is desirable when it is preferred to the null gamble. According to this connection, we will start by introducing the notion of *signed-desirable gamble* as a primary notion, and we will derive from it the concept of *signed-preference relation.* This last concept will be the generalization of the notion of statistical preference to the theory of imprecise probabilities.

Definition 2. *Consider a coherent set of desirable gambles $\mathscr{D}$ in $\mathscr{L}$. We will say that a gamble $X \in \mathscr{L}$ is* signed-desirable *if the gamble*

$$sgn(X) = 1_{X>0} - 1_{X<0}$$

belongs to $\mathscr{D}$. (In the above expression, sgn denotes the well know "sign function" and 1_B denotes the indicator of the subset B.)

In words, a gamble X is signed-desirable when you are disposed to give up the gamble $1_{X<0}$ (it means, paying one probability currency unit if X takes a negative value) in return for the gamble $1_{X>0}$ (receiving 1 unit if X takes a -strictly- positive value.)

Remark 1. Analogously to Definition 2, we can introduce the notions of *signed-almost desirable gamble*, as a gamble X satisfying the restriction $\mathrm{sgn}(X) \in \mathscr{D}^-$ and *signed- strictly desirable* as a gamble satisfying the condition $\mathrm{sgn}(X) \in \mathscr{D}^+$, where $\mathscr{D}^-$ and $\mathscr{D}^+$ respectively denote coherent families of almost/strict desirable gambles. We will use the respective notations $X \in \mathscr{D}_S^-$ and $X \in \mathscr{D}_S^+$.

Proposition 2. *Consider a coherent set of desirable gambles $\mathscr{D}$, and the associated sets of almost/strict desirable gambles, respectively denoted $\mathscr{D}^-$ and $\mathscr{D}^+$. Then:*

- *The family of signed-desirable gambles $\mathscr{D}_S$ satisfies axioms D1 to D3.*
- *The family of signed-almost desirable gambles $\mathscr{D}_S^-$ satisfies D1', D2, D3 and D5.*
- *The family of signed-strict desirable gambles $\mathscr{D}_S^-$ satisfies D1 to D3, and D6.*

None of the above sets of gambles satisfies axiom D4 of additivity. It is a key axiom to identify coherent sets of (almost desirable) gambles with coherent lower previsions in the theory of imprecise probabilities. The notion of lower prevision extends the concept of expectation in (classical) probability theory. In the next section, we will associate sets of signed-desirable gambles with lower medians.

Based on the above definition of signed-desirability, we can derive the following three partial preference orderings.

Definition 3. *Consider a coherent set of desirable gambles $\mathscr{D}$ in $\mathscr{L}$. A gamble X is said to be* signed -preferred *to another gamble Y if $X-Y$ is signed-desirable, i.e., if $X-Y \in \mathscr{D}_S$.*

The notions of *signed-almost preference* and *signed-strict preference* can be introduced analogously, referring to the membership of the gamble $X-Y$ to the respective sets $\mathscr{D}_S^-$ and $\mathscr{D}_S^+$. In the next proposition, we will show that the above preference partial orderings generalize the notion of statistical preference.

Proposition 3. *Let P be a linear prevision and let us respectively denote by $\mathscr{D}^-$ and $\mathscr{D}^+$ the sets of gambles $\mathscr{D}^- = \{X : P(X) \geq 0\}$ and $\mathscr{D}^+ = \{X : P(X) > 0,$ or $[X \geq 0$ and $X \neq 0]\}$. Then, for a pair of gambles X and Y:*

- *$X \geq_{SP} Y$ if and only if $X-Y \in \mathscr{D}_S^-$.*
- *$X >_{SP} Y$ if and only if $X-Y \in \mathscr{D}_S^+$.*

The above result states that almost signed-preference generalizes statistical preference and signed-strict preference generalizes strict statistical preference. The notion of signed-preference is in between the two, and it has no counterpart within the classical theory of probability. The distinction between almost desirability and desirability becomes important within the theory of imprecise probabilities. For instance, different coherent sets of gambles inducing the same credal set propagate different information about conditioning, as it is illustrated in [1, 6], for instance. It will be a matter of future study whether the distinction between signed-almost desirability and signed-(plain) desirability is also of importance or not.

4 Behavioral Interpretation of the Median

According to the definitions introduced in the last section, a coherent set of desirable gambles determines a set of signed-desirable gambles. Now, let us start from signed-desirability as a primary notion and consider the lower prevision of X:

$$\underline{P}_{\mathscr{D}_{\mathscr{S}}}(X) = \sup\{c : X - c \in \mathscr{D}_S\}$$

It is interpreted as a threshold for the desirability in the following sense: for any strictly lower quantity $c < \underline{P}_S(X)$, you are disposed to pay some fixed quantity (say 1 probability currency unit) if $X < c$ holds, in return for the same quantity if $X > c$ occurs, because you have stronger beliefs on the event $X > c$ than on $X > c$. For any strictly higher quantity, you are not. We can give a dual interpretation, as a threshold for the desirability of $c - X$ to the infimum:

$$\overline{P}_{\mathscr{D}_{\mathscr{S}}}(X) = \inf\{c : c - X \in \mathscr{D}_S\}.$$

The next result connects the above definitions with the classical notion of median. It is parallel to the connection existing between pairs of lower and upper previsions of a gamble and the bounds of its expectations, when we range the probability measures in the credal set.

Theorem 1. *Let $\mathscr{P}$ be a credal set and let be $\mathscr{D}^+$ the coherent set of strict desirable gambles:*

$$\mathscr{D}^+ = \{X \in \mathscr{L} : P(X) > 0 \ \forall P \in \mathscr{P} \ \textit{or} \ [X \geq 0 \ \textit{and} \ X \neq 0]\}.$$

Given a linear prevision, and an arbitrary gamble X, let $\mathrm{Me}_P(X)$ *denote the interval of the medians of X,*

$$\mathrm{Me}_P(X) = \{x : P(1_{X \geq x}) \geq 0.5 \ \textit{and} \ P(1_{X \leq x}) \geq 0.5\}.$$

Then the following equalities hold:

$$\sup\{c : \mathrm{sgn}(X-c) \in \mathscr{D}^+\} = \inf \cup_{P \in \mathscr{P}} \mathrm{Me}_P(X) \ \textit{and}$$

$$\inf\{c : \mathrm{sgn}(c-X) \in \mathscr{D}^+\} = \sup \cup_{P \in \mathscr{P}} \mathrm{Me}_P(X).$$

According to the last theorem, we can introduce the notions of lower and upper median as follows:

Definition 4. *Let $\mathscr{D}^+ \subset \mathscr{L}$ be a coherent set of strict desirable gambles. The* lower median *of an arbitrary gamble $X \in \mathscr{L}$ is defined as the quantity*

$$\underline{\mathrm{Me}}(X) = \sup\{c : \mathrm{sgn}(X-c) \in \mathscr{D}^+\}.$$

Analogously, the upper median *of X is defined as the quantity*

$$\overline{\mathrm{Me}}(X) = \inf\{c : \mathrm{sgn}(c-X) \in \mathscr{D}^+\}.$$

In the general theory of imprecise probabilities, there is a well know connection between the value of the lower prevision of a gamble and its desirability: a gamble is almost-desirable if and only if its lower prevision is non negative. Furthermore, if the lower prevision is strictly positive, then it is strictly desirable. In the next result we will show a parallel connection between the value of the lower median and the sign-desirability of a gamble:

Proposition 4. *Consider a coherent set of desirable gambles $\mathscr{D}$ and let $\mathscr{P}_{\mathscr{D}}$ the associated credal set. Let $\mathscr{D}^-$ (resp. $\mathscr{D}^+$) denote the coherent sets of almost-(resp. strict-)desirable gambles derived from it. The following implications hold:*

$$\underline{\mathrm{Me}}(X) > 0 \Rightarrow \mathrm{sgn}(X) \in \mathscr{D}^+ \Rightarrow \mathrm{sgn}(X) \in \mathscr{D} \Rightarrow \mathrm{sgn}(X) \in \mathscr{D}^- \Rightarrow \underline{\mathrm{Me}}(X) \geq 0.$$

As a consequence of the above result, when we restrict to a single probability, the statistical preference of a random variable X over another one Y is very closely related to the sign of the median of their difference:

Corollary 1. *Let $(\Omega, \mathscr{A}, P)$ be an arbitrary probability space and let (X,Y) be a random vector defined on it. Let $Me_P(X-Y)$ denote the set of medians of $X-Y$, i.e.,*

$$Me_P(X-Y) = \{x : P(X-Y \geq x) \geq 0.5 \ \textit{and} \ P(X-Y \leq x) \geq 0.5\}.$$

Then, the following implications hold:

$$\inf Me_P(X-Y) > 0 \Rightarrow X >_{SP} Y \Rightarrow X \geq_{SP} Y \Rightarrow \inf Me_P(X-Y) \geq 0$$

Corollary 2. *Let (X,Y) be random vector with absolutely continuous distribution. Then:*

- $X \geq_{SP} Y$ *if and only if* $\inf Me_P(X-Y) \geq 0$
- $X >_{SP} Y$ *if and only if* $\inf Me_P(X-Y) > 0$

We easily derive from the above result that X is statistically preferred to Y if and only if the expectation of X is greater than the expectation of Y, when the difference $X-Y$ is absolutely continuous and it has a symmetric distribution.

5 Concluding Remarks

We have extended the concept of median to Imprecise Probabilities, and provided it with a behavioral meaning. We have also introduced the notion of (almost)-signed preference as a generalization of the so-called statistical preference. X is said to be signed-preferred to Y when the gamble $\mathrm{sgn}(X-Y)$ is desirable, and therefore $\underline{P}(1_{X-Y>0} - 1_{Y-X>0}) \geq 0$. The last condition is weaker than the condition $\underline{P}(1_{X-Y>0}) \geq \overline{P}(1_{Y-X>0})$, which simultaneously extends statistical preference, and the preference relation considered in [4, 5]. In the future, we will investigate further connections between both extensions.

Acknowledgements. This work has been supported by the FEDER-MEC Grants MTM2007-61193, TIN2007-67418-C03-03 and TIN2008-06681-C06-04.

References

1. Couso, I., Moral, S.: Sets of Desirable Gambles and Credal Sets. In: Procceddings of the 6th International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK (2009)
2. De Schuymer, B., De Meyer, H., De Baets, B.: Cycle-transitive comparison of independent random variables. J. Multivariate Anal. 96, 352–373 (2005)
3. De Schuymer, B., De Meyer, H., De Baets, B., Jenei, S.: On the cycle-transitivity of the dice model. Theory Decis 54, 261–285 (2003)
4. Sánchez, L., Couso, I., Casillas, J.: Modeling Vague Data with Genetic Fuzzy Systems under a Combination of Crisp and Imprecise Criteria. In: Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Multicriteria Decision Making, MCDM 2007, Honolulu, Hawaii (2007)
5. Sánchez, L., Couso, I., Casillas, J.: Genetic Learning of Fuzzy Rules based on Low Quality Data. Fuzzy Sets Syst. 160, 2524–2552 (2009)
6. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
7. Walley, P.: Towards a unified theory of imprecise probability. Internat. J. Approx. Reason. 24, 125–148 (2000)

Functional Classification and the Random Tukey Depth. Practical Issues

Juan A. Cuesta-Albertos and Alicia Nieto-Reyes

Abstract. Depths are used to attempt to order the points of a multidimensional or infinite dimensional set from the "center of the set" to the "outer of it". There are few definitions of depth which are valid in the functional case. One of them is the so-called random Tukey depth, which is based on some randomly chosen one-dimensional projections and thus varies (randomly) from computation to computation. Some theoretical properties of this depth are well-known, but it has not yet been studied from a practical point of view. The aim of this paper is to analyze its behavior in classification problems, the interest of this study being increased by the random character of the depth. To do this, we compare the performance of the random Tukey depth in a real data set with the results obtained with the López-Pintado and Romo depths.

Keywords: Random Tukey depth, Functional depth, Random projections, Functional classification.

1 Introduction

Given a probability P defined in a multidimensional or infinite-dimensional space $\mathscr{X}$, a depth attempts to order the points in $\mathscr{X}$ from the "center (of P)" to the "outer (of P)". Obviously, this problem includes data sets if we consider P as the empirical distribution associated to the data set at hand.

In the multidimensional setting, the first definition of depth was established by Mahalanobis (in [12]). This definition is based on the well-known Mahalanobis distance. If μ and Σ are, respectively, the mean and covariance matrix of P, then, the Mahalanobis depth of x with respect to P is

Juan A. Cuesta-Albertos and Alicia Nieto-Reyes
Departamento de Matemáticas, Estadística y Computación,
Universidad de Cantabria, Spain
e-mail: juan.cuesta@unican.es,alicia.nieto@unican.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 123–130.
springerlink.com

$$D_H(x,P) := \frac{1}{1+(x-\mu)^t\Sigma^{-1}(x-\mu)},\ x \in \mathbb{R}^p.$$

From this starting point, subsequent definitions of depth (see [9]) clarified that depths as well as having some robustness properties, are a highly flexible tool for handling nonparametrically statistical problems involving testing, classification, descriptive statistics,... This, in turn, has led to the study of the possibility of introducing depths in the functional setting. However, most of the known multidimensional depths cannot be generalized to the functional case because the dimension of the space under consideration plays a key role in them, or alternatively, because of the associated computational difficulties. For instance, the computation of the Tukey depth (a precise definition appears in (1)) is unfeasible for dimensions as low as eight if the sample size is only 100.

As far as we know, some definitions of depth valid for functional spaces have been proposed in [5], [6], [7] and [11]. In this paper, we are particularly interested in the so-called random Tukey depth which was studied in [3, 4] because these papers leave some practical issues open. Our goal here is to make a first attempt to show how these gaps can be filled when handling classification problems.

Let us begin with some definitions. Apart from its lack of robustness, the Mahalanobis depth has some flaws: it is not defined if the mean or the covariance matrix does not exist and it treats P as symmetric (because points at the same Mahalanobis distance from the mean have the same depth). A reasonable way to overcome these problems in the one-dimensional case could be to define the depth of the point x with respect to P by

$$D_1(x,P) := \min\{P(-\infty,x], P[x,\infty)\}$$

which is a monotone transformation of the Mahalanobis depth if μ and Σ exist and P is symmetric, thus providing the same order of the points.

The *Tukey depth* was introduced in [16] and can be defined as follows. Let P be a probability on $\mathbb{R}^p$ and $v \in \mathbb{R}^p$. If Π_v denotes the projection on the one-dimensional subspace generated by v and P_v the one-dimensional marginal of P on the same subspace, then, the Tukey depth of x with respect to P is

$$D_T(x,P) := \inf\{D_1(\Pi_v(x), P_v) : v \in \mathbb{R}^p\}. \tag{1}$$

The computational problems we mentioned above, led the authors of [3] to introduce the random Tukey depth, which is a random approximation of the Tukey depth. In [3], the following generalization to Hilbert spaces was proposed:

Definition 1. *Let $\mathscr{X}$ be a separable Hilbert space, P be a probability distribution on $\mathscr{X}$, ν be a Gaussian distribution with non-degenerated marginals on $\mathscr{X}$ and $v_1,\dots,v_k$ be i.i.d. random vectors with distribution ν. The random Tukey depth of $x \in \mathscr{X}$ with respect to P based on k random vectors chosen with ν is*

$$D_{T,k,\nu}(x,P) := \min\{D_1(\Pi_{\nu_i}(x), P_{\nu_i}) : i = 1,\ldots,k\}.$$

The random Tukey depth was used in [3], in the finite dimensional case, to handle several testing problems and, in addition, it was shown there that this depth has some useful properties in the infinite and finite cases. In particular, it was shown that in the infinite dimensional case, it satisfies most of the requirements of the definition stated in [8] and formalized in [17] for a statistical depth.

However, in [3] nothing is said about the influence that the selection of ν and k might have in practice. The aim of this paper is to make a preliminary analysis of these issues from the point of view of a classification problem, and, at the same time, to compare the results obtained with the random Tukey depth with those provided in [10] with the depths proposed in [11].

The situation we have chosen to carry out this comparison is the supervised classification problem which was carried out in [10]. In this paper, the authors analyze a data set consisting of the growth curves of a sample of 39 boys and 54 girls, the aim being to classify them, by sex, using just this information. We represent the data in Figure 1.

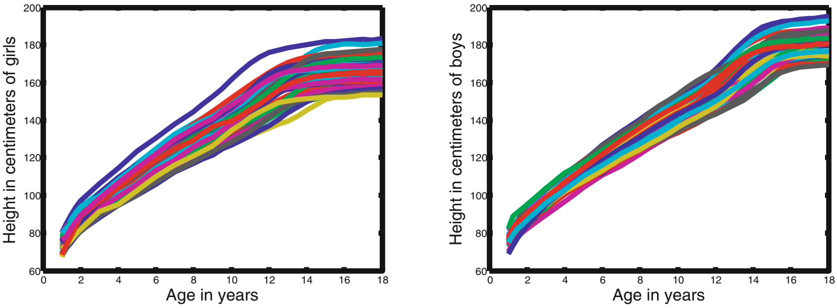


Fig. 1 Growth curves of 54 girls (left-hand side) and 39 boys (right-hand side) measured 31 times each between 1 and 18 years of age.

Heights were measured in centimeters 31 times in the period from one to eighteen years. In the period from one to two years, the measures were taken every three months, in the period from three to seven years one time a year and, finally, in the period from eight to eighteen years two times a year. The data are in the file growth.zip, downloaded from `ftp://ego.psych.mcgill.ca/pub/ramsay/FDAfuns/Matlab`. On this web-page, some notes that make use of the data can also be found. These notes were designed to accompany the books [13, 14]. In addition, these data are used in the recent book [15].

It is well-known that when handling this kind of data, it is useful to consider not only the growth curve but also accelerations of height (see, for instance, [13]). However, we only consider here the growth curves, as did [10], because our interest lies in comparing our results with those obtained by them.

It should be noted that the distribution ν which appears in Definition 1 does not need to be Gaussian. In fact, as shown in [1], any dissipative distribution works here. Thus, in the finite dimensional case, the uniform distribution on the unit sphere may be enough. Regrettably, in the functional setting, there is no distribution like this which can be taken as a reference. Although some papers have already appeared using random projections (in the finite and in the infinite dimensional cases), as far as we know, except for a small comment in [2] in the finite dimensional case, none of them has paid attention to the problem of the precise selection of ν.

A preliminary step in addressing this question is given in Section 2, where we also comment on the selection of the number of vectors used in the definition of the random Tukey depth. Then, in Section 3 we compare the results obtained with the random Tukey depth with those obtained in [10].

2 Distribution and Number of Vectors for the Random Tukey Depth in Practice

In order to analyze the effect of the selection of ν in the random Tukey depth in classification problems, the idea is to analyze the same data using two strategies: firstly, one that does not admit variations in ν, i.e. ν is a fixed distribution. Secondly, one that selects ν from a parametric family of distributions, thus making it possible to chose the parameters which determine ν in a data-dependent way.

The parametric family we handle has two real parameters $a \geq 0$ and $c \geq 0$, and is defined forthwith. Let us assume that we are in a two-class classification problem and that we have two training samples $\mathbb{X} = \{X_1(t), ..., X_n(t)\}$ and $\mathbb{Y} = \{Y_1(t), ..., Y_m(t)\}$, where $t \in [0, T]$. First, compute the point-wise median in both samples: $m_{\mathbb{X}}(t)$, and $m_{\mathbb{Y}}(t)$, $t \in [0, T]$. Then, given a, c let $\nu = S_{a,c}$ be the solution of the of the following stochastic differential equation

$$S_{a,c}(0) = c \text{ and } dS_{a,c}(t) = |m_X(t) - m_Y(t)|^a dB(t),$$

where B is a standard Brownian motion.

The fixed distribution that we compare with is the standard Brownian motion, which is the member of the family corresponding to the case $a = c = 0$.

In the following section, we choose $a \in \{0, 1\}$. Note that when $a = 0$ the difference between the functions m_X and m_Y has no influence on ν. The constant c specifies the initial value for the solution. We have tried the values $c = 0, 1, 5$. The reason for introducing c is that the Brownian motion always starts at 0 and is continuous, thus erasing the differences in the early states of the process.

In practice, we will assume that the trajectories have been measured in the same finite set of values $t_1 < \ldots < t_h$. Then, given a and c, to simulate the random trajectories we have taken

$$S_{a,c}(t_1) = c$$
$$S_{a,c}(t_i) = S_{a,c}(t_{i-1}) + |m_X(t_i) - m_Y(t_i)|^a Z_i, \ i = 2,...,31,$$

where Z_i, $i = 2,...,h$, are independent random variables with distribution $N(0, t_i - t_{i-1})$.

Concerning k, in [3] the authors carry out some simulations to select k in the finite dimensional case for dimensions ranging from $p = 2,8,50$ and several sample sizes. Those results suggest that high values for k are not required. The results that follow in Section 3 have been obtained by selecting $k \in \{1,...,100\}$. Although the upper bound for k might be considered too low, we have repeated the process replacing 100 by 1,000 and the results obtained have been similar.

We propose the use of leave-one-out cross validation to choose the right value of k, as well as those of a and c when required.

3 The Procedure in Practice

As stated, in this section, we compare the results of classifying the heights data set when employing the random Tukey depth with those obtained with the depths proposed in [11]. To do this, we have repeated the study made in [10] with three differences:

1. Most importantly, we have replaced the functional depths handled there with the random Tukey depth.
2. In [10], the authors consider the curves as elements in $L^1[0,1]$, which is not possible here, because we need a separable Hilbert space. Thus, we have taken $\mathbb{H} = L^2[0,1]$.
3. In [10], the authors smoothed the original data using a spline basis. We have omitted this step because it is not necessary for our method.

Regarding item 2., remember that the heights were measured 31 times on times $t_i \in [1,18]$, $i = 1,...,31$ where

$$t_i = 3/4 + i/4 \text{ for } i = 1,...,5,$$
$$t_i = i - 3 \text{ for } i = 6,...,10$$
$$t_i = 2.5 + i/2 \text{ for } i = 11,...,31.$$

If $i = 2,\ldots,30$, then the observation $X(t_i)$ represents the height of the individual in the interval $((t_i + t_{i-1})/2, (t_{i+1} + t_i)/2)$. Taking into account that, in the last part of the study, the measurements were taken every half a year, we can assume that $X(t_{31})$ is valid for the period $(17.5, 18.5)$. Finally, it seems safe to assume that the $X(t_1)$ is not valid for representing previous heights. Therefore, we can assume that the interval in which the measurements have been taken is $[1, 18.5]$. In consequence, first, we need to modify the time in order to transform the interval $[1, 18.5]$ into $[0, 1]$ and, then, we can employ properties of the Rieman integral to make the approximation

$$< X, s_{a,c} > = \int_0^1 X(17.5u+1)s_{a,c}(17.5u+1)du \approx \sum_{i=1}^{31} X(t_i)s_{a,c}(t_i)\Delta_i,$$

where $s_{a,c}$ is drawn with distribution $S_{a,c}$ and Δ_i denotes the length of the interval associated to the point t_i. Then, if we define $t_0 = 1$ and $t_{32} = 18.5$, we have

$$\Delta_i = (t_{i+1} - t_{i-1})/35, \; i = 1,...,31.$$

In [10], the authors consider three possibilities for splitting the sample into training and validation sets. For the sake of brevity, we split the sample using only leave-one-out cross-validation.

Let us briefly explain how the whole process works. Note that we have a sample of size 93. Therefore, we have repeated 100 times the following: for each observation in the sample, we consider the training sample composed of the remaining 92 observations. Then, we have generated at random 100 vectors with each of the distributions of the random variables $S_{a,c}$ for $a = 0,1$ and $c = 0,1,5$, which gives 6 different samples of random directions with size 100 each.

Firstly, we have focused our attention on the $S_{0,0}$ distribution. Here we only have to select the value of k. As stated previously, this value is chosen by leave-one-out cross-validation applied to the remaining sample with 92 observations. Henceforth, this procedure is called $S_{0,0}$.

Moreover, we have applied the procedure allowing variations in a and c. Here, also using leave-one-out cross-validation, we have chosen the best combination of k, a and c. Henceforth, this procedure is denoted by $S_{a,c}$. Note that in this case, it may occur that the chosen a and c satisfy $a = c = 0$; thus, the $S_{a,c}$ procedure should give better results than $S_{0,0}$.

The results of the comparison appear in Table 1, which includes the obtained failure rates using the three methods proposed in [10] when applied to the random Tukey depth and to the depths proposed in [11]. These methods are: distance to the trimmed mean ($\mathrm{M}_{\alpha,\beta}$), weighted average distance (AM) and trimmed weighted average distance (TAM). We have chosen $\alpha = \beta = 0.2$ as in [10]. The depths handled in [10] are the band depth determined by three different curves (DS3), by four different curves (DS4) and the generalized band depth (DGS). Their error rates are contained in the last three columns of Table 1 and have been taken from Tables 1-3 in [10]. The previous two columns of Table 1 concern the random Tukey depth. The first includes the failure rates when using the procedure $S_{0,0}$ and the second when using $S_{a,c}$.

According to Table 1, if we employ the $M_{\alpha,\beta}$ method, the random Tukey depth with the procedure $S_{0,0}$ works worse than the other depths and, when coupled with $S_{a,c}$ performs similarly to the $DS3$ and $DS4$ depths but worse than the DGS. However, for the AM and TAM methods, the random Tukey depth provides better results than the depths used in [10] when we take the

Table 1 Rates of mistakes when classifying the growth curves by sex for the shown methods and depths.

Classification	Random Tukey		Depths proposed in [10]		
method	$S_{0,0}$	$S_{a,c}$	DS3	DS4	DGS
$\mathrm{M}_{\alpha,\beta}$	.1858	.1825	.1828	.1828	.1613
AM	.1403	.1368	.2473	.2473	.1935
TAM	.1542	.1430	.2436	.2436	.1690

standard Brownian motion and even better when parameters a, c in $S_{a,c}$ are chosen with cross-validation.

The medians of the number of random vectors used have been 1 for each of the three methods with $S_{0,0}$. In the case of $S_{a,c}$, the median of the number of random vectors has been 2 for the $\mathrm{M}_{\alpha,\beta}$ method and 1 for both of the other two methods.

4 Conclusions

The Tukey depth is one of the best-behaved multidimensional depths but it cannot be used in the functional setting. However, the random Tukey depth, which approximates it in multidimensional spaces, does work in functional settings.

The definition of the random Tukey depth involves choosing a distribution. We have seen how in practice the behavior of this functional depth varies depending on the chosen distribution. Specifically, its performance increases when the distribution is data driven.

Furthermore, to compute the random Tukey depth, a finite number of vectors have to be drawn with this chosen distribution. This number is of great importance since the computational time needed to compute the random Tukey depth depends on it. In [3] it was seen that this number is low in multidimensional spaces and, in view of our experience, it seems that it is also surprisingly low when dealing with functional data.

References

1. Cuesta-Albertos, J.A., del Barrio, T., Fraiman, R., Matrán, C.: The random projection method in goodness of fit for functional data. Comput. Statist. Data Anal. 51(10), 4814–4831 (2007)
2. Cuesta-Albertos, J.A., Febrero-Bande, M.: Multiway ANOVA for Functional Data. Test (to appear, 2010)
3. Cuesta-Albertos, J.A., Nieto-Reyes, A.: The random Tukey depth. Comput. Statist. Data Anal. 52(11), 4979–4988 (2008)

4. Cuesta-Albertos, J.A., Nieto-Reyes, A.: A random functional depth. In: Dabo-Niang, S., Ferraty, F. (eds.) Functional and Operational Statistics, pp. 121–126. Springer, New-York (2008)
5. Cuevas, A., Febrero, M., Fraiman, R.: Robust estimation and classification for functional data via projection-based depth notions. Comput. Statist. 22(3), 481–496 (2007)
6. Cuevas, A., Fraiman, R.: On depth measures and dual statistics. A methodology for dealing with general data. J. Multivariate Anal. 100(4), 753–766 (2009)
7. Fraiman, R., Muniz, G.: Trimmed means for functional data. Test 10(2), 419–440 (2001)
8. Liu, R.Y.: On a notion of data depth based on random simplices. Ann. Statist. 18, 405–414 (1990)
9. Liu, R.Y., Parelius, J.M., Singh, K.: Multivariate analysis by data depth: descriptive statistics, graphics and inference. Ann. Statist. 27(3), 783–858 (1999)
10. López-Pintado, S., Romo, J.: Depth-based classification for functional data. In: Liu, R., Serfling, R., Souvaine, D.L. (eds.) Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications. DIMACS Series, vol. 72, pp. 103–119. American Mathematical Society, Providence (2006)
11. López-Pintado, S., Romo, J.: On the concept of depth for functional data. J. Amer. Statist. Assoc. 104(486), 718–734 (2009)
12. Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Natl. Inst. Science 12, 49–55 (1936)
13. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (1997)
14. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer, New York (2002)
15. Ramsay, J.O., Hooker, G., Graves, S.: Functional Data Analysis with R and MATLAB. Springer, New York (2009)
16. Tukey, J.W.: Mathematics and picturing of data. In: Proc. ICM, Vancouver, vol. 2, pp. 523–531 (1975)
17. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Statist. 28(2), 461–482 (2000)

On Concordance Measures and Copulas with Fractal Support

E. de Amo, M. Díaz Carrillo, and J. Fernández-Sánchez

Abstract. Copulas can be used to describe multivariate dependence structures. We explore the rôle of copulas with fractal support in the study of association measures.

1 General Introduction and Motivation

Copulas are of interest because they link joint distributions to their marginal distributions. Sklar [12] showed that, for any real-valued random variables X_1 and X_2 with joint distribution H, there exists a copula C such that $H(u,v) = C(F_1(u), F_2(v))$, where F_1 and F_2 denote the cumulative (or margin) distributions of X_1 and X_2, respectively. If the marginals are continuous, then the copula is unique. Notice that it is also true the converse implication of Sklar's Theorem. In fact, we may link any univariate distributions with any copula in order to obtain a valid joint distribution function. An implication of Sklar Theorem is that the dependence among X_1 and X_2 is fully described by the associated copula. Indeed, most conventional dependence measures can be explicitly expressed in terms of the copula, and they are designed to capture certain aspects of dependence or association between random variables.

On the other hand, all the examples of singular copulas we have found in the literature are supported by sets with Hausdorff dimension 1. However, it is implicit in some papers, for example in [11], that the well-known examples of Peano and Hilbert curves provide self-similar copulas with fractal support, since the Hausdorff dimension of their graphs is 3/2.

E. de Amo and J. Fernández-Sánchez
Departamento de Álgebra y Análisis Matemático, Universidad of Almería, Spain
e-mail: edeamo@ual.es

M. Díaz Carrillo
Departamento de Análisis Matemático, Universidad of Granada, Spain
e-mail: madiaz@ugr.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 131–138.
springerlink.com

Recently, Fredricks et al. [7], using an iterated function system, constructed families of copulas whose supports are fractals. In particular, they give sufficient conditions for the support of a self-similar copula to be a fractal whose Hausdorff dimension is between 1 and 2.

In [1], the authors prove that a necessary and sufficient condition for a copula to be the independence (or product) copula Π is that the pair of measure preserving transformations representing the copula be independent as random variables; and a general constructive method for representation of copulas in terms of measure preserving transformations is given. Specifically, we study the copulas introduced by Fredricks et al. in [7], through two representation number systems we construct *ad hoc*. This paper is devoted to study these copulas in depth.

Firstly, we give an example of copula with a support with a Lebesgue measure 0, and a Hausdorff dimension 2. Moreover, we study the coefficients of upper and lower tail dependence of these copulas. Finally, we explore some well-known measures of dependence, namely Kendall's τ, Spearman's ρ, and Gini's index γ. The results we prove coincide with those regard to the independence copula Π.

2 Preliminaries about Copulas with Fractal Support

This section contains background information and useful notation.

(1) Let $\mathbb{I}$ be the closed unit interval $[0,1]$ and let $\mathbb{I}^2 = \mathbb{I} \times \mathbb{I}$ be the unit square. For an introduction to copulas see, for example, [4] or [9].

(2) A *transformation matrix* is a matrix T with nonnegative entries, for which the sum of the entries is 1 and none row or column has zero as entry everywhere.

Following [7], we recall that each transformation matrix T determines a subdivision of $\mathbb{I}^2$ into subrectangles $R_{ij} = [p_{i-1}, p_i] \times [q_{j-1}, q_j]$, where p_i (respect. q_j) denotes the sum of the entries in the first i columns (respect. j rows) of T. For a transformation matrix T and a copula C, $T(C)$ denotes the copula that, for each (i,j), spreads mass on R_{ij} in the same way in which C spreads mass on $\mathbb{I}^2$.

Theorem 2 in [7] shows that for each transformation matrix $T \neq [1]$, there is an unique copula C_T such that $T(C_T) = C_T$.

(3) Let T be a transformation matrix, and let us consider the following conditions:

(i) T has at least one zero entry.

(ii) For each non-zero entry of T, the row and column sums through that entry are equal.

(iii) There is at least one row or column of T with two nonzero entries.

Theorem 3 in [7] shows that if T is a transformation matrix satisfying condition (i) , then C_T is singular (that is, its support has Lebesgue measure zero or $\mu_{C_T} \equiv \mu^s_{C_T}$, where $\mu^s_{C_T}$ is the singular measure given by the Lebesgue Decomposition Theorem). See, for example, [9] or [10].

We say that a copula C is *invariant* if $C = C_T$ for some transformation matrix T. An invariant copula C_T is said to be *self-similar* if T satisfies condition (ii).

Theorem 6 in [7] shows that the support of a self-similar copula C_T, with T satisfying (i) and (iii), is a fractal whose Hausdorff dimension is between 1 and 2.

(4) A mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is called a *contracting similarity* (or a similarity transformation) of ratio r $(0 < r < 1)$ if $\|F(x) - F(y)\| = r\|x - y\|$, for all $x, y \in \mathbb{R}^n$. A similarity transforms subsets of $\mathbb{R}^n$ into geometrically similar sets. The invariant set (or attractor) for a finite family of similitaries is said to be a *self-similar set*. Theorem 4 in [7] shows that the support of the copula C_T is the invariant set for a system of similarities obtained from partitions of $\mathbb{I}$ which are determined by T. (See (2) above.)

For an introduction to the techniques of fractal representation by iterated function systems (IFS) the reader can see [5] or [6].

(5) Finally, we recall that the notion of an IFS may be extended to define invariant measures supported by the attractor of the system. Explicitly, let $\{F_1, ..., F_m\}$ be an IFS on $K \subset \mathbb{R}^n$ and $p_1, \ldots, p_m$ be probabilities or mass ratios, with $p_i > 0$ for all i and $\sum_{i=1}^m p_i = 1$. A measure μ is said to be *self-similar* if for some p_i and F_i, $\mu(A) = \sum_{i=1}^m p_i \mu\left(F_i^{-1}(A)\right)$ and any Borel set A. The existence of such measure is ensured. (See [6, Th.2.8].)

3 A Singular Copula with Fractal Support and Hausdorff Dimension 2

Now, we consider the family of transformation matrices

$$T_r = \begin{pmatrix} r/2 & 0 & r/2 \\ 0 & 1-2r & 0 \\ r/2 & 0 & r/2 \end{pmatrix}$$

with $r \in \left]0, \frac{1}{2}\right[$. According to **(2)** and **(3)** above, $\left\{C_r = C_{T_r} : r \in \left]0, \frac{1}{2}\right[\right\}$ is a family of copulas whose support is a fractal with a Hausdorff dimension in the interval $]1,2[$.

Precisely, Fredricks et al. [7, Th. 1] proved that if $s \in]1,2[$, then there exists a copula $C = C_{r(s)}$ satisfying that the Hausdorff dimension of the support of C is s. We denote it by S_r. It is clear that there exist singular functions whose support is of Hausdorff dimension 1; and that there are no copulas whose support is of Hausdorff dimension smaller than 1, as well.

On the other hand, as far as we know, there is no a singular copula whose fractal support is, exactly, of Hausdorff dimension 2. We now show an example of copula with these properties.

Theorem 1. *Let $J_i = \left[\frac{1}{i+1}, \frac{1}{i}\right]$, with $i \in \mathbb{Z}^+$, and the copula $C^i = C_{r\left(\frac{2i-1}{i}\right)}$. Then, the ordinal sum of $\left\{C^i\right\}_{i\in\mathbb{Z}^+}$ with respect to $\{J_i\}_{i\in\mathbb{Z}^+}$, that is,*

$$C(u,v) = \begin{cases} \frac{1}{i+1} + \frac{1}{i(i+1)} C^i\left(i(i+1)\left(u - \frac{1}{i+1}\right), i(i+1)\left(v - \frac{1}{i+1}\right)\right), & (u,v) \in J_i^2 \\ \min(u,v), & \textit{otherwise,} \end{cases}$$

is a singular copula, and its support is of Hausdorff dimension 2.

Proof. The way we have defined C ensures that it is a copula (see [3] or [9]).

Out of the J_i^2 squares, the copula C does not have associated mass distribution. On each square J_i^2, the similarity

$$F_i(u,v) \to \left(\frac{1}{i+1} + \frac{u}{i(i+1)}, \frac{1}{i+1} + \frac{v}{i(i+1)}\right)$$

spreads the mass distribution in the support of C^i. It is straightforward (via the definition of C) that $\frac{\partial^2 C^i}{\partial u \partial v}(u,v) = 0$ out of that set.

The function F_i is a similarity, and hence, bi-Lipschitz (i.e., a bijective Lipschitz function whose inverse function is also Lipschitz). Therefore, it preserves the Hausdorff dimension (see [6, Cor. 2.4]). As a consequence,

$$\dim_{\mathscr{H}} F_i\left(S_{r\left(\frac{2i-1}{i}\right)}\right) = \frac{2i-1}{i}.$$

But, the Hausdorff dimension of the set $\cup_i F_i\left(S_{r((2i-1)/i)}\right)$ is the supremum of the above numbers; that is, 2. Therefore, C is a singular copula whose support is of Hausdorff dimension 2. □

4 Tail Dependence for C_r

Copulas are useful to model tail dependence, that is, the dependence that arises between random variables from extreme observations.

Let X and Y be continuous random variables with distribution functions F and G, respectively, and associated copula C. We study the coefficients of upper and lower tail dependence of X and Y (their definition can be found in [8] or [9]).

Theorem 2. *Let $r \in \left]0, \frac{1}{2}\right[$. Given the copula C_r, we have*

$$\lambda_U(C_r) = \lambda_L(C_r) = 0.$$

Proof. The symmetry of the mass distribution of the measure μ_{C_r} ensures that, if they exist, the two values are the same. We study the case λ_L.

The self-similarity of the measure provides that $C_r(r^n, r^n) = \left(\frac{r}{2}\right)^n$ for $n \in \mathbb{Z}^+$. Let $u \in]0,1[$. Then, there exists n satisfying $r^{n+1} < u \leq r^n$. Therefore,

$$\frac{C_r(r^{n+1},r^{n+1})}{r^n} \leq \frac{C_r(u,u)}{u} \leq \frac{C_r(r^n,r^n)}{r^{n+1}};$$

that is,

$$\frac{r}{2^{n+1}} \leq \frac{C_r(u,u)}{u} \leq \frac{1}{2^n r}.$$

But, in this case, if $u \to 0^+$, then $n \to \infty$; and there exists the limit

$$\lambda_L = \lim_{u\to 0^+} \frac{C_r(u,u)}{u} = 0. \qquad \square$$

5 Concordance Measures for C_r

We study several association measures that mesh the probability of concordance between random variables with a given copula. For a review of concordance measures and the rôle played by the copulas in the study of dependence or association between random variables, see [9, Chap. 5].

Let us recall that concordance or discordance are basic when introducing association measures. Formally, two ordered pairs of real numbers, (x_1,y_1) and (x_2,y_2), are concordant if $(x_1-x_2)(y_1-y_2)>0$. They are discordant if they are not concordant.

Let (X_1,Y_1) and (X_2,Y_2) be two continuous random pairs with the same marginal distribution functions, and associated copulas C_1 and C_2, respectively. A *concordance function* is defined by

$$Q(C_1,C_2) = P[(X_1-X_2)(Y_1-Y_2)>0] - P[(X_1-X_2)(Y_1-Y_2)<0].$$

Moreover, if the above continuous random pairs (X_1,Y_1) and (X_2,Y_2) are independent then

$$Q(C_1,C_2) = 4\int_{\mathbb{I}^2} C_2(x,y)d\mu_{C_1}(x,y) - 1.$$

Definition 1. *Let (X,Y) be a continuous random pair with associated copula C. The value $Q(C,C)$ is a measure of association called the Kendall's τ of (X,Y). Moreover, the value $3Q(C,\Pi)$ is a measure of association called the Spearman's ρ of (X,Y). And, the value $Q(C,M)+Q(C,W)$ is another measure of association for (X,Y) called the Gini's γ.*

For computational purposes we have the following result (see **(5)** above):

Lemma 1. *Let $K \subset \mathbb{R}^n$ and let μ be a self-similar measure associated to the family of similarity transformations $\{F_1,\ldots,F_m\}$ with respective mass ratios $\{p_1,\ldots,p_m\}$. Then, for any continuous function $g: K \to \mathbb{R}$ and any $k, (1 \leq k \leq m)$, we have*

$$\int_{F_k(K)} g(x)d\mu(x) = p_k \int_K g(F_k(x))d\mu(x).$$

Proof. The map F_k is a self-similarity transformation, hence, it is an isomorphism between measurable spaces. As a consequence, there exists a natural bijection from the step functions on K to $F_k(K)$ (considering the induced σ-algebra, in both cases). The measures of the measurable sets A and $F_k(A)$ are proportional with ratio p_k, therefore, and the statement is true in the case that g is a step function. Density arguments establish that the statement is also true for all integrable functions. □

(6) An immediate consequence from the above lemma is the following useful expression:

$$\int_K g(x)d\mu(x) = \sum_{k=1}^{m} p_k \int_K g(F_k(x))d\mu(x).$$

Now, by applying **(6)** and using **(2)** and **(3)**, we can express the concordance in terms of the family of copulas C_{T_r}.

Proposition 1. *Given the copula $C_{T_r} = C_r$ of parameter $r \in \left]0, \frac{1}{2}\right[$, the following equalities hold:*

a) $\int_{[0,1]^2} \max(x+y-1,0)dC_r(x,y) = \frac{1-r}{8-10r}$
b) $\int_{[0,1]^2} \min(x,y)dC_r(x,y) = \frac{3-4r}{8-10r}$
c) $\int_{[0,1]^2} xydC_r(x,y) = 1/4$
d) $\int_{[0,1]^2} C_r(x,y)\ dC_r(x,y) = 1/4$

Proof. *a*) Let us decompose the integral as a sum on five regions in the unit square. The measure μ_{C_r} is self-similar; and therefore:

$$\begin{aligned}\int_{\mathbb{I}^2} W(x,y)d\mu_{C_r}(x,y) &= \frac{r}{2}\int_{\mathbb{I}^2} W(rx,ry)d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2} W(rx+1-r,ry)d\mu_{C_r} \\ &+\frac{r}{2}\int_{\mathbb{I}^2} W(rx,ry+1-r)d\mu_{C_r} \\ &+\frac{r}{2}\int_{\mathbb{I}^2} W(rx+1-r,ry+1-r)d\mu_{C_r} \\ &+(1-2r)\int_{\mathbb{I}^2} W((1-2r)x+r,(1-2r)y+r)d\mu_{C_r} \\ &= \frac{r}{2}\int_{\mathbb{I}^2} 0\ d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2} rW(x,y)d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2} rW(x,y)d\mu_{C_r} \\ &+\frac{r}{2}\int_{\mathbb{I}^2} (rx+ry+1-2r)d\mu_{C_r} + (1-2r)^2\int_{\mathbb{I}^2} W(x,y)d\mu_{C_r} \\ &= \left(r^2+(1-2r)^2\right)\int_{\mathbb{I}^2} W(x,y)d\mu_{C_r}(x,y) + \frac{r(1-r)}{2};\end{aligned}$$

and working out the integral, we have the statement.

b) We proceed as in the above case.

c) Here,

$$\begin{aligned}
\int_{\mathbb{I}^2} xyd\mu_{C_r}(x,y) = {}& \frac{r}{2}\int_{\mathbb{I}^2} rxryd\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2}(rx+1-r)ryd\mu_{C_r} \\
&+\frac{r}{2}\int_{\mathbb{I}^2}(rx)(ry+1-r)d\mu_{C_r} \\
&+\frac{r}{2}\int_{\mathbb{I}^2}(rx+1-r)(ry+1-r)d\mu_{C_r} \\
&+(1-2r)\int_{\mathbb{I}^2}((1-2r)x+r)((1-2r)y+r)d\mu_{C_r} \\
= {}& 2r^3\int_{\mathbb{I}^2} xyd\mu_{C_r} + (1-2r)^3\int_{\mathbb{I}^2} xyd\mu_{C_r} \\
&+\left(2r^2(1-r)+2(1-2r)^2r\right)\int_{\mathbb{I}^2} yd\mu_{C_r} \\
&+\frac{r(1-r)^2}{2}+(1-2r)r^2 \\
= {}& \left(2r^3+(1-2r)^3\right)\int_{\mathbb{I}^2} xyd\mu_{C_r}(x,y)+\frac{3r}{2}-3r^2+\frac{3r^3}{2};
\end{aligned}$$

and the statement follows.

d) We decompose the integral as in the first case:

$$\begin{aligned}
\int_{\mathbb{I}^2} C_r(x,y)d\mu_{C_r}(x,y) = {}& \frac{r}{2}\int_{\mathbb{I}^2} C_r(rx,ry)d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2} C_r(rx+1-r,ry)d\mu_{C_r} \\
&+\frac{r}{2}\int_{\mathbb{I}^2} C_r(rx,ry+1-r)d\mu_{C_r} \\
&+\frac{r}{2}\int_{\mathbb{I}^2} C_r(rx+1-r,ry+1-r)d\mu_{C_r}+ \\
&+(1-2r)\int_{\mathbb{I}^2} C_r((1-2r)x+r,(1-2r)y+r)d\mu_{C_r} \\
= {}& \frac{r}{2}\int_{\mathbb{I}^2}\frac{r}{2}C_r(x,y)d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2}\frac{r}{2}y+\frac{r}{2}C_r(x,y)d\mu_{C_r} \\
&+\frac{r}{2}\int_{\mathbb{I}^2}\frac{r}{2}x+\frac{r}{2}C_r(x,y)d\mu_{C_r} + \frac{r}{2}\int_{\mathbb{I}^2}\frac{r}{2}+1-2r+\frac{r}{2}(x+y) \\
&+\frac{r}{2}C_r(x,y)d\mu_{C_r} + (1-2r)\int_{\mathbb{I}^2}\frac{r}{2}+(1-2r)C_r(x,y)d\mu_C \\
= {}& \left(r^2+(1-2r)^2\right)\int_{\mathbb{I}^2} C_r(x,y)d\mu_{C_r}(x,y)+\frac{3}{4}r^2+r(1-2r);
\end{aligned}$$

and the result follows. □

Corollary 1. *Kendall's* τ, *Spearman's* ρ, *and Gini's* γ *are zero for all* $r \in \left]0,\frac{1}{2}\right[$.

The independence copula Π is the limit case when $r \to 1/2$. The values we obtain for these association measures and for the copula Π are the same. As a consequence, there is no monotone dependence in any degree for these measures, that is the same case for Π.

Acknowledgements. We are grateful to Juan Quesada for his careful reading and useful suggestions that improved the paper.

References

1. de Amo, E., Díaz Carrillo, M., Fernández-Sánchez, J.: Measure-Preserving Functions and the Independence Copula. Mediterr. J. Math. 8(4) (to appear, 2010) doi:10.1007/s00009-010-0073-9
2. de Amo, E., Díaz Carrillo, M., Fernández-Sánchez, J.: Copulas and associated fractal sets (submitted for publication, 2010)
3. Durante, F., Saminger, S., Sarkoci, P.: Rectangular patchwork for bivariate copulas and tail dependence. Comm. Statist. Theory Methods 38, 2515–2527 (2009)
4. Durante, F., Sempi, C.: Copula theory: an introduction. In: Durante, F., Härdle, W., Jaworki, P., Rychlik, T. (eds.) Proceedings of Workshop on Copula Theory and its Applications. Lecture Notes in Statistics. Springer, Dordrecht (in press, 2010)
5. Edgar, G.A.: Measure, Topology, and Fractal Geometry. Springer Undergraduate Texts in Mathematics, Heidelberg (1990)
6. Falconer, K.J.: Fractal Geometry. Mathematical Foundations and Applications, 2nd edn. John Wiley & Sons, London (2003)
7. Fredricks, G.A., Nelsen, R.B., Rodríguez-Lallena, J.A.: Copulas with fractal supports. Insur. Math. Econ. 37(1), 42–48 (2005)
8. Joe, H.: Multivariate Models and Dependence Concepts. Chapman & Hall, London (1997)
9. Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer Series in Statistics. Springer, New York (2006)
10. Rudin, W.: Real and Complex Analysis, 3rd edn. McGraw-Hill Book Co., New York (1987)
11. Sagan, H.: Space Filling Curves. Springer, New York (1994)
12. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges, vol. 8, pp. 229–231. Publ. Inst. Statist. Univ., Paris (1959)

Factorisation Properties of the Strong Product

Gert de Cooman, Enrique Miranda, and Marco Zaffalon

Abstract. We investigate a number of factorisation conditions in the framework of sets of probability measures, or coherent lower previsions, with finite referential spaces. We show that the so-called strong product constitutes one way to combine a number of marginal coherent lower previsions into an independent joint lower prevision, and we prove that under some conditions it is the only independent product that satisfies the factorisation conditions.

Keywords: Coherent lower previsions, Epistemic independence, Strong independence, Factorisation.

1 Introduction

In this paper, we investigate some relationships between the formalist and epistemic approaches to independence in a generalised setting that allows probabilities to be imprecisely specified. By formalist approach, we mean a way to construct an independent joint from given marginals that is based on requiring the joint to satisfy a number of mathematical properties, such as factorisation. An epistemic approach, on the other hand, uses judgements of equality between the marginal and conditional probability models to construct a joint from the marginals.

Gert de Cooman
SYSTeMS Research Group, Ghent University, 9052 Zwijnaarde Belgium
e-mail: `gert.decooman@ugent.be`

Enrique Miranda
Dep. of Statistics and O.R., University of Oviedo, 33007 Oviedo, Spain
e-mail: `mirandaenrique@uniovi.es`

Marco Zaffalon
IDSIA, 6928 Manno (Lugano), Switzerland
e-mail: `zaffalon@idsia.ch`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 139–147.
springerlink.com

We will consider a finite number of logically independent variables X_n assuming values in respective finite sets $\mathscr{X}_n$, $n \in N$, where N denotes a finite index set. In an epistemic approach, we want to express that these variables are independent, in the sense that learning the values of some of them will not affect beliefs about the remaining ones. We base our analysis on the theory of *coherent lower previsions*, which are lower expectation functionals equivalent to closed convex sets of probability mass functions. In the case of precise probability, we refer to an expectation functional as a *linear prevision*. We give a succinct introduction to the theory of coherent lower previsions in Section 2.

The real work begins in Section 3, where, in the formalist spirit, we introduce a number of factorisation conditions for coherent lower previsions, and establish relationships between them. In Section 4, we investigate a specific and fairly popular method for combining marginal coherent lower previsions into a joint lower prevision: the strong product. We show that this method satisfies all the formalist factorisation properties from Section 3, as well as two independence notions of an epistemic bent, called epistemic many-to-one and epistemic many-to-many independence. Moreover, we show that in certain cases the strong product is the only functional with these properties. Due to limitations of space, we have omitted the proofs of the main results.

2 Coherent Lower Previsions

We start with a brief introduction to the notions of the theory of coherent lower previsions; we refer to [8] for an in-depth study, and to [6] for a survey.

Consider a finite space $\mathscr{X}$. A gamble on $\mathscr{X}$ is a real-valued map $f\colon \mathscr{X} \to \mathbb{R}$. The set of all gambles on $\mathscr{X}$ is denoted by $\mathscr{L}(\mathscr{X})$. A *linear prevision* on $\mathscr{L}(\mathscr{X})$ is the expectation operator with respect to a probability on $\mathscr{X}$. A *coherent lower prevision* $\underline{P}$ on $\mathscr{L}(\mathscr{X})$ is the lower envelope of a closed and convex set of linear previsions, which we denote by $\mathscr{M}(\underline{P})$. One particular instance is the *vacuous* lower prevision with respect to a subset A of $\mathscr{X}$, given by $\underline{P}^A(f) = \min_{\omega \in A} f(\omega)$ for all gambles f on $\mathscr{X}$.

Next, consider a number of random variables X_n, $n \in N$, taking values in the respective finite sets $\mathscr{X}_n$. For every subset J of N, we denote by X_J the tuple of random variables (with one component for each $j \in J$) that takes values in the Cartesian product $\mathscr{X}_J = \times_{j \in J} \mathscr{X}_j$. We denote by $\mathscr{L}(\mathscr{X}_J)$ the set of all gambles on $\mathscr{X}_J$. We will frequently use the simplifying device of identifying a gamble f_J on $\mathscr{X}_J$ with its *cylindrical extension* to $\mathscr{X}_N$, which is the gamble f_N defined by $f_N(x_N) = f_J(x_J)$ for all $x_N \in \mathscr{X}_N$, where x_J is the element of $\mathscr{X}_J$ *consistent* with x_N (consistency means here that the components x_j of x_J and x_N coincide for all $j \in J$).

Given two disjoint subsets O and I of N, we define a *conditional lower prevision* $\underline{P}_{O \cup I}(\cdot|X_I)$ as a special two-place function. For any $x_I \in \mathscr{X}_I$, $\underline{P}_{O \cup I}(\cdot|x_I)$ is a real-valued functional on the set $\mathscr{L}(\mathscr{X}_{O \cup I})$ of all gambles on $\mathscr{X}_{O \cup I}$. For any gamble f on $\mathscr{X}_{O\cup I}$, $\underline{P}_{O\cup I}(f|x_I)$ is the *lower prevision of f, conditional on* $X_I = x_I$. Moreover, the object $\underline{P}_{O\cup I}(f|X_I)$ is considered as a gamble on $\mathscr{X}_I$ that assumes the value $\underline{P}_{O\cup I}(f|x_I)$ in x_I.

We define, for any gamble f on $\mathscr{X}_{O\cup I}$, the *$\mathscr{X}_I$-support* $S_I(f)$ of f as $S_I(f) = \left\{x_I \in \mathscr{X}_I \colon \mathbb{I}_{\{x_I\}} f \neq 0\right\}$. Then a number of conditional *linear* previsions $P_{O_j\cup I_j}(\cdot|X_{I_j})$ defined on the sets of gambles $\mathscr{L}(\mathscr{X}_{O_j\cup I_j})$, $j = 1,\ldots,m$ are called *coherent* if for all $f_j \in \mathscr{L}(\mathscr{X}_{O_j\cup I_j})$, $j = 1,\ldots,m$, there is some $j^* \in \{1,\ldots,m\}$, $x \in S_{I_{j^*}}(f_{j^*})$ such that:

$$\left[\sum_{j=1}^{m}\left(f_j - P_{O_j\cup I_j}(f_j|X_{I_j})\right)\right](x_N) \geq 0$$

for some $x_N \in \mathscr{X}_N$ consistent with x. A number of conditional lower previsions $\underline{P}_{O_j\cup I_j}(\cdot|X_{I_j})$ on $\mathscr{L}(\mathscr{X}_{O_j\cup I_j})$, $j = 1,\ldots,m$ are called *coherent* if and only if they are the lower envelopes of some collection $\left\{P^{\lambda}_{O_j\cup I_j}(\cdot|X_{I_j}) \colon \lambda \in \Lambda\right\}$ of coherent conditional linear previsions. In that case, they also satisfy in particular the property of *weak coherence*, which in this context holds if and only if there is some coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ such that $\underline{P}_N(\mathbb{I}_{\{x_{I_j}\}}[f - \underline{P}_{O_j\cup I_j}(f|x_{I_j})]) = 0$ for all $x_{I_j} \in \mathscr{X}_{I_j}$ and all gambles f on $\mathscr{X}_{O_j\cup I_j}$, $j \in \{1,\ldots,m\}$.

Finally, for any non-empty $R \subseteq N$, we denote by $\underline{P}_R$ (and by $\underline{P}_r$ if $R = \{r\}$) the $\mathscr{X}_R$-marginal of a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$, given by $\underline{P}_R(f) = \underline{P}_N(f)$ for all gambles $f \in \mathscr{L}(\mathscr{X}_R)$.

3 Factorisation Conditions

We begin our discussion by introducing a number of generalisations of the notion of an independent product of linear previsions. We have used the first of them in the context of our research on credal networks [2].

Definition 1. *Consider a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$. We call this lower prevision*

1. factorising *if for all $o \in N$ and all non-empty $I \subseteq N \setminus \{o\}$, all $g \in \mathscr{L}(\mathscr{X}_o)$ and all non-negative $f_i \in \mathscr{L}(\mathscr{X}_i)$, $i \in I$, $\underline{P}_N(f_I g) = \underline{P}_N(f_I \underline{P}_N(g))$, where $f_I = \prod_{i\in I} f_i$.*
2. strongly factorising *if $\underline{P}_N(fg) = \underline{P}_N(f\underline{P}_N(g))$ for all $g \in \mathscr{L}(\mathscr{X}_O)$ and $f \in \mathscr{L}(\mathscr{X}_I)$, $f \geq 0$, where I and O are any disjoint proper subsets of N.*

Our notion of factorisation when restricted to lower probabilities and events, is called *strict factorisation* in [7].

Next, we come to a property that V. Kuznetsov [5] first drew attention to:

Definition 2. *Consider a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$. We call this lower prevision*

1. Kuznetsov *if* $\underline{\overline{P}}_N(\prod_{n\in N} f_n) = \boxtimes_{n\in N}\underline{\overline{P}}_n(f_n)$ *for all* $f_n \in \mathscr{L}(\mathscr{X}_n)$, $n \in N$.
2. strongly Kuznetsov *if* $\underline{\overline{P}}_N(fg) = \underline{\overline{P}}_I(f) \boxtimes \underline{\overline{P}}_O(g)$ *for all* $g \in \mathscr{L}(\mathscr{X}_O)$ *and all* $f \in \mathscr{L}(\mathscr{X}_I)$, *where* $\boldsymbol{I}$ *and* $\boldsymbol{O}$ *are any disjoint proper subsets of* N.

Here $\boxtimes$ *is the (commutative and associative) interval product defined by:*

$$\begin{aligned}[a,b]\boxtimes[c,d] &= \{xy\colon x\in[a,b] \text{ and } y\in[c,d]\}\\ &= [\min\{ac,ad,bc,bd\},\max\{ac,ad,bc,bd\}]\end{aligned}$$

for all $a \leq b$ *and* $c \leq d$ *in* $\mathbb{R}$, *and* $\underline{\overline{P}}(f)$ *is the interval* $[\underline{P}(f),\overline{P}(f)]$.

There are the following general relationships between these properties:

Proposition 1. *Consider a coherent lower prevision* $\underline{P}_N$ *on* $\mathscr{L}(\mathscr{X}_N)$. *Then*

$$\begin{array}{ccc} \underline{P}_N \textit{ is strongly Kuznetsov} & \Rightarrow & \underline{P}_N \textit{ is strongly factorising} \\ \Downarrow & & \Downarrow \\ \underline{P}_N \textit{ is Kuznetsov} & \Rightarrow & \underline{P}_N \textit{ is factorising.} \end{array}$$

What about the converse implications? We show in Example 2 that factorisation is not equivalent to being Kuznetsov, and that strong factorisation is not equivalent to being strongly Kuznetsov. In Example 3, we give an instance of a lower prevision that is factorising but not strongly factorising.

4 The Strong Product

The remainder of this paper is devoted to the study of a particular product of coherent lower previsions, called the *strong product*, which also appears under the name *type-1 product* [8, Section 9.3.5]. Our name for it seems to go back to Cozman [1]. If we have coherent lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, then [8, Section 9.3.5] their strong product $\underline{S}_N = \times_{n\in N}\underline{P}_n$ is defined by

$$\begin{aligned}\underline{S}_N(f) &= \inf\{\times_{n\in N}P_n(f)\colon (\forall n\in N)P_n\in\mathscr{M}(\underline{P}_n)\} &\quad (1)\\ &= \inf\{\times_{n\in N}P_n(f)\colon (\forall n\in N)P_n\in ext(\mathscr{M}(\underline{P}_n))\} &\quad (2)\end{aligned}$$

for every $f \in \mathscr{L}(\mathscr{X}_N)$, where $\times_{n\in N}P_n$ is the usual independent product of the considered linear previsions. The strong product of lower previsions satisfies the following marginalisation and associativity properties.

Proposition 2. *Consider coherent lower previsions* $\underline{P}_n$ *on* $\mathscr{L}(\mathscr{X}_n)$, $n \in N$.
1. *For any non-empty subset* R *of* N, $\underline{S}_R$ *is the* $\mathscr{X}_R$*-marginal of* $\underline{S}_N$;
2. $ext(\mathscr{M}(\underline{S}_N)) = \{\times_{n\in N}P_n\colon (\forall n\in N)P_n\in ext(\mathscr{M}(\underline{P}_n))\}$;
3. *For any partition* N_1 *and* N_2 *of* N, $\underline{S}_N = \underline{S}_{N_1} \times \underline{S}_{N_2}$.

We can deduce from the second statement that the infima in Eqs. (1) and (2) are actually minima. This allows us to deduce that the strong product of lower previsions satisfies all the conditions introduced in Section 3.

Proposition 3. *The strong product $\underline{S}_N$ is strongly Kuznetsov, and therefore also Kuznetsov, strongly factorising and factorising.*

This generalises a result established for the case of two variables by Cozman [1]. It also guarantees that the strong product satisfies the weak law of large numbers established in [3].

As a next step, we establish a tighter relationship between the strong product and the epistemic approach to independence. Consider two disjoint proper subsets I and O of N. We say that a subject judges that X_I *is epistemically irrelevant to* X_O when he assumes that learning which value X_I assumes in $\mathscr{X}_I$ will not affect his beliefs about X_O. We say that a subject judges the variables X_n, $n \in N$ to be *epistemically many-to-many independent* when he judges for any disjoint proper subsets I and O of N that X_I is epistemically irrelevant to X_O. If our subject has a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$, and he makes such an assessment, then he can infer from his joint model $\underline{P}_N$ a family of conditional models

$$\mathscr{I}(\underline{P}_N) = \{\underline{P}_{O\cup I}(\cdot|X_I) \colon I \text{ and } O \text{ disjoint proper subsets of } N\},$$

where $\underline{P}_{O\cup I}(\cdot|X_I)$ is a coherent lower prevision on $\mathscr{L}(\mathscr{X}_{O\cup I})$ that is given by:

$$\underline{P}_{O\cup I}(h|x_I) = \underline{P}_N(h(\cdot,x_I)) \text{ for all } h \in \mathscr{L}(\mathscr{X}_{O\cup I}) \text{ and all } x_I \in \mathscr{X}_I.$$

Definition 3. *A coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ is called* many-to-many independent *if it is coherent with the family of conditional lower previsions $\mathscr{I}(\underline{P}_N)$, and in that case it is also called a* many-to-many independent product *of its marginal coherent lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$.*

In a similar way, we say that a subject judges the variables X_n, $n \in N$ to be *epistemically many-to-one independent* when he assumes that learning the value of any number of these variables will not affect his beliefs about any single other. If our subject has a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$, and he makes such an assessment, then he can infer from his joint model $\underline{P}_N$ a family of conditional models

$$\mathscr{N}(\underline{P}_n, n \in N) = \left\{\underline{P}_{\{o\}\cup I}(\cdot|X_I) \colon o \in N \text{ and } I \subseteq N \setminus \{o\}\right\},$$

where $\underline{P}_{\{o\}\cup I}(\cdot|X_I)$ is a coherent lower prevision on $\mathscr{L}(\mathscr{X}_{\{o\}\cup I})$ given by:

$$\underline{P}_{\{o\}\cup I}(h|x_I) = \underline{P}_N(h(\cdot,x_I)) = \underline{P}_o(h(\cdot,x_I))$$

for all gambles h on $\mathscr{X}_{\{o\}\cup I}$ and all $x_I \in \mathscr{X}_I$.

Definition 4. *A coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$ is called* many-to-one independent *if it is coherent*[1] *with the family $\mathscr{N}(\underline{P}_n, n \in N)$, and in that case it is also called a* many-to-one independent product *of its marginal coherent lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$.*

[1] Actually, thanks to [4, Proposition 10], weak coherence suffices here.

A basic coherence result [8, Theorem 7.1.6] states that taking lower envelopes of a family of coherent conditional lower previsions again produces coherent conditional lower previsions. Using this, we can deduce that there always is at least one many-to-many (and therefore also many-to-one) independent product: the strong product.

Proposition 4. *Consider arbitrary coherent lower previsions $\underline{P}_n$ on $\mathscr{L}(\mathscr{X}_n)$, $n \in N$. Then their strong product $\times_{n\in N}\underline{P}_n$ is a many-to-many and many-to-one independent product of the marginal lower previsions $\underline{P}_n$.*

The strong product is not in general the only many-to-one (or many-to-many) independent product of given marginals: there usually are an infinity of them. The smallest is called the *independent natural extension*. We have studied it in detail in another paper [4]. It does not coincide in general with the strong product [8, Section 9.3.4]. Neither is the strong product generally the greatest many-to-one independent product of its marginals:

Example 1. Consider $\mathscr{X}_1 = \mathscr{X}_2 = \{0,1\}$, and let $\underline{P}_1$ and $\underline{P}_2$ be the vacuous lower previsions on $\mathscr{L}(\mathscr{X}_1)$ and $\mathscr{L}(\mathscr{X}_2)$, respectively. Then the strong product $\underline{S}_{\{1,2\}} = \underline{P}_1 \times \underline{P}_2$ is the vacuous lower prevision on $\mathscr{L}(\mathscr{X}_{\{1,2\}})$.

Let $\underline{Q}_{\{1,2\}}$ be the vacuous lower prevision relative to $\{(0,0),(1,1)\}$, which clearly strictly dominates the strong product $\underline{S}_{\{1,2\}}$. To see that it is also a many-to-one independent product of the marginals $\underline{P}_1$ and $\underline{P}_2$, it suffices [cf. footnote 1] to show that $\underline{Q}_{\{1,2\}}(\mathbb{I}_{\{x_1\}}[g_2 - \underline{P}_2(g_2)]) = 0$ for all $x_1 \in \mathscr{X}_1$ and all $g_2 \in \mathscr{L}(\mathscr{X}_2)$ [the case $x_2 \in \mathscr{X}_2$ and all $g_1 \in \mathscr{L}(\mathscr{X}_1)$ is symmetric]. But $\underline{Q}_{\{1,2\}}(\mathbb{I}_{\{x_1\}}[g_2 - \underline{P}_2(g_2)])$ is equal to

$$\min\left\{\mathbb{I}_{\{x_1\}}(0)[g_2(0) - \underline{P}_2(g_2)], \mathbb{I}_{\{x_1\}}(1)[g_2(1) - \underline{P}_2(g_2)]\right\} = 0,$$

since both $\mathbb{I}_{\{x_1\}}(0)[g_2(0) - \underline{P}_2(g_2)]$ and $\mathbb{I}_{\{x_1\}}(1)[g_2(1) - \underline{P}_2(g_2)]$ are non-negative, and at least one of these numbers is zero. ◊

We have shown in [4, Proposition 13] that when $N = \{1,2\}$ and one of the marginals is vacuous, the strong product coincides with the independent natural extension and is therefore the smallest independent product of the given marginals. Example 1 shows it is not the only many-to-one independent product if one of the marginals is vacuous. Yet, it is the only one *factorising*:

Proposition 5. *Let $\underline{P}_1^{A_1}$ be the vacuous lower prevision on $\mathscr{L}(\mathscr{X}_1)$ relative to the non-empty set $A_1 \subseteq \mathscr{X}_1$, and let $\underline{P}_2$ be any coherent lower prevision on $\mathscr{L}(\mathscr{X}_2)$. Then any factorising product $\underline{P}$ of these marginals satisfies*

$$\underline{P}(f) = (\underline{P}_1^{A_1} \times \underline{P}_2)(f) = \min_{x_1 \in A_1} \underline{P}_2(f(x_1,\cdot)) \text{ for all gambles } f \in \mathscr{L}(\mathscr{X}_{\{1,2\}}).$$

We now come to the notion of external additivity:

Definition 5. *Consider a coherent lower prevision $\underline{P}_N$ on $\mathscr{L}(\mathscr{X}_N)$. It is called* externally additive *if for all non-empty $R \subseteq N$ and all gambles f_r on $\mathscr{X}_r$, $r \in R$,*

$\underline{P}_N(\sum_{r\in R} f_r) = \sum_{r\in R} \underline{P}_N(f_r)$, *and* strongly externally additive *if* $\underline{P}_N(f+g) = \underline{P}_N(f) + \underline{P}_N(g)$ *for all* $f \in \mathscr{L}(\mathscr{X}_I), g \in \mathscr{L}(\mathscr{X}_O)$, *where* I *and* O *are any disjoint proper subsets of* N.

Clearly, strong external additivity implies external additivity. Cozman calls the latter *summation independence*, and shows [1, Theorem 1] that the strong product is externally additive for the case of two variables. We generalise this by proving that the strong product is strongly externally additive.

Proposition 6. *Consider arbitrary coherent lower previsions* $\underline{P}_n$, $n \in N$. *Then their strong product* $\underline{S}_N$ *is strongly externally additive.*

We have established in [4, Theorem 5] that the independent natural extension is strongly factorising. We now show that it is not Kuznetsov in general:

Example 2. Consider random variables X_1, X_2 assuming values in $\{0,1\}$, and let their marginal lower previsions be given by

$$\underline{P}_j(f_j) = \frac{1}{2} f_j(0) + \frac{2}{5} f_j(1) + \frac{1}{10} \min\{f_j(0), f_j(1)\} \text{ for all } f_j \in \mathscr{X}_j$$

for $j = 1,2$ (these are linear-vacuous mixtures, and hence coherent [8, Section 2.9.2]). Consider the gambles $f = \mathbb{I}_{\{0\}} - \mathbb{I}_{\{1\}}$ on $\mathscr{X}_1$ and $g = \mathbb{I}_{\{0\}} - \mathbb{I}_{\{1\}}$ on $\mathscr{X}_2$. Then $\underline{P}_1(f) = \underline{P}_2(g) = 0$ and $\overline{P}_1(f) = \overline{P}_2(g) = 1/5$.

As a consequence, $\underline{\overline{P}}_1(f) \boxtimes \underline{\overline{P}}_2(g) = [0, 1/25]$, whereas their independent natural extension assumes [8, Example 9.3.4] the value $\underline{E}_{\{1,2\}}(fg) = -1/11$. This shows that the independent natural extension $\underline{E}_{\{1,2\}}$, which is factorising, is not Kuznetsov. Moreover, in this example where $N = \{1,2\}$, factorisation is equivalent to strong factorisation, and being Kuznetsov to being strongly Kuznetsov. ♢

A convex combination of many-to-one independent products of the same marginals is again a many-to-one independent product of these marginals [4, Proposition 8]. A similar result holds for factorising or Kuznetsov lower previsions. We use these ideas to construct the following counterexample, which shows that a many-to-one independent product is not necessarily many-to-many, and that a factorising lower prevision need not be strongly factorising.

Example 3. Let $N = \{1,2,3\}$. Consider random variables X_j assuming values in $\mathscr{X}_j = \{0,1\}$ for $j = 1,2,3$. Let the corresponding marginal lower previsions be given by

$$\underline{P}_j(f_j) = \frac{1}{2} f_j(0) + \frac{2}{5} f_j(1) + \frac{1}{10} \min\{f_j(0), f_j(1)\} \text{ for all } f_j \in \mathscr{L}(\mathscr{X}_j)$$

for $j = 1,2,3$. Let $\underline{E}_N$ be their independent natural extension and $\underline{S}_N$ their strong product, and define $\underline{Q}_N$ on $\mathscr{L}(\mathscr{X}_N)$ as $\underline{Q}_N = 1/2(\underline{E}_N + \underline{S}_N)$. It follows from Propositions 3, 4 and [4, Proposition 8] that $\underline{Q}_N$ is factorising and a many-to-one independent product. We are going to prove that $\underline{Q}_N$ is neither a many-to-many independent product nor strongly factorising.

Consider the conditional lower prevision $\underline{Q}_N(\cdot|X_3)$ defined from the joint lower prevision $\underline{Q}_N$ using the epistemic irrelevance of X_3 to $X_{\{1,2\}}$. In order to show that $\underline{Q}_N$ is not a many-to-many independent product, it suffices to show that it is not weakly coherent with $\underline{Q}_N(\cdot|X_3)$. Consider the event A that $X_1 = X_2$, and the corresponding indicator gamble $g = \mathbb{I}_A$ on $\mathscr{X}_{\{1,2\}}$. It follows from [8, Example 9.3.4] that $\underline{E}_N(A) = 5/11$ and $\underline{S}_N(A) = 1/2$, so $\underline{Q}_N(A) = 21/44$. Let $x_3 = 0$. Since both $\underline{E}_N$ and $\underline{S}_N$ are strongly factorising (by Proposition 3 and [4, Theorem 5]), we see that $\underline{E}_N(\mathbb{I}_{\{x_3\}}[g - \underline{Q}_N(g)]) = -3/220$ whereas $\underline{S}_N(\mathbb{I}_{\{x_3\}}[g - \underline{Q}_N(g)]) = 1/88$, and then $\underline{Q}_N(\mathbb{I}_{\{x_3\}}[g - \underline{Q}_N(g)]) = -1/440 < 0$. This shows that $\underline{Q}_N$ is not weakly coherent with $\underline{Q}_N(\cdot|X_3)$, and also that it is not strongly factorising. ◊

5 Conclusions

The strong product satisfies all factorisation properties introduced in this paper, and it is a many-to-many independent product of the given marginals. In this sense, it satisfies more factorisation properties than the independent natural extension we studied in another paper [4], because the latter need not be Kuznetsov. Topics of future research could be the generalisation of these results towards infinite spaces as well as the study of the sets of all independent products in some interesting particular cases.

Acknowledgements. Work supported by SBO project 060043 of the IWT-Vlaanderen, by projects TIN2008-06796-C04-01 and MTM2007-61193, and by Swiss NSF grants n. 200020-116674/1 and 200020-121785/1.

References

1. Cozman, F.: Constructing sets of probability measures through Kuznetsov's independence condition. In: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA 2001, Ithaca, NY, USA, pp. 104–111. Shaker Publishing, Maastricht (2001)
2. De Cooman, G., Hermans, F., Antonucci, A., Zaffalon, M.: Epistemic irrelevance in credal nets: the case of imprecise Markov trees. Internat. J. Approx. Reason (2010) (accepted for publication, extended version of a paper in the proceedings of ISIPTA 2009)
3. De Cooman, G., Miranda, E.: Weak and strong laws of large numbers for coherent lower previsions. J. Stat. Plann. Inference 138(8), 2409–2432 (2008)
4. De Cooman, G., Miranda, E., Zaffalon, M.: Independent natural extension. In: Hüllermeier, E., Kruse, R., Hoffman, F. (eds.) Computational Intelligence for Knowledge-Based Systems Design. LNCS, vol. 6178, pp. 737–746. Springer, Heidelberg (2010)
5. Kuznetsov, V.: Interval Statistical Methods. Radio i Svyaz Publ. (1991) (in Russian)

6. Miranda, E.: A survey of the theory of coherent lower previsions. Internat. J. Approx. Reason. 48(2), 628–658 (2008)
7. Vicig, P.: Epistemic independence for imprecise probabilities. Internat. J. Approx. Reason. 24(3), 235–250 (2000)
8. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)

Hadamard Majorants for the Convex Order and Applications

Jesús de la Cal, Javier Cárcamo, and Luis Escauriaza

Abstract. The problem of establishing Hadamard-type inequalities for convex functions on d-dimensional convex bodies ($d \geq 2$) translates into the problem of finding appropriate majorants of the involved random vector for the usual convex order. In this work, we use a stochastic approach based on the Brownian motion to establish a multidimensional version of the classical Hadamard inequality. The main result is closely related to the Dirichlet problem and is applied to obtain inequalities for harmonic functions on general convex bodies.

Keywords: Convex order, Hadamard inequalities, Convex functions, Brownian motion, Dirichlet problem, Harmonic functions.

1 Introduction

It is well known that, for every real convex function f on the interval $[a,b]$, we have

Jesús de la Cal
Departamento de Matemática Aplicada y Estadística e Investigación Operativa, Facultad de Ciencia y Tecnología, Universidad del País Vasco, Apartado 644, 48080 Bilbao, Spain
e-mail: `jesus.delacal@ehu.es`

Javier Cárcamo
Departamento de Matemáticas, Facultad de Ciencias,
Universidad Autónoma de Madrid, 28049 Madrid, Spain
e-mail: `javier.carcamo@uam.es`

Luis Escauriaza
Departamento de Matemáticas, Facultad de Ciencia y Tecnología,
Universidad del País Vasco, Apartado 644, 48080 Bilbao, Spain
e-mail: `luis.escauriaza@ehu.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 149–154.
springerlink.com

$$\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x \le \frac{f(a)+f(b)}{2}. \tag{1}$$

This is the celebrated *Hadamard inequality.* In probabilistic words, it says that

$$\mathrm{E}f(\xi) \le \mathrm{E}f(\xi^*), \qquad f \in \mathscr{C}_{\mathrm{cx}}, \tag{2}$$

where E denotes mathematical expectation, ξ (respectively, ξ^*) is a random variable having the uniform distribution on the interval $[a,b]$ (respectively, on the set $\{a,b\}$), and $\mathscr{C}_{\mathrm{cx}}$ is the set of all real convex functions on $[a,b]$; another way of expressing (2) is

$$\xi \le_{\mathrm{cx}} \xi^*,$$

where $\le_{\mathrm{cx}}$ stands for the so called *convex order* of random variables (see [7] and [10]).

There is an extensive literature devoted to develop applications of the inequality (1), as well as to discuss its extensions, by considering other measures, other kinds of convexity, or higher dimensions. An account of many of such realizations is given in [4].

In this paper, which follows the spirit of [2] and [3], we consider multidimensional analogues of (2), where $[a,b]$ is replaced by a (nonempty) d-dimensional compact convex set $K \subset \mathbb{R}^d$ ($d \ge 2$), ξ is an arbitrary integrable random vector taking values in K, and $\mathscr{C}_{\mathrm{cx}}$ is the set of all real continuous convex functions on K. In this multidimensional setting, we can distinguish two different problems, according to whether the role of $\{a,b\}$ is played by the set of extreme points of K, to be denoted by K^*, or by the boundary of K, to be denoted by K_*.

- *Strong problem*: Find an H*-majorant of ξ, that is, (the probability distribution of) a random vector ξ^* taking values in K^*, such that

$$\mathrm{E}f(\xi) \le \mathrm{E}f(\xi^*), \qquad f \in \mathscr{C}_{\mathrm{cx}}. \tag{3}$$

- *Weak problem*: Find an H$_*$-majorant of ξ, that is, (the probability distribution of) a random vector ξ_* taking values in K_*, such that

$$\mathrm{E}f(\xi) \le \mathrm{E}f(\xi_*), \qquad f \in \mathscr{C}_{\mathrm{cx}}. \tag{4}$$

Since $K^* \subset K_*$, each solution to the strong problem is also a solution to the weak one, and both problems coincide when $K^* = K_*$ (for instance, if K is a closed ball in l_p, with $1 < p < \infty$).

The following theorem is a specific version, in the setting of finite-dimensional spaces, of a more general result established by Niculescu [8] on the basis of Choquet's theory [9].

Theorem 1. *Every K-valued random vector ξ has at least one H^*-majorant.*

This interesting result leaves open the problem of finding such an H*-majorant, a necessary task in order to achieve concrete inequalities of the Hadamard type and other related results.

Remark 1. Observe that (3) (respectively, (4)) implies that $\mathrm{E}\xi = \mathrm{E}\xi^*$ (respectively, $\mathrm{E}\xi = \mathrm{E}\xi_*$), since affine functions are convex.

Remark 2. As it was pointed out in [2], the distributions of ξ_* and ξ^* are not necessarily unique. They depend on the geometric structure of K and on the probability distribution of ξ.

Example 1. In the 1-dimensional case, the strong problem coincides with the weak one. Further, taking into account Theorem 1 and Remark 1, given a random variable ξ defined in the interval $[a,b]$, it is easy to find the *unique* (in the sense of distribution) H*-majorant,

$$\xi^* = \begin{cases} a & \text{with probability } \frac{b-\mathrm{E}\xi}{b-a}, \\ b & \text{with probability } \frac{\mathrm{E}\xi-a}{b-a}. \end{cases}$$

This H*-majorant generates the Hadamard-type inequality

$$\mathrm{E}f(\xi) \le \frac{b-\mathrm{E}\xi}{b-a}f(a) + \frac{\mathrm{E}\xi-a}{b-a}f(b), \qquad f \in \mathscr{C}_{\mathrm{cx}},$$

which was first obtained by Fink [5] (see also [7, Example 1.10.5]).

In this paper, we deal with the weak problem described earlier. That is, given a K-valued random vector ξ, where K is a compact convex set of $\mathbb{R}^d$, we obtain a majorant for the convex order of ξ concentrated on the boundary of K, K_*. This is done in the next section. In Section 3, we show that the main result is closely related to the solution of the Dirichlet problem on K. Therefore, we derive some inequalities for the harmonic functions on K for which the restriction to the boundary is a convex function. Some multidimensional Hadamard-type inequalities are also obtained by using this new approach.

2 Main Result

Let $K \subset \mathbb{R}^d$ be a (nonempty) d-dimensional compact convex set and let ξ be a K-valued random vector. The main result of this section, Theorem 2 below, provides an explicit H$_*$-majorant of ξ. To find such a majorant, we note that given two integrable random vectors ξ and η such that $\xi \le_{\mathrm{cx}} \eta$, the Strassen's theorem (see for instance [10, Theorem 7.A.1., p. 324]) ensures that there exist two random vectors $\hat{\xi}$ and $\hat{\eta}$, defined on the same probability space, such that $\hat{\xi}$ and $\hat{\eta}$ have the same distribution as ξ and η, respectively, and $\{\hat{\xi}, \hat{\eta}\}$ is a martingale, that is,

$$\mathrm{E}[\hat{\eta} \mid \hat{\xi}] = \hat{\xi}, \quad \text{a.s.}$$

Therefore, to find an H_*-majorant of the vector ξ, we can construct a continuous time martingale $\{\xi_t\}_{t\geq 0}$ starting from ξ (i.e., $\xi_0 = \xi$) and stopped at a random time τ on the border of K (i.e., $\xi_\tau \in K_*$). We use the Brownian motion as a natural continuous time martingale connecting ξ with ξ_τ (see Figure 1).

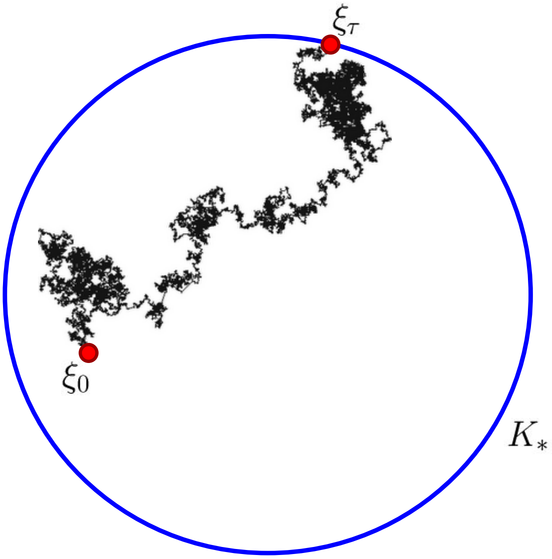


Fig. 1 Construction of an H_*-majorant through the Brownian motion on the disk

Therefore, to achieve our main result we need the Brownian motion. For general notions and results concerning this topic, we refer to [6].

We recall that a d-dimensional Brownian motion is an $\mathbb{R}^d$-valued stochastic process $\{\xi_t : t \geq 0\}$ having the following properties:

(a) it has stationary independent increments,
(b) for all $s \geq 0$ and $t > 0$, $\xi_{s+t} - \xi_s$ has the Gaussian distribution with density

$$g_t(x) = (2\pi t)^{-d/2} e^{-|x|^2/2t}, \qquad x \in \mathbb{R}^d$$

($|\cdot|$ being the Euclidean norm),
(c) with probability 1, it has continuous paths.

The random variable ξ_0 gives the (random) starting point of the process. The process $\{\xi_t - \xi_0 : t \geq 0\}$ is a Brownian motion starting at 0, which is called standard Brownian process. Such a process is defined on some probability space $(\Omega, \mathscr{F}, \mathrm{P})$, and it is well-known that it is a (continuous) martingale with respect to the right-continuous filtration $\{\mathscr{F}_t^+ : t \geq 0\}$ given by

$$\mathscr{F}_t^+ := \bigcap_{s>t} \mathscr{F}_s, \qquad t \geq 0,$$

where, for $s > 0$, $\mathscr{F}_s$ is the sub-σ-field of $\mathscr{F}$ generated by $\{\xi_t : 0 \leq t \leq s\}$.

Our main result is stated as follows. We use the standard convention that the infimum of the empty set is $+\infty$, and we denote by K° the interior of K. Also, $\xi =_{\mathrm{st}} \xi'$ stands for the fact that the random vectors ξ and ξ' have the same probability distribution.

Theorem 2. *Let $K \subset \mathbb{R}^d$ be a (nonempty) d-dimensional compact convex set, let ξ be a K-valued random vector, and let $\{\xi_t : t \geq 0\}$ be a d-dimensional Brownian motion such that $\xi_0 =_{\mathrm{st}} \xi$. Then, the random time τ given by*

$$\tau := \inf\{t \geq 0 : \xi_t \notin K^\circ\}$$

is a stopping time with respect to the filtration $\{\mathscr{F}_t^+ : t \geq 0\}$ that fulfills

$$\mathrm{P}(\tau < \infty) = 1, \tag{5}$$

and the random vector ξ_τ is an H_*-majorant of ξ.

3 Applications: Harmonic Functions and the Dirichlet Problem

It is well known that the stopped Brownian motion that appears in Theorem 2 is connected with the solution of the Dirichlet problem. Since K is convex, K° is obviously a regular domain, in the sense of [6, p. 394]. Then (see the last reference, or [1, p. 90]), for each $g \in \mathscr{C}(K_*)$, the Dirichlet problem

$$\begin{cases} \Delta u = 0 & \text{on } K^\circ \\ u = g & \text{on } K_*, \end{cases}$$

has a unique solution in $\mathscr{C}^2(K^\circ) \cap \mathscr{C}(K)$, to be denoted by Hg, which is given by

$$Hg(x) = \mathrm{E}[g(\xi_\tau) \mid \xi_0 = x], \qquad x \in K, \tag{6}$$

(where $\mathrm{E}[\cdot \mid \cdot]$ denotes conditional expectation). We therefore have

$$\mathrm{E}g(\xi_\tau) = \mathrm{E}[Hg(\xi_0)] = \mathrm{E}[Hg(\xi)]$$

(the last equality because $\xi_0 =_{\mathrm{st}} \xi$), and this yields the following corollary (where we write Hf instead of $Hf_{|K_*}$).

Corollary 1. *For each $f \in \mathscr{C}_{cx}$, the upper Hermite-Hadamard inequality $\mathrm{E}f(\xi) \leq \mathrm{E}f(\xi_\tau)$ can be written in the form*

$$\mathrm{E}f(\xi) \leq \mathrm{E}[Hf(\xi)].$$

This result holds true for each K-valued random vector ξ. Therefore, on taking $\xi \equiv x$ ($x \in K$), we conclude the following.

Corollary 2. *For each $f \in \mathscr{C}_{cx}$, we have*

$$f \leq Hf.$$

Example 2. Let K be the closed unit Euclidean ball in $\mathbb{R}^d$. Using the previous ideas it is possible to generate a multidimensional version of the classical Hadamard inequality (1). When $f \in \mathscr{C}_{\mathrm{cx}}$ and ξ has the uniform distribution on K, we obtain

$$\frac{1}{\mathrm{Vol}(K)} \int_K f(x)\,\mathrm{d}x \leq \frac{1}{\sigma(K_*)} \int_{K_*} f(y)\,\sigma(\mathrm{d}y),$$

where σ is the surface measure on K_*. This result was already found in [2] by using a different approach.

Acknowledgements. This research was supported by the Spanish Government, Projects MTM2008-06281-C02-02 and MTM2007-62186, and by the Basque Government, Project 00390966N. Also, Project MTM2008-06281-C02-02 supports the assistance of J. Cárcamo to the conference SMPS 2010.

We thank two anonymous referees for the careful reading of the paper and for pointing to us some typos of the previous version.

References

1. Bass, R.F.: Probabilistic Techniques in Analysis. Springer, New York (1995)
2. de la Cal, J., Cárcamo, J.: Multidimensional Hermite-Hadamard inequalities and the convex order. J. Math. Anal. Appl. 324, 248–261 (2006)
3. de la Cal, J., Cárcamo, J., Escauriza, L.: A general multidimensional Hermite-Hadamard type inequality. J. Math. Anal. Appl. 356, 659–663 (2009)
4. Dragomir, S.S., Pearce, C.E.M.: Selected Topics on Hermite-Hadamard Inequality and Applications. Victoria University, Melbourne (2000), `http://www.staff.vu.edu.au/rgmia/monographs.asp`
5. Fink, A.M.: A best possible Hadamard inequality. Math. Inequal. Appl. 1, 223–230 (1998)
6. Kallenberg, O.: Foundations of Modern Probability. Springer, New York (1997)
7. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. Wiley, New York (2002)
8. Niculescu, C.P.: The Hermite-Hadamard inequality for convex functions of a vector variable. Math. Inequal. Appl. 5, 619–623 (2002)
9. Phelps, R.R.: Lectures on Choquet's Theorem, 2nd edn. Lecture Notes in Math, vol. 1757. Springer, Berlin (2001)
10. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer Series in Statistics. Springer, New York (2006)

How to Avoid LEM Cycles in Mutual Rank Probability Relations

K. De Loof, B. De Baets, and H. De Meyer

Abstract. The mutual rank probability (MRP) relation of a poset of size $n \geq 9$ can contain linear extension majority (LEM) cycles. We experimentally derive minimum cutting levels for MRP relations of posets of size $n \leq 13$ such that the crisp cut relation is cycle-free.

1 Introduction

In the probability space consisting of the set of linear extensions of a given poset P equipped with the uniform probability measure, the concept of the mutual rank probability (MRP) relation of a poset appears naturally. It plays an important role from an application [7] as well as from a theoretical [4, 12, 17] point of view. Although the study of the type of transitivity exhibited by MRP relations has received considerable attention [4, 12, 19, 21], this transitivity remains far from characterized. It is, however, well known that MRP relations are in general not weakly stochastic transitive [4], allowing for the occurrence of linear extension majority (LEM) cycles in the MRP relation of posets of size $n \geq 9$.

Quite some attention has been given to LEM cycles in literature. Examples of posets with LEM cycles are given in [1, 11, 12, 13, 14, 16, 18], frequency estimates for LEM cycles have been reported in [15, 17], and the occurrence of LEM cycles in certain subclasses of posets has been studied in [2, 9]. Moreover, in previous work [8], the present authors have succeeded in counting the

K. De Loof and H. De Meyer
Department of Applied Mathematics and Computer Science,
Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium
e-mail: karel.deloof@ugent.be

B. De Baets
Department of Applied Mathematics, Biometrics and Process Control,
Ghent University, Coupure links 653, B-9000 Gent, Belgium

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 155–163.
springerlink.com

posets of size $n \leq 13$ with LEM cycles. Besides the fact that the existence of LEM cycles is an intriguing phenomenon in its own right, a better understanding of LEM cycles might help in the ongoing quest to characterize the transitivity of MRP relations. In the present paper we focus on the determination of minimum cutting levels at which the MRP relation becomes free of cycles.

2 Posets, MRP Relations and LEM Cycles

A binary relation $\leq_P$ on a set P is called an *order relation* if it is reflexive ($x \leq_P x$), antisymmetric ($x \leq_P y$ and $y \leq_P x$ imply $x =_P y$) and transitive ($x \leq_P y$ and $y \leq_P z$ imply $x \leq_P z$). A *linear order relation* $\leq_P$ is an order relation in which every two elements are comparable ($x \leq_P y$ or $y \leq_P x$). If $x \leq_P y$ and $x \neq y$, we write $x <_P y$. If neither $x \leq_P y$ nor $x \geq_P y$, we say that x and y are *incomparable* and write $x \,||_P\, y$. A couple $(P, \leq_P)$, where P is a set of objects and $\leq_P$ is an order relation on P, is called a partially ordered set or *poset* for short. The *size* of a poset $(P, \leq_P)$ is defined as the cardinality of P. A poset of size n will be called an n-element poset for short. The poset $(P, \leq_P^\top)$ for which $y \leq_P^\top x$ if and only if $x \leq_P y$ for all $x,y \in P$ is called the *dual poset* of $(P, \leq_P)$.

The binary relation $\prec_P$, for which it holds that $(x,y) \in \prec_P$ if and only if $x <_P y$ and there exists no $z \in P$ such that $x <_P z <_P y$, is called the *covering relation* of $(P, \leq_P)$. The covering relation $\prec_P$ of a poset $(P, \leq_P)$ can be conveniently represented by a so-called *Hasse diagram* where a sequence of connected lines upwards from x to y is present if and only if $x <_P y$. Examples of representations of posets by such Hasse diagrams can be found in the appendix of this paper.

Let Q be a set and R and S two binary relations on Q. If $R \subset S$, then (Q,S) is called an extension of (Q,R). A *linear extension* of a poset $(P, \leq_P)$ is an extension $(P, \leq_L)$ for which $\leq_L$ is a linear order relation. The *mutual rank probability* $\mathrm{p}(x > y)$ of two elements x and y of a poset $(P, \leq_P)$ is defined as the probability that $x >_L y$ in a linear extension $(P, \leq_L)$ that has been sampled uniformly at random from the set of linear extensions of $(P, \leq_P)$. Stated differently, it is the number of linear extensions of $(P, \leq_P)$ in which $x >_L y$, divided by the number of linear extensions of $(P, \leq_P)$. *The mutual rank probability (MRP) relation* M_P is the $[0.1]$-valued binary relation on P defined by $M_P(x,y) = \mathrm{p}(x > y)$ for all $x,y \in P$ where $x \neq y$ and $M_P(x,x) = 1/2$ for all $x \in P$. Note that M_P is a so-called reciprocal relation since $M_P(x,y) + M_P(y,x) = 1$.

The *linear extension majority (LEM) relation* [20] of a poset P is the binary relation $\succ_{\mathrm{LEM}}$ on P such that $x \succ_{\mathrm{LEM}} y$ if $\mathrm{p}(x > y) > \mathrm{p}(y > x)$. Due to the reciprocity of the MRP relation, it is equivalent to define $x \succ_{\mathrm{LEM}} y$ if $\mathrm{p}(x > y) > 1/2$. It is well known [10] that the LEM relation $\succ_{\mathrm{LEM}}$ can contain cycles, *i.e.* subsets $\{x_1, x_2, \ldots, x_m\}$ of elements of P such that $x_1 \succ_{\mathrm{LEM}} x_2 \succ_{\mathrm{LEM}} \cdots \succ_{\mathrm{LEM}} x_m \succ_{\mathrm{LEM}} x_1$, and thus is not transitive. These cycles are referred to as *LEM cycles* on m elements, or *m-cycles* for short.

The *strict δ-cut* with $\delta \in [1/2,1[$ of a reciprocal relation Q defined on a set A is the crisp relation Q^δ defined by

$$Q^\delta(x,y) = \begin{cases} 1 \ , \ \text{if } Q(x,y) > \delta \, , \\ 0 \ , \ \text{otherwise.} \end{cases}$$

We define the *minimum cutting level* δ_m as the smallest number such that for any finite poset the strict δ_m-cut of the corresponding MRP relation is free of l-cycles, with $l \leq m$.

3 Minimum Cutting Levels for Posets of Size $n \leq 13$

The present authors have shown in [6] that the MRP relation can be computed using the lattice of ideals representation of a poset without necessitating the enumeration of all linear extensions. This approach is ideally suited for obtaining the MRP relation of posets of size $n \leq 13$. A combination of the poset generation algorithm of Brinkmann and McKay [3] and the algorithm to compute the MRP relation for a given poset enabled us to obtain exact counts of LEM cycles for posets on up to 13 elements [8]. We adapted this algorithm to keep track of the minimum cutting level δ_m^n and of all posets requiring this cutting level δ_m^n such that all mutual rank probabilities in an m-cycle are greater than or equal to δ_m^n. In Table 1 these minimum cutting levels δ_m^n to avoid m-cycles in n-element posets are shown.

Since one can trivially construct a poset of size $n+1$ from a poset of size n with an equal minimum cutting level by adding an element which is either smaller than, larger than or incomparable to the given n elements, the minimum cutting levels δ_m^n are monotone in n. Therefore, a minimum cutting level to avoid m-cycles avoids all LEM cycles of length $l \leq m$. In Table 1 one can observe that for $n = 11$ no higher cutting level for avoiding 4-cycles is found than for $n = 10$ since $\delta_4^{11} = \delta_4^{10}$, and similarly it is found that $\delta_4^{13} = \delta_4^{12}$.

In Figures 1-14 the posets requiring the non-trivial minimum cutting levels indicated in boldface in Table 1 are depicted by their Hasse diagrams. Note that the dual of a poset has an equal minimum cutting level, and is therefore not shown. However, four depicted posets are identical to their dual posets (Figures 1, 5, 6 and 14). It is also interesting to mention that some posets

Table 1 Minimum cutting level δ_m^n to avoid m-cycles in posets of size $n = 9,\ldots,13$ for $m = 3,\ldots,7$.

$n\backslash m$	3	4	5	6	7
9	**0.50314465**	0.5	0.5	0.5	0.5
10	**0.50396825**	**0.50284900**	0.5	0.5	0.5
11	**0.50619469**	0.50284900	0.5	0.5	0.5
12	**0.50735039**	**0.50866575**	**0.50039788**	**0.50242592**	0.5
13	**0.50886687**	0.50866575	**0.50289997**	**0.50246440**	**0.50018080**

have multiple LEM cycles with an identical cutting level, while others have LEM cycles of different lengths. The 9-element poset in Figure 1, for example, has three 3-cycles with identical probabilities, while for the 12-element poset in Figure 5, aside from the cycle with length 3, a 4-cycle is present, since it holds that $\mathrm{p}(5>7)=\mathrm{p}(7>8)=6184/12244$. The 12-element poset in Figure 7 is quite remarkable in this respect, since aside from the 5-cycle, cycles of length 3, 4 and 6 are present. The poset in Figure 14 even has cycles of length 3, 4, 5, 6 and 7. Furthermore, the poset in Figure 8 also has a 3-cycle, the poset in Figure 11 has cycles of length 3 and 4, and the posets in Figures 12 and 13 both have 3-cycles. For some cutting levels multiple posets, aside from their dual versions, are found. This is the case for the posets of size 13 in Figures 9 and 10 which attain the minimum cutting level for 3-cycles. The same is true for 6-cycles in Figures 12 and 13.

4 Conclusion

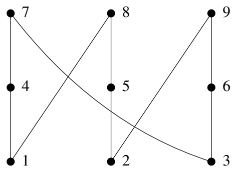
One of the aims of this experiment was to find common properties for posets with LEM cycles or to see a common structure emerging in the posets requiring the minimum cutting level. However, to our surprise the posets have little in common. Possibly due to the fact that the posets are still very limited in size no common (sub)structures can yet be observed for increasing size. The symmetrical and relatively simple structure of the 12-element poset in Figure 6 requiring the minimum cutting level δ_4 inspired us to try to generalize it and to find a lower bound for δ_4 as sharp as possible for increasing poset size. We have reported on these results elsewhere [5].

References

1. Aigner, M.: Combinatorial Search. Wiley-Teubner, Chichester (1988)
2. Brightwell, G., Fishburn, P., Winkler, P.: Interval orders and linear extension cycles. Ars Combin. 36, 283–288 (1993)
3. Brinkmann, G., McKay, B.: Posets on up to 16 points. Order 19, 147–179 (2002)
4. De Baets, B., De Meyer, H., De Loof, K.: On the cycle-transitivity of the mutual rank probability relation of a poset. Fuzzy Sets Syst., doi:10.1016/j.fss.2010.05.005
5. De Loof, K., De Baets, B., De Meyer, H.: Cycle-free cuts of mutual rank probability relations. Discrete Appl. Math. (submitted for publication, 2010)
6. De Loof, K., De Meyer, H., De Baets, B.: Exploiting the lattice of ideals representation of a poset. Fund. Inform. 71, 309–321 (2006)
7. De Loof, K., De Baets, B., De Meyer, H.: A hitchhiker's guide to poset ranking. Comb. Chem. High Throughput Screen. 11, 734–744 (2008)
8. De Loof, K., De Baets, B., De Meyer, H.: Counting linear extension majority cycles in partially ordered sets on up to 13 elements. Comput. Math. Appl. 59(4), 1541–1547 (2010)

9. Ewacha, K., Fishburn, P., Gehrlein, W.: Linear extension majority cycles in height-1 orders. Order 6, 313–318 (1990)
10. Fishburn, P.: On the family of linear extensions of a partial order. J. Combin. Theory Ser. B 17, 240–243 (1974)
11. Fishburn, P.: On linear extension majority graphs of partial orders. J. Combin. Theory Ser. B 21, 65–70 (1976)
12. Fishburn, P.: Proportional transitivity in linear extensions of ordered sets. J. Combin. Theory Ser. B 41, 48–60 (1986)
13. Fishburn, P., Gehrlein, W.: A comparative analysis of methods for constructing weak orders from partial orders. J. Math. Sociol. 4, 93–102 (1975)
14. Ganter, B., Hafner, G., Poguntke, W.: On linear extensions of ordered sets with a symmetry. Discrete Math 63, 153–156 (1987)
15. Gehrlein, W.: Frequency estimates for linear extension majority cycles on partial orders. RAIRO Oper. Res. 25, 359–364 (1991)
16. Gehrlein, W.: The effectiveness of weighted scoring rules when pairwise majority rule cycles exist. Math. Social Sci. 47, 69–85 (2004)
17. Gehrlein, W., Fishburn, P.: Linear extension majority cycles for partial orders. Ann. Oper. Res. 23, 311–322 (1990)
18. Gehrlein, W., Fishburn, P.: Linear extension majority cycles for small ($n \leq 9$) partial orders. Comput. Math. Appl. 20, 41–44 (1990)
19. Kahn, J., Yu, Y.: Log-concave functions and poset probabilities. Combinatorica 18, 85–99 (1998)
20. Kislitsyn, S.: Finite partially ordered sets and their associated sets of permutations. Matematicheskiye Zametki 4, 511–518 (1968)
21. Yu, Y.: On proportional transitivity of ordered sets. Order 15, 87–95 (1998)

Appendix: Posets Requiring Minimum Cutting Levels δ_m^n

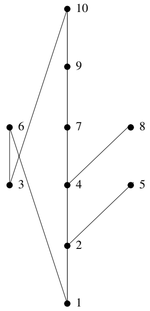


$$\mathrm{p}(7>8)=\mathrm{p}(8>9)=\mathrm{p}(9>7)=\frac{720}{1431}\approx 0.50314465$$

$$\mathrm{p}(4>5)=\mathrm{p}(5>6)=\mathrm{p}(6>4)=\frac{720}{1431}\approx 0.50314465$$

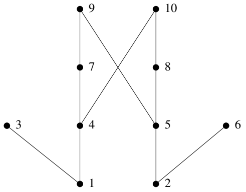
$$\mathrm{p}(1>2)=\mathrm{p}(2>3)=\mathrm{p}(3>1)=\frac{720}{1431}\approx 0.50314465$$

Fig. 1 9-element poset with a LEM cycle requiring the minimal cutting level δ_3^9.



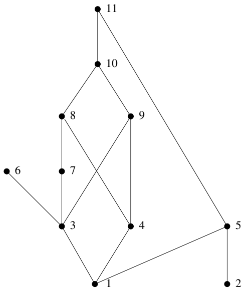
$$\mathrm{p}(8>6)=\mathrm{p}(6>9)=\frac{508}{1008}\approx 0.50396825$$
$$\mathrm{p}(9>8)=\frac{512}{1008}$$

Fig. 2 10-element poset with a LEM cycle requiring the minimal cutting level δ_3^{10}.



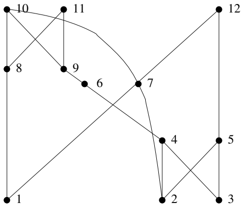
$$\mathrm{p}(7>3)=\mathrm{p}(3>8)=\mathrm{p}(8>6)=\mathrm{p}(6>7)$$
$$=\frac{1765}{3510}\approx 0.50284900$$

Fig. 3 10-element poset with a LEM cycle requiring the minimal cutting level δ_4^{10}.



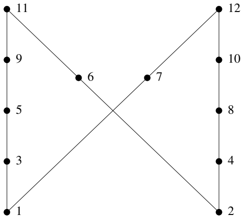
$$\mathrm{p}(5>8)=\frac{1146}{2260}$$
$$\mathrm{p}(8>6)=\frac{1144}{2260}\approx 0.50619469$$
$$\mathrm{p}(6>5)=\frac{1145}{2260}$$

Fig. 4 11-element poset with a LEM cycle requiring the minimal cutting level δ_3^{11}.



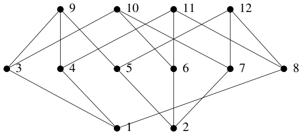
$$\mathrm{p}(8>6)=\mathrm{p}(6>5)=\frac{6214}{12244}$$
$$\mathrm{p}(5>8)=\frac{6212}{12244}\approx 0.50735039$$

Fig. 5 12-element poset with a LEM cycle requiring the minimal cutting level δ_3^{12}.



$$\begin{aligned}\mathrm{p}(5>7)=\mathrm{p}(7>8)=\mathrm{p}(8>6)=\mathrm{p}(6>5)\\=\frac{7396}{14540}\approx 0.50866575\end{aligned}$$

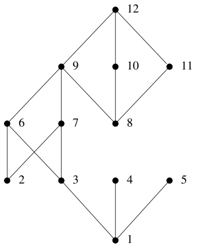
Fig. 6 12-element poset with a LEM cycle requiring the minimal cutting level δ_4^{12}.



$$\mathrm{p}(5>4)=\mathrm{p}(4>3)=\frac{60400}{120640}$$

$$\mathrm{p}(3>6)=\mathrm{p}(6>8)=\mathrm{p}(8>5)=\frac{60368}{120640}\approx 0.50039788$$

Fig. 7 12-element poset with a LEM cycle requiring the minimal cutting level δ_5^{12}.

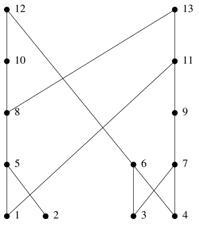


$$\mathrm{p}(7>4)=\mathrm{p}(6>5)=\frac{46392}{92336}\approx 0.50242592$$

$$\mathrm{p}(4>10)=\mathrm{p}(5>11)=\frac{46560}{92336}$$

$$\mathrm{p}(10>6)=\mathrm{p}(11>7)=\frac{46850}{92336}$$

Fig. 8 12-element poset with a LEM cycle requiring the minimal cutting level δ_6^{12}.

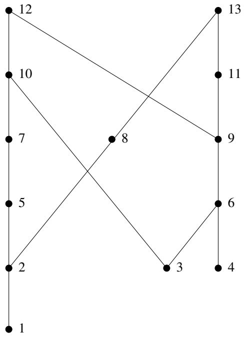


$$\mathrm{p}(6>8)=\frac{12240}{24022}$$

$$\mathrm{p}(8>9)=\frac{12262}{24022}$$

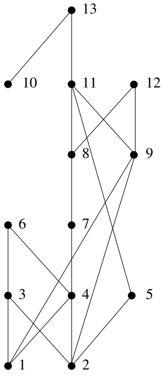
$$\mathrm{p}(9>6)=\frac{12224}{24022}\approx 0.50886687$$

Fig. 9 First 13-element poset with a LEM cycle requiring the minimal cutting level δ_3^{13}.



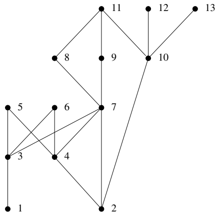
$$p(7>8)=\frac{6112}{12011}\approx 0.50886687$$
$$p(8>9)=\frac{6120}{12011}$$
$$p(9>7)=\frac{6131}{12011}$$

Fig. 10 Second 13-element poset with a LEM cycle requiring the minimal cutting level δ_3^{13}.



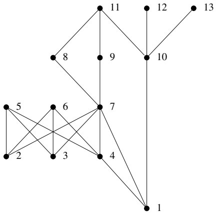
$$p(10>9)=\frac{33871}{67242}$$
$$p(9>5)=\frac{33916}{67242}$$
$$p(5>7)=\frac{33816}{67242}\approx 0.50289997$$
$$p(7>3)=\frac{33834}{67242}$$
$$p(3>10)=\frac{34151}{67242}$$

Fig. 11 13-element poset with a LEM cycle requiring the minimal cutting level δ_5^{13}.



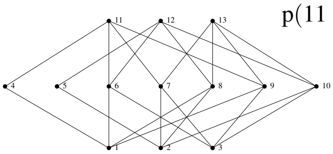
$$p(12>9)=p(13>8)=\frac{66354}{131472}$$
$$p(9>6)=p(8>5)=\frac{66060}{131472}\approx 0.50246440$$
$$p(6>13)=p(5>12)=\frac{66306}{131472}$$

Fig. 12 First 13-element poset with a LEM cycle requiring the minimal cutting level δ_6^{13}.



$$\begin{aligned} p(12 > 9) = p(13 > 8) &= \frac{132708}{262944} \\ p(9 > 6) = p(8 > 5) &= \frac{132120}{262944} \approx 0.50246440 \\ p(6 > 13) = p(5 > 12) &= \frac{132612}{262944} \end{aligned}$$

Fig. 13 Second 13-element poset with a LEM cycle requiring the minimal cutting level δ_6^{13}.



$$\begin{aligned} p(11 > 6) = p(6 > 8) = p(4 > 11) &= \frac{268352}{536510} \approx 0.50018080 \\ p(8 > 10) = p(5 > 4) &= \frac{268384}{536510} \\ p(10 > 12) = p(12 > 5) &= \frac{268465}{536510} \end{aligned}$$

Fig. 14 13-element poset with a LEM cycle requiring the minimal cutting level δ_7^{13}.

Functional Inequalities Characterizing the Frank Family of Copulas

Hans De Meyer and Bernard De Baets

Abstract. Given a random vector with components that are pairwisely coupled by means of a same commutative copula C, we analyze the transitivity of the reciprocal relation obtained from the pairwise comparison of these components. The transitivity of this reciprocal relation can be elegantly described within the cycle-transitivity framework if the commutative copula C satisfies a countably infinite family of (functional) inequalities. Each functional inequality uniquely characterizes the Frank family of copulas. Finally, we highlight the transitivity results for a random vector whose coupling structure is captured by an extended Frank m-copula.

Keywords: Copulas, Frank family, Functional inequality, Reciprocal relations, Transitivity.

1 Introduction

Many methods can be established for the comparison of the components (random variables, r.v.) of a random vector $(X_1,\ldots,X_n)$, as there are many ways to extract useful information from the joint cumulative distribution function (c.d.f.) $F_{X_1,\ldots,X_n}$ that characterizes the random vector.

A simplification consists in restricting the comparison strategy to methods that aim at comparing the r.v. two by two. We have recently put forward

Hans De Meyer
Department of Applied Mathematics and Computer Science, Ghent University,
B-9000 Gent, Belgium
e-mail: hans.demeyer@ugent.be

Bernard De Baets
Department of Applied Mathematics, Biometrics and Process Control,
Ghent University, B-9000 Gent, Belgium
e-mail: bernard.debaets@ugent.be

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 165–172.
springerlink.com

such a method [8]. We associate to a random vector a reciprocal relation Q, which can be regarded as a graded preference relation. The cornerstone for computing the reciprocal relation Q is the copula C_{ij} that joins the one-dimensional marginal c.d.f. F_{X_i} and F_{X_j} into the bivariate marginal c.d.f. F_{X_i,X_j}, i.e. $F_{X_i,X_j} = C_{ij}(F_{X_i}, F_{X_j})$. Note that the copulas should not be the same for all pairs of r.v. For a collection of independent r.v., however, they all coincide with the product $T_{\mathbf{P}}(x,y) = xy$.

We have analyzed the case where all copulas C_{ij} are the same but not necessarily equal to the product, neither to the greatest copula $T_{\mathbf{M}}(x,y) = \min(x,y)$, the minimum operator, nor to the smallest copula $T_{\mathbf{L}}(x,y) = \max(x+y-1,0)$, also known as the Łukasiewicz t-norm. Note that in the latter case the pairwise couplings should be considered as purely artificial, as no n-copula ($n \geq 3$) exists such that all 2-copulas contained in it are equal to $T_{\mathbf{L}}$. Our analysis has revealed that the reciprocal relations generated by these couplings possess transitivity properties that can be nicely characterized [4, 7, 8, 9, 10].

The concept of transitivity is unique for crisp relations, but for reciprocal relations there is a whole range of transitivity properties. Sometimes it is possible to capture the transitivity in the form of a type of stochastic transitivity or a type of T-transitivity, with T a t-norm, well known from the theory of fuzzy relations, but mostly these types prove insufficient to deal with the transitivity of reciprocal relations. Instead, we have developed a new framework, called the cycle-transitivity framework, that allows to characterize the types of transitivity that arise in the present investigation [3, 5].

As a by-product of our investigations, we have laid bare an infinite family of functional inequalities, each of which characterizes the Frank family of copulas [12]. In the past, many investigations were aimed at finding the solution(s) of functional equations in the space of uniform distribution functions [1]. The functional equation of Frank [11], the Frank equation for short, is perhaps the best known example. This equation, however, does not characterize uniquely the Frank family of copulas, as it also has as solutions the ordinal sums of Frank copulas. Note that Frank copulas and ordinal sums of Frank copulas are more often regarded as t-norms [13] and that in this context the Frank equation has even been studied for the more general class of uninorms [2]. The fact that a sharper characterization of a family of copulas can be acquired by means of a functional inequality, rather than by means of a functional equation, has to our knowledge, not been recognized before.

In the next section, we briefly summarize the concept of cycle-transitivity. In Section 3 we recall the method used to compare r.v. and the way a reciprocal relation is generated from it. We investigate its transitivity when the r.v. are coupled with the same copula and derive an infinite family of inequalities which the copula should satisfy. Section 4 emphasizes the role of the Frank family of copulas as unique solutions of the family of inequalities. Finally, Section 5 is concerned with the transitivity of the reciprocal relation generated by a random vector whose coupling structure is described by a Frank m-copula.

2 Cycle-Transitivity of Reciprocal Relations

Reciprocal ($[0,1]$-valued binary relations Q satisfying $Q(a,b)+Q(b,a)=1$, provide a convenient tool for expressing the result of the pairwise comparison of a set of alternatives. Recently, we have presented a general framework for studying the transitivity of reciprocal relations, encompassing various types of T-transitivity and stochastic transitivity [3, 5].

Recall that a fuzzy relation R on A is an $A^2 \to [0,1]$ mapping that expresses the degree of relationship between elements of A. For such relations, the concept of T-transitivity is very natural.

Definition 1. *Let T be a t-norm. A fuzzy relation R on A is called T-transitive if for any $(a,b,c) \in A^3$ it holds that $T(R(a,b),R(b,c)) \leq R(a,c)$.*

Though the semantics of reciprocal relations and fuzzy relations are different, the concept of T-transitivity is sometimes formally applied to reciprocal relations as well. However, more often the transitivity properties of reciprocal relations can be characterized as of one of various kinds of stochastic transitivity [3].

In the cycle-transitivity framework [5], for a reciprocal relation Q on A, the quantities

$$\alpha_{abc} = \min(Q(a,b),Q(b,c),Q(c,a))\,,\ \beta_{abc} = \mathrm{med}(Q(a,b),Q(b,c),Q(c,a))\,,$$

$$\gamma_{abc} = \max(Q(a,b),Q(b,c),Q(c,a))\,,$$

are defined for all $(a,b,c) \in A^3$. Obviously, $\alpha_{abc} \leq \beta_{abc} \leq \gamma_{abc}$. Also, the notation $\Delta = \{(x,y,z) \in [0,1]^3 \,|\, x \leq y \leq z\}$ will be used.

Definition 2. *A function $U : \Delta \to \mathbb{R}$ is called an upper bound function if it satisfies:*

(i) $U(0,0,1) \geq 0$ and $U(0,1,1) \geq 1$;
(ii) for any $(\alpha,\beta,\gamma) \in \Delta$:

$$U(\alpha,\beta,\gamma) + U(1-\gamma,1-\beta,1-\alpha) \geq 1\,.$$

Definition 3. *A reciprocal relation Q on A is called cycle-transitive w.r.t. an upper bound function U if for any $(a,b,c) \in A^3$ it holds that*

$$\alpha_{abc} + \beta_{abc} + \gamma_{abc} - 1 \leq U(\alpha_{abc},\beta_{abc},\gamma_{abc})\,.$$

For two upper bound functions such that $U_1 \leq U_2$, it clearly holds that cycle-transitivity w.r.t. U_1 implies cycle-transitivity w.r.t. U_2. It is clear that $U_1 \leq U_2$ is not a necessary condition for the latter implication to hold. Two upper bound functions U_1 and U_2 will be called *equivalent* if for any $(\alpha,\beta,\gamma) \in \Delta$ it holds that $\alpha+\beta+\gamma-1 \leq U_1(\alpha,\beta,\gamma)$ is equivalent to $\alpha+\beta+\gamma-1 \leq U_2(\alpha,\beta,\gamma)$.

The different types of fuzzy and stochastic transitivity can be reformulated in the cycle-transitivity framework and are then characterized by an upper bound function $U(\alpha,\beta,\gamma)$.

Proposition 1. *A reciprocal relation Q is T-transitive, with T a 1-Lipschitz t-norm (or equivalently, an associative copula), if and only if Q is cycle-transitive w.r.t. to the upper bound function U_T given by*

$$U_T(\alpha,\beta,\gamma) = \alpha+\beta-T(\alpha,\beta)\,.$$

In fact, many examples of reciprocal relations we have encountered in our research on the comparison of random variables are neither fuzzy nor stochastic transitive, but have a type of transitivity that can be nicely expressed as an instance of cycle-transitivity. In the present study we will encounter the weaker counterparts of T-transitivity obtained by replacing in the expression for U_T α by β and β by γ.

Definition 4. *A reciprocal relation Q that is cycle transitive w.r.t. to the upper bound function U_{WT} defined by*

$$U_{WT}(\alpha,\beta,\gamma) = \beta+\gamma-T(\beta,\gamma)\,,$$

with T a 1-Lipschitz t-norm, is called weak T-transitive.

Weak $T_{\mathbf{M}}$-transitivity is also known as partial stochastic transitivity, weak $T_{\mathbf{P}}$-transitivity as dice-transitivity or weak product transitivity, whereas weak $T_{\mathbf{L}}$-transitivity is equivalent to $T_{\mathbf{L}}$-transitivity.

3 Generating Transitive Reciprocal Relations from Random Vectors

An immediate way of comparing two r.v. is to consider the probability that the first one takes a greater value than the second one. Proceeding along this line of thought, a random vector $(X_1,X_2,\ldots,X_n)$ generates a reciprocal relation.

Definition 5. *Given a random vector $(X_1,X_2,\ldots,X_n)$, the binary relation Q defined by*

$$Q(X_i,X_j) = \mathrm{Prob}\{X_i > X_j\} + \frac{1}{2}\mathrm{Prob}\{X_i = X_j\}$$

is a reciprocal relation.

Since the copulas C_{ij} that couple the univariate marginal c.d.f. into the bivariate marginal c.d.f. can be different from another, the analysis of the reciprocal relation and in particular the identification of its transitivity properties appear rather cumbersome. It is nonetheless possible to state in general, without making any assumptions on the bivariate c.d.f., that the reciprocal relation

Q generated by an arbitrary random vector always shows some minimal form of transitivity.

Proposition 2. [4] *The reciprocal relation Q generated by a random vector is $T_{\mathbf{L}}$-transitive.*

Our further interest is to study the situation where momentarily abstraction is made that the r.v. are components of a random vector, and all bivariate c.d.f. are enforced to depend in the same way upon the univariate c.d.f., in other words, we consider the situation of all copulas being the same.

To get insight in what kind of transitivity properties one might expect in general, the present authors have previously unravelled three particular cases, namely the case of the product copula $T_{\mathbf{P}}$, and the cases of the two extreme copulas, the minimum operator $T_{\mathbf{M}}$ and the Łukasiewicz t-norm $T_{\mathbf{L}}$, respectively related to a presumed but not-necessarily existing comonotonic and countermonotonic pairwise dependence of the r.v. [15]. From these studies the following results can be reported.

Proposition 3. [8, 10] *The reciprocal relation Q generated by a collection of independent random variables (i.e. pairwisely coupled by $T_{\mathbf{P}}$) is dice-transitive (weak $T_{\mathbf{P}}$-transitive).*

Proposition 4. [7, 9] *The reciprocal relation Q generated by a collection of random variables pairwisely coupled by $T_{\mathbf{M}}$ is $T_{\mathbf{L}}$-transitive.*

Proposition 5. [7, 9] *The reciprocal relation Q generated by a collection of random variables pairwisely coupled by $T_{\mathbf{L}}$ is partially stochastic transitive (weak $T_{\mathbf{M}}$-transitive).*

We further considered the case where all C_{ij} are the same copula C. It then turns out that the transitivity of the generated reciprocal relation Q can only be captured as a type of cycle-transitivity when the copula C fulfills a countably infinite family of conditions. These conditions are presented in the form of inequalities. Hence, we require C to be a solution of an infinite system of inequalities.

Proposition 6. [4] *Let C be a commutative copula such that for any $k > 1$ and for all $0 \leq x_1 \leq x_2 \leq \cdots \leq x_k \leq 1$ and $0 \leq y_1 \leq y_2 \leq \cdots \leq y_k \leq 1$, it holds that*

$$\begin{aligned}&\sum_i C(x_i, y_i) - \sum_j C(x_{k-2j}, y_{k-2j-1}) - \sum_j C(x_{k-2j-1}, y_{k-2j})\\ &\leq C\Bigg(x_k + \sum_j C(x_{k-2j-2}, y_{k-2j-1}) - \sum_j C(x_{k-2j}, y_{k-2j-1}),\\ &\qquad y_k + \sum_j C(x_{k-2j-1}, y_{k-2j-2}) - \sum_j C(x_{k-2j-1}, y_{k-2j})\Bigg), \qquad (1)\end{aligned}$$

where the sums extend over all integer values that lead to meaningful indices of x and y. Then the reciprocal relation Q generated by a collection of random variables pairwisely coupled by C is cycle-transitive w.r.t. to the upper bound function U^C defined by:

$$U^C(\alpha,\beta,\gamma) = \max(\beta + C(1-\beta,\gamma), \gamma + C(\beta, 1-\gamma))\,.$$

If C is stable, i.e. $C(x,y)+1-C(1-x,1-y) = x+y$ for all $(x,y) \in [0,1]^2$, then

$$U^C(\alpha,\beta,\gamma) = \beta + C(1-\beta,\gamma) = \gamma + C(\beta,1-\gamma))\,.$$

Note that symmetrical ordinal sums of Frank copulas are stable [14].

4 Solving the Family of Inequalities

It is natural to ask whether commutative copulas that fulfil (1) can be characterized in an alternative way. However, they are not necessarily satisfied for any stable commutative copula, as is illustrated by the following example of a symmetrical ordinal sum of two Frank copulas.

Example 1. Let C be the commutative copula defined by

$$C(x,y) = \begin{cases} \frac{1}{3} + \max(x+y-1, 0) & , \text{ if } (x,y) \in [1/3,2/3]^2\,, \\ \min(x,y) & , \text{ elsewhere}\,. \end{cases}$$

It is the ordinal sum $\langle 1/3, 2/3, T_{\mathbf{L}} \rangle$ with $T_{\mathbf{L}}$ linearly rescaled to the square $[1/3,2/3]^2$. It is easily verified that C is stable (as it is a symmetrical ordinal sum of Frank copulas [14]). Let $x_1 = y_1 = 1/4$ and $x_2 = y_2 = 3/4$. For $n = 2$, the left-hand side of (1) becomes $C(1/4,1/4)+C(3/4,3/4)-C(1/4,3/4)-C(3/4,1/4) = 1/4+3/4-1/4-1/4 = 1/2$, while the right-hand side evaluates to $C(x_2 - C(x_2,y_1), y_2 - C(x_1,y_2)) = C(1/2,1/2) = 1/3$, showing that (1) does not hold for $n = 2$ and for all $0 \le x_1 \le x_2 \le 1$ and $0 \le y_1 \le y_2 \le 1$.

In [4], we conjectured that for the Frank copulas themselves conditions (1) are always satisfied but only recently we were able to give a complete proof [6]. Moreover, the Frank copulas are the only associative copulas that are solution of all inequalities separately. Hence, each inequality uniquely characterizes the Frank family of copulas. Special attention is drawn on the first one ($k = 2$), which we call 'the Frank inequality'.

Proposition 7. *Let C be an associative copula. The following statements are equivalent:*

(i) For any $0 \le x \le x' \le 1$ and $0 \le y \le y' \le 1$, it holds that

$$C(x,y) + C(x',y') - C(x,y') - C(x',y) \le C(x' - C(x',y), y' - C(x,y'))\,;$$

(ii) C is a member of the Frank family of copulas.

Note that the left-hand side of (i) is the C-volume of the rectangle $[x,x'] \times [y,y']$ and the condition states that this C-volume is bounded from above by some well-defined C-dependent quantity, i.e. the C-volume of the rectangle $[0,x'-C(x',y)] \times [0,y'-C(x,y')]$.

Proposition 7 can be extended to the other inequalities contained in 1.

Proposition 8. *Let C be an associative copula. The following statements are equivalent:*

(i) C satisfies all inequalities contained in (1);
(ii) C satisfies any one of the inequalities contained in (1);
(iii) C is a member of the Frank family of copulas.

The following result is now immediate.

Proposition 9. *The reciprocal relation Q generated by a collection of random variables pairwisely coupled by the Frank copula $T^{\mathbf{F}}_\lambda$ is cycle-transitive w.r.t. the upper bound function $U^{\mathbf{F}}_\lambda$ given by:*

$$U^{\mathbf{F}}_\lambda(\alpha,\beta,\gamma) = \beta + T^{\mathbf{F}}_\lambda(1-\beta,\gamma) = \beta+\gamma-T^{\mathbf{F}}_{1/\lambda}(\beta,\gamma),$$

otherwise stated, Q is weak $T^{\mathbf{F}}_{1/\lambda}$-transitive.

In the above transition, we have used the fact that $T^{\mathbf{F}}_\lambda(1-x,y) = y - T^{\mathbf{F}}_{1/\lambda}(x,y)$. Since for $\lambda \leq \lambda'$ it holds that $T^{\mathbf{F}}_\lambda \geq T^{\mathbf{F}}_{\lambda'}$, it also follows that $U^{\mathbf{F}}_\lambda \geq U^{\mathbf{F}}_{\lambda'}$. Therefore, the lower the value of λ when the r.v. are coupled by $T^{\mathbf{F}}_\lambda$, the weaker the type of transitivity exhibited by the probabilistic relation generated by these r.v. In particular, the strongest type of transitivity is encountered when coupling by $T_{\mathbf{L}}$ (i.e. partial stochastic transitivity), the weakest when coupling by $T_{\mathbf{M}}$ (i.e. $T_{\mathbf{L}}$-transitivity).

5 Reciprocal Relations Generated by Frank m-Copulas

So far, we have considered collections of r.v. that are pairwisely coupled by the same Frank copula. The obtained transitivity results can be easily extended to the case of random vectors with m components, such that the components are pairwisely coupled by the same Frank copula. It is known that such random vectors exist for certain values of the λ-parameter.

Definition 6. *For any $m \geq 2$ and any $\lambda \in]0,1]$, the m-ary function $C^{m\mathbf{F}}_\lambda : [0,1]^m \to [0,1]$, defined by*

$$C^{m\mathbf{F}}_\lambda(x_1,x_2,\ldots,x_m) = \log_\lambda\left[1+\frac{(\lambda^{x_1}-1)(\lambda^{x_2}-1)\ldots(\lambda^{x_m}-1)}{(\lambda-1)^{m-1}}\right], \qquad (2)$$

is an m-copula, called Frank m-copula.

Note that the definition can be slightly extended to cover the parameter values $\lambda \in]0, s_m]$, where s_m is an m-dependent upper bound. For instance, $s_3 = 2$, $s_4 = 3 - \sqrt{3}$, $s_5 = 2(3 - \sqrt{6})$, ..., $s_9 = 1.00438$, $\lim_{k\to\infty} s_k = 1$.

Proposition 10. *The reciprocal relation Q generated by a random vector with m components and coupled by $C_\lambda^{m\mathbf{F}}$ is weak $T_{1/\lambda}^{\mathbf{F}}$-transitive.*

References

1. Alsina, C.: Some functional equations in the space of uniform distribution functions. Aequationes Math. 22, 153–164 (1981)
2. Calvo, T., De Baets, B., Fodor, J.: The functional equations of Frank and Alsina for uninorms and nullnorms. Fuzzy Sets Syst. 120, 385–394 (2001)
3. De Baets, B., De Meyer, H.: Transitivity frameworks for reciprocal relations: cycle-transitivity versus FG-transitivity. Fuzzy Sets Syst. 152, 249–270 (2005)
4. De Baets, B., De Meyer, H.: On the cycle-transitive comparison of artificially coupled random variables. Internat. J. Approx. Reason. 47, 306–322 (2008)
5. De Baets, B., De Meyer, H., De Schuymer, B., Jenei, S.: Cyclic evaluation of transitivity of reciprocal relations. Soc. Choice Welf. 26, 217–238 (2006)
6. De Meyer, H., De Baets, B.: The Frank inequality (in preparation, 2010)
7. De Meyer, H., De Baets, B., De Schuymer, B.: On the transitivity of the comonotonic and countermonotonic comparison of random variables. J. Multivariate Anal. 98, 177–193 (2007)
8. De Schuymer, B., De Meyer, H., De Baets, B.: Cycle-transitive comparison of independent random variables. J. Multivariate Anal. 96, 352–373 (2005)
9. De Schuymer, B., De Meyer, H., De Baets, B.: Extreme copulas and the comparison of ordered lists. Theory and Decision 62, 195–217 (2007)
10. De Schuymer, B., De Meyer, H., De Baets, B., Jenei, S.: On the cycle-transitivity of the dice model. Theory and Decision 54, 261–285 (2003)
11. Frank, M.: On the simultaneous associativity of $F(x,y)$ and $x + y - F(x,y)$. Aequationes Math. 19, 194–226 (1979)
12. Genest, C.: Frank's family of bivariate distributions. Biometrika 74, 549–555 (1987)
13. Klement, E., Mesiar, R., Pap, E.: Triangular Norms. Trends in Logic—Studia Logica Library, vol. 8. Kluwer Academic Publishers, Dordrecht (2000)
14. Klement, E., Mesiar, R., Pap, E.: Invariant copulas. Kybernetika 38, 275–285 (2002)
15. Nelsen, R.: An Introduction to Copulas, 2nd edn. Springer, Heidelberg (2006)

Recent Developments in Censored, Non-Markov Multi-State Models

Jacobo de Uña-Álvarez

Abstract. Nonparametric estimation of transition probabilities for a censored multi-state model is traditionally performed under a Markov assumption. However, this assumption may (and will) fail in some applications, leading to the inconsistency of the time-honoured Aalen-Johansen estimator. In such a case, alternative (non-Markov) estimators are needed. In this work we review some recent developments in this area. We also review the key problem of testing if a given (censored) multi-state model is Markov, giving modern ideas for the construction of an omnibus test statistic.

1 Introduction

A multi-state model is a model for a stochastic process $\{X(t), t \geq 0\}$ allowing individuals to move along a finite number of states. At each time point t, $X(t)$ denotes the state occupied by a representative individual in a homogeneous population, and let $\{X_r(t), t \geq 0\}$, $r = 1, ..., n$, be a collection of n trajectories (or histories) corresponding to n subjects randomly sampled from the target population. In this setup, much effort has been made to estimate the so-called transition probabilities

$$p_{ij}(s,t) = P\left(X(t) = j | X(s) = i\right) \tag{1}$$

where i and j are two states, and $0 \leq s < t$. The obvious estimator for $p_{ij}(s,t)$ is the empirical transition probability

$$p_{ij,n}(s,t) = \frac{\sum_{r=1}^{n} I\left(X_r(t) = j, X_r(s) = i\right)}{\sum_{r=1}^{n} I\left(X_r(s) = i\right)} \tag{2}$$

which is simply the proportion of observed transitions from i to j in the time interval $[s,t]$. In practice, this estimator is typically unavailable because of

Jacobo de Uña-Álvarez
Department of Statistics and OR, University of Vigo, Spain
e-mail: jacobo@uvigo.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 173–179.
springerlink.com

censoring. [1] introduced a nonparametric estimator for $p_{ij}(s,t)$ under censoring. The Aalen-Johansen estimator has become the standard tool for estimating the transition probabilities in a nonparametric way. However, it is constructed on the basis that the underlying process is Markovian, and its consistency can not be ensured in general.

To be more precise, applications of multi-state models to biomedical data have shown that the Markov assumption is sometimes violated. A stochastic process is said to be Markovian when, given the present, the future evolution does not depend on the past. Consider as an illustrative example the PROVA trial of bleeding episodes and mortality in liver cirrhosis in [2]. In this example, a three-state model is used to represent the individuals' histories; this model allows for three possible transitions in a progressive way: from healthy (no bleeding) to bleeding, from bleeding to death, and directly from healthy to death. This model is usually named 'illness-death model', and it is progressive in the sense that past states can not be visited again. [2] provided evidence (in agreement to previous studies) on the fact that the mortality is markedly increased shortly after the bleeding episode. This means that subjects in state 'bleeding' may (and will) have a different prognosis according to their entry times in that state (i.e., the past history is important for the future, so the process is not Markov).

There has been some investigation oriented to analyze the properties of the Aalen-Johansen estimator when the Markov assumption fails. For example, Aalen et al. [1] and Datta and Satten [3] established the consistency of the Aalen-Johansen estimator of the stage occupation probabilities $P_j(t) = P(X(t) = j)$ for a non-Markov process, while Glidden [4] developed confidence bands for such an estimator. More recently, however, Meira-Machado et al. [5] showed that, in general, the Aalen-Johansen estimator of $p_{ij}(s,t)$ may be dramatically biased if the Markov assumption is not fulfilled. The practical conclusion is that one should asses the Markovianity of the process before using the Aalen-Johansen estimator for estimation and inference purposes. And, if there is evidence of non-Markovianity, some alternative estimators should be used.

In this work we review some recent developments in nonparametric estimation of transition probabilities in non-Markov multi-state models. It is assumed that the available trajectories can be right-censored by a potential censoring time that is independent of the process. The available estimators are given in Section 2. In Section 3, we consider the problem of testing if a given process is Markov, reviewing the traditional approach and giving some modern alternative ideas too.

2 Non-Markov Transition Probabilities

For the best of our knowledge, Meira-Machado et al. [5] proposed for the first time nonparametric estimators for the transition probabilities of a censored non-Markov multi-state model. These authors considered the progressive illness-death (or disability) model, which consists in three different states (1='healthy', 2='diseased', and 3='dead') and the three possible transitions 1→2, 2→3, and 1→3. Put Z and T for the sojourn time in state 1 and the total survival time of the process (that is, the time up to reaching the absorbing state 3) respectively. It is seen in [5] that the transition probabilities are probabilities involving (Z,T), and hence the question becomes how the joint distribution function of (Z,T) can be consistently estimated under censoring. Here we briefly present their ideas, with a slightly different notation to simplify things.

The available information is represented by $\left(\widetilde{Z},\widetilde{T},\Delta_1,\Delta\right)$, where $\widetilde{Z}$ and $\widetilde{T}$ stand for the censored versions of Z and T, and Δ_1 and Δ are their respective censoring indicators. Note that the individual is observed to pass through state 2 if and only if $\widetilde{Z}<\widetilde{T}$, and in such a case $\widetilde{Z}$ is uncensored. Let $\left\{\left(\widetilde{Z}_i,\widetilde{T}_i,\Delta_{1i},\Delta_i\right),1\le i\le n\right\}$ be an iid sample of $\left(\widetilde{Z},\widetilde{T},\Delta_1,\Delta\right)$, and let $W_i=\frac{\Delta_i}{n-R_i+1}\prod_{R_j<R_i}\left[1-\frac{\Delta_j}{n-R_j+1}\right]$ be the Kaplan-Meier weight attached to $\widetilde{T}_i$ (here $R_i=Rank(\widetilde{T}_i)$). With this notation, any functional of the form $S(\varphi)=E\left[\varphi\left(Z,T\right)\right]$ is estimated by $S_n(\varphi)=\sum_{i=1}^n W_i\varphi\left(\widetilde{Z}_i,\widetilde{T}_i\right)$. By noting that

$$p_{11}(s,t)=\frac{P(t<Z)}{P(s<Z)},\quad p_{13}(s,t)=\frac{P\left(s<Z,T\le t\right)}{P(s<Z)}=\frac{E\left[\varphi_{s,t}\left(Z,T\right)\right]}{P(s<Z)},\tag{3}$$

where $\varphi_{s,t}\left(u,v\right)=I\left(u>s,v\le t\right)$, and $p_{12}(s,t)=1-p_{11}(s,t)-p_{13}(s,t)$, we have the following estimators for $\left\{p_{1j}(s,t),j=1,2,3\right\}$:

$$\widehat{p}_{11}(s,t)=\frac{\widehat{S}_Z(t)}{\widehat{S}_Z(s)},\quad \widehat{p}_{13}(s,t)=\frac{1}{\widehat{S}_Z(s)}\sum_{i=1}^n W_i\varphi_{s,t}\left(\widetilde{Z}_i,\widetilde{T}_i\right),\tag{4}$$

and $\widehat{p}_{12}(s,t)=1-\widehat{p}_{11}(s,t)-\widehat{p}_{13}(s,t)$, where $\widehat{S}_Z(.)$ is the Kaplan-Meier estimator of the survival function of Z. Similarly, since

$$p_{23}\left(s,t\right)=\frac{P\left(Z\le s,s<T\le t\right)}{P\left(Z\le s<T\right)}=\frac{E\left[\widetilde{\varphi}_{s,t}\left(Z,T\right)\right]}{P(T>s)-P(s<Z)}\tag{5}$$

where $\widetilde{\varphi}_{s,t}\left(u,v\right)=I\left(u\le s,s<v\le t\right)$, and $p_{22}(s,t)=1-p_{23}(s,t)$, one can introduce the following estimators for $\left\{p_{2j}(s,t),j=2,3\right\}$:

$$\widehat{p}_{23}\left(s,t\right)=\frac{1}{\widehat{S}_T(s)-\widehat{S}_Z(s)}\sum_{i=1}^n W_i\widetilde{\varphi}_{s,t}\left(\widetilde{Z}_i,\widetilde{T}_i\right)\quad\text{and}\quad\widehat{p}_{22}(s,t)=1-\widehat{p}_{23}(s,t),\tag{6}$$

where $\widehat{S}_T(.)$ is the Kaplan-Meier estimator of the survival function of T.

In the uncensored case, all the involved Kaplan-Meier weights reduce to $1/n$, and hence the estimators introduced along (4) and (6) collapse to (2).

Under censoring, the consistency of $\widehat{S}_Z(.)$ and $\widehat{S}_T(.)$ follows because the censoring is independent of the process; on the other hand, Theorem 1 in [5] ensures the consistency of $\widehat{p}_{13}(s,t)$ and $\widehat{p}_{23}(s,t)$, which involve multivariate Kaplan-Meier integrals in the sense of [7]. When the process is Markov, the estimators (4)-(6) are less efficient than their Aalen-Johansen counterparts. However, when the Markov condition is violated, these estimators are preferred since Aalen-Johansen can be systematically biased. Both estimators (Markov, non-Markov) were compared in simulated and practical settings; [5] used the PROVA trial data as an illustrative example, while [6] provided a nice comparison of the methods in a trial on breast cancer.

The idea in [5] for the construction of non-Markov transition probabilities in the illness-death model can be generalized to any other progressive multistate model. Certainly, since each given transition probability is a function of the sojourn times in the several existing states, the Kaplan-Meier weights pertaining to the time up to reaching the final absorbing state can be used to construct consistent empirical weighted averages. Details on this are provided in [8].

Doubtless, the main drawback of the non-Markov estimators is their large variance, mainly in the heavily censored case. In order to mitigate this problem, some presmoothing of the censoring indicators Δ_i can be performed. By 'presmoothing' it is meant that each Δ_i is replaced by some fit to the binary regression $m(z,t) = P\left(\Delta = 1|\widetilde{Z} = z, \widetilde{T} = t\right)$ before the Kaplan-Meier weights W_i are computed. In [10] this idea was applied in the scope of the three-state progressive model (which is just an illness-death model with forbidden transition 1→3) to introduce a new estimator of the joint distribution of the sojourn times. Also, [11] consider presmoothed non-Markov transition probabilities for the illness-death model. The main consequence of this approach is that non-Markov estimators with improved variance can be obtained. In practice, the presmoothing function $m(z,t)$ is estimated by fitting some parametric model or via nonparametric regression methods.

3 Testing the Markov Assumption

In this section, we consider the problem of checking the Markov assumption in practice. Note that this issue is relevant, since (as discussed above) the Aalen-Johansen estimator may be inconsistent when the Markov condition is violated. For simplicity of exposure, consider the progressive illness-death model and let $\lambda(t|s)$ be the hazard rate of T at time $t \geq s$ conditionally on $Z = s$ and $Z < T$. Note that the Markov assumption states that the value of $\lambda(t|s)$ does not depend on s. This is typically tested via a proportional hazards specification $\lambda(t|s) = \lambda_0(t)e^{\beta s}$. Then, the null hypothesis representing the Markov condition is $H_0 : \beta = 0$. The model can be fitted (and a test performed) by standard methods from the cases with an uncensored Z. For the PROVA trial data, the estimated coefficient was $\widehat{\beta} = 0.00526$ (s.e. $= 0.00167$), and the likelihood ratio test gave a p-value of $p = 0.000434$ thus rejecting the Markov condition.

In practice, the model $\lambda(t|s) = \lambda_0(t)e^{\beta s}$ may not be appropriated due to several reasons. First, the linear predictor βs can not cope in general with other type of effects. To illustrate this, consider the three-state progressive model with $\log(T_2) = f(Z) + \varepsilon$ where $T_2 = T - Z$ and where ε is an error term independent of Z. In this case we get $\lambda(t|s) = \lambda_0((t-s)e^{-f(s)})e^{-f(s)}$ where λ_0 stands for the hazard of $W = e^{\varepsilon}$. Take a extreme-value distribution for ε (so λ_0 becomes constant), $Z \sim U[0,2]$ and $f(s) = (s-1)^2$. Then, the test for $\beta = 0$ under the linear specification $\lambda(t|s) = \lambda_0(t)e^{\beta s}$ is expected to have low (or even no) power. Second, the proportional hazards assumption may fail; this is the case, for example, when $\log(T_2) = f(Z) + \varepsilon$ and $W = e^{\varepsilon}$ does not follow a Weibull distribution. Of course, this may influence the performance of the test. In Table 1 we report the proportion of rejection of this test at level $\alpha = 0.05$ for several sample sizes n among 1,000 Monte Carlo trials in these two situations (we take $f(s) = 0$ and $\varepsilon \sim N(0,1)$ in the second case, labeled as No PH; in this case, the Markov assumption does not hold because T_2 is not exponentially distributed). We see that, in these two cases, the classical test exhibits a very poor power. Of course, in the first simulated scenario (labeled as PH) power could be increased through a more flexible specification of the predictor; in the second model, however, there is not a clear solution to the lack of power of the test.

In Table 1, we have also reported the results pertaining to a new method of testing. The new method is based on the fact that, under the Markov assumption, the variables T and Z are independent conditionally on the event $A_t = \{Z \le t < T\}$, for each given $t > 0$. More specifically, we have performed a test of no correlation between T and Z conditionally on A_t with $t = 2$. The choice $t = 2$ is interesting because it guarantees the maximum expected sample size (i.e. the largest $P(A_t)$) under the two simulated models. We see in Table 1 that this new idea may lead to a more powerful test. Besides, we have seen in the simulations that there is some negative correlation between the p-values of the classical test and those of the new approach, indicating that both testing procedures are able to detect different type of alternatives. Hence, a complementary use of both approaches could be recommended in practice.

Clearly, the combination of several t values should help to increase the power of the new method of testing. Moreover, by considering a whole set of t-values one could explore the variation of the pertaining p-values. To illustrate this, consider the significance trace $\{p(t) : t \in [t_0, t_1]\}$, where $p(t)$ stands for the p-value of the suggested correlation test, when conditioning on A_t. In Figure 1 we depict this curve for the simulated model No PH in Table 1 with $n = 500$, $t_0 = 1$ and $t_1 = 3$; the given curve is indeed the first quartile of the p-values along the 1,000 Monte Carlo simulations. Roughly speaking, the information in this Figure is that (a) in more than 25% of the cases the trace is able to reject the Markovianity of the process (recall however the

Table 1 Proportion of rejection at level $\alpha = 0.05$ along 1,000 Monte Carlo simulations of sample size n for two non-Markov three-state progressive models: classical method vs. new approach ($t = 2$).

	PH		No PH	
n	Classical	New	Classical	New
100	.069	.119	.073	.097
250	.096	.159	.089	.157
500	.091	.193	.099	.248
1000	.122	.267	.148	.378

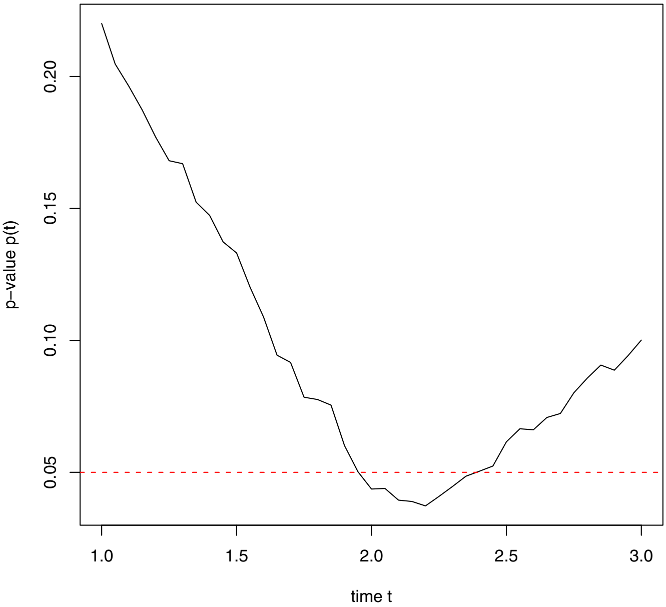


Fig. 1 Significance trace for No PH model with $n = 500$: first quartile along 1,000 Monte Carlo simulations.

poor power of the classical method in this case, see Table 1), and that (b) the greatest evidence against the null is achieved around $t = 2.2$.

Of course, although we have used the Pearson correlation coefficient to implement the new method, it can be adapted in an obvious manner to be based on other measures of association too. A key issue here is how to incorporate the censoring effects in the definition of the test statistic. This is not obvious at all. In [9], an omnibus test statistic which compares the joint distribution function of (Z,T) to the product of marginals conditioning on each A_t was introduced, accounting for censoring effects. However, the performance of this test in practice is still unexplored, and this seems to be a very promising field of research.

The problem of testing the Markov assumption has been discussed here for the illness-death model (and for the three-state progressive model) for the sake of conciseness. In general, one will be interested in testing that the entry time to the present state (and other measured covariates in the individual's history) is unrelated to the future hazard. At the end, this type of assumptions can be formalized in a simple manner so the methods reviewed here (or obvious modifications) still apply.

Acknowledgements. The author thanks P.K. Andersen for providing the PROVA trial data. Research funded by projects MTM2008-03129, PGIDIT07PXIB300191PR, and by the INBIOMED project (DXPCTSUG, Ref. 2009/063).

References

1. Aalen, O., Johansen, S.: An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. Scand. J. Stat. 5, 141–150 (1978)
2. Andersen, P.K., Esbjerj, S., Sorensen, T.I.A.: Multistate models for bleeding episodes and mortality in liver cirrhosis. Stat. Med. 19, 587–599 (2000)
3. Datta, S., Satten, G.A.: Validity of the AalenJohansen estimators of stage occupation probabilities and Nelson Aalen integrated transition hazards for non-Markov models. Stat. Probab. Lett. 55, 403–411 (2001)
4. Glidden, D.: Robust inference for event probabilities with non-Markov event data. Biometrics 58, 361–368 (2002)
5. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C.: Nonparametric estimation of transition probabilities in a non-Markov illness-death model. Lifetime Data Anal. 12, 325–344 (2006)
6. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., Andersen, P.K.: Multi-state models for the analysis of time-to-event data. Statist. Meth. Med. Research 18, 195–222 (2009)
7. Stute, W.: Consistent estimation under random censorship when covariables are present. J. Multivariate Anal. 45, 89–103 (1993)
8. de Uña-Álvarez, J.: Estimación no paramétrica en modelos multi-estado no-Markovianos. In: Actas del XXX Congreso Nacional de la SEIO, Valladolid, Spain (2007a)
9. de Uña-Álvarez, J.: Testing that a multi-state model is Markov: new methods. In: Gomes, M.I., Pestana, D., Silva, P. (eds.) Abstracts of the 56th Session of the ISI, Lisbon, Portugal (2007b)
10. de Uña-Álvarez, J., Amorim, A.P.: A semiparametric estimator of the bivariate distribution function for censored gap times. Discussion Papers in Statistics and OR 09/03, University of Vigo (2009)
11. de Uña-Álvarez, J., Amorim, A.P., Meira-Machado, L.: Presmoothing the transition probabilities in the illness-death model (in preparation, 2010)

Maximum Likelihood from Evidential Data: An Extension of the EM Algorithm

Thierry Denœux

Abstract. We consider the problem of statistical parameter estimation when the data are uncertain and described by belief functions. An extension of the Expectation-Maximization (EM) algorithm, called the Evidential EM (E^2M) algorithm, is described and shown to maximize a generalized likelihood function. This general procedure provides a simple mechanism for estimating the parameters in statistical models when observed data are uncertain. The method is illustrated using the problem of univariate normal mean and variance estimation from uncertain data.

Keywords: Belief functions, Dempster-Shafer theory, Statistical inference, Uncertain data.

1 Introduction

In statistics, observations of random quantities are usually assumed to be either precise or imprecise, i.e., set-valued. The latter situation occurs, e.g., in the case of censored data, where an observation is only known to belong to a set, usually an interval. The Expectation-Maximization (EM) algorithm [4, 8] has proved to be a powerful mechanism for performing maximum likelihood parameter estimation from such incomplete data.

There are situations, however, where the observations are not only imprecise, but also *uncertain*, i.e., partially reliable [1]. Consider, e.g., a classification problem in which objects in a population belong to one and only one group. Let $\mathscr{X}$ be the finite set of groups, and X be the group of an object randomly drawn from the population. In some applications, realizations x of X are not known with certainty. Rather, an expert provides a subjective

Thierry Denœux
Heudiasyc, Université de Technologie de Compiègne, CNRS, Compiègne, France
e-mail: tdenoeux@hds.utc.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 181–188.
springerlink.com

assessment of x (a process known as *labeling*). This assessment may take the form of a subset $A \subseteq \mathscr{X}$, a probability distribution p on $\mathscr{X}$ or, more generally, a mass function m on $\mathscr{X}$, i.e., a function $m : 2^{\mathscr{X}} \to [0,1]$. It must be stressed that, in this example, the data generation process can be decomposed into two components: a random component, which generates a realization x from X, and a non random component, which produces a mass function m that models the expert's partial knowledge of x.

If this process is repeated n times independently, the data takes the form of n mass functions $m_1,\ldots,m_n$, considered as a partial specification of an unknown realization $x_1,\ldots,x_n$ of an i.i.d. random sample $X_1,\ldots,X_n$. We will refer to such data as *evidential data*. If a parametric model is postulated for X, how can the method of maximum likelihood be extended to handle such data? This is the problem considered in this paper. A generalization of the likelihood function will be proposed, and an extension of the EM algorithm, called the evidential EM (E^{2}M) algorithm, will be introduced for its maximization.

We may note that, in the special case where each mass functions m_i is consonant, the data can be equivalently represented as n possibility distribution $\widetilde{x}_1,\ldots,\widetilde{x}_n$, which constitutes a *fuzzy random sample*. The problem of statistical inference from fuzzy data, which has received a lot of attention in the past few years [5, 6], is thus a special case of the problem considered here.

Early attempts to adapt the EM algorithm to evidential data, in the special case of mixture models with evidential class labels, were presented in [10, 7]. A rigorous solution to this problem, which is a special case of the general method presented in this paper, was introduced in [2].

The rest of the paper is organized as follows. The EM algorithm will first be recalled in Section 2. The extension of the likelihood function and the E^{2}M algorithm will then be introduced in Sections 3 and 4, respectively. Section 5 will demonstrate the application of this algorithm to the problem of univariate normal mean and variance estimation using uncertain data.

2 The EM Algorithm

The EM algorithm is a broadly applicable mechanism for computing MLEs from incomplete data, in situations where ML estimation would be straightforward if complete data were available [4]. Formally, we assume the existence of two sample spaces $\mathscr{X}$ and $\mathscr{Y}$, and a many-to-one mapping φ from $\mathscr{X}$ to $\mathscr{Y}$. The observed (incomplete) data $\mathbf{y}$ are a realization from $\mathscr{Y}$, while the corresponding $\mathbf{x}$ in $\mathscr{X}$ is not observed and is only known to lie in the set

$$\mathscr{X}(\mathbf{y}) = \varphi^{-1}(\mathbf{y}) = \{\mathbf{x} \in \mathscr{X} \,|\, \varphi(\mathbf{x}) = \mathbf{y}\}.$$

Vector $\mathbf{x}$ is referred to as the *complete data* vector. It is a realization from a random vector $\mathbf{X}$ with p.d.f. $g_c(\mathbf{x};\boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = (\Psi_1,\ldots,\Psi_d)'$ is a vector of unknown parameters with parameter space $\boldsymbol{\Omega}$. The observed data likelihood $L(\boldsymbol{\Psi})$ is related to $g_c(\mathbf{x};\boldsymbol{\Psi})$ by

$$L(\boldsymbol{\Psi}) = \int_{\mathscr{X}(\mathbf{y})} g_c(\mathbf{x};\boldsymbol{\Psi}) d\mathbf{x}. \tag{1}$$

The EM algorithm approaches the problem of maximizing the observed-data log likelihood $\log L(\boldsymbol{\Psi})$ by proceeding iteratively with the complete-data log likelihood $\log L_c(\boldsymbol{\Psi}) = \log g_c(\mathbf{x};\boldsymbol{\Psi})$. Each iteration of the algorithm involves two steps called the expectation step (E-step) and the maximization step (M-step).

The E-step requires the calculation of

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) = \mathbb{E}_{\boldsymbol{\Psi}^{(q)}}\left[\log L_c(\boldsymbol{\Psi})|\mathbf{y}\right],$$

where $\boldsymbol{\Psi}^{(q)}$ denotes the current fit of $\boldsymbol{\Psi}$ at iteration q, and $\mathbb{E}_{\boldsymbol{\Psi}^{(q)}}$ denotes expectation using the parameter vector $\boldsymbol{\Psi}^{(q)}$.

The M-step then consists in maximizing $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$ with respect to $\boldsymbol{\Psi}$ over the parameter space $\boldsymbol{\Omega}$. The E- and M-steps are iterated until the difference $L(\boldsymbol{\Psi}^{(q+1)}) - L(\boldsymbol{\Psi}^{(q)})$ becomes smaller than some arbitrarily small amount.

3 Generalized Likelihood Function

Let us now consider the more complex situation where the relationship between the observed and complete spaces is uncertain, so that observed data $\mathbf{y}$ can no longer be associated with certainty to a unique subset of $\mathscr{X}$. This situation will be formalized as follows.

Let us assume the existence of a set Θ of interpretations, one and only one of which holds, and a probability measure Pr on Θ. If $\mathbf{y}$ has been observed and $\theta \in \Theta$ is the true interpretation, then the complete data $\mathbf{x}$ is known to belong to $\mathscr{X}(\mathbf{y},\theta) \subseteq \mathscr{X}$. Having observed $\mathbf{y}$, the probability measure Pr is carried to $2^{\mathscr{X}}$ by the mapping $\theta \to \mathscr{X}(\mathbf{y},\theta)$, which defines a Dempster-Shafer mass function m on $\mathscr{X}$. For simplicity, we will assume from now on that Θ is finite: $\Theta = \{\theta_1,\ldots,\theta_K\}$, in which case m is a discrete mass function with focal sets $\mathscr{X}_k = \mathscr{X}(\mathbf{y},\theta_k)$ and masses $m_k = m(\mathscr{X}_k) = \Pr(\{\theta_k\})$ for $k = 1,\ldots,K$.

With the same notations as in the previous section, the observed data likelihood may now be defined as:

$$L(\boldsymbol{\Psi}) = \sum_{k=1}^{K} m_k \int_{\mathscr{X}_k} g_c(\mathbf{x};\boldsymbol{\Psi}) d\mathbf{x} = \int_{\mathscr{X}} g_c(\mathbf{x};\boldsymbol{\Psi}) \left(\sum_{k=1}^{K} m_k 1_{\mathscr{X}_k}(\mathbf{x})\right) d\mathbf{x}$$
$$= \int_{\mathscr{X}} g_c(\mathbf{x};\boldsymbol{\Psi}) pl(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\boldsymbol{\Psi}}\left[pl(\mathbf{X})\right], \tag{2}$$

where $pl : \mathscr{X} \to [0,1]$ is the contour function associated to m.

The generalized likelihood of $\boldsymbol{\Psi}$ is thus equal to the expectation of the plausibility contour function, with respect to the probability distribution $g_c(\mathbf{x};\boldsymbol{\Psi})$. We can remark that, when m is consonant, the contour function can be seen as the membership function of a fuzzy subset of $\mathscr{X}$: $L(\boldsymbol{\Psi})$ is then the probability of that fuzzy subset, according to Zadeh's definition of the probability of a fuzzy event [11].

In the more general setting of belief functions, $L(\boldsymbol{\Psi})$ has another interpretation that will now be explained. Let $g_c(\cdot|m;\boldsymbol{\Psi}) = m \oplus g_c(\cdot;\boldsymbol{\Psi})$ denote the p.d.f. obtained by combining m with the complete data p.d.f. $g_c(\cdot;\boldsymbol{\Psi})$ using Dempster's rule [3, 9]:

$$g_c(\mathbf{x}|m;\boldsymbol{\Psi}) = \frac{g_c(\mathbf{x};\boldsymbol{\Psi})pl(\mathbf{x})}{\int_{\mathscr{X}} g_c(\mathbf{u};\boldsymbol{\Psi})pl(\mathbf{u})d\mathbf{u}} = \frac{g_c(\mathbf{x};\boldsymbol{\Psi})pl(\mathbf{x})}{L(\boldsymbol{\Psi})}. \tag{3}$$

The normalizing constant $L(\boldsymbol{\Psi})$ at the denominator of the above expression is equal to one minus the degree of conflict between m and $g_c(\mathbf{x};\boldsymbol{\Psi})$. Consequently, maximizing $L(\boldsymbol{\Psi})$ amounts to *minimizing the conflict* between the observations (represented by m) and the parametric model $g_c(\cdot;\boldsymbol{\Psi})$.

The expression of the observed data likelihood (2) can often be simplified by making independence assumptions. Let us assume that the observed data $\mathbf{x} = (x_1,\ldots,x_n)$ is a realization from a random vector $\mathbf{X} = (X_1,\ldots,X_n)$. In many applications, we can make the following assumptions:

A1: Stochastic independence of the r.v. $X_1,\ldots,X_n$:

$$g_c(\mathbf{u};\boldsymbol{\Psi}) = \prod_{i=1}^{n} g_c(u_i;\boldsymbol{\Psi}), \quad \forall \mathbf{u} = (u_1,\ldots,u_n) \in \mathscr{X}.$$

A2: The plausibility contour function $pl(\mathbf{x})$ can be written as

$$pl(\mathbf{u}) = \prod_{i=1}^{n} pl_i(u_i), \quad \forall \mathbf{u} = (u_1,\ldots,u_n) \in \mathscr{X},$$

where pl_i is the contour function corresponding to the marginal mass function m_i on x_i.

It should be noted that Assumption A2 is totally unrelated to A1: it is not a property of the random variables $X_1 \ldots, X_n$, but of the uncertain observation process. It is actually a weaker form of the *cognitive independence* assumption, as defined by Shafer [9].

Under Assumptions A1 and A2, the observed data log likelihood can be written as a sum of n terms:

$$\log L(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \mathbb{E}_{\boldsymbol{\Psi}}\left[pl_i(X_i)\right].$$

4 The Evidential EM Algorithm

To maximize function $L(\boldsymbol{\Psi})$ defined by (2), we propose to adapt the EM algorithm as follows. Let the E-step now consist in the calculation of the expectation of $\log L_c(\boldsymbol{\Psi})$ with respect to $g_c(\cdot|m;\boldsymbol{\Psi}^{(q)})$ defined by (3):

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) = \mathbb{E}_{\boldsymbol{\Psi}^{(q)}}\left[\log L_c(\boldsymbol{\Psi})|m\right] = \frac{\int \log(L_c(\boldsymbol{\Psi}))g_c(\mathbf{x};\boldsymbol{\Psi}^{(q)})pl(\mathbf{x})d\mathbf{x}}{L(\boldsymbol{\Psi}^{(q)})}. \quad (4)$$

The M-step is unchanged and requires the maximization of $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$ with respect to $\boldsymbol{\Psi}$. The E^{2}M algorithm alternately repeats the E- and M-steps above until the increase of observed-data likelihood becomes smaller than some threshold. The following theorem shows that E^{2}M algorithm inherits the monotonicity property of the EM algorithm, which ensures convergence provided the sequence of incomplete-data likelihood values remains bounded from above.

Theorem 1. *Any sequence $L(\boldsymbol{\Psi}^{(q)})$ for $q = 0,1,2,\ldots$ of likelihood values obtained using the E^{2}M algorithm is non decreasing, i.e., it verifies*

$$L(\boldsymbol{\Psi}^{(q+1)}) \geq L(\boldsymbol{\Psi}^{(q)}), \quad \forall q. \quad (5)$$

Proof. The proof is similar to that of Dempster et al. [4]. □

To conclude this section, we may note that the p.d.f. $g_c(\mathbf{x}|m;\boldsymbol{\Psi})$ and, consequently, the E^{2}M algorithm depend only on the contour function $pl(\mathbf{x})$ and they are unchanged if $pl(\mathbf{x})$ is multiplied by a constant. Consequently, the results are unchanged if m is converted into a probability distribution by normalizing the contour function.

5 Normal Mean and Variance Estimation

To illustrate the above algorithm, let us assume that the complete data $\mathbf{x} = (x_1,\ldots,x_n) \in \mathscr{X} = \mathbb{R}^n$ is a realization from an i.i.d. random sample from a univariate normal distribution $\mathscr{N}(\mu,\sigma^2)$. The parameter vector is thus $\boldsymbol{\Psi} = (\mu,\sigma)$. The observed data has the form $\mathbf{y} = (\mathbf{y}_1,\ldots,\mathbf{y}_n)$ with $\mathbf{y}_i = (w_i,\alpha_i)$, where w_i is an estimate of x_i (provided, e.g., by a sensor), and $\alpha_i \in [0,1]$ is a degree of confidence in that estimation. For each $\mathbf{y}_i$, there are two interpretations θ_{i1} and θ_{i2}. Under interpretation θ_{i1}, we admit that $x_i = w_i$; under interpretation θ_{i2}, we know only that $\mathbf{x}_i \in \mathbb{R}$. The probability for interpretation θ_{i1} to be correct is α_i, which can thus be interpreted as a degree of reliability of the piece of information $\mathbf{y}_i$. The induced mass function m_i on $\mathbb{R}$ is defined by

$$m_i(\{w_i\}) = \alpha_i, \quad m_i(\mathbb{R}) = 1-\alpha_i.$$

The corresponding contour function is defined by

$$pl_i(x) = \alpha_i\delta(x-w_i) + 1 - \alpha_i$$

for all $x \in \mathbb{R}$, where $\delta(\cdot)$ is the Dirac Delta function.

Let $g_c(\cdot;\mu,\sigma)$ denote the normal p.d.f. with mean μ and standard deviation σ. The observed data log likelihood is

$$\log L(\mu,\sigma) = \sum_{i=1}^{n} \log\left(\int_{-\infty}^{\infty} g_c(x;\mu,\sigma) pl_i(x) dx\right) = \sum_{i=1}^{n} \log\left(\alpha_i g_c(w_i;\mu,\sigma) + 1 - \alpha_i\right),$$

which is to be maximized with respect to μ and σ.

The complete data log likelihood is

$$\log L_c(\mu,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + \mu^2\right).$$

Consequently,

$$Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\beta_i^{(q)} - 2\mu\sum_{i=1}^{n}\gamma_i^{(q)} + \mu^2\right), \quad (6)$$

where $\gamma_i^{(q)}$ and $\beta_i^{(q)}$ denote, respectively, the expectations of X and X^2 with respect to the conditional probability distribution

$$g_c(\cdot|m_i;\boldsymbol{\Psi}^{(q)}) = g_c(\cdot;\mu^{(q)},\sigma^{(q)}) \oplus m_i$$

defined by

$$g_c(x|m_i;\boldsymbol{\Psi}^{(q)}) = \frac{g_c(x;\boldsymbol{\Psi}^{(q)}) pl_i(x)}{\int_{-\infty}^{+\infty} g_c(u;\boldsymbol{\Psi}^{(q)}) pl_i(u) du} = \frac{g_c(x;\boldsymbol{\Psi}^{(q)})\left[\alpha_i \delta_{w_i}(x) + (1-\alpha_i)\right]}{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) + 1 - \alpha_i}.$$

The following equalities thus hold:

$$\gamma_i^{(q)} = \frac{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) w_i + (1-\alpha_i)\mu^{(q)}}{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) + 1 - \alpha_i} \quad (7)$$

and

$$\beta_i^{(q)} = \frac{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) w_i^2 + (1-\alpha_i)\left[\left(\mu^{(q)}\right)^2 + \left(\sigma^{(q)}\right)^2\right]}{\alpha_i g_c(w_i;\boldsymbol{\Psi}^{(q)}) + 1 - \alpha_i}. \quad (8)$$

The maximum of $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$ defined by (6) is obtained for the following values of μ and σ:

$$\mu^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i^{(q)} \quad (9)$$

and

$$\sigma^{(q+1)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \beta_i^{(q)} - \left(\mu^{(q+1)}\right)^2}. \tag{10}$$

In E-step of the E^{2}M algorithm for this problem thus consists in the calculation of $\gamma_i^{(q)}$ and $\beta_i^{(q)}$ for all i using (7) and (8), respectively. The M-step then updates the estimates of μ and σ using (9) and (10). The algorithm stops when the relative increase of the observed data likelihood becomes less than some threshold ε.

Example 1. To illustrate the application of the above algorithm to a situation where data are unreliable, we considered the following experiments. Random samples of size $n = 100$ were drawn from a standard normal distribution. For each realization x_i, a number α_i was drawn from the uniform distribution $\mathscr{U}_{[0,1]}$. With probability α_i, w_i was defined as x_i, and with probability $1-\alpha_i$ it was set to $x_i + \varepsilon_i$, with $\varepsilon_i \sim \mathscr{N}(0, s^2)$. Parameters μ and σ were estimated using the E^{2}M algorithm based on the data (w_i, α_i), $i = 1, \ldots, n$. The experiment was repeated $N = 100$ times and mean squared errors on μ and σ were computed. The results are shown in Figure 1. Our approach was compared with the simple strategy that consists in estimating μ and σ by the sample mean and standard deviation of the w_i for all i such that $\alpha_i \geq c$, for different choices of c. We can see that the E^{2}M algorithm is much more robust than this simple reference method. Further experiments involving comparisons with more sophisticated alternative estimators are under way.

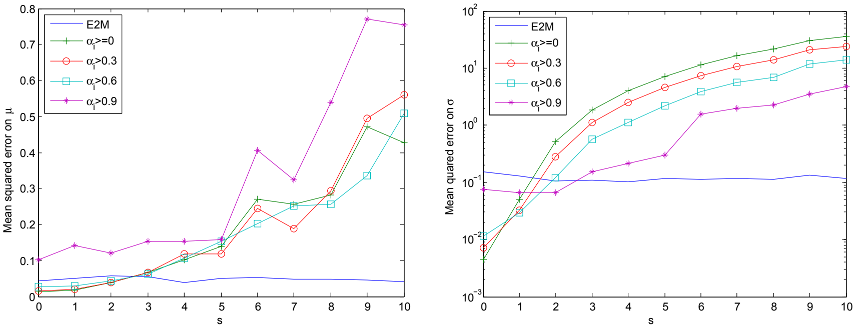


Fig. 1 Mean squared errors on μ (left) and σ (right, logarithmic y scale) as functions of the noise standard deviation s for the E^{2}M algorithm and alternative methods (see details in text).

6 Conclusion

An iterative procedure for estimating the parameters in a statistical model using evidential data has been proposed. This procedure, which generalizes the EM algorithm, minimizes the degree of conflict between the uncertain observations and the parametric model. It provides a general mechanism for statistical inference when the observed data are uncertain. It remains an open problem to determine the conditions under which the obtained estimator is consistent. This is the topic of on-going research.

References

1. Aggarwal, C.C., Yu, P.S.: A survey of uncertain data algorithms and applications. IEEE Trans. Knowl. Data Eng. 21(5), 609–623 (2009)
2. Côme, E., Oukhellou, L., Denœux, T., Aknin, P.: Learning from partially supervised data using mixture models and belief functions. Pattern Recogn. 42(3), 334–348 (2009)
3. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Statist. 38, 325–339 (1967)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38 (1977)
5. Denœux, T., Masson, M.-H., Hébert, P.-A.: Nonparametric rank-based statistics and significance tests for fuzzy data. Fuzzy Sets Syst. 153, 1–28 (2005)
6. Gebhardt, J., Gil, M.A., Kruse, R.: Fuzzy set-theoretic methods in statistics. In: Slowinski, R. (ed.) Fuzzy sets in decision analysis, operations research and statistics, pp. 311–347. Kluwer Academic Publishers, Boston (1998)
7. Jraidi, I., Elouedi, Z.: Belief classification approach based on generalized credal EM. In: Mellouli, K. (ed.) ECSQARU 2007. LNCS (LNAI), vol. 4724, pp. 524–535. Springer, Heidelberg (2007)
8. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
9. Shafer, G.: A mathematical theory of evidence. Princeton University Press, Princeton (1976)
10. Vannoorenberghe, P., Smets, P.: Partially supervised learning by a credal EM approach. In: Godó, L. (ed.) ECSQARU 2005. LNCS (LNAI), vol. 3571, pp. 956–967. Springer, Heidelberg (2005)
11. Zadeh, L.A.: Probability measures of fuzzy events. J. Math. Anal. Appl. 10, 421–427 (1968)

A Decision Rule for Imprecise Probabilities Based on Pair-Wise Comparison of Expectation Bounds

Sébastien Destercke

Abstract. There are many ways to extend the classical expected utility decision rule to the case where uncertainty is described by (convex) probability sets. In this paper, we propose a simple new decision rule based on the pairwise comparison of lower and upper expected bounds. We compare this rule to other rules proposed in the literature, showing that this new rule is both precise, computationally tractable and can help to boost the computation of other, more computationally demanding rules.

Keywords: Maximality, Hurcwitz criterion, E-admissibility, Lower previsions, Γ-maximin.

1 Introduction

We are concerned here with the problem of making a decision d, which may be taken from a set of N available decisions $\mathscr{D} = \{d_1, \ldots, d_N\}$. Usually, this decision is not chosen arbitrarily, i.e., it should be the best possible in the current situation. In our case, the benefits that an agent would gain by taking decision d_i depend on a variable X and the knowledge we have about its value. We assume here that the true value of X is uncertain, that it takes its value on a finite domain $\mathscr{X}$ and that the benefit (or gain, reward) of choosing d_i can be modelled by a real-valued and bounded utility function $U_i : \mathscr{X} \to \mathbb{R}$, with $U_i(x)$ the gain of choosing action d_i when x is the value of X. The problem of decision making is then to select, based on this information, the decisions that are optimal, i.e. are likely to gives the best possible gain.

When uncertainty on X is (or can be) modelled by a probability distribution $p : \mathscr{X} \to [0,1]$, many authors (for example De Finetti [2]) have argued

Sébastien Destercke
UMR IATE, Campus Supagro, 34060 Montpellier, France
e-mail: sebastien.destercke@irsn.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 189–197.
springerlink.com

that the optimal decision $\overline{d} \in \mathscr{D}$ should be the one maximising the expected utility, i.e., $\overline{d}_{\mathbb{E}_p} = \arg\max_{d_i \in \mathscr{D}} \mathbb{E}_p(U_i) = \sum_{x \in X} U_i(x)p(x)$. Thus, selecting the optimal decision in the sense of expected utility comes down to considering the complete (pre-)order induced by expected utility, here denoted by $\leq_{\mathbb{E}}$, over decisions in $\mathscr{D}$ ($d_i \leq_{\mathbb{E}} d_j$ if $\mathbb{E}_p(U_i) \leq \mathbb{E}_p(U_j)$), and to choose the decision which is not dominated by others (Given a partial order $\leq$ on $\mathscr{D}$, we say that d dominates d' if $d' \leq d$). In the sequel, we will say that a decision d is optimal w.r.t. an order $\leq$, or a decision rule, if it is non-dominated in the order induced by this decision rule.

However, it may happen that our uncertainty about the value of X cannot be modelled by a single probability, for the reason that not enough information is available to identify the probability $p(x)$ of every element $x \in \mathscr{X}$. In such a case, convex sets of probabilities, here called credal sets [5] (which are formally equivalent to coherent lower previsions [9]), have been proposed as an uncertainty representation allowing us to model information states going from full ignorance to precise probabilities, thus coping with insufficiencies in our information. Formally, they encompass most of the uncertainty representations that integrate the notion of imprecision (e.g., belief functions, possibility distributions, ...). To select optimal decisions in this context, it is necessary to extend the expected utility criterion, as the expected utility $\mathbb{E}(U)$ is no longer precise and becomes a bounded interval $[\underline{\mathbb{E}}(U), \overline{\mathbb{E}}(U)]$. In the past decades, several such extensions, based on the evaluations of expectation bounds rather than of precise expected values, have been proposed (see Troffaes [6] for a concise and recent review). Roughly speaking, two kinds of generalisations are possible: either using a combination of the lower and upper expectation bounds to induce a complete (pre-)order between decisions, reaching a unique optimal decision, or relaxing the need of a complete order and extending expected utility criterion to obtain a partial (pre-)order between decisions. In this latter cases, there may be several optimal decisions, the inability to select between them reflecting the imprecision in our information.

In this paper, we propose and explore a new decision rule of the latter kind, based on a pair-wise comparison of lower and upper expectation bounds. This rule, which has not been studied before in the framework of imprecise probabilities (to our knowledge), is quite simple and computationally tractable. Section 2 recalls the imprecise probabilistic framework as well as the existing decision rules. We then present in Section 3 the new rule and compare it to existing rules. We will show that this rule is (surprisingly) precise when compared to other rules inducing partial pre-orders between decisions.

2 Imprecise Probabilities and Decision Rules

We consider that our information and uncertainty regarding the value of a variable X is modelled by a credal set $\mathscr{P}$. Given a function $U_i : \mathscr{X} \to \mathbb{R}$ over the space $\mathscr{X}$, the lower and upper expectations $\underline{\mathbb{E}}_{\mathscr{P}}(U_i), \overline{\mathbb{E}}_{\mathscr{P}}(U_i)$ of U_i are such that

$$\underline{\mathbb{E}}_{\mathscr{P}}(U_i) = \inf_{p \in \mathscr{P}} \mathbb{E}_p(U_i) \qquad \overline{\mathbb{E}}_{\mathscr{P}}(U_i) = \sup_{p \in \mathscr{P}} \mathbb{E}_p(U_i)$$

In Walley's [9] behavioural interpretation of imprecise probabilities, $\underline{\mathbb{E}}_{\mathscr{P}}(U_i)$ is interpreted as the maximum buying price an agent would be ready to pay for U_i, associated to decision d_i. Conversely, $\overline{\mathbb{E}}_{\mathscr{P}}(U_i)$ is interpreted as the minimum selling price an agent would be ready to receive for U_i. These two expectation bounds are dual, in the sense that, for any real-valued bounded function f over $\mathscr{X}$, we have $\underline{\mathbb{E}}(f) = -\overline{\mathbb{E}}(-f)$.

When proposing a decision rule based on lower and upper expectations $\underline{\mathbb{E}}, \overline{\mathbb{E}}$, a basic requirement is that this decision rule should reduce to the classical expected utility rule when $\mathscr{P}$ reduces to a single probability distribution. Still, there are many ways to do so, providing $\mathscr{D}$ with a complete or a partial (pre-)order. In the former case, there is a unique optimal non-dominated decision, while in the latter there may be a set of such non-dominated decisions. We will review the most commonly used approaches, dividing them according to the kind of order they induce on $\mathscr{D}$.

Example 1. In order to illustrate our purpose, let us consider the same example as Troffaes [6]. Consider a coin that can either fall on head (H) or tails (T), thus $\mathscr{X} = \{H, T\}$, with our uncertainty given as $p(H) \in [0.28; 0.7]$ and $p(T) \in [0.3; 0.72]$. Different decisions and their pay-off in case of landing on Heads or Tails are summarized in Table 1, together with the lower and upper expectations reached by each decision.

Table 1 Example 1 possible decisions and expectation bounds.

$\mathscr{D}$		U_i	H	T	$\underline{\mathbb{E}}$	$\overline{\mathbb{E}}$
d_1		U_1	4	0	1.12	2.8
d_2		U_2	0	4	1.2	2.88
d_3	→	U_3	3	2	2.28	2.7
d_4		U_4	$1/2$	3	1.25	2.3
d_5		U_5	$47/20$	$47/20$	2.35	2.35
d_6		U_6	$41/10$	$-3/10$	0.932	2.78

2.1 Rules Inducing a Complete Order

Let us start with the rules pointing to a unique optimal decision.

Γ-maximin. The Γ-maximin rule [3], denoted by $\leq_{\underline{\mathbb{E}}}$, consists in replacing the expected value with the lower expectation. The optimal decision under this rule is such that

$$\overline{d}_{\leq_{\underline{\mathbb{E}}}} = \arg\max_{d_i \in \mathscr{D}} \underline{\mathbb{E}}_{\mathscr{P}}(U_i).$$

This rule correspond to a pessimistic attitude, since it consists in maximizing the worst possible expected gain. In example 1, $\overline{d}_{\leq_{\underline{\mathbb{E}}}} = d_5$.

***Γ*-maximax.** The optimistic version of the Γ-maximin, denoted by $\leq_{\overline{\mathbb{E}}}$ and consisting in selecting as optimal the decision that maximises the expected outcome is such that

$$\overline{d}_{\leq_{\overline{\mathbb{E}}}} = \arg\max_{d_i \in \mathscr{D}} \overline{\mathbb{E}}_{\mathscr{P}}(U_i).$$

In example 1, $\overline{d}_{\leq_{\overline{\mathbb{E}}}} = d_2$.

Hurcwitz Criterion. Hurcwitz criterion in imprecise probabilities [4], denoted here by $\leq_\alpha$, consists in choosing a so-called pessimism index $\alpha \in [0,1]$, and to induce an order where the behaviour of the decision maker range from fully pessimistic ($\alpha = 1$) to fully optimistic ($\alpha = 0$). Once a pessimistic index α has been chosen, Hurwictz rule is such that $d_i \leq_\alpha d_j$ whenever $\alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_i) + (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_j) + (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_j)$, and the optimal decision $\overline{d}_{\leq_\alpha}$ under this rule is

$$\overline{d}_{\leq_\alpha} = \arg\max_{d_i \in \mathscr{D}} \ \alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_i) + (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_i).$$

Γ-maximin and -maximax are respectively retrieved when $\alpha = 1$ and $\alpha = 0$. In Example 1, the set of optimal decisions $\overline{d}_{\leq_\alpha}$ that can be reached by different values of α is $\{d_2, d_3, d_5\}$

Note that determining optimal decisions for these three criteria requires N comparisons and at most $2N$ computations of expectation bounds.

2.2 Rules Inducing a Partial Order

The other alternative when extending expected utility criterion is to let drop off the assumption that the order on the decisions has to be complete. That is, to allow the order to be partial and to possibly induce a set of optimal decisions rather than a single one. Three rules following this way have been proposed up to now.

Interval dominance. A first natural extension to the comparison of precise expectations to the case of interval-valued expectations is the interval dominance order $\leq_{\mathscr{I}}$ such that $d_i \leq_{\mathscr{I}} d_j$ if and only if $\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \underline{\mathbb{E}}_{\mathscr{P}}(U_j)$. That is, d_j dominates d_i if the expected gain of d_j is at least as great as the one of d_i. The resulting set of non-dominated (or optimal) decisions is denoted by $\overline{\mathscr{D}}_{\mathscr{I}}$ and is such that

$$\overline{\mathscr{D}}_{\mathscr{I}} = \{d \in \mathscr{D} | \nexists d' \in \mathscr{D}, d \leq_{\mathscr{I}} d'\}.$$

Computing $\overline{\mathscr{D}}_{\mathscr{I}}$ requires the computation of $2N$ expectations and $2N$ comparisons. For Example 1, we have $\overline{\mathscr{D}}_{\mathscr{I}} = \{d_1, d_2, d_3, d_5, d_6\}$. As we can see, this rule has the advantage to be computationally efficient, but is also very imprecise.

Maximality. When expectations are precise, we have $d_i \geq_{\mathbb{E}} d_j$ whenever $\mathbb{E}_p(U_i) \geq \mathbb{E}_p(U_j)$ or, equivalently, whenever $\mathbb{E}_p(U_i - U_j) \geq 0$. The notion of maximality consists in extending this notion by inducing a pre-order $\geq_{\mathscr{M}}$ such that $d_i >_{\mathscr{M}} d_j$ whenever $\underline{\mathbb{E}}_{\mathscr{P}}(U_i - U_j) > 0$. In Walley's interpretation, $\underline{\mathbb{E}}_{\mathscr{P}}(U_i - U_j) > 0$ means that we are ready to pay a positive price to exchange U_i for U_j, hence that decision d_i is preferred to decision d_j. The resulting set of optimal decisions $\overline{\mathscr{D}}_{\mathscr{M}}$ is such that

$$\overline{\mathscr{D}}_{\mathscr{M}} = \{d \in \mathscr{D} | \nexists d' \in \mathscr{D}, d \leq_{\mathscr{M}} d'\}.$$

Computing $\overline{\mathscr{D}}_{\mathscr{M}}$ requires the computation of $N^2 - N$ lower expectations and $N^2 - N$ comparisons. For Example 1, we have $\overline{\mathscr{D}}_{\mathscr{M}} = \{d_1, d_2, d_3, d_5\}$.

E-admissibility. Robustifying the expected utility criterion when uncertainty is modelled by sets of probabilities can simply be done by selecting as optimal those decisions that are optimal w.r.t. classical expected utility for at least one probability measure of $\mathscr{P}$. In this case, the set of optimal decision $\overline{\mathscr{D}}_{\mathscr{E}}$ is such that

$$\overline{\mathscr{D}}_{\mathscr{E}} = \{d \in \mathscr{D} | \exists p \in \mathscr{P} \text{ s.t. } \overline{d}_{\mathbb{E}_p} = d\}.$$

Utkin and Augustin [7] have proposed algorithms that allow computing $\overline{\mathscr{D}}_{\mathscr{E}}$ by solving N linear programs whose complexity is slightly higher than the ones usually associated to the computation of a lower expectation. For Example 1, we have $\overline{\mathscr{D}}_{\mathscr{E}} = \{d_1, d_2, d_3\}$. Both E-admissibility and Maximality give more precise statements than Interval dominance, but their computational burden is also higher (hence, they are more difficult to use in complex problems).

3 The New Decision Rule

The rules presented in the previous section consist, for most of them, in comparing numeric values (expectation bounds) to determine which decisions are dominated by others and are therefore non-optimal. Other ways to order interval-valued numbers can therefore be considered and studied as potential decision rules. One such ordering that has not be studied in imprecise probability theory (as far as we know) is the one where an interval $[a, b]$ is considered as lower than $[c, d]$ if $a \leq c$ and $b \leq d$. This comes down to a pair-wise comparison of the interval bounds.

Using this ordering, we therefore propose a new decision rule, that we call *Interval bound dominance* ($\mathscr{I}\mathscr{B}$-dominance for short), denoted by $\leq_{\mathscr{I}\mathscr{B}}$, and defined as follows

Definition 1 (Interval bound dominance). *Given a credal set $\mathscr{P}$ and two decisions $d_i, d_j \in \mathscr{D}$, $d_i \leq_{\mathscr{I}\mathscr{B}} d_j$ whenever $\underline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \underline{\mathbb{E}}_{\mathscr{P}}(U_i)$ and $\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \overline{\mathbb{E}}_{\mathscr{P}}(U_i)$ ($d_i <_{\mathscr{I}\mathscr{B}} d_j$ when at least one of the two inequalities is strict).*

Note that, as for the rules of Section 2.2, the order $\leq_{\mathscr{I}\mathscr{B}}$ is partial and induces a set of optimal decisions. The set of optimal decisions $\mathscr{D}_{\mathscr{I}\mathscr{B}}$ resulting from this decision rule is such that

$$\overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}} = \{d \in \mathscr{D} | \nexists d' \in \mathscr{D}, d \leq_{\mathscr{I}\mathscr{B}} d'\}.$$

In Example 1, we have $\mathscr{D}_{\mathscr{I}\mathscr{B}} = \{d_2, d_3, d_5\}$, which is different from any set obtained with other decision rules of Section 2.2.

Computing the set $\overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$ requires the computation of $2N$ expectation bounds (the same as for computing $\overline{\mathscr{D}}_{\mathscr{I}}$) and $2N$ comparisons at most. It is therefore as computationally efficient as the interval dominance criterion, and can be more precise (see Example 1). Actually, we will show that it is always at least as precise.

Let us now study the relation of this new decision rule with previous ones. First, we will show that the $\mathscr{I}\mathscr{B}$ decision rule is coherent with the rules inducing a complete order between decisions, before processing to the rules inducing a partial order.

3.1 Relations with Complete Ordering Rules

Let us first start with Γ-maximin and Γ-maximax. As indicates the next proposition, we can easily show that the $\mathscr{I}\mathscr{B}$ decision rule considers as optimal the decisions selected by these two rules.

Proposition 1. *The two optimal decisions $\overline{d}_{\leq_{\underline{\mathbb{E}}}}$ and $\overline{d}_{\leq_{\overline{\mathbb{E}}}}$ in the sense of Γ-maximin and Γ-maximax are also optimal in the sense of $\mathscr{I}\mathscr{B}$ dominance, that is*

$$\{\overline{d}_{\leq_{\underline{\mathbb{E}}}}, \overline{d}_{\leq_{\overline{\mathbb{E}}}}\} \subseteq \overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$$

Proof. We will only prove $\overline{d}_{\leq_{\underline{\mathbb{E}}}} \in \overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$, proof for $\overline{d}_{\leq_{\overline{\mathbb{E}}}}$ being similar. Let $\overline{d}_{\leq_{\underline{\mathbb{E}}}} = d_i$, as by definition there are no decision $d_j \in \mathscr{D}$ such that $\underline{\mathbb{E}}(U_i) < \underline{\mathbb{E}}(U_j)$, this means that there are no decision that $\mathscr{I}\mathscr{B}$-dominates d_i, hence $\overline{d}_{\leq_{\underline{\mathbb{E}}}} \in \overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$. □

The next proposition shows that $\mathscr{I}\mathscr{B}$ decision rule can also be seen as a robustification of Hurwictz criterion.

Proposition 2. *Let d_i, d_j be two different decisions. Then, $d_i \leq_{\mathscr{I}\mathscr{B}} d_j$ if and only if $d_i \leq_\alpha d_j$ for every $\alpha \in [0,1]$*

Proof. Let us first prove the "if" part. Since $d_i \leq_\alpha d_j$ for every α, if we consider $\alpha = 1$ and $\alpha = 0$ we respectively have that $\underline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq_1 \underline{\mathbb{E}}_{\mathscr{P}}(U_j)$ and $\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq_0 \overline{\mathbb{E}}_{\mathscr{P}}(U_j)$. These two inequalities leading to $d_i \leq_{\mathscr{I}\mathscr{B}} d_j$.

Let us now concentrate on the "only if" part. $d_i \leq_{\mathscr{I}\mathscr{B}} d_j$ means that $\underline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \underline{\mathbb{E}}_{\mathscr{P}}(U_j)$ and $\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \overline{\mathbb{E}}_{\mathscr{P}}(U_j)$ (these two inequalities covering the case where $\alpha = 0$ and $\alpha = 1$). Hence, for any value $\alpha \in (0,1)$, we also have $\alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_j)$ and $(1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_j)$. Summing left and right-hand sides of each equations, we have $\alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_i) + (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \alpha\underline{\mathbb{E}}_{\mathscr{P}}(U_j) + (1-\alpha)\overline{\mathbb{E}}_{\mathscr{P}}(U_j)$, hence $d_i \leq_{\mathscr{I}\mathscr{B}} d_j$ implies $d_i \leq_\alpha d_j$ for any α. □

The $\mathscr{I}\mathscr{B}$ decision rule can thus be seen as a decision rule where a decision dominates another if and only if it dominates it under all different pessimistic/optimistic attitudes, thus safeguarding the decision maker against the need to commit into such an attitude in a first analysis. Actually, it looks possible that $\overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$ contains all actions that are optimal in the Hurcwitz sense for some value of α, as is the case in the example. Let us now study the relations with the rules inducing a partial ordering.

3.2 Relations with Partial Ordering Rules

The next proposition indicates that Interval dominance implies $\mathscr{I}\mathscr{B}$ dominance.

Proposition 3. *Given a decision set $\mathscr{D}$ and a credal set $\mathscr{P}$, we have $\overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}} \subseteq \overline{\mathscr{D}}_{\mathscr{B}}$, with the inclusion being usually strict.*

Proof. We need to show that if a decision d_i is not optimal w.r.t. $\leq_{\mathscr{I}}$, then it is also not optimal w.r.t. $\leq_{\mathscr{I}\mathscr{B}}$. If d_i is not optimal w.r.t. $\leq_{\mathscr{I}}$, it means that there is a decision d_j such that $d_i <_{\mathscr{I}} d_j$, hence that $\overline{\mathbb{E}}_{\mathscr{P}}(U_i) < \underline{\mathbb{E}}_{\mathscr{P}}(U_j)$. Since $\underline{\mathbb{E}}_{\mathscr{P}}(U_i) \leq \overline{\mathbb{E}}_{\mathscr{P}}(U_i) < \underline{\mathbb{E}}_{\mathscr{P}}(U_j) \leq \overline{\mathbb{E}}_{\mathscr{P}}(U_j)$, this implies $d_i <_{\mathscr{I}\mathscr{B}} d_j$. □

The next result concerns the relation of $\mathscr{I}\mathscr{B}$ decision rule with maximality.

Proposition 4. *Given a decision set $\mathscr{D}$ and a credal set $\mathscr{P}$, we have $\overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}} \subseteq \overline{\mathscr{D}}_{\mathscr{M}}$, with the inclusion being usually strict.*

Proof. Let us show that if a decision d_i is not optimal w.r.t. $\leq_{\mathscr{M}}$, then it will also be non-optimal w.r.t. $\leq_{\mathscr{I}\mathscr{B}}$. If $d_i \notin \overline{\mathscr{D}}_{\mathscr{M}}$, then it means $\exists d_j$ s.t. $\underline{\mathbb{E}}(U_j - U_i) > 0$. Using the properties of lower expectations (see Walley [9, Ch. 2]), we have $\underline{\mathbb{E}}(U_j) + \overline{\mathbb{E}}(-U_i) \geq \underline{\mathbb{E}}(U_j - U_i)$. Using this inequality and the duality between lower and upper expectations, we have $\underline{\mathbb{E}}(U_j) + \overline{\mathbb{E}}(-U_i) = \underline{\mathbb{E}}(U_j) - \underline{\mathbb{E}}(U_i) > 0$, hence $\underline{\mathbb{E}}(U_j) > \underline{\mathbb{E}}(U_i)$. Similarly, we have that $\overline{\mathbb{E}}(U_j) + \underline{\mathbb{E}}(-U_i) \geq \underline{\mathbb{E}}(U_j - U_i)$. Using the same reasoning and duality, we have $\overline{\mathbb{E}}(U_j) - \overline{\mathbb{E}}(U_i) > 0$, meaning that $\overline{\mathbb{E}}(U_j) > \overline{\mathbb{E}}(U_i)$. Hence, $d_i <_{\mathscr{M}} d_j$ implies $d_i <_{\mathscr{I}\mathscr{B}} d_j$, and $d_i \notin \overline{\mathscr{D}}_{\mathscr{I}\mathscr{B}}$. □

This proposition tells us, among other things, that $\mathscr{I}\mathscr{B}$-dominance can be used as a quick estimate of an inner approximation of the set $\overline{\mathscr{D}}_{\mathscr{M}}$, while interval dominance can be used to estimate an outer approximation of this set. This means that both interval dominance and $\mathscr{I}\mathscr{B}$-dominance, which present a low computational complexity when compared to maximility, can be used to reduce drastically the number of required computations to evaluate $\overline{\mathscr{D}}_{\mathscr{M}}$. In the example, only two decisions that are in $\overline{\mathscr{D}}_{\mathscr{I}}$ but not in $\mathscr{D}_{\mathscr{I}\mathscr{B}}$ would need to be verified: $\{d_1, d_6\}$.

Concerning E-admissibility and $\mathscr{I}\mathscr{B}$-dominance, it is easy to see, from the example, that none imply the other, since the set of optimal actions under these rules only overlap (and their union is the set $\overline{\mathscr{D}}_{\mathscr{M}}$). Figure 1

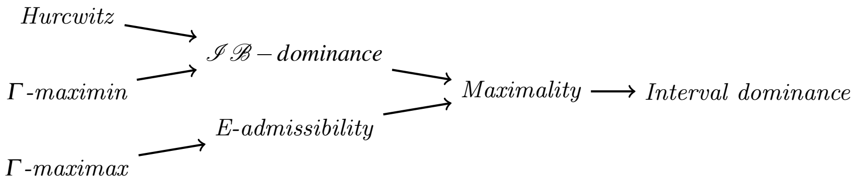


Fig. 1 Relations between decision rules: $A \to B$ means that a decision optimal in the sense of A is also optimal in the sense of B

recalls [6] and summarises the implications relation between the different rules, integrating $\mathscr{I}\mathscr{B}$-dominance into it. Roughly speaking, the figure goes from the most precise decision rules (left) to the most imprecise (right).

4 Conclusion

In this paper, we have proposed a simple new decision rule for imprecise probabilities, based on expectation bound pair-wise comparison, and have studied its relation with other existing decision rules. The interest of this rule is that it remains in the spirit of an imprecise probabilistic approach, since less information will lead to a larger set of optimal decisions, but is both computationally tractable and less conservative than most other rules. Another interesting fact is that this rule implies maximality (i.e. $\mathscr{I}\mathscr{B}$ optimal decisions are also maximal). Therefore, if not used for itself, the $\mathscr{I}\mathscr{B}$ decision rule can boost the computational tractability of $\overline{\mathscr{D}}_{\mathscr{M}}$, using it in conjunction with interval dominance to reduce the number of decision whose optimality under maximality criterion must be checked.

The next step is to evaluate to which extent this decision rule can improve the results of some tasks such as classification [10], and if it is consistent with a dynamic programming approach when dynamics enters the picture [1].

References

1. de Cooman, G., Troffaes, M.C.M.: Dynamic programming for deterministic discrete-systems with uncertain gain. Internat. J. Approx. Reason. 39, 257–278 (2004)
2. Finetti, B.: Theory of probability, vol 1(2). Wiley, NY (1974) (translation of 1970 book)
3. Gilboa, I., Schmeidler, D.: Maxmin expected utility with non-unique prior. J. Math. Econom. 18(2), 141–153 (1989)
4. Jaffray, J.Y., Jeleva, M.: Information processing under imprecise risk with the Hurwicz criterion. In: Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2007, Prague, Czech Republic, pp. 233–254 (2007)
5. Levi, I.: The Enterprise of Knowledge. MIT Press, London (1980)

6. Troffaes, M.: Decision making under uncertainty using imprecise probabilities. Internat. J. Approx. Reason. 45, 17–29 (2007)
7. Utkin, L., Augustin, T.: Powerful algorithms for decision making under partial prior information and general ambiguity attitudes. In: Proceedings of the Fourth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2005, pp. 349–358. Carnegie Mellon University, Pittsburgh (2005)
8. von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)
9. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall, New York (1991)
10. Zaffalon, M.: The naive credal classifier. Imprecise probability models and their applications (Ghent, 1999). J. Statist. Plann. Inference 105(1), 5–21 (2002)

Handling Bipolar Knowledge with Credal Sets

Sébastien Destercke

Abstract. How to represent and handle bipolar information has recently received a lot of attention. Being bipolar means that the information has a positive and negative part. In this paper, we consider asymmetric bipolar information (i.e. situations where positive and negative information are unrelated and should be processed separately). We propose a framework to represent and handle it with so-called credal sets, i.e., with convex sets of probability distributions. We also provide some illustrative examples.

Keywords: Bipolarity, Imprecise probabilities, Information fusion.

1 Introduction

Bipolarity consists in differentiating between positive and negative information. This information usually concerns either evidences about the true value assumed by an ill-known variable or preferences expressed by some agents. In this paper, we are concerned with the first type of information. One can consider at least three different types of bipolarity (See [9] for more details). The first one, called symmetric univariate, models bipolarity by the use of an univariate scale and can be represented by the means of classical probability measures. The second one, called symmetric bivariate, handles two separate unipolar scales (positive and negative) that refer to the same information and are usually linked by some duality relation. Lower and dual upper previsions [16], whose expressiveness is equivalent to credal sets, are examples of such kind of bipolarity, as well as other models encompassed by this representation (lower/upper probabilities, belief functions, possibility distributions).

The last type of bipolarity, coined as asymmetric or heterogeneous, is the one addressed in this paper. Such bipolarity is used when considering two

Sébastien Destercke
UMR IATE, Campus Supagro, 34060 Montpellier, France
e-mail: sebastien.destercke@irsn.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 199–207.
springerlink.com

unrelated kinds of information that have to be processed in parallel: one constraining the possible values of a variable (negative information), the other exhibiting what is likely to be observed (positive information). The first kind of information corresponds (for example) to constraints, physical laws, expert opinions, while examples, observations and measurements are instances of the second type. Note that the two kinds of information are effectively unrelated (for instance, an expert may judge as possible a value that will never be observed), hence the need for asymmetry. Also, some psychological studies [3] support the fact that the brain processes differently positive and negative information.

Notions of bipolarity have been declined in a number of frameworks: multi-criteria decision making [11], conflict resolution in argumention [1], uncertainty and preferences representation in possibility theory [9]. In this paper, we propose a framework to model, represent and treat bipolar information when uncertainty is modelled by convex sets of probabilities, here called credal sets [12], which constitute very generic uncertainty models. The idea behind this framework is quite simple: we propose to represent each corpus of positive and negative information as two separate credal sets, and then to conjunctively merge them in a single credal set. We also propose some solutions to deal with conflicting negative and positive information. After recalling the basics of credal sets, Section 2 presents our proposal. Section 3 then provides some illustrative examples, using the popular imprecise probabilistic representations that are p-boxes and probability intervals.

2 Handling Bipolar Information with Credal Sets

In this paper, we consider that information regarding a variable X assuming its values on a space $\mathscr{X}$ made of mutually exclusive elements is modelled by the means of a credal set $\mathscr{P}$. Let us denote by $\mathscr{L}(\mathscr{X})$ the set of real-valued bounded functions on $\mathscr{X}$. Given a function $f \in \mathscr{L}(\mathscr{X})$, one can compute the lower and upper expectations $\underline{E}_{\mathscr{P}}(f), \overline{E}_{\mathscr{P}}(f)$ induced by $\mathscr{P}$ such that

$$\underline{E}_{\mathscr{P}}(f) = \inf_{p \in \mathscr{P}} E_p(f) \quad \overline{E}_{\mathscr{P}}(f) = \sup_{p \in \mathscr{P}} E_p(f),$$

where p is a probability distribution over $\mathscr{X}$ and $E_p(f)$ the expected value of f w.r.t. p. These two values are dual, in the sense that $\underline{E}_{\mathscr{P}}(f) = -\overline{E}_{\mathscr{P}}(-f)$. Thanks to this duality, one can only work with one of the two mappings (usually $\underline{E}$).

Alternatively, one can start from a lower mapping $\underline{P} : \mathscr{K} \to \mathbb{R}$ from a subset $\mathscr{K} \subseteq \mathscr{L}(\mathscr{X})$, and consider the induced credal set $\mathscr{P}(\underline{P})$ such that

$$\mathscr{P}(\underline{P}) = \{p \in \mathbb{P}_{\mathscr{X}} | (\forall f \in \mathscr{K})(E_p(f) \geq \underline{P}(f))\}.$$

with $\mathbb{P}_{\mathscr{X}}$ the set of all probability mass functions over $\mathbb{P}_{\mathscr{X}}$. In his theory of lower previsions [16], Walley starts from the mapping $\underline{P}$ that he calls lower prevision. He interprets $\underline{P}(f)$ as the supremum buying price for the uncertain reward f. A lower prevision $\underline{P}$ is then said to avoid sure loss iff $\mathscr{P}(\underline{P}) \neq \emptyset$, and to be coherent if the lower expectation $\underline{E}_{\mathscr{P}(\underline{P})}(f) = \underline{P}(f)$ coincides with $\underline{P}$ for every $f \in \mathscr{K}$ (i.e., $\underline{P}$ is the lower envelope of $\mathscr{P}(\underline{P})$). He also shows that coherent lower previsions and credal sets have the same expressive power (in the sense that any credal can be identified by a unique lower prevision, and vice versa). Given a credal set $\mathscr{P}$, its lower (resp. upper) probability of an event A, denoted by $\underline{P}_{\mathscr{P}}(A)$ (resp. $\overline{P}_{\mathscr{P}}(A)$), corresponds to the lower (resp. upper) expectation of the indicator function $\mathbf{1}_{(A)}$ of the event , that takes value one on A and zero elsewhere. By duality, we have $\underline{P}_{\mathscr{P}}(A) = 1 - \overline{P}_{\mathscr{P}}(A^c)$.

Credal sets are very general uncertainty models, in the sense that they encompass most of the other known uncertainty models, in particular both necessity measures of possibility theory [7] and belief measures of evidence theory [13] correspond to particular classes of lower probabilities inducing specific credal sets.

2.1 Collecting and Representing Bipolar Information

As what is done in possibility theory [9] and evidence theory [14], we propose to model positive and negative information by using two separate models of our chosen framework. That is, positive information is modelled by a credal set $\mathscr{P}^+$, while negative information is modelled by another credal set $\mathscr{P}^-$.

Negative information ($\mathscr{P}^-$): Negative information expresses constraints about the value X can assume. It rules out possible values of X, considering them impossible or less likely than others (expert opinions are an example of such information). The *negative credal set* $\mathscr{P}^-$ corresponding to such information will typically be induced by a collection of expectation bounds over a set of chosen functions[1] $f_1, \ldots, f_k \in \mathscr{L}(\mathscr{X})$, in the form

$$\underline{P}(f_i) \leq \sum_{n=1}^{N} f_i(x_n) p(x_n) \leq \overline{P}(f_i) \tag{1}$$

Note that pieces of negative information are treated conjunctively, in the sense that we consider the credal set induced by all constraints (1) at once. This means that the more we accumulate negative information, the more precise is $\mathscr{P}^-$. We assume here that $\mathscr{P}^- \neq \emptyset$ (i.e., the lower prevision $\underline{P}$ given by Eq. (1) avoids sure loss).

Positive information ($\mathscr{P}^+$): Positive information consists in a set of M observations (experiments), coming in the form of data in our case. To obtain a *positive credal set* $\mathscr{P}^+$ from these data, one can use a model or a learning process. For instance, multinomial data can be associated to the well-known

[1] For example, functions corresponding to some chosen events, moments such as the mean value.

Imprecise Dirichlet model [2]. Again, positive information is accumulated conjunctively, since the more data we have, the more precise is $\mathscr{P}^+$. This is due to the fact that $\mathscr{X}$ is made of mutually exclusive elements, meaning that observing a value more often makes the observation of others less likely.

Note that there are cases where either positive or negative information should be combined disjunctively instead of conjunctively. Smets [14], when combining reasons to believe and reasons not to believe, proposes a rule that combines disjunctively negative information and conjunctively positive information. He works at a different level from ours, since we work directly with knowledge about variables, and not with evidences from which this knowledge is inferred. In their possibilistic approach, Dubois and Prade [9] also work directly with knowledge about variables, but propose to combine positive information disjunctively and negative information conjunctively. However, their proposal concerns variables taking their values on a conjunctive space $\mathscr{X}$, i.e., the true value of X can be several values of $\mathscr{X}$ (in their example, the opening hours of a museum). In that case, it appears natural to combine disjunctively positive information, as observing a particular value does not make the others less likely.

2.2 Merging Bipolar Information

Once negative and positive information have been collected, it is desirable to combine them into a unique credal set. This unique credal set should be non-empty (i.e., consistent) and more precise than the positive and negative credal sets considered separately. Given these requirements, it seems natural to merge them through a conjunctive combination operator, namely to consider as our final information the merged credal set $\mathscr{P}^{+\cap-} := \mathscr{P}^+ \cap \mathscr{P}^-$, when this intersection is not empty.

When positive and negative information conflict with each other (i.e., $\mathscr{P}^{+\cap-} = \emptyset$), it is desirable to restore consistency through some revision process. As in [9], we propose to weaken one type of information to restore consistency. Given a parameter $\varepsilon \in [0,1]$ and a credal set $\mathscr{P}$, let us first define the ε-discounted credal set $\mathscr{P}_\varepsilon$ as

$$\mathscr{P}_\varepsilon = \{\varepsilon p_{\mathscr{P}} + (1-\varepsilon)p | p_{\mathscr{P}} \in \mathscr{P}, p \in \mathbb{P}_{\mathscr{X}}\}. \tag{2}$$

When dealing with bipolar knowledge, observations are usually judged more reliable than negative information, thus it seems more reasonable to weaken $\mathscr{P}^-$ rather than $\mathscr{P}^+$. A solution to restore consistency is to consider the minimal value ε^* such that $\mathscr{P}^-_{\varepsilon^*}$ is consistent with $\mathscr{P}^+$, i.e.,

$$\varepsilon^* = \min\{\varepsilon \in [0,1] | \mathscr{P}^-_{\varepsilon^*} \cap \mathscr{P}^+ \neq \emptyset\} \tag{3}$$

and then take $\mathscr{P}^-_{\varepsilon^*} \cap \mathscr{P}^+$ as our final state of knowledge. However, as the above revision can lead to a very precise final information state, one may consider some value $\varepsilon \leq \varepsilon^*$. The same revision process can be applied to $\mathscr{P}^+$.

Even if this strategy makes more sense when bipolar information represent preferences [9], it could also be used in knowledge representation when data reliability is questionable.

2.3 Revising Knowledge with new Pieces of Information

Another case where differentiating positive and negative information rather than directly considering the merged representation $\mathscr{P}^{+\cap-}$ is useful is the case when one receives new pieces of information to be incorporated into its knowledge. For example, consider new negative information, possibly provided by an additional (reliable) expert, and modelled as a credal set $\mathscr{P}^-_{new}$. The information conveyed by $\mathscr{P}^-_{new}$ should be first added to $\mathscr{P}^-$, e.g., by computing $\mathscr{P}^{'-} = \mathscr{P}^-_{new} \cap \mathscr{P}^-$, before merging negative and positive information in a single representation. Note that making this distinction can be important, as $\mathscr{P}^-_{new}$ may be non-conflicting with $\mathscr{P}^-$ (i.e., $\mathscr{P}^-_{new} \cap \mathscr{P}^- \neq \emptyset$), while it may be conflicting with the current positive and negative information taken together (i.e., $\mathscr{P}^-_{new} \cap \mathscr{P}^+ \cap \mathscr{P}^- = \emptyset$).

3 Illustrative Examples

Let us now provide some illustrative examples of the proposed way to deal with bipolar knowledge. The examples concern two popular imprecise probabilistic models: p-boxes [10] and probability intervals [4].

3.1 p-Boxes

A p-box $[\underline{F},\overline{F}]$ defined on the (here discretized) real line $\mathbb{R}$ is a pair of lower and upper cumulative distributions describing our uncertainty about the value of a variable. They consists in lower and upper probabilities given over events of the type $(-\infty,x]$, inducing a credal set $\mathscr{P}_{[\underline{F},\overline{F}]}$ such that

$$\mathscr{P}_{[\underline{F},\overline{F}]} = \{p \in \mathbb{P}_{\mathbb{R}} | \forall x \in \mathbb{R},\, \underline{F}(x) \leq F_p(x) = P([-\infty,x]) \leq \overline{F}(x)\},$$

where F_p is the cumulative distribution of p.

Positive information. Following [10], it is possible to derive a p-box from a limited set of observations $(x_1,\dots,x_m)$ by using Kolmogorov-Smirnov confidence limits to define bounds around the empirical distribution F_m, thus making no assumption about the distribution form. The distribution F_m is defined as

$$F_m(x) = \begin{cases} 0 & \text{for} \quad x \leq x_{(1)} \\ i/n & \text{for} \quad x_{(i)} \leq x \leq x_{(i+1)} \\ 1 & \text{for} \quad x_{(m)} \leq x \end{cases}$$

where $x_{(i)}$ are the ordered sampled values. Given the samples and a confidence level $\alpha \in [0,1]$, one can use KS confidence limits to obtain a p-box $[\underline{F}_m, \overline{F}_m]$ such that

$$\underline{F}_m = \max(0, F_m - D_m(\alpha)) \quad \text{and} \quad \overline{F}_m = \min(1, F_m + D_m(\alpha))$$

We denote by $\mathscr{P}^{+}_{[\underline{F},\overline{F}]}$ the credal set obtained from this positive information.

Negative information. Negative information forming p-boxes usually comes from experts evaluating some percentiles for a set of fixed values. We denote by $\mathscr{P}^{-}_{[\underline{F},\overline{F}]}$ the credal set induced by negative information.

Merging. In the particular case of p-boxes, the credal set $\mathscr{P}[\underline{F},\overline{F}]^{-\cap+} = \mathscr{P}[\underline{F},\overline{F}]^{+} \cap \mathscr{P}[\underline{F},\overline{F}]^{-}$ is also induced by a p-box $[\underline{F},\overline{F}]^{-\cap+}$ such that

$$[\underline{F},\overline{F}]^{-\cap+} = [\max\{\underline{F}^{-},\underline{F}^{+}\}, \min\{\overline{F}^{-},\overline{F}^{+}\}].$$

In case of conflict, applying Eq. (2) does not usually result in a credal set induced by a p-box. However, given a value ε, the p-box $[\underline{F},\overline{F}]^{\varepsilon}$ such that $\underline{F}^{\varepsilon} = \varepsilon \underline{F}$ and $\overline{F}^{\varepsilon} = \varepsilon \overline{F} + 1 - \varepsilon$ induces an outer approximation of $\mathscr{P}_{\varepsilon}$.

Example 1. Assume $X \in [0,16]$. 10 samples (1; 1.5; 3; 3.5; 4; 6; 10; 11; 14; 15) provide an empirical cumulative distribution. For a confidence level of 0.95, the value $D_{10}(0.95) = 0.40925$. An expert also provides its opinion about the probabilities that the variable value is lower than values $4, 8, 12$, in the form of the following lower and upper bounds: $[0, 0.2]$, $[0.1, 0.3]$, $[0.5, 0.7]$. Figure 1 displays the p-boxes $[\underline{F},\overline{F}]^{+}$ and $[\underline{F},\overline{F}]^{-}$ resulting from these two types of information as well as the merging result.

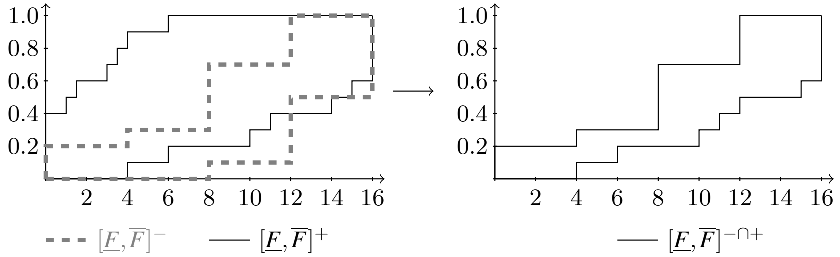


Fig. 1 Illustrative example: p-boxes

3.2 Probability Intervals

Probability intervals [4] are a set of lower and upper probabilistic bounds given over singletons $x \in \mathscr{X}$. They can be described by a set $L = \{[l(x), u(x)] | x \in \mathscr{X}\}$ of intervals. They induce a credal set $\mathscr{P}_L$ such that

$$\mathscr{P}_L = \{p \in \mathbb{P}_{\mathscr{X}} | \forall x \in \mathscr{X}, l(x) \leq p(x) \leq u(x)\}.$$

Necessary and sufficient conditions for probability intervals to induce a non-empty credal set are provided by [4]. They can be summarized by the conditions that, $\forall x \in \mathscr{X}$,

$$u(x) + \sum_{y \in \mathscr{X} \setminus x} l(y) \leq 1 \text{ and } l(x) + \sum_{y \in \mathscr{X} \setminus x} u(y) \geq 1$$

Positive information. There are mutliple models to compute confidence bounds on multinomial data with a limited number of samples. This can be done, for instance, by considering statistical confidence intervals over multinomial data [6] or by using the so-called Imprecise Dirichlet Model (IDM) [2]. Here, we consider the IDM. Let $\{x_1, \ldots, x_N\}$ be an arbitrary indexing of elements of $\mathscr{X}$, M the total number of observations, m_k the number of times x_k has been observed, and s a positive real value determining the quickness of convergence of the IDM. Then, the probability intervals derived from the IDM are such that, for x_k, $k = 1, \ldots, N$

$$l(x_k) = \frac{m_k}{m+s} \qquad \text{and} \qquad u(x_k) = \frac{m_k + s}{m+s}. \tag{4}$$

We denote by L^+ the obtained probability intervals, and $\mathscr{P}_L^+$ the induced credal set.

Negative information. As for p-boxes, negative information can be provided by some experts or by a propagation through a model (e.g., a credal network [5]). We denote by L^- the obtained probability intervals, and $\mathscr{P}_L^-$ the induced credal set.

Merging. The credal set $\mathscr{P}_L^{-\cap+} = \mathscr{P}_L^+ \cap \mathscr{P}_L^-$ is again induced by a probability interval $L^{+\cap-}$ which is such that, $\forall x \in \mathscr{X}$,

$$l^{+\cap-}(x) = \max\{l^+(x), l^-(x), 1 - \sum_{y \in \mathscr{X} \setminus x} u^+(y), 1 - \sum_{y \in \mathscr{X} \setminus x} u^-(y)\}$$
$$u^{+\cap-}(x) = \min\{u^+(x), u^-(x), 1 - \sum_{y \in \mathscr{X} \setminus x} l^+(y), 1 - \sum_{y \in \mathscr{X} \setminus x} l^-(y)\}.$$

Also note that the result of Eq (2), when applied to probability intervals L, result in a credal set still induced by probability intervals L^ε such that, $\forall x \in \mathscr{X}$,

$$l^\varepsilon(x) = \varepsilon l(x) \text{ and } u^\varepsilon(x) = \varepsilon u(x) + 1 - \varepsilon$$

Example 2. We consider a 3-elements space $\mathscr{X} = \{x_1, x_2, x_3\}$ on which are defined our probability intervals. The observed samples are such that $m = 8$ with $m_1 = 1, m_2 = 7, m_3 = 0$. To model positive information, we use the IDM with a parameter $s = 2$ and apply Eq. (4) to obtain the probability intervals L^+ such that

$$u^+(x_1) = 0.3, u^+(x_2) = 0.9, u^+(x_3) = 0.2\,; l^+(x_1) = 0.1, l^+(x_2) = 0.7, l^+(x_3) = 0.$$

Negative information is assumed to be an expert opinion given as a set L^- such that

$$u^-(x_1) = 0.4, u(x_2)^- = 0.5, u(x_3)^- = 0.3\,; l^-(x_1) = 0.2, l(x_2)^- = 0.4, l(x_3)^- = 0.$$

In this case, negative and positive information are conflicting ($u(x_2)^- \leq l(x_2)^+$), as $\mathscr{P}_L^+ \cap \mathscr{P}_L^- = \emptyset$. Using Eq. (3), we obtain $\varepsilon^* = 0.6$ and $L^-_{\varepsilon^*}$ such that

$$u^-(x_1) = 0.64, u(x_2)^- = 0.7, u(x_3)^- = 0.58\,; l^-(x_1) = 0.12, l(x_2)^- = 0.24, l(x_3)^- = 0.$$

Finally giving the merged structure $L^{+\cap-}_{\varepsilon^*}$

$$u^-(x_1) = 0.3, u(x_2)^- = 0.7, u(x_3)^- = 0.18\,; l^-(x_1) = 0.12, l(x_2)^- = 0.7, l(x_3)^- = 0$$

which indeed gives a very precise evaluation of the uncertainty of having $X = x_2$.

4 Conclusion

We have proposed a framework to handle bipolar asymmetric information in the framework of imprecise probabilities, when this information concerns knowledge about the value of a given variable. The proposal is illustrated with some credal sets induced by specific probability bounds often used in practice. This work is a first step towards the modelling and handling of bipolar information within the recent theory of imprecise probabilities. It still has to be compared in a deeper way with other approaches made in possibility theory and evidence theory, possibly by making sense of the concept of guaranteed possibility [8] or of commonality function in the context of imprecise probabilities.

Another interesting problem is how to handle bipolarity when credal sets or lower previsions are used not to express uncertainty but imprecise preferences or utilities. An idea would be to consider the alternative model of desirable gambles, recently considered as a solution to multicriteria decision problems [15].

References

1. Amgoud, L., Cayrol, C., Lagasquie-Schiex, M., Livet, P.: On bipolarity in argumentation frameworks. Int. J. Intell. Syst. 23, 1062–1093 (2008)
2. Bernard, J.M.: An introduction to the imprecise Dirichlet model. Internat. J. Approx. Reason. 39, 123–150 (2008)
3. Cacioppo, J., Bernston, G.: The affect system: architecture and operating characteristics. Curr. Dir. Psychol. Sci. 8, 133–137 (1999)
4. de Campos, L., Huete, J., Moral, S.: Probability intervals: a tool for uncertain reasoning. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 2, 167–196 (1994)
5. Cozman, F.: Credal networks. Artificial Intelligence 120, 199–233 (2000)

6. Denoeux, T.: Constructing belief functions from sample data using multinomial confidence regions. Internat. J. Approx. Reason. 42, 228–252 (2006)
7. Dubois, D., Prade, H.: Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York (1988)
8. Dubois, D., Prade, H.: Interval-valued fuzzy sets, possibility theory and imprecise probability. In: Proceedings of International Conference in Fuzzy Logic and Technology, EUSFLAT 2005, Barcelona, Spain (2005)
9. Dubois, D., Prade, H.: An overview of the asymmetric bipolar representation of positive and negative information in possibility theory. Fuzzy Sets Syst. 160, 1355–1366 (2009)
10. Ferson, S., Ginzburg, L., Kreinovich, V., Myers, D., Sentz, K.: Constructing probability boxes and Dempster-Shafer structures. Tech. rep., Sandia National Laboratories (2003)
11. Grabisch, M., Labreuche, C.: Bi-capacities I: definition, Möbius transform and interaction. II: The choquet integral. Fuzzy Sets Syst. 151, 211–259 (2005)
12. Levi, I.: The Enterprise of Knowledge. MIT Press, London (1980)
13. Shafer, G.: A mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
14. Smets, P.: The canonical decomposition of a weighted belief. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 1995, Montréal, Qu'ebec, Canada, pp. 1896–1901 (1995)
15. Utkin, L.: Multi-criteria decision making with a special type of information about importance of groups of criteria. In: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK, pp. 411–420 (2009)
16. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall, New York (1991)

Coherent Upper Conditional Previsions and Their Integral Representation with Respect to Hausdorff Outer Measures

Serena Doria

Abstract. In a metric space, a necessary and sufficient condition is given for a coherent upper conditional prevision to be uniquely represented by the Choquet integral with respect to the upper conditional probability defined by its Hausdorff outer measure.

1 Introduction

Separately coherent upper conditional previsions [8] are functionals on a linear space of bounded random variables satisfying the axioms of coherence. In this paper a model of separately coherent upper conditional previsions is proposed in a metric space. They are defined by the Choquet integral with respect to the Hausdorff outer measure if the conditioning event has positive and finite Hausdorff outer measure in its dimension. Otherwise if the conditioning event has Hausdorff outer measure in its dimension equal to zero or infinity they are defined by a 0-1 valued finitely, but not countably, additive probability. If the conditioning event B has positive and finite Hausdorff outer measure in its dimension s then the upper conditional prevision defined on the linear space of all bounded random variables on B is proven to be a functional, which is monotone, submodular, comonotonically additive and continuous from below. Moreover, let $\mathbf{L}(B)$ be a linear lattice of bounded random variables on B containing all constants. It is proven that a sufficient condition for a coherent upper conditional prevision defined on $\mathbf{L}(B)$ to be uniquely represented as Choquet integral with respect to the upper conditional probability defined by the Hausdorff s-dimensional outer measure is to be monotone, submodular, comonotonically additive and continuous from below.

Serena Doria
Department of Sciences, University G. d'Annunzio, Chieti-Pescara, Italy
e-mail: s.doria@dst.unich.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 209–216.
springerlink.com

2 Coherent Upper Conditional Previsions Defined with Respect to Hausdorff Outer Measures

Separately coherent upper conditional previsions $\overline{P}(\cdot|B)$ are functionals, defined on a linear space of bounded random variables, satisfying the axioms of coherence [8].

Definition 1. *Let* $(\boldsymbol{\Omega},d)$ *be a metric space and let* $\boldsymbol{B}$ *be a partition of* $\boldsymbol{\Omega}$*. For every* $B \in \boldsymbol{B}$ *let* $\boldsymbol{L}(B)$ *be a linear space of bounded random variables defined on* B*. Then separately coherent upper conditional previsions are functionals* $\overline{P}(\cdot|B)$ *defined on* $\boldsymbol{L}(B)$*, such that the following conditions hold for every* X *and* Y *in* $\boldsymbol{L}(B)$ *and every strictly positive constant* λ*:*

1) $\overline{P}\ (X|B) \leq sup(X|B)$*;*
2) $\overline{P}(\lambda\ X|B) = \lambda\ \overline{P}(X|B)$ *(positive homogeneity);*
3) $\overline{P}(X+Y)|B) \leq \overline{P}(X|B) + \overline{P}(Y|B)$*;*
4) $\overline{P}(B|B) = 1$*.*

In this section coherent upper conditional previsions are defined by the Choquet integral with respect to the Hausdorff outer measures if the conditioning event B has positive and finite Hausdorff outer measure in its dimension; if the conditioning event B has Hausdorff outer measure in its dimension equal to zero or infinity they are defined by a 0-1 valued finitely, but not countably, additive probability.

Let $(\boldsymbol{\Omega},d)$ be a metric space. The diameter of a non empty set U of $\boldsymbol{\Omega}$ is defined as $|U| = \sup\{d(x,y) : x,y \in U\}$ and if a subset A of $\boldsymbol{\Omega}$ is such that $A \subset \bigcup_i U_i$ and $0 < |U_i| < \delta$ for each i, the class $\{U_i\}$ is called a δ-cover of A. Let s be a non-negative number. For δ >0 we define $h_{s,\delta}(A) = \inf \sum_{i=1}^{\infty} |U_i|^s$, where the infimum is over all δ-covers $\{U_i\}$. The *Hausdorff s-dimensional outer measure* [6] of A, denoted by $h^s(A)$, is defined as $h^s(A) = \lim_{\delta \to 0} h_{s,\delta}(A)$. This limit exists, but may be infinite, since $h_{s,\delta}(A)$ increases as δ decreases because less δ-covers are available. The *Hausdorff dimension* of a set A, $dim_H(A)$, is defined as the unique value, such that $h^s(A) = \infty$ if $0 \leq s < dim_H(A)$ and $h^s(A) = 0$ if $dim_H(A) < s < \infty$. We can observe that if $0 < h^s(A) < \infty$ then $dim_H(A) = s$, but the converse is not true. Hausdorff outer measures are *metric* outer measures, that is $h^s(E \cup F) = h^s(E) + h^s(F)$ whenever E and F are *positively separated*, i.e. $d(E,F) = \inf\{d(x,y) : x \in E, y \in F\} > 0$. A subset A of $\boldsymbol{\Omega}$ is called *measurable* with respect to the outer measure h^s if it decomposes every subset of $\boldsymbol{\Omega}$ additively, that is if $h^s(E) = h^s(A \cap E) + h^s(E - A)$ for all sets $E \subseteq \boldsymbol{\Omega}$. All Borel subsets of $\boldsymbol{\Omega}$ are measurable with respect to any metric outer measure [4, Theorem 1.5]. So every Borel subset of $\boldsymbol{\Omega}$ is measurable with respect to every Hausdorff outer measure h^s since Hausdorff outer measures are metric. The restriction of h^s to the σ-field of h^s-measurable sets, containing the σ-field of the Borel sets, is called *Hausdorff s-dimensional measure*. In particular the Hausdorff 0-dimensional measure is the counting measure and the Hausdorff 1-dimensional measure is the Lebesgue measure. The Hausdorff s-dimensional measures are *modular* on the Borel σ-field, that is $h^s(A \cup B) + h^s(A \cap B) = h^s(A) + h^s(B)$ for every pair of Borelian sets A and B; so that [2, Proposition 2.4] the Hausdorff outer measures are *submodular* $(h^s(A \cup B) + h^s(A \cap B) \leq h^s(A) + h^s(B))$. Moreover Hausdorff outer measures are *continuous from below* [4, Lemma 1.3], that is for any increasing sequence of sets $\{A_i\}$ we have $\lim_{i \to \infty} h^s(A_i) = h^s(\lim_{i \to \infty} A_i)$.

Let μ: $S \to \overline{\Re_+} = \Re_+ \cup \{+\infty\}$ be a monotone set function defined on S properly contained in $\wp(\Omega)$ and $X : \Omega \to \overline{\Re} = \Re \cup \{-\infty, +\infty\}$ an arbitrary function on Ω then the set function $G_{\mu,X}(x) = \mu\{\omega \in \Omega : X(\omega) > x\}$ is decreasing and it is called *decreasing distribution function* of X with respect to μ. Denote by μ^* and μ_* respectively the outer and inner measure of μ. A function X: $\Omega \to \overline{\Re}$ is called upper μ*-measurable* if $G_{\mu^*,X}(x) = G_{\mu_*,X}(x)$. Given an upper μ-measurable function X: $\Omega \to \overline{\Re}$ with decreasing distribution function $G_{\mu,X}(x)$, the Choquet integral of X with respect to μ [2] is defined by $\int X d\mu = \int_{-\infty}^{0} (G_{\mu,X}(x) - \mu(\Omega))dx + \int_0^\infty G_{\mu,X}(x)dx$ if $\mu(\Omega) < +\infty$.

If X is bounded and $\mu(\Omega) = 1$ then the Choquet integral is given by $\int X d\mu = \int_{\inf X}^{\sup X} G_{\mu,X}(x)dx + \inf X$.

Theorem 1. *Let (Ω, d) be a metric space and let $\boldsymbol{B}$ be a partition of Ω. For every $B \in \boldsymbol{B}$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let $\boldsymbol{L}(B)$ be the class of all bounded random variables on $\boldsymbol{B}$. Moreover, let m be a 0-1 valued finitely, but not countably, additive probability on $\wp(B)$ such that a different m is chosen for each B. Then for each $B \in \boldsymbol{B}$ the functionals $\overline{P}(X|B)$ defined on $\boldsymbol{L}(B)$ by*

$$\overline{P}(X|B) = \frac{1}{h^s(B)} \int_B X dh^s \qquad if \qquad 0 < h^s(B) < \infty$$

and by

$$\overline{P}(X|B) = m(XB) \qquad if \qquad h^s(B) = 0, \infty$$

are separately coherent upper conditional previsions.

Proof. Since $\mathbf{L}(B)$ is a linear space we have to prove that, for every $B \in \mathbf{B}$ $\overline{P}(X|B)$ satisfies conditions 1), 2), 3), 4) of Definition 1.

If B has finite and positive Hausdorff outer measure in its dimension s then $\overline{P}(X|B) = \frac{1}{h^s(B)} \int_B X dh^s$, so properties 1) and 2) are satisfied since they hold for the Choquet integral [2, Proposition 5.1]. Property 3) follows from the Subadditivity Theorem [2, Theorem 6.3] since Hausdorff outer measures are monotone, submodular and continuous from below. Property 4) holds since $\overline{P}(B|B) = \frac{1}{h^s(B)} \int_B dh^s = 1$. If B has Hausdorff outer measure in its dimension equal to zero or infinity we have that the class of all coherent (upper) previsions on $\mathbf{L}(B)$ is equivalent to the class of 0-1 valued additive probabilities defined on $\wp(B)$ then $\overline{P}(X|B) = m(XB)$. Then properties 1), 2), 3) are satisfied since m is a 0-1 valued finitely additive probability on $\wp(B)$. Moreover since a different m is chosen for each B we have that $\overline{P}(B|B) = m(B) = 1$. □

The unconditional upper prevision is obtained as a particular case when the conditioning event is $\boldsymbol{\Omega}$. Upper conditional probabilities are obtained when only 0-1 valued random variables are considered; they have been defined in [3]:

Theorem 2. *Let $(\boldsymbol{\Omega},d)$ be a metric space and let $\boldsymbol{B}$ be a partition of $\boldsymbol{\Omega}$. For every $B \in \boldsymbol{B}$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let m be a 0-1 valued finitely, but not countably, additive probability on $\wp(B)$ and a different m is chosen for each B. Then, for each $B \in \boldsymbol{B}$, the functions defined on $\wp(B)$ by*

$$\overline{P}(A|B) = \frac{h^s(AB)}{h^s(B)} \quad if \quad 0 < h^s(B) < \infty$$

and by

$$\overline{P}(A|B) = m(AB) \quad if \quad h^s(B) = 0, \infty$$

are separately coherent upper conditional probabilities.

Let $B \in \mathbf{B}$ be a set with positive and finite Hausdorff outer measure in its dimension s. Denote by h^s the s-dimensional Hausdorff outer measure and for every $A \in \wp(B)$ by $\mu_B^*(A) = \overline{P}(A|B) = \frac{h^s(AB)}{h^s(B)}$ the upper conditional probability defined on $\wp(B)$. From Theorem 1 we have that the upper conditional prevision $\overline{P}(\cdot|B)$ is a functional defined on $\mathbf{L}(B)$ with values in $\Re$ and the upper conditional probability μ_B^* integral represents $\overline{P}(X|B)$ since $\overline{P}(X|B) = \int X d\mu_B^* = \frac{1}{h^s(B)} \int X dh^s$. The number $\frac{1}{h^s(B)}$ is a normalizing constant. A class of bounded random variables is called a *lattice* if it is closed under point-wise maximum $\vee$ and point-wise minimum $\wedge$. In the following theorem it is proven that, if the conditioning event has positive and finite Hausdorff outer measure in its dimension s and $\mathbf{L}$(B) is a linear lattice of bounded random variables defined on B, necessary conditions for the functional $\overline{P}(X|B)$ to be represented as Choquet integral with respect to the upper conditional probability μ_B^*, i.e. $\overline{P}(X|B) = \frac{1}{h^s(B)} \int X dh^s$, are that $\overline{P}(X|B)$ is monotone, comonotonically additive, submodular and continuous from below.

Theorem 3. *Let $(\boldsymbol{\Omega},d)$ be a metric space and let $\boldsymbol{B}$ be a partition of $\boldsymbol{\Omega}$. For every $B \in \boldsymbol{B}$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let $\boldsymbol{L}(B)$ be the class of all bounded random variables defined on B. If the conditioning event B has positive and finite Hausdorff s-dimensional outer measure in its dimension then the upper conditional prevision $\overline{P}(\cdot|B)$ defined on $\boldsymbol{L}(B)$ as in Theorem 1 satisfies the following properties:*

i) $X \leq Y$ implies $\overline{P}(X|B) \leq \overline{P}(Y|B)$ (monotonicity);
ii) if X and Y are comonotonic, i.e.$(X(\omega_1) - X(\omega_2))(Y(\omega_1) - (Y(\omega_2)) \geq 0$ $\forall \omega_1, \omega_2 \in B$, then $\overline{P}(X+Y|B) = \overline{P}(X|B) + \overline{P}(Y|B)$ (comonotonic additivity);
iii)$\overline{P}(X \vee Y|B) + \overline{P}(X \wedge Y|B) \leq \overline{P}(X|B) + \overline{P}(Y|B)$ (submodularity);
iv)$lim_{n\to\infty}\overline{P}(X_n|B) = \overline{P}(X|B)$ if X_n is an increasing sequence of random variables converging to X (continuity from below).

Proof. Since the conditioning event B has positive and finite Hausdorff outer measure in its dimension s then the functional $\overline{P}(\cdot|B)$ is defined on $\mathbf{L}(B)$ by the Choquet integral with respect to the upper conditional probability $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$; so conditions *i*) and *ii*) are satisfied because they are properties of the Choquet integral [2, Proposition 5.2]. Condition *iii*) is equivalent to require that the monotone set function that represents the functional $\overline{P}(\cdot|B)$ is submodular and it is satisfied since Hausdorff outer measures are submodular. Moreover every s-dimensional Hausdorff measure is continuous from below then from the Monotone Convergence Theorem [2, Theorem 8.1] we have that the functional $\overline{P}(\cdot|B)$ is continuous from below, that is condition iv). □

3 Uniqueness of the Representing Set Function for Coherent Upper Conditional Previsions

If the conditioning event B has positive and finite Hausdorff measure in its dimension there is the problem of determining conditions, which assure that a coherent upper conditional prevision $\overline{P}(\cdot|B)$ can be represented by the Choquet integral with respect to a monotone set function and to determine the interval of monotone set functions which represent $\overline{P}(\cdot|B)$.

The representation of coherent lower previsions as Choquet integrals with respect to supermodular lower probabilities has been studied in [1]. In the quoted paper a representation result for exact n-monotone functionals in terms of Choquet integrals has been proven. But the result does not address the uniqueness of the representing function. Moreover 2-monotone lower probabilities are investigated in [7].

Given a family $\mathbf{L}$ of functions $X : \Omega \to \overline{\Re}$ and a functional $\Gamma : \mathbf{L} \to \overline{\Re}$ we say that Γ can be represented as Choquet integral with respect to a monotone set function μ on $\wp(\Omega)$ if $\Gamma(X) = \int X d\mu$. In Denneberg [2, Chapter 13], representation theorems for functionals with minimal requirements on the domain are examined. Let $\mathbf{L}$ be a class of random variables such that

a) $X \geq 0$ for all $X \in \mathbf{L}$ (non negativity);
b) $aX, X \wedge a, X - X \wedge a \in \mathbf{L}$ if $X \in \mathbf{L}$, $a \in \Re^+$;
c) $X \wedge Y, X \vee Y$ if $X, Y \in \mathbf{L}$ (lattice property).

In [2, Proposition 13.5] it is proven that if a functional Γ, defined on the domain $\mathbf{L}$, is monotone, comonotonically additive, submodular and continuous from below then Γ is representable as Choquet integral with respect to a monotone, submodular set function which is continuous from below. Furthermore all set functions on $\wp(\Omega)$ with these properties agree on the set system of weak upper level sets $M = \{\{X \geq x\} \,|X \in \mathbf{L}, x \in \Re_+\}$. The uniqueness of the representing set function [2, Lemma 13.1] is due to the fact that the function $\Gamma(X \wedge x)$ determines the distribution function $G_{\mu,X}$ of an upper μ-measurable and positive random variable X with respect to any set function μ representing Γ; it occurs since $G_{\mu,X} = \frac{d}{dx}\Gamma(X \wedge x)$ for $X \in \mathbf{L}$ and for all $x \in \Re^+$ of continuity for $G_{\mu,X}$. If μ is continuous from below then $G_{\mu,X}$ is right continuous and it is the derivative from the right of $\Gamma(Y \wedge x)$ for every point $x \in \Re^+$. If the domain $\mathbf{L}$ is a linear lattice containing all constants this result can be extended to every bounded random variable. In fact since X is bounded, there exists a constant k such that $Y = X - k \in \mathbf{L}$ and $Y = X - k \geq 0$ so that $G_{\mu,Y} = \frac{d}{dx}\Gamma(Y \wedge x)$.

In the next theorem a sufficient condition is given such that a coherent upper conditional prevision is uniquely represented as Choquet integral with respect to the upper conditional probability μ_B^* defined by Hausdorff outer measure. It is proven that if the conditioning event B has positive and finite Hausdorff outer measure in its dimension s and the coherent upper conditional prevision $\overline{P}(\cdot|B)$ is monotone, comonotonically additive, submodular and continuous from below then the upper conditional probability μ_B^* defined by the s-dimensional Hausdorff outer measure h^s is the unique monotone set function on the set system of weak upper level sets $M = \{\{X \geq x\} | X \in \mathbf{L}(B), x \in \Re\}$, which is submodular, continuous from below and representing $\overline{P}(\cdot|B)$ as Choquet integral. That is for every monotone set function β on $\wp(B)$, which is submodular, continuous from below and represents $\overline{P}(\cdot|B)$ we have that $\overline{P}(X|B) = \int_B X d\beta = \int_B X d\mu_B^* = \frac{1}{h^s(B)}\int_B X dh^s$ for every bounded random variable X.

Theorem 4. *Let (Ω, d) be a metric space and let $\boldsymbol{B}$ be a partition of Ω. For every $B \in \boldsymbol{B}$ denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. Let $\boldsymbol{L}(B)$ be a linear lattice of bounded random variables on B containing all constants. If B has positive and finite Hausdorff outer measure in its dimension and the coherent upper conditional prevision $\overline{P}(\cdot|B)$ on $\boldsymbol{L}(B)$ satisfies the properties i), ii), iii), iv) then $\overline{P}(\cdot|B)$ is representable as Choquet integral with respect to a monotone, submodular set function which is continuous from below. Furthermore all monotone set functions on $\wp(B)$ with these properties agree on the set system of weak upper level sets $M = \{\{X \geq x\} | X \in \boldsymbol{L}(B), x \in \Re\}$ with the upper conditional probability $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$ for $A \in \wp(B)$. Let β be any monotone set function on $\wp(B)$, which is submodular, continuous from below and such that represents $\overline{P}(\cdot|B)$ as Choquet integral. Then the following equalities hold*

$$\overline{P}(X|B) = \int_B X d\beta = \int_B X d\mu_B^* = \frac{1}{h^s(B)}\int_B X dh^s.$$

Proof. $\mathbf{L}(B)$ is a linear lattice containing all constants so we can assume that property a) is true because otherwise since X is bounded there exists a constant k such that $X - k \in \mathbf{L}$(B) and $X - k \geq 0$. Moreover conditions b) and c) are satisfied. So from Proposition 13.5 of [2] we obtain that the functional $\overline{P}(\cdot|B)$ is representable by a monotone, submodular, continuous from below set function and all set functions with these properties agree on the set system of weak upper level sets $M = \{\{X \geq x\} | X \in \mathbf{L}(B), x \in \Re\}$. Every s-dimensional Hausdorff outer measure is monotone, submodular and continuous from below so, if B has positive and finite Hausdorff outer measure in its dimension then the monotone set function $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$ defined on $\wp(B)$ by the s-dimensional Hausdorff measure represents the functional $\overline{P}(\cdot|B)$. Moreover all monotone set functions on $\wp(B)$ which are submodular, continuous from below and represent the functional $\overline{P}(\cdot|B)$ agree on the set system of weak upper level sets with the upper conditional probability $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$. Denote by β any monotone set function on $\wp(B)$, which is submodular, continuous from below and such that represents $\overline{P}(\cdot|B)$ as Choquet integral. Then μ_B^* and β agree on the set system of weak upper level sets M and $G_{\mu_B^*,X}(x) = G_{\beta,X}(x)$. Moreover $\mu_B^*(B) = \beta(B) = 1$. Since every X belonging to $\mathbf{L}(B)$ is bounded the following equalities hold:

$$\overline{P}(X|B) = \int_B X d\beta = \int_{\inf X}^{\sup X} G_{\beta,X}(x)dx + \inf X$$

$$= \int_{\inf X}^{\sup X} G_{\mu_B^*,X}(x)dx + \inf X = \int_B X d\mu_B^* = \frac{1}{h^s(B)}\int_B X dh^s.$$

□

The same result can be obtained if the coherent upper conditional probabilities μ_B^* and β are defined on the class S properly contained in $\wp(B)$ and $\mathbf{L}(B)$ is a linear lattice of bounded upper S-measurable random variables on B containing all constants.

Given a monotone set function β in Greco [5] a definition of measurability for positive functions with respect to a class S of subsets of Ω is given with the aim to determine the functions X such that the Choquet integral $\int X d\beta$ depends only on the values of β on S.

Definition 2. *[5, p.165] A positive random variable X is S-measurable if and only if $\int X d\beta = \int X d\alpha$, where α, β are monotone set functions defined on $\wp(\Omega)$ such that $\alpha(A) = \beta(A)$ for every set A in S. A random variable X is S-measurable if X^+ and X^- are S-measurable where $X^+ = X \vee 0$ and $X^- = (-X) \vee 0$.*

The previous definition is proven [5, Theorem 1] to be equivalent to the following condition 5):

$\forall a, b \in \Re, a < b$ there exists a set $H \in S$ so that $\{X > a\} \supset H \supset \{X > b\}$.

In Denneberg [2, p.49] a random variable X is defined to be upper S-measurable if it is upper μ-measurable ($G_{\mu^*,X}(x) = G_{\mu_*,X}(x)$) for any monotone set function μ on S. Condition 5) is a necessary and sufficient condition [2, Proposition 4.2] for upper S-measurability of a random variable X. In particular X is upper S-measurable if the upper set system $M_X = \{\{X \geq x\}, x \in \Re\}$, is contained in S. If S is a σ-field and M_X and M_{-X} are contained in S then we have the classical condition of measurability of functions.

Theorem 5. *Let (Ω, d) be a metric space and let $\boldsymbol{B}$ be a partition of Ω. For every $B \in \boldsymbol{B}$ let $\boldsymbol{L}(B)$ be a linear lattice of bounded random variables* *on B containing all constants. Let $\overline{P}(\cdot|B)$ be a coherent upper conditional prevision satisfying i), ii), iii), iv). Let S be a subclass properly contained in $\wp(B)$ such that it contains the set system of weak upper level sets $M = \{\{X \geq x\} | X \in \boldsymbol{L}(B); x \in \Re\}$. Denote by s the Hausdorff dimension of the conditioning event B and by h^s the Hausdorff s-dimensional outer measure. If $0 < h^s(B) < +\infty$ define $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$, for every $A \in S$; let β be a coherent upper probability on S, which is submodular, continuous from below and such that represents $\overline{P}(\cdot|B)$ as Choquet integral. Then the following equalities hold:*

$$\overline{P}(X|B) = \int X d\beta = \int X d\mu_B^* = \frac{1}{h^s(B)} \int X dh^s.$$

4 Conclusions

In this paper coherent upper conditional previsions are characterized in a metric space as Choquet integrals with respect to the upper conditional probabilities defined by the Hausdorff outer measures. Let B be a conditioning event with positive and finite Hausdorff outer measure in its dimension s; a coherent upper conditional prevision $\overline{P}(X|B)$ defined on a linear lattice $\mathbf{L}(B)$ of bounded random variables on B containing all constants, is proven to be monotone, comonotonically additive, submodular and continuous from below if and only if it is representable as the Choquet integral with respect to the upper conditional probability $\mu_B^*(A) = \frac{h^s(AB)}{h^s(B)}$, defined on $\wp(B)$ by the Hausdorff s-dimensional outer measure h^s.

References

1. de Cooman, G., Troffaes, M., Miranda, E.: n-Monotone exact functional. J. Math. Anal. Appl. 347(1), 133–146 (2008)
2. Denneberg, D.: Non-additive measure and integral. Kluwer Academic Publishers, Dordrecht (1984)
3. Doria, S.: Probabilistic independence with respect to upper and lower conditional probabilities assigned by Hausdorff outer and inner measures. Internat. J. Approx. Reason. 46, 617–635 (2007)
4. Falconer, K.J.: The geometry of fractals sets. Cambridge University Press, Cambridge (1986)
5. Greco, G.: Sur la mesurabilité d'une fonction numérique par rapport à une famille d'ensembles. Rend. Sem. Mat. Univ. Padova 65, 21–42 (1981)
6. Rogers, C.A.: Hausdorff measures. Cambridge University Press, Cambridge (1970)
7. Walley, P.: Coherent lower (and upper) probabilities. Statistics Research Report, University of Warwick (1981)
8. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)

Statistical Inference with Belief Functions and Possibility Measures: A Discussion of Basic Assumptions

Didier Dubois and Thierry Denœux

Abstract. This paper reconsiders the problem of statistical inference from the standpoint of evidence theory and possibility theory. The Generalized Bayes theorem due to Smets is described and illustrated on a small canonical example. Critiques addressed to this model are discussed as well as the robust Bayesian solution. Finally, the proposal made by Shafer to exploit likelihood information in terms of consonant belief function within the scope of possibility theory is reconsidered. A major objection to this approach, due to a lack of commutativity between combination and conditioning, is circumvented by assuming that the set of hypotheses or parameter values is rich enough.

Keywords: Statistical inference, Belief function possibility theory, Likelihood principle.

1 Introduction

Let X be a space of observations. Given a probabilistic parametric model $P_\theta, \theta \in \Theta$, interpreted as a conditional probability $P(\cdot|\theta)$, a set of independent observations $x_1,\ldots,x_k$ obtained in the same conditions and a subjective prior probability $P_{sub}(\theta)$, Bayes' theorem in probability theory prescribes that a posterior probability on Θ can computed as $P(\theta|x_1,\ldots,x_k) \propto P_{sub}(\theta)\prod_{i=1}^{k} P(x_i|\theta)$, where $P(x_i|\theta)$ is the likelihood function. A recurring question in statistical inference is: what information does observed data provide about a probabilistic model when no prior probability is supplied and Bayes' theorem cannot

Didier Dubois
IRIT, CNRS and Université de Toulouse, France
e-mail: dubois@irit.fr

Thierry Denœux
HEUDIASYC, CNRS and Université de Technologie de Compiègne, France
e-mail: thierry.denoeux@hds.utc.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 217–225.
springerlink.com

be applied? What to say on the basis of observations, when only likelihood information is available?

In this paper, we review some statistical inference methods that can be proposed in the setting of belief functions and possibility theory. Both settings have the merit of not requiring prior knowledge when learning from data. We try to provide a clear presentation of Smets' Generalized Bayes theorem without prior, laying bare the assumptions. Then we discuss the related literature in probability theory. We study to what extent a similar approach makes sense in possibility theory.

2 The Generalized Bayes Theorem for Belief Functions

The problem of inferring knowledge from likelihood functions has been addressed by Philippe Smets in his 1978 thesis [11] in the setting of belief functions, given several observations forming a finite set X and a non-binary parameter space Θ. The Generalized Bayes Theorem (GBT) computes a non-trivial uncertainty measure on the parameter space from parameterized belief functions on X even if no prior knowledge about the parameter is available. If there is some prior information, it can be used. Bayes' theorem is retrieved in the special case where belief functions are probability measures and a prior probability distribution on Θ is given. The GBT has been applied to classification problems [4]. It is interesting to study what are its underlying assumptions and under which conditions it can be applied to statistical inference; of interest is how it compares with other approaches.

Let X be a frame of discernment. An uncertain body of evidence is defined by means of a mass function m which is a probability distribution over the power set 2^X. In particular, $\sum_{E\subseteq X} m(E) = 1$. The mass $m(E)$ is the probability mass that could be allocated to some element of E but is not by lack of information. The quantity $m(\emptyset)$ represents a degree of internal conflict, and according to Smets, may suggest the idea that the truth may lie outside X (open world assumption). For simplicity, we assume $m(\emptyset) = 0$ (closed world assumption). The following notions are useful in the sequel:

- The degree of belief is $bel(A) = \sum_{E\subseteq A} m(E)$;
- The degree of plausibility is $pl(A) = \sum_{\emptyset\neq E\cap A} m(E) = 1 - bel(\bar{A})$, where $\bar{A}$ is the complement of A;
- Standard (normalized) conditioning :

$$pl(A|B) = \frac{pl(A\cap B)}{pl(B)}; \quad bel(A|B) = \frac{bel(A\cup \bar{B}) - bel(\bar{B})}{1 - bel(\bar{B})};$$

- Conjunctive merging $\bigcirc\!\!\!\!\cap$: $(m_1 \bigcirc\!\!\!\!\cap m_2)(C) = \sum_{A,B,A\cap B=C} m_1(A)m_2(B)$;
- Dempster rule of combination $\oplus$: It consists in renormalizing $m_1 \bigcirc\!\!\!\!\cap m_2$ dividing it by $1-(m_1 \bigcirc\!\!\!\!\cap m_2)(\emptyset)$, which makes sense under a closed-world assumption.

Given a family $\{bel_X(\cdot|\theta), \theta \in \Theta\}$ of belief functions (supposed to be normalized), parameterized by θ, the ballooning extension (or conditional embedding) of $bel_X(\cdot|\theta)$ into $X \times \Theta$ is the least committed belief function whose conditional on θ is $bel_X(\cdot|\theta)$. It consists in assigning each mass $m_X(E|\theta)$ to the subset $E \cup \overline{\{\theta\}} \subseteq X \times \Theta, \forall E \subseteq X$. On $X \times \Theta$, the ballooning extension is such that $bel^{\theta}(E \cup \overline{\{\theta\}}) = bel_X(E|\theta)$ (assuming $pl(E \cup \overline{\{\theta\}}) = 1, \forall \theta \in \Theta$).

The inference problem can then be stated as follows: Given a set of parametric belief functions $bel_X(\cdot|\theta)$, $\theta \in \Theta$, and some observation $x \in X$, compute $bel_{\Theta}(\cdot|x)$. It is assumed that for $T \subseteq \Theta, pl_X(x|T)$ is a function of elementary likelihoods $pl_X(x|\theta), pl_X(x|\overline{\{\theta\}})$, $\theta \in T$. Computing the posterior belief function $bel_{\Theta}(\cdot|x)$ goes as follows, given a **finite** parameter space Θ and a set of parametric belief functions $bel_X(\cdot|\theta), \theta \in \Theta$:

1. **Conditional embedding** of each $bel_X(\cdot|\theta)$ in $X \times \Theta$ (ballooning);
2. **Conjunctive merging** of the embedded belief functions $bel^{\theta}, \theta \in \Theta$ on $X \times \Theta$;
3. **Conditioning** of the result on the observation x;
4. **Marginalizing** on Θ.

The use of the conjunctive merging rule in step 2 assumes that the belief functions $bel_X(\cdot|\theta), \theta \in \Theta$ have been inferred from distinct sets of empirical data obtained from independent sources. Moreover, this step comes down to applying to $T = \Theta$ the disjunctive combination rule to the conditional belief functions $bel_X(\cdot|\theta)$: $bel_X(A|T) = \prod_{\theta \in T} bel_X(A|\theta), \forall A \subseteq X$. Finally, after marginalization, posterior plausibility functions $pl_{\Theta}(T|A)$ are proportional to $1 - \prod_{\theta \in T}(1 - pl_X(A|\theta)), \forall T \subseteq \Theta$.

The problem has been extended to n independent observations $x_1, \ldots, x_n$ in $\{x, \bar{x}\}^n$ [12]. The GBT has a nice commutativity property. One may compute $bel_{X^n}(x_1, \ldots x_n|\theta)$, conjunctively combining $bel_X(\cdot|\theta)$, perform a conditional embedding on $X^n \times \Theta$, then get the posterior belief function $bel_{\Theta}(\theta|x_1, \ldots, x_n)$. It is equivalent to computing n posterior belief functions $bel_{\Theta}(\theta|x_i)$ and get the same $bel_{\Theta}(\theta|x_1, \ldots, x_n)$ by Dempster's rule of combination of these $bel_{\Theta}(\theta|x_i)$. In other words the following identity holds: $bel_{\Theta}(\cdot|x_1, \ldots, x_n) = bel_{\Theta}(\cdot|x_1) \oplus \ldots \oplus bel_{\Theta}(\cdot|x_n)$.

3 Computing the Posterior Belief Function from Likelihoods

Suppose only a finite number of frequentist likelihood functions $\{P(\cdot|\theta_i), i = 1, \ldots, k\}$, are available, ***and each one comes from a different population.*** The procedure described in the previous section then specializes as follows:

1. *Conditional embedding* of $P(\cdot|\theta_i)$ over $X \times \Theta$ into belief functions bel^i: the associated mass function is defined by $m^i(\bar{\theta}_i \cup \{x\}) = m^i(\{(\theta_i, x)\} \cup (\overline{\{\theta_i\}} \times X)) = P(\{x\}|\theta_i), x \in X$; bel^i on $X \times \Theta$ has a vacuous marginal on Θ and yields $P(\cdot|\theta_i)$ back when conditioned on θ_i.

2. *Conjunctive merging* of the bel^{θ}'s on $X \times \Theta$. This step comes down to assigning mass $\prod_{i=1,\dots,k} P(x_{j_i}|\theta_i)$ to the set $\bigcap_{i=1,\dots,k}\{(\theta_i,x_{j_i})\} \cup (\overline{\{\theta_i\}} \times X) = \bigcup_{i=1,\dots,k}\{(\theta_i,x_{j_i})\}$. Let ϕ be the mapping assigning observation x_{j_i} to each θ_i. We can write $m(\phi)$ for $m(\bigcup_{i=1,\dots,k}\{(\theta_i,x_{j_i})\})$.
3. *Conditioning* m *on the observation* x. Then $pl_{\Theta}(\theta|x) = \frac{\sum_{\phi:\phi(\theta)=x} m(\phi)}{\sum_{\theta\in\Theta}\sum_{\phi\in X^{\Theta}:\phi(\theta)=x} m(\phi)}$.

The simplest example of the problem is a simple space $\mathscr{S} = \{x,\bar{x}\} \times \{\theta,\bar{\theta}\}$ with two possible mutually exclusive hypotheses $\Theta = \{\theta,\bar{\theta}\}$, and two possible mutually exclusive observations $\{x,\bar{x}\}$. The available knowledge consists in the two likelihood values $a = P(x|\theta) > b = P(x|\bar{\theta})$. And it is assumed that x is observed.

For this example (actually studied by Shafer [10]), conditional embedding comes down to defining $m_1(x \cup \bar{\theta}) = a, m_1(\bar{x} \cup \bar{\theta}) = 1-a$, and likewise: $m_2(x \cup \theta) = b, m_2(\bar{x} \cup \theta) = 1-b$. Conjunctive merging yields $m(x) = ab; m(\bar{x}) = (1-a)(1-b)$; $m((x \cap \theta) \cup (\bar{x} \cap \bar{\theta})) = a(1-b); m((x \cap \bar{\theta}) \cup (\bar{x} \cap \theta)) = a(1-b)$.

The following results are obtained if x is observed:

$$bel_{\Theta}(\theta|x) = \frac{pl(x) - pl(x \cap \bar{\theta})}{pl(x)} = \frac{a(1-b)}{a+b-ab}; bel_{\Theta}(\bar{\theta}|x) = \frac{b(1-a)}{a+b-ab}. \quad (1)$$

It is natural that $bel_{\Theta}(\theta|x)$ should be all the higher as $P(x|\theta)$ is close to 1 and $P(x|\bar{\theta})$ is low. In particular

1. $bel_{\Theta}(\theta|x) = 1$ if and only if $P(x|\theta) = 1$ and $P(x|\bar{\theta}) = 0$;
2. $bel_{\Theta}(\theta|x) = 0 = bel_{\Theta}(\bar{\theta}|x)$ if and only if $P(x|\theta) = P(x|\bar{\theta}) = 0$ or $= 1$;
3. If $a = b$ then $0 \leq bel_{\Theta}(\theta|x) = bel_{\Theta}(\bar{\theta}|x) \leq 1/4$.

Shafer [10] extended this example to n observations of the form x or $\bar{x}$. He showed that for large values of $n, bel(\theta|x_1,\dots,x_n) + bel(\bar{\theta}|x_1,\dots,x_n) \approx 1$ and that the posterior beliefs agree at the limit with the Bayesian solution with uniform prior.

A different approach applies sensitivity analysis to Bayes rule, varying the unknown prior probability. This approach is popular in the robust Bayesian community where some prior information is supposed to be available in the form of a suitable family of probability functions (see Whitcomb [14] for a bibliography). The sensitivity analysis approach and the GBT presuppose different assumptions: In the former, no information on the dependence between the two items $a = P(x|\theta)$ and $b = P(x|\bar{\theta})$ is assumed; but in case of total ignorance on the prior, the resulting posterior is unknown and no information is gained from observing x. But the GBT assumes cognitive independence between two distinct populations or sources that provide each likelihood function. This is what makes the posterior belief function non-trivial. A number of other approaches to the *no prior* problem come down to selecting a "reasonable" probability measure on $\mathscr{S}$ in the set $\mathscr{P} = \{P, a = P(x|\theta) > b = P(x|\bar{\theta})\}$, induced by the likelihood values, for instance applying the maximum likelihood principle, i.e., maximizing $P(x)$ (which is not so good as it results in $P(\theta|x) = 1$). Several such approaches are reviewed by Dubois, Gilio and Kern-Isberner [7]: maximal entropy, Shapley value, uniform prior, etc.

Alternatively, one may keep the likelihood values upon observing x as $\lambda(\theta) = a, \lambda(\bar{\theta}) = b$ and view them as a measures of confidence, as strongly advocated by frequentist statisticians after Fisher and Edwards [8]; however, this approach may be considered as lacking formal foundations, all the more so as this school of thought never considers extending such uncertainty measures from elementary parameter values to disjunctions thereof.

4 Critiques of the GBT

There are several situations where the GBT is questionable, as discussed by Shafer [10]. Moreover, some authors like Walley [13] have criticized it as not satisfying the strong likelihood principle.

The binomial example. Consider the case of a coin such that $P(x|\theta) = \theta \in \Theta = [0,1]$ is the probability of getting a tail (x), to be learned from observations. We now have an uncountable infinite family of conditional belief functions such that $bel_X(x|\theta) = \theta, bel_X(\bar{x}|\theta) = 1-\theta$, $\theta \in [0,1]$. *The assumption that these belief functions have been obtained from distinct sets of data is no longer tenable, as this would imply an infinite quantity of information*! A way to circumvent this problem could be to discretize the domain Θ into $\Theta' = \{\theta_1, \ldots, \theta_k\}$, with $\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$. However, the k belief functions $bel_X(\cdot|\theta_i)$ for $i = 1, \ldots, k$ are now linked by the following relations: $bel_X(x|\theta_i) \leq bel_X(x|\theta_j)$ whenever $\theta_i \leq \theta_j$. Consequently, they cannot be independent. As noted by Shafer [10], "the choice of a belief function analysis depends on the nature of the evidence for the model, not just on the model itself".

The fiducial example. Shafer also considers the case of a measuring instrument with errors. Let $\theta \in \Theta$ be the unknown quantity and $x \in X$ be the measured value. It is supposed that $\Theta = X = \mathbb{N}$. Suppose we know the symmetric probability distribution P of errors $e = x - \theta$. This probability distribution can be viewed as a belief function on $X \times \Theta$, letting $m(\{(x,\theta) : e = |x-\theta|\}) = P(e)$. The projection of this belief function on X is $bel(x|\theta) = P(x-\theta)$, i.e., it is additive and coincides with $P(x|\theta)$. But the same holds for the projection of this belief function on Θ, since $P(x|\theta) = P(\theta|x) = P(x-\theta)$. If $\Theta = X = \{0,1\}$ and $x = \theta + e$ modulo 2, assuming $P(0) = a, P(1) = b = 1-a$, we find that $bel(\theta = 1|x = 1) = a$, which differs from the value obtained with the GBT if $b = 1-a$, that is $\frac{a^2}{1-a-a^2}$. Again in this case the two likelihood functions are related.

The strong likelihood principle. In the statistical literature, likelihood functions are considered to live on a ratio scale. Edwards [8] considers the likelihood function $\lambda(\theta)$ to be proportional to $P(x|\theta)$, the proportionality constant being arbitrary. In particular, no comparison of likelihood of hypotheses across data sets, say $\lambda_1(\theta) = P(x_1|\theta)$ and $\lambda_2(\theta) = P(x_2|\theta)$ is considered meaningful; only likelihood ratios $\frac{P(x|\theta_2)}{P(x|\theta_1)}$ make sense. Moreover, the likelihood principle states that all the information that is provided by the data x concerning the relative merits of two hypotheses θ_1 and θ_2 is contained in the likelihood ratio of these hypotheses. Hence the invariance property, recalled by Walley [13], here stated in terms of belief functions: Let f be the function such that $bel_\Theta(\cdot|x_1, \ldots x_n) = f(P(x_i|\theta), i = 1, \ldots n, \theta \in \Theta)$. Then, for all real values $c > 0$, $f(P(x_i|\theta), i = 1, \ldots n, \theta \in \Theta) = f(c \cdot P(x_i|\theta), i = 1, \ldots n, \theta \in \Theta)$. It is clear that the GBT violates this property, as well as some other inference techniques recalled in Section 3. However the Bayesian inference method does satisfy this strong likelihood principle. Walley essentially shows that, when the initial information takes the form of likelihood functions $P(x_i|\theta)$, enforcing the strong likelihood principle to the GBT leads to a probabilistic posterior belief function where the plausibility of each singleton $\theta \in \Theta$ is proportional to $P(x|\theta)^\alpha$ for some $\alpha > 0$. So it comes down to working with a Bayesian approach under uniform priors, up to a rescaling of the likelihood functions.

Is the strong likelihood principle a *sine qua non* condition for statistical inference? It can be questioned. First, there seems to be a clash of intuitions between this principle and the frequentist approach based on a fixed amount of observations N. Suppose $P(x|\theta)$ derives from the result of experiments that yield $N(x\theta) = n_1, N(\bar{x}\theta) = n_2, N(x\bar{\theta}) = n_3, N(\bar{x}\bar{\theta}) = n_4$ with $N = \sum_{i=1}^4 n_i$. Then $P(x|\theta) = a = \frac{n_1}{n_1+n_2}$ and $P(x|\bar{\theta}) = b = \frac{n_3}{n_3+n_4}$. Hence $\frac{n_1}{a} + \frac{n_3}{b} = N$, so that multiplying a and b by positive constant c clearly implies dividing N by c. In such a situation, claiming the invariance of the likelihood under positive scalar multiplication comes down to considering the statistical validity of the joint probability distribution on $X \times \Theta$ as not being affected by the number N of outcomes.

Another reason for questioning the strong likelihood principle is that if we extend the likelihood $\lambda(\theta) = cP(x|\theta)$ of elementary hypotheses, viewed as a representation of uncertainty about θ, to disjunctions of hypotheses, the corresponding set-function Λ should obey the laws of possibility measures [3, 6] in the absence of probabilistic prior, namely, the following properties look reasonable for such a set-function Λ:

- The properties of probability theory enforce $\forall T \subseteq \Theta, \Lambda(T) \leq \max_{\theta \in T} \lambda(\theta)$;
- A set-function representing likelihood should be monotonic with respect to inclusion: If $\theta \in T, \Lambda(T) \geq \lambda(\theta)$;
- Keeping the same scale as probability functions, we assume $\Lambda(\Theta) = 1$.

Then it is clear that $\lambda(\theta) = \frac{P(x|\theta)}{\max_{\theta \in \Theta} P(x|\theta)}$ and $\Lambda(T) = \max_{\theta \in T} \lambda(\theta)$, i.e., the extended likelihood function is a possibility measure, and the coefficient c is then fixed. We recover Shafer's proposal of a consonant belief function induced by likelihood information [9].

5 Statistical Inference in Possibility Theory

It is interesting to see if the same approach as the GBT can be carried out in the more restrictive setting of possibility theory, where only consonant belief functions are used. Suppose conditional possibility distributions $\{\pi(\cdot|\theta), \theta \in \Theta\}$ in the unit interval are available. The consonant conditional embedding consists in defining possibility distributions π_θ on $X \times \Theta$ as $\pi_\theta(x, \theta_i) = \pi(x|\theta)$ if $\theta_i = \theta$ and 1 otherwise. It is clear that the projection of π_θ on Θ is vacuous, i.e., $\max_{x \in X} \pi_\theta(x, \theta_i) = 1, \forall \theta_i \in \Theta$. Combining all these $\pi_\theta(\cdot,\cdot)$ conjunctively by means of any t-norm just yields the joint possibility distribution $\pi(x,\theta) = \pi(x|\theta)$. By conditioning on observation x, it yields $\pi_\Theta(\theta|x) = \frac{\pi(x|\theta)}{\max_{\theta' \in \Theta} \pi(x|\theta')}$.

In case of n observations x_i, we are faced again with two procedures to compute $\pi(\theta|x_1, \ldots, x_n)$: either combine the resulting conditional possibilities $\pi_\Theta(\theta|x_i)$, using an appropriate t-norm $\star$; or combine first the possibilistic likelihoods as $\pi(x_1, \ldots, x_n|\theta)$ and condition next. It is clear that these two procedures are not equivalent since $\star_{i=1\ldots n} \frac{\pi(x_i|\theta)}{\max_{\theta' \in \Theta} \pi(x_i|\theta')} \neq \frac{\star_{i=1\ldots n} \pi(x_i|\theta)}{\max_{\theta' \in \Theta} \star_{i=1\ldots n} \pi(x_i|\theta')}$. This difficulty is the cause of the rejection of this technique by Shafer himself [10]. In fact it is easy to see that a sufficient condition for these two approaches coinciding is that

$$\max_{\theta' \in \Theta} \pi(x|\theta') = 1, \forall x \in X.$$

This property, previously laid bare in [5], can be called the Hypothesis Exhaustivity Assumption (HEA). It means that the distribution $\pi(x|\theta)$ is a normalized possibility distribution on Θ as much as it is on X. This situation is similar to the one for probabilistic likelihood functions in the fiducial case. This is an assumption about Θ stating that for any piece of evidence $x \in X$, at least one hypothesis θ is not in conflict with x, i.e., $\forall x, \exists \theta, \pi(x|\theta) = 1$. It will hold if Θ is large enough to explain all observations. Aickin [1] seems to have rediscovered it and calls $\pi(x|\theta)$ *committed to the model.* An example where such an assumption is verified is the following: Suppose lower probability bounds $0 < a_{x\theta} \leq P(x|\theta)$ are available. They can be viewed as conditional necessity values $N(\{x\}|\theta) = a_{x\theta}, \theta \in \Theta$. Now, $N(\{x\}|\theta) = a_{x\theta} > 0$ is equivalent to $\pi(x|\theta) = 1$, $\pi(x'|\theta) = 1 - a_{x\theta}$ for $x' \neq x$. The HEA on Θ now means that for each $x \in X$ there is a constraint of the form $0 < a_{x\theta} \leq P(x|\theta)$ for some $\theta \in \Theta$, so that this observation is totally possible, under some assumption θ. Let $\Theta(x) = \{\theta \in \Theta, P(x|\theta) \geq a_{x\theta} > 0\}$ be the set of hypotheses that may tentatively explain x. The HEA says $\forall x \in X, \Theta(x) \neq \emptyset$. Note that $\Theta(x) = \{\theta \in \Theta, \pi(x|\theta) = 1\}$, so that $\forall x \in X, \max_{\theta' \in \Theta} \pi(x|\theta') = 1$ holds.

Let us now consider the properties of possibilistic inference in this case:

- If lower bounds on likelihoods are viewed as unrelated items of possibilistic information, we can combine possibility degrees via product in case of a sequence of observations $x_1, \ldots x_n$: $\pi(\theta|x_1, \ldots x_n) = \prod_{i=1,\ldots,n} \pi(x_i|\theta) = \prod_{i:\theta \notin \Theta(x_i)} (1 - a_{x_i\theta})$. It means that we can all the more certainly rule out assumption θ as there are more observations for which θ is not a plausible explanation.

- $N(\theta|x_1,\ldots x_n) = 1 - \max_{\theta' \neq \theta} \prod_{i=1,\ldots,n} \pi(x_i|\theta') > 0$ only if $\forall \theta' \neq \theta$, $\exists x_i$, $\pi(x_i|\theta') < 1$, that is: $\forall \theta' \neq \theta, \exists x_i : \theta' \notin \Theta(x_i)$. It means that:
  - We become more and more certain about θ as long as all hypotheses other than θ fail to plausibly explain one of the observations.
  - We have no longer any certainty at all about θ, if $\theta' \in \bigcup_{i=1,\ldots,n} \Theta(x_i)$, for some $\theta' \neq \theta$, i.e., some hypothesis other than θ can explain the whole set of observations.

In other words this form of statistical inference looks as reasonable as can be.

6 Conclusion

It is clearly interesting from both theoretical and practical points of view to reconsider the statistical inference methodology outside the Bayesian framework, beyond a mere sensitivity analysis method as done by robust statisticians, when only likelihood functions, or even only bounds on them are available and prior probabilities are not assigned. In particular, it is clear that the inference technique should depend on what kind of information is available and on the way it is acquired. One situation where likelihood functions can be exploited in a non-trivial way is when these likelihoods come from separate populations for each parameter value. More generally, some additional assumption is needed to complement the pure likelihood information. This paper has reviewed a number of techniques to that effect, whereby the notion of conditioning at work in learning schemes of probabilistic inference is extended to other theories of uncertainty. It seems that possibility theory may play a key role in the development of simple inference techniques under poor information, especially as an approximation of more complex methods, due to the close connections between likelihoods and possibility distributions. A more extensive account of the literature is needed so as to encompass alternative approaches based on imprecise probabilities such as the imprecise Dirichlet model [2]. It is useful to re-examine, in the light of the GBT and the possibilistic inference scheme, Bayesian objections against classical likelihood-based inference techniques, which have often been developed in an ad hoc way with no relations to new uncertainty theories.

References

1. Aickin, M.: Connecting Dempster-Shafer belief functions with likelihood-based inference. Synthese 123(3), 347–364 (2000)
2. Bernard, J.-M.: The imprecise Dirichlet model. Internat. J. Approx. Reason. 50(2), 201–203 (2009)
3. Coletti, G., Scozzafava, R.: Coherent conditional probability as a measure of uncertainty of the relevant conditioning events. In: Nielsen, T.D., Zhang, N.L. (eds.) ECSQARU 2003. LNCS (LNAI), vol. 2711, pp. 407–418. Springer, Heidelberg (2003)

4. Denœux, T., Smets, P.: Classification using belief functions: the relationship between the case-based and model-based approaches. IEEE Trans. Syst. Man Cybern. B 36(6), 1395–1406 (2006)
5. Dubois, D., Prade, H.: On the combination of evidence in various mathematical frameworks. In: Flamm, J., Luisi, T. (eds.) Reliability Data Collection and Analysis, pp. 213–241. Kluwer Acad. Publ., Dordrecht (1992)
6. Dubois, D.: Possibility theory and statistical reasoning. Comput. Statist. Data Anal. 51(1), 47–69 (2006)
7. Dubois, D., Gilio, A., Kern-Isberner, G.: Probabilistic abduction without priors. Internat. J. Approx. Reason. 47(3), 333–351 (2008)
8. Edwards, W.F.: Likelihood. Cambridge University Press, Cambridge (1972)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
10. Shafer, G.: Belief Functions and Parametric Models. J. Roy. Statist. Soc. Ser. B 44, 322–352 (1982)
11. Smets, P.: Un modèle mathématico-statistique simulant le processus du diagnostic médical. Université Libre de Bruxelles, Brussels, Belgium (1978)
12. Smets, P.: Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. Internat. J. Approx. Reason. 9(1), 1–35 (1993)
13. Walley, P.: Belief Function Representations of Statistical Evidence. Ann. Statist. 15, 1439–1465 (1987)
14. Whitcomb, K.: Quasi-Bayesian analysis using imprecise probability assessments and the generalized Bayes rule. Theory Decis. 58, 209–238 (2005)

Representation of Exchangeable Sequences by Means of Copulas

Fabrizio Durante and Jan-Frederik Mai

Abstract. Given a sequence $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ of exchangeable continuous random variables, it is proved that the joint distribution function of every finite subset of random variables belonging to $\mathbf{X}$ is fully described by means of a suitable bivariate copula and a univariate distribution function.

Keywords: Copula, Exchangeability.

1 Introduction

Given a family $\mathbf{X} = \{X_i\}_{i\in\mathscr{J}}$ of real-valued random variables (=r.v.'s) defined on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, it is well known that several properties of $\mathbf{X}$ can be expressed in terms of the class $\mathscr{H}$ that contains the joint distribution functions (=d.f.'s) of all the finite subfamilies of $\mathbf{X}$, $\mathscr{H} = \{H_A\}$, where A is a finite set of indices in $\mathscr{J}$, $A = \{i_1, i_2, \dots, i_k\}$, and $H_A : \mathbb{R}^{|A|} \to \mathbb{R}$ is the d.f. of $(X_{i_1}, X_{i_2}, \dots, X_{i_k})$, $H_A(x_1, \dots, x_k) = \mathbb{P}(X_{i_1} \le x_1, \dots, X_{i_k} \le x_k)$.

Moreover, since *Sklar's Theorem* [19], it is known that the d.f. H of every continuous random vector $(X_1, \dots, X_n)$ can be uniquely represented in terms of the univariate marginal d.f.'s F_i, $i \in \{1, \dots, n\}$, and a suitable $C_n \colon \mathbb{I}^n \to \mathbb{I}$ ($\mathbb{I} := [0,1]$), called *copula*, in the following way:

$$H(x_1, x_2, \dots, x_n) = C_n(F_1(x_1), F_2(x_2), \dots, F_n(x_n)), \tag{1}$$

Fabrizio Durante
School of Economics and Management, Free University of Bozen-Bolzano,
Bolzano, Italy
e-mail: fabrizio.durante@unibz.it

Jan-Frederik Mai
HVB–Institute for Mathematical Finance, Technische Universität München,
Garching, Germany
e-mail: mai@tum.de

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 227–232.
springerlink.com

for every $x_1, x_2, \dots, x_n$ in $\mathbb{R}$. We recall that a copula is an n–dimensional d.f. having univariate marginals uniformly distributed on $\mathbb{I}$. Basic examples of copulas are: the independence copula $\Pi_n(\mathbf{x}) = x_1 x_2 \cdots x_n$, and the comonotonicity copula $M_n(\mathbf{x}) = \min\{x_1, x_2, \dots, x_n\}$. See, for example, [9, 10, 15].

Therefore, every family of continuous r.v.'s $\mathbf{X} = \{X_i\}_{i \in \mathscr{J}}$ can be uniquely expressed in terms of the couple $(\mathscr{F}_{\mathbf{X}}, \mathscr{C}_{\mathbf{X}})$, where $\mathscr{F}_{\mathbf{X}} = \{F_i\}_{i \in \mathscr{J}}$ is the family formed by the (univariate) d.f.'s associated with each X_i and $\mathscr{C}_{\mathbf{X}}$ contains the copulas that are associated with all finite subsets of $\{X_i\}_{i \in \mathscr{J}}$, in such a way that, if H is the joint d.f. of $(X_{i_1}, X_{i_2}, \dots, X_{i_n})$, then H can be expressed in the form (1), where C_n is in $\mathscr{C}_{\mathbf{X}}$ and $F_{i_1}, F_{i_2}, \dots, F_{i_n}$ are in $\mathscr{F}_{\mathbf{X}}$. This representation was adopted, for example, in [5] in order to describe a Markov process (see also [12]).

In this short note, we aim at giving a representation of the same type for an *exchangeable sequence* of continuous r.v.'s, i.e. for a sequence $\mathbf{X} = (X_n)_{n \in \mathbb{N}}$ of r.v.'s such that the d.f. of every finite subset of k $(k \geq 1)$ of these r.v.'s depends only upon k and not on the particular subset (see [11] for more details).

2 The Representation

Given a random vector $(X_1, \dots, X_n)$ with joint d.f. H and continuous univariate marginals $F_1, \dots, F_n$, its associated copula C_n is actually the d.f. of the random vector $(F_1(X_1), \dots, F_n(X_n))$, and, hence, C_n can be recovered from the d.f. H of $(X_1, \dots, X_n)$ by taking, for all $\mathbf{u} \in \mathbb{I}^n$,

$$C_n(u_1, \dots, u_n) = H(F_1^{\leftarrow}(u_1), \dots, F_n^{\leftarrow}(u_n)),$$

where $F^{\leftarrow}(y) = \inf\{x \in \mathbb{R} \mid F(x) \geq y\}$ denotes the quantile inverse of any univariate d.f. F.

Every copula C_n is a Lipschitz function (with constant 1) and admits partial derivatives $\frac{\partial C_n}{\partial u_i} = \partial_i C_n$ almost everywhere on $\mathbb{I}^n$. If C_n is the copula of the continuous random vector $(X_1, \dots, X_n)$, then, similarly to [5], it can be proved that

$$\partial_j C_n(F(x_1), \dots, F(x_{j-1}), F_j(X_j), F(x_{j+1}), \dots, F(x_n))$$

is a version of $\mathbb{P}\left(\cap_{i \neq j}\{X_i \leq x_i\} \mid X_j\right) := \mathbb{E}\left(\mathbf{1}_{\{X_i \leq x_i, i \neq j\}} \mid X_j\right)$.

Following [16], $\mathbf{X} = (X_n)_{n \in \mathbb{N}}$ is an exchangeable sequence of real-valued r.v.'s if, and only if, there exists a real-valued r.v. Λ such that $X_1, X_2, \dots$ are conditionally independent and identically distributed (briefly, i.i.d.) given Λ.

Now, starting with this last fact, we state our main result.

Theorem 1. *Let $(X_n)_{n \in \mathbb{N}}$ be an exchangeable sequence of continuous r.v.'s. Then there exist a one–dimensional d.f. F and a 2–copula A such that, the joint d.f. H_n of every subset of $n \geq 2$ r.v.'s from the sequence may be represented, for all $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, as*

$$H_n(x_1,x_2,\dots,x_n) = C_n(F(x_1),F(x_2),\dots,F(x_n)), \tag{2}$$

where the copula C_n is given, for all $(u_1,\dots,u_n)\in\mathbb{I}^n$, by

$$C_n(u_1,\dots,u_n) = \int_0^1 \frac{\partial A(u_1,t)}{\partial t}\cdot\frac{\partial A(u_2,t)}{\partial t}\cdots\frac{\partial A(u_n,t)}{\partial t}\,dt. \tag{3}$$

Proof. Given the exchangeable sequence $(X_n)_{n\in\mathbb{N}}$, as said before, there exists a r.v. Λ with d.f. L, such that the r.v.'s X_n are conditionally i.i.d. given Λ (see, e.g., [16]). Therefore, there is a family $(G_\lambda)_{\lambda\in\mathbb{R}}$ of d.f.'s such that, for all $n\in\mathbb{N}$ and for all $(x_1,\dots,x_n)\in\mathbb{R}^n$,

$$\mathbb{P}(X_1\le x_1,\dots,X_n\le x_n\,|\,\Lambda) = G_\Lambda(x_1)\cdots G_\Lambda(x_n).$$

Without loss of generality (since we are only interested in statements in distribution) we may assume that the r.v.'s $(X_n)_{n\in\mathbb{N}}$ and Λ are defined on a probability space $(\Omega,\mathscr{A},\mathbb{P})$ in the following canonical manner, see [2, p. 12-13].

- Let U and $V_1,V_2,\dots,$ be i.i.d. r.v.'s on $(\Omega,\mathscr{A},\mathbb{P})$ that are uniformly distributed on $\mathbb{I}$.
- Define $\Lambda := L^{\leftarrow}(U)$, where $L^{\leftarrow}$ denotes the generalized inverse of L, i.e. Λ has distribution function L (see, for instance, [3, Theorem 2]).
- For each $n\in\mathbb{N}$ define the r.v. X_n as a function of Λ and V_n via $X_n := G_\Lambda^{\leftarrow}(V_n)$; i.e. conditioned on Λ, X_n has d.f. G_Λ, or, conditioned on U, X_n has d.f. $G_{L^{\leftarrow}(U)}$.

For each n, the copula C_n of $(X_1,\dots,X_n)$ coincides with the joint distribution function of $\big(F(X_1),\dots,F(X_n)\big)$ since F is continuous. Hence, using the canonical construction above as well as continuity of F (which implies that $F^{\leftarrow}$ is strictly increasing and $F^{\leftarrow}\circ F(X_n)$ is equal in distribution to X_n by [14, p. 495, Proposition A.3-4]), it holds that

$$\mathbb{P}\big(F(X_1)\le u_1,F(X_2)\le u_2,\dots,F(X_n)\le u_n\mid U\big) = \prod_{i=1}^n G_{L^{\leftarrow}(U)}\big(F^{\leftarrow}(u_i)\big).$$

Now let A be the joint distribution function of $\big(F(X_n),U\big)$. Notice that such A does not depend on n by conditional independence; in fact:

$$\begin{aligned}\mathbb{P}\big(F(X_n)\le x,U\le u\big) &= \mathbb{E}[\mathbb{P}\big(F(X_n)\le x,U\le u\mid U\big)]\\ &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}_{\big\{F\big(G^{\leftarrow}_{L^{\leftarrow}(U)}(V_n)\big)\le x\big\}}\,\Big|\,U\Big]\,\mathbf{1}_{\{U\le u\}}\Big]\\ &\overset{(*)}{=} \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}_{\big\{F\big(G^{\leftarrow}_{L^{\leftarrow}(z)}(V_n)\big)\le x\big\}}\Big]\Big|_{z=U}\mathbf{1}_{\{U\le u\}}\Big]\end{aligned}$$

$$\begin{aligned}&= \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}_{\big\{F\big(G^{\leftarrow}_{L^{\leftarrow}(z)}(V_1)\big)\le x\big\}}\Big]\Big|_{z=U}\mathbf{1}_{\{U\le u\}}\Big]\\ &= \mathbb{P}\big(F(X_1)\le x,U\le u\big),\end{aligned}$$

where equality $(*)$ follows from [7, Example 1.5, page 224]. Then, for almost all $u_i \in \mathbb{I}$, one has

$$G_{L^{\leftarrow}(U)}\big(F^{\leftarrow}(u_i)\big) = \mathbb{P}\big(F(X_i) \le u_i \mid U\big) = \frac{\partial}{\partial t} A(u_i,t)\bigg|_{t=U}.$$

Thus, the copula C_n of any sequence of n r.v.'s in $(X_n)_{n\in\mathbb{N}}$ can be represented via (3). □

The copula C_n given by (3) is called the *n–product* of A. In [5], the authors considered an operation $*$ on the class of 2–copulas given, for any 2–copulas A and B, by

$$(A*B)(u_1,u_2) = \int_0^1 \frac{\partial A(u_1,t)}{\partial t} \cdot \frac{\partial B(t,u_2)}{\partial t}\,\mathrm{d}t.$$

It is easy to show that the 2–product of A coincides with the copula given by $A*A^T$, where $A^T(u_1,u_2) = A(u_2,u_1)$ for every $(u_1,u_2) \in \mathbb{I}^2$.

Theorem 1 can be used in order to construct a sequence of exchangeable r.v.'s by using only a univariate d.f. and a 2–copula. The procedure runs as follows:

1. assign in any manner a 2–copula A and a d.f. F;
2. for $n > 1$, set C_n the n–product of A given by (3);
3. set $H_1 = F$ and $H_n = C_n(F,F,\dots,F)$ for every $n \ge 2$;
4. apply Daniell-Kolmogorov Theorem [17] to $\mathscr{H} = \{H_n\}_{n\in\mathbb{N}}$ in order to obtain a sequence of exchangeable r.v.'s $\mathbf{X} = (X_n)_{n\in N}$ such that every joint d.f. of any finite subset of $\mathbf{X}$ is in $\mathscr{H}$.

A sequence of i.i.d. r.v.'s can be constructed, for example, by taking any univariate d.f. F and $A = \Pi_2$. Notice that different 2–copulas can produce the same sequence. For instance, the mapping $W_2\colon \mathbb{I}^2 \to \mathbb{I}$ given by $W_2(u_1,u_2) = \max\{u_1+u_2-1,0\}$ is a 2–copula such that $W_2 * W_2 = M_2 * M_2 = M_2$, therefore, the sequences generated by M_2 and by W_2 can be associated with the same family of finite-dimensional d.f.'s.

Some interesting consequences can be derived from Theorem 1.

We recall that a random vector $(X_1,\dots,X_n)$ is *infinitely extendible* if it is the first segment of a sequence of exchangeable r.v.'s (see [18, 20]). The following corollary gives a representation for any copula that is associated with an infinitely extendible random vector.

Corollary 1. *Let $(X_1,\dots,X_n)$ be an exchangeable random vector with (symmetric) copula C. Then $(X_1,\dots,X_n)$ is infinitely extendible if, and only if, C is the n–product of some copula A.*

If the 2–copula C_2 of (X_1,X_2) is symmetric and *idempotent* with respect to the operation $*$, i.e. $C_2 * C_2 = C_2$ (see [1, 5]), then (X_1,X_2) is infinitely extendible. Generally, if a family of symmetric bivariate copulas (like Fréchet family and FGM family) is closed with respect to the $*$-operation (and, hence, with respect to the 2-product operation), then its members can be used for constructing an infinitely extendible random pair.

Example 1. Corollary 1 is illustrated with bivariate Cuadras-Augé copulas, defined, for every $\alpha \in \mathbb{I}$, by $C_\alpha(u,v) = \min\{u,v\} \max\{u,v\}^{1-\alpha}$ (see [4]). Let (X_1,X_2) be a random vector with identical continuous marginals and 2–copula C_α. Given the bivariate Marshall-Olkin copula $A(u,v) = \min\{u,u^\alpha v\}$, it follows easily that

$$C_\alpha(u,v) = \int_0^1 \frac{\partial}{\partial t} A(u,t) \frac{\partial}{\partial t} A(v,t)\,dt.$$

Hence (X_1,X_2) is infinitely extendible. Moreover, for $n \geq 2$ it holds that

$$\int_0^1 \prod_{i=1}^{n} \frac{\partial}{\partial t} A(u_i,t)\,dt = u_{[1]} \prod_{i=2}^{n} u_{[i]}^{1-\alpha}, \tag{4}$$

where $u_{[1]} \leq u_{[2]} \leq \ldots \leq u_{[n]}$ denotes the components of $(u_1,\ldots,u_n) \in \mathbb{I}^n$, rearranged in increasing order. Copulas of type (4) have been considered in [6, 13].

The following result, instead, was proved in [2, Lemma 4.11] (see also [8]) and admits now another proof.

Corollary 2. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of exchangeable r.v.'s. If X_i and X_j are independent for every $i,j \in \mathbb{N}$, $i \neq j$, then $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. r.v.'s.*

Proof. In view of Theorem 1, given a sequence $(X_n)_{n\in\mathbb{N}}$ of exchangeable r.v.'s, there exists a 2–copula A such that the 2–product of A, denoted by B, is the copula of the random pair (X_i,X_j) for every $i,j \in \mathbb{N}$, $i \neq j$. In particular, if X_i and X_j are independent, then $B = A * A^T = \Pi_2$. But, in general, $A * \Pi_2 = \Pi_2$, which yields that $(A * A^T) - (A * \Pi_2) = 0$. Thus, for every $(u_1,u_2) \in \mathbb{I}^2$, $\partial_t A(t,u_2) = u_2$ for almost all $t \in \mathbb{I}$, viz. A has linear section in the first component being the second fixed. Thus, $A = \Pi_2$ and the copula of every subset of n r.v.'s of the sequence is Π_n. □

Note that, as well known, for a finite vector of exchangeable r.v.'s, pairwise independence does not imply independence. Consider, for example, the trivariate vector (U_1,U_2,U_3) whose d.f. H is given on $\mathbb{I}^3$ by:

$$H(u_1,u_2,u_3) = u_1u_2u_3(1+\theta(1-u_1)(1-u_2)(1-u_3)),$$

for a suitable $\theta \in [-1,1]$ (see, e.g., [15, Example 3.31]).

Acknowledgements. The first author acknowledges the support of School of Economics and Management (Free University of Bozen-Bolzano) via the project "Multivariate dependence models".

References

1. Albanese, A., Sempi, C.: Countably generated idempotent copulas. In: Soft methodology and random information systems, Adv. Soft Comput., pp. 197–204. Springer, Berlin (2004)
2. Aldous, D.J.: Exchangeability and related topics. In: École d'été de probabilités de Saint-Flour, XIII—1983. Lecture Notes in Math., vol. 1117, pp. 1–198. Springer, Berlin (1985)
3. Angus, J.E.: The probability integral transform and related results. SIAM Rev. 36(4), 652–654 (1994)
4. Cuadras, C.M., Augé, J.: A continuous general multivariate distribution and its properties. Comm. Statist. A—Theory Methods 10(4), 339–353 (1981)
5. Darsow, W.F., Nguyen, B., Olsen, E.T.: Copulas and Markov processes. Illinois J. Math. 36(4), 600–642 (1992)
6. Durante, F., Quesada-Molina, J.J., Úbeda-Flores, M.: On a family of multivariate copulas for aggregation processes. Inform. Sci. 177(24), 5715–5724 (2007)
7. Durrett, R.: Probability: theory and examples, 2nd edn. Duxbury Press, Belmont (1996)
8. Hu, T.C.: On pairwise independent and independent exchangeable random variables. Stochastic Anal. Appl. 15(1), 51–57 (1997)
9. Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds.): Copula Theory and its Applications. Lecture Notes in Statistics - Proceedings. Springer, Dortrecht (2010)
10. Joe, H.: Multivariate models and dependence concepts. Monographs on Statistics and Applied Probability, vol. 73. Chapman & Hall, London (1997)
11. Kallenberg, O.: Probabilistic symmetries and invariance principles. In: Probability and its Applications. Springer, New York (2005)
12. Lagerås, A.N.: Copulas for Markovian dependence. Bernoulli (in press, 2010)
13. Mai, J.F., Scherer, M.: Lévy-Frailty copulas. J. Multivariate Anal. 100(7), 1567–1585 (2009)
14. McNeil, A.J., Frey, R., Embrechts, P.: Quantitative risk management. Concepts, techniques and tools. Princeton Series in Finance. Princeton University Press, Princeton (2005)
15. Nelsen, R.B.: An introduction to copulas, 2nd edn. Springer Series in Statistics. Springer, New York (2006)
16. Olshen, R.: A note on exchangable sequences. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 28, 317–321 (1973/1974)
17. Rogers, L.C.G., Williams, D.: Diffusions, Markov processes, and martingales. Cambridge Mathematical Library, vol. 1. Cambridge University Press, Cambridge (2000), (Foundations, Reprint of the second (1994) edition)
18. Scarsini, M., Verdicchio, L.: On the extendibility of partially exchangeable random vectors. Statist. Probab. Lett. 16(1), 43–46 (1993)
19. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)
20. Spizzichino, F.: Extendibility of symmetric probability distributions and related bounds. In: Exchangeability in probability and statistics, Rome, pp. 313–320. North-Holland, Amsterdam (1981/1982)

Area-Level Time Models for Small Area Estimation of Poverty Indicators

M.D. Esteban, D. Morales, A. Pérez, and L. Santamaría

Abstract. Small area parameters usually take the form $h(y)$, where y is the vector containing the values of all units in the domain and h is a linear or nonlinear function. If h is not linear or the target variable is not normally distributed, then the unit-level approach has no standard procedure and each case should be treated with a specific methodology. Area-level linear mixed models can be generally applied to produce new estimates of linear and non linear parameters because direct estimates are weighted sums, so that the assumption of normality may be acceptable. In this communication we treat the problem of estimating small area non linear parameters, with special emphasis on the estimation of poverty indicators. For this sake, we borrow strength from time by using area-level linear time models. We consider two time-dependent area-level models, empirically investigate their behavior and apply them to estimate poverty indicators in the Spanish Living Conditions Survey.

1 Area-Level Linear Time Model

In small area estimation samples are drawn from a finite population, but estimations are required for subsets (called small areas or domains) where the effective sample sizes are too small to produce reliable direct estimates. An estimator of a small area parameter is called direct if it is calculated just with the sample data coming from the corresponding small area. Thus, the lack of sample data from the target small area affects seriously the accuracy of the direct estimators, and this fact has given rise to the development of new tools for obtaining more precise estimates. See a description of this theory in the monograph of Rao ([4]).

Area-level models relate direct estimates of small area means to area-level auxiliary variables. The idea is to borrow strength from other domains, related

M.D. Esteban, D. Morales, A. Pérez, and L. Santamaría
Universidad Miguel Hernández de Elche, Spain
e-mail: d.morales@umh.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 233–237.
springerlink.com

variables, past time instants and correlations, in order to produce new model-based estimates. In this work we consider the model

$$y_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}, \quad d = 1,\dots,D, \quad t = 1,\dots,m_d, \tag{1}$$

where y_{dt} is a direct estimator of the indicator of interest for area d and time instant t, $\mathbf{x}_{dt}$ is a vector containing the aggregated (population) values of p auxiliary variables, the random vectors $(u_{d1},\dots,u_{dm_d})$, $d = 1,\dots,D$, are i.i.d. AR(1), with variance and auto-correlation parameters σ_u^2 and ρ respectively, the errors e_{dt}'s are independent $N(0,\sigma_{dt}^2)$ with known σ_{dt}^2's, and the u_{dt}'s and the e_{dt}'s are independent. In the applications to real data we may also consider a simpler model obtained by restricting model (1) to $\rho = 0$. Model (1) is related to the model of Rao and Yu [3] in the sense that u_d is substituted by u_{dt} to take into account the area-by-time variability through specific random effects.

In matrix notation, model (1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2}$$

where

$$\mathbf{y} = \underset{1\le d\le D}{\mathrm{col}}(\mathbf{y}_d), \quad \mathbf{y}_d = \underset{1\le t\le m_d}{\mathrm{col}}(y_{dt}), \quad \mathbf{u} = \underset{1\le d\le D}{\mathrm{col}}(\mathbf{u}_d), \quad \mathbf{u}_d = \underset{1\le t\le m_d}{\mathrm{col}}(u_{dt}),$$

$$\mathbf{e} = \underset{1\le d\le D}{\mathrm{col}}(\mathbf{e}_d), \quad \mathbf{e}_d = \underset{1\le t\le m_d}{\mathrm{col}}(e_{dt}), \quad \mathbf{X} = \underset{1\le d\le D}{\mathrm{col}}(\mathbf{X}_d), \quad \mathbf{X}_d = \underset{1\le t\le m_d}{\mathrm{col}}(\mathbf{x}_{dt}),$$

$$\mathbf{x}_{dt} = \underset{1\le k\le p}{\mathrm{col}'}(x_{dtk}), \quad \boldsymbol{\beta} = \underset{1\le k\le p}{\mathrm{col}}(\beta_k), \quad \mathbf{Z} = \mathbf{I}_{M\times M}, \quad M = \sum_{d=1}^{D} m_d.$$

We assume that $\mathbf{u} \sim N(\mathbf{0},\mathbf{V}_u)$ and $\mathbf{e} \sim N(\mathbf{0},\mathbf{V}_e)$ are independent with covariance matrices

$$\mathbf{V}_u = \sigma_u^2\Omega(\rho), \quad \Omega(\rho) = \underset{1\le d\le D}{\mathrm{diag}}(\Omega_d(\rho)), \quad \mathbf{V}_e = \underset{1\le d\le D}{\mathrm{diag}}(\mathbf{V}_{ed}), \quad \mathbf{V}_{ed} = \underset{1\le t\le m_d}{\mathrm{diag}}(\sigma_{dt}^2),$$

where the variances σ_{dt}^2 are known and

$$\Omega_d = \Omega_d(\rho) = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \dots & \rho^{m_d-2} & \rho^{m_d-1} \\ \rho & 1 & \ddots & & \rho^{m_d-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{m_d-2} & & \ddots & 1 & \rho \\ \rho^{m_d-1} & \rho^{m_d-2} & \dots & \rho & 1 \end{pmatrix}_{m_d\times m_d}.$$

The BLU estimators and predictors of $\boldsymbol{\beta}$ and $\mathbf{u}$ are

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \text{and} \qquad \widehat{\mathbf{u}} = \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}),$$

where $\operatorname{var}(\mathbf{y}) = \mathbf{V} = \sigma_u^2 \underset{1\le d\le D}{\operatorname{diag}}(\Omega_d(\rho)) + \mathbf{V}_e = \underset{1\le d\le D}{\operatorname{diag}}(\sigma_u^2\Omega_d(\rho) + \mathbf{V}_{ed}) = \underset{1\le d\le D}{\operatorname{diag}}(\mathbf{V}_d)$. The model is fitted by using the residual maximum likelihood method and $\mu_{dt} = \mathbf{x}_{dt}\boldsymbol{\beta} + u_{dt}$ is predicted with the empirical bet linear unbiased predictor (EBLUP) $\widehat{\mu}_{dt} = \mathbf{x}_{dt}\widehat{\boldsymbol{\beta}} + \widehat{u}_{dt}$. If we do not take into account the error, e_{dt}, this is equivalent to predict $y_{dt} = \mathbf{a}'\mathbf{y}$, where $\mathbf{a} = \underset{1\le \ell\le D}{\operatorname{col}}(\delta_{d\ell}\mathbf{a}_\ell)$ and $\mathbf{a}_\ell = \underset{1\le k\le m_\ell}{\operatorname{col}}(\delta_{tk})$. The population mean $\overline{Y}_{dt}$ is estimated by means of $\widehat{\overline{Y}}_{dt}^{eblup} = \widehat{\mu}_{dt}$. Following Prasad and Rao [2], see also Rao [4] or Jiang and Lahiri [1], the mean squared error (MSE) of $\widehat{\overline{Y}}_{dt}^{eblup}$ takes the form

$$MSE(\widehat{\overline{Y}}_{dt}^{eblup}) = g_1(\theta) + g_2(\theta) + g_3(\theta),$$

where $\theta = (\sigma_u^2, \rho)$,

$$\begin{aligned}
g_1(\theta) &= \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a},\\
g_2(\theta) &= [\mathbf{a}'\mathbf{X} - \mathbf{a}'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{V}_e^{-1}\mathbf{X}]\mathbf{Q}[\mathbf{X}'\mathbf{a} - \mathbf{X}'\mathbf{V}_e^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{a}],\\
g_3(\theta) &\approx \operatorname{tr}\left\{(\nabla\mathbf{b}')\mathbf{V}(\nabla\mathbf{b}')'E\left[(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)'\right]\right\}
\end{aligned}$$

and $\mathbf{Q} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$, $\mathbf{T} = \mathbf{V}_u - \mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{V}_u$, $\mathbf{b}' = \mathbf{a}'\mathbf{Z}\mathbf{V}_u\mathbf{Z}'\mathbf{V}^{-1}$. The estimator of $MSE(\widehat{\overline{Y}}_{dt}^{eblup})$ is

$$mse(\widehat{\overline{Y}}_{dt}^{eblup}) = g_1(\hat{\theta}) + g_2(\hat{\theta}) + 2g_3(\hat{\theta}). \tag{3}$$

2 Estimation of Poverty Indicators

Let us consider a finite population P_t partitioned into D domains P_{dt} at time period t, and denote their sizes by N_t and N_{dt}, $d = 1,\dots,D$. Let z_{dtj} be an income variable measured in all the units of the population and let z_t be the poverty line, so that units with $z_{dtj} < z_t$ are considered as poor at time period t. The main goal of this section is to estimate the poverty incidence (proportion of individuals under poverty) and the poverty gap in Spanish domains. These two measures belongs to the family

$$Y_{\alpha;dt} = \frac{1}{N_{dt}}\sum_{j=1}^{N_{dt}} y_{\alpha;dtj}, \quad \text{where } y_{\alpha;dtj} = \left(\frac{z_t - z_{dtj}}{z_t}\right)^{\alpha} I(z_{dtj} < z_t), \tag{4}$$

$I(z_{dtj} < z_t) = 1$ if $z_{dtj} < z_t$ and $I(z_{dtj} < z_t) = 0$ otherwise. The proportion of units under poverty in the domain d and period t is thus $Y_{0;dt}$ and the poverty gap is $Y_{1;dt}$.

We use data from the Spanish Living Conditions Survey (SLCS) corresponding to years 2004-2006 with sample sizes 44648, 37491, 34694 respectively. The SLCS is the Spanish version of the "European Statistics on Income and Living Conditions" (EU-SILC), which is one of the statistical operations that have been harmonized for EU countries. We consider $D=104$ domains obtained by crossing 52 provinces with 2 sexes. The SLCS does not produce official estimates at the domain level (provinces × sex), but the analogous direct estimator of the total $Y_{dt}=\sum_{j=1}^{N_{dt}} y_{dtj}$ is

$$\hat{Y}_{dt}^{dir}=\sum_{j\in S_{dt}} w_{dtj}\, y_{dtj}.$$

where S_{dt} is the domain sample at time period t and the w_{dtj}'s are the official calibrated sampling weights which take into account for non response. In the particular case $y_{dtj}=1$, for all $j\in P_{dt}$, we get the estimated domain size

$$\hat{N}_{dt}^{dir}=\sum_{j\in S_{dt}} w_{dtj}.$$

Using this quantity, a direct estimator of the domain mean $\bar{Y}_{dt}$ is $\bar{y}_{dt}=\hat{Y}_{dt}^{dir}/\hat{N}_{dt}^{dir}$. The direct estimates of the domain means are used as responses in the area-level time model. The design-based variances of these estimators can be approximated by

$$\hat{V}_{\pi}(\hat{Y}_{dt}^{dir})=\sum_{j\in S_{dt}} w_{dtj}(w_{dtj}-1)\left(y_{dtj}-\bar{y}_{dt}\right)^{2} \quad \text{and} \quad \hat{V}_{\pi}(\bar{y}_{dt})=\hat{V}\left(\hat{Y}_{dt}^{dir}\right)/\hat{N}_{dt}^{2}.$$

As we are interested in the cases $y_{dtj}=y_{\alpha;dtj}$, $\alpha=0,1$, we select the direct estimates of the poverty incidence and poverty gap at domain d and time period t (i.e. $\bar{y}_{0;dt}$ and $\bar{y}_{1;dt}$ respectively) as target variables for the time dependent area-level models. The considered auxiliary variables are the known domain means of the category indicators of the following variables:

- INTERCEPT: First auxiliary variable is equal to one.
- AGE: Age groups for the intervals ≤ 15, $16-24$, $25-49$, $50-64$ and ≥ 65.
- EDUCATION: Highest level of education completed, with 4 categories for Less than primary education level, Primary education level, Secondary education level and University level.
- CITIZENSHIP: with 2 categories for Spanish and Not Spanish.
- LABOR: Labor situation with 4 categories for Below 16 years, Employed, Unemployed and Inactive.

The Poverty Threshold is fixed as the 60% of the median of the normalized incomes in Spanish households. The total number of normalized household members is

$$H_{dtj}=1+0.5(H_{dtj\geq 14}-1)+0.3H_{dtj<14}$$

where $H_{dtj\geq 14}$ is the number of people aged 14 and over and $H_{dtj<14}$ is the number of children aged under 14. The normalized net annual income of a household is obtained by dividing its net annual income by its normalized size. The Spanish poverty thresholds (in euros) in 2004-06 are $z_{2004} = 6098.57$, $z_{2005} = 6160.00$ and $z_{2006} = 6556.60$ respectively. These are the z_t-values used in the calculation of the direct estimates of the poverty incidence and gap.

We consider the linear model $\overline{y}_{dt} = \bar{\mathbf{X}}_{dt}\boldsymbol{\beta} + u_{dt} + e_{dt}$, $d = 1,\dots,D$, where $\overline{y}_{dt} = \hat{Y}_{dt}^{dir}/\hat{N}_{dt}^{dir}$, $\sigma_{dt}^2 = \hat{V}_\pi(\overline{y}_{dt})$ and $\bar{\mathbf{X}}_d$ is the $1\times p$ vector containing the population (aggregated) mean values of all the categories (except the last one) of the explanatory variables. Random effects errors are assumed to follow the distributional assumptions of model (1). Obtained EBLUP estimates of % poverty proportions $p_d = 100\cdot\hat{Y}_{0;d,2006}^{eblup1}$ and poverty gaps $g_d = 100\cdot\hat{Y}_{1;d,2006}^{eblup1}$ are presented in the Figure 1.

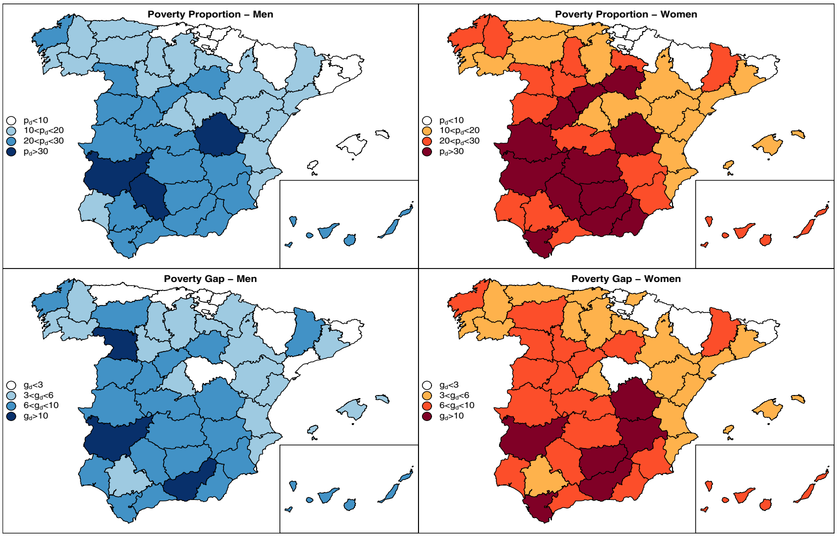


Fig. 1 Estimates of Spanish poverty proportions (top) and gaps (bottom) for men (left) and women (right) in 2006.

References

1. Jiang, J., Lahiri, P.: Mixed model prediction and small area estimation. Test 15, 1–96 (2006)
2. Prasad, N.G.N., Rao, J.N.K.: The estimation of the mean squared error of small-area estimators. J. Amer. Statist. Assoc. 85, 163–171 (1990)
3. Rao, J.N.K., Yu, M.: Small area estimation by combining time series and cross sectional data. Canad. J. Statist. 22, 511–528 (1994)
4. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)

Flood Analysis: On the Automation of the Geomorphological-Historical Method

Elena Fernández, Miguel Fernández, Soledad Anadón,
Gil González-Rodríguez, and Ana Colubi

Abstract. Different methods to assess the flood return period are available in the literature. The hydrological-hydraulic approaches, among the best-known quantitative methods, oversimplify the complex characteristics of the fluvial systems. Additionally, they rely on data that are usually criticized because of their low quality and representativity. In contrast, the semi-quantitative approach based on geomorphological and historical information has lead to more realistic and promising results in pilot studies. This approach is based on highly informative field data providing valuable knowledge which can be used to test the aforementioned quantitative approaches. The aim of this work is to analyze the kind of information that is required to apply the latter method and to explore the possibilities of its automation.

Keywords: Flood frequency, Geomorphological-historical information, Imprecise data, Supervised classification.

1 Introduction

Floods are one of the most common hazards in Europe. They are causing nowadays large losses. To reduce such losses, it is essential to improve the assessment of the return period of these events. The approach that has traditionally been employed to estimate the return period is based on hydrological-hydraulic models. Nevertheless, this approach is being more and more criticized due to the unrealistic assumptions that it requires and the poor

Elena Fernández, Miguel Fernández, Soledad Anadón, and Ana Colubi
INDUROT, University of Oviedo, 33600 Mieres, Spain
e-mail: elena,soledad,miguelfa,colubi@indurot.uniovi.es

Gil González-Rodríguez
European Center for Soft Computing, 33600 Mieres, Spain
e-mail: gil.gonzalez@softcomputing.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 239–246.
springerlink.com

results that are obtained when the available data are scarce [7, 9]. These limitations make the hydrological-hydraulic models not always suitable for scarcely populated mountain zones where not enough reliable data have historically been recorded [6]. An increasing number of authors are suggesting the need of considering complementary information [3, 10]. In this sense, the Spanish Authorities have approved a program to elaborate a National Cartography System of Flooding Areas combining different methodologies.

The geomorphological-historical method [2] has shown to lead to more realistic results in recent studies developed in North Spain [4, 7]. Nevertheless, the employed information is very heterogeneous, has different degrees of reliability and precision and the final combination to assess the return period is made by expert criteria. In order to guarantee the objectivity of the approach a systematic analysis of the information and an automation of the final assessment is required. The usual flooding categories are those established by the EU Flood Directive, namely, low, medium and high flooding probability (respectively associated with return periods of about 500, 100 and 10 year). The final aim will be to obtain an automatic classification rule from a supervised experiment which has to be properly designed. In this work, some results obtained by pilot studies are discussed.

2 Analyzing the Flood Frequency with Geomorphological-Historical Information

Geomorphological and historical information is collected both in field and office work. In [4] historical data obtained from documentary sources and riverside inhabitants interviews are used to define an index of the flood magnitude. The index is based on 5 partial indicators. Namely, the discharge measure, the event magnitude according to the interviews, the proportion of interviewees mentioning the event, the flooded area percentage and the proportion of other documentary sources mentioning the event.

Given that the effect of some of the considered indicators cannot be precisely evaluated, intervals and fuzzy sets reflecting the imprecision are employed. The imprecision varies depending on different factors, so different ways of obtaining the intervals as a function of those factors are introduced. On the other hand, the importance and/or reliability of the indicators is different according to the expert criteria. Thus, the synthetic index gathering the information of all the indicators is computed as a weighted (interval-valued) mean of the valid data.

Once the information of the different indicators is computed and merged into the synthetic interval-valued index, a representation as the one in Fig. 1 is obtained. This index allows us to complete the flood chronology by deduction. For instance, although no historical information was obtained for 1993 in a given unit, it can be deduced because it is known that the unit was flooded

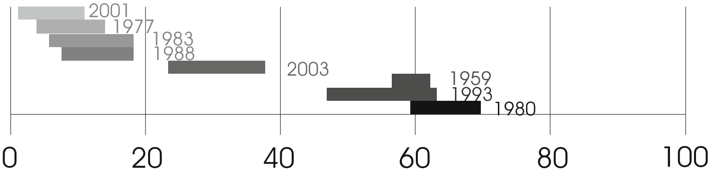


Fig. 1 Interval values for the synthetic index measuring the magnitude of each event.

in 2003: since it is known that a smaller event flooded the area, the larger event had to flood that unit too.

After reconstructing the series of historical floods, a lower bound for the return period can be obtained. The flood probability can be estimated through the flood frequency in the considered period. Since such a period is just a sample, the underlying stochastic variability can be considered. Specifically, confidence intervals based on the score method are proposed to be computed in [4], due to the performance of this method for small sample sizes. Time non-stationarity could also be considered to improve the results in this approach.

It is clear that having documentary references to all the historical floods is not feasible, even in the current information society. This is especially critical in sparsely populated areas. For this reason, it is proposed to complement this information with geomorphological data.

If a given unit is frequently flooded, visible geomorphic evidences can be found by the experts [8, 9]. Thus, observing the presence/absence of morphologies such as those in Fig. 2 provides us with highly valuable information about the flood frequency. Specifically, they allow to identify the high frequently flooded plains, which is essential for the flood hazard management. The shortcoming of this kind of data is that they do not allow to determine high return periods. Additionally, a certain degree of expert knowledge is required in order to identify the morphologies.

Nevertheless, there are other indicators, as the height of the river bank which may also supply information about the flooding frequency. For instance, in Fig. 3 a flooding plain with low river bank is shown. In contrast, Figure 4 displays the opposite situation. If they refer to the same stretch of the river, it is clear that the first unit may be easily flooded. According to the experts, the corresponding water cross-section to reach the flooding plain, and the surface of the drainage basin of each flooding plain are other quantitative variables to be considered. To quantify all these indicators, Digital Elevation Models (DEMs) are frequently used when available. In this particular study, DEMs were available with a 1-meter pixel resolution and up to milimetric precision for the height values.

Pilot studies have shown that some of those variables (height and cross-section) are statistically related to the probability of belonging to the class of high/medium/low frequency flood determined by expert criteria. However,

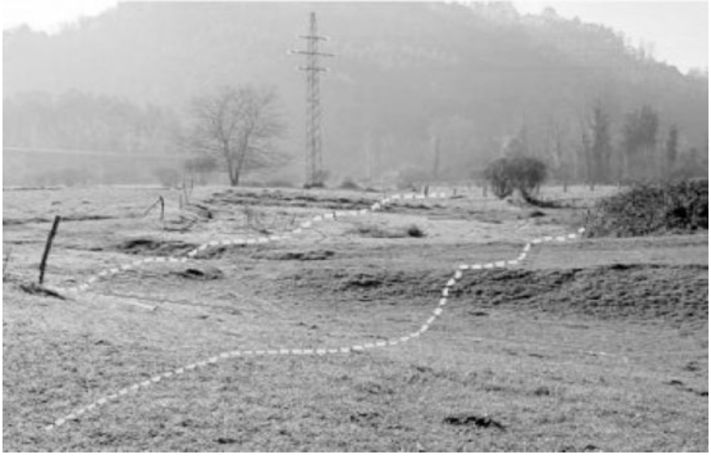


Fig. 2 Geomorphological evidences.

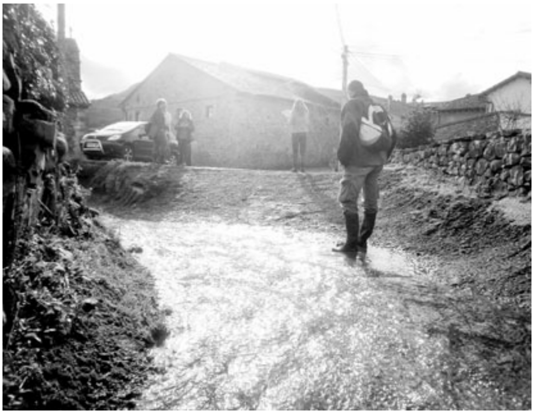


Fig. 3 Flood plain with low height.

considering only the height leads to results almost as good as those obtained by taking into account more variables (see Table 1).

From Table 1, we can conclude that DEM height measurements are very valuable for classifying between Medium and Low frequent flooding plains. Unfortunately, this accuracy is partially lost when the critical classification problem between High and Medium classes is considered. In this case, it seems that the combination with the cross-section information improves the results. This lack of accuracy is probably connected with the lower reliability of DEM measurements due to the usual abundance of vegetation in the High and Medium frequency flooding plains. Consequently, we consider that obtaining more reliable measures of the height is an essential task.

Table 1 Leave-One-Out percentage of right classification for discriminant analysis based on DEM measurements.

	High/Medium	Medium/Low	High/Medium/Low
Height	78.9%	93.3%	84.6%
Cross-section	78.9%	63.3%	64.1%
Surface	00.0%	68.8%	37.2%
Height and cross-section	84.2%	90.0%	87.2%

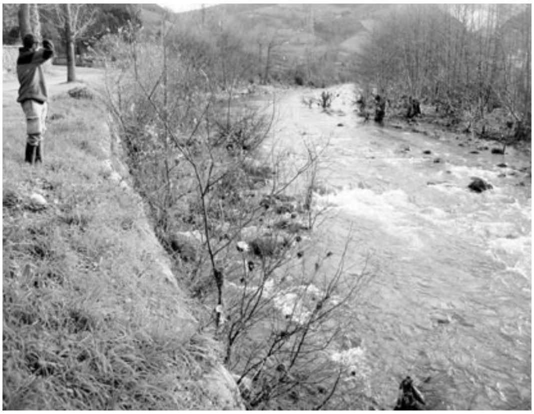


Fig. 4 Flood plain with high height.

Measuring exactly the height in field work would be too expensive. Nevertheless, obtaining a subjective valuation from the field researchers by a simple visual inspection is very easy. On the contrary, the cross-section and surface measurements require more expensive tools, as DEM or 1:2000-scale topographic maps.

Following the approach in [5], a fuzzy scale allowing to capture, not only the perception but also the uncertainty of the field researcher is proposed. Specifically, in the pilot study carried out, the field researchers were asked to collect their valuation of the height by means of trapezoidal fuzzy sets. A trapezoidal fuzzy set T is characterized by a $[0,1]-$valued function defined on $S \subset \mathbb{R}$ assuming positive values over an interval $[a,b]$, called 0-level, the value 1 over an interval $[c,d]$, called $1-$level, linearly increasing between a and c and linearly decreasing between b and d (see an example in Fig. 5). For each $x \in S$, $T(x)$ represents the degree of compatibility of the perception of the expert with the assertion "the height is x".

Thus, the experts are asked to choose the $0-$level of the fuzzy set as the smaller interval that they would not completely discard as containing the "true" Height value, whereas the $1-$cut is to be chosen as the interval of

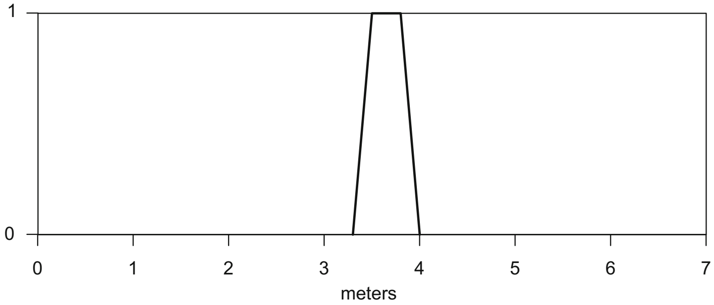


Fig. 5 Example of fuzzy perception of the height.

Table 2 Leave-One-Out percentage of right classification for discriminant analysis based on fuzzy field valuations.

	High/Medium	Medium/Low	High/Medium/Low
X_1 =Inf T_0	89.5%	83.3%	79.5%
X_2 =Inf T_0	89.5%	83.3%	82.1%
X_3 =Sup T_0	94.7%	83.3%	84.6%
X_4 =Sup T_0	89.5%	83.3%	82.1%
X_1,X_2,X_3,X_4	84.2%	90.0%	84.6%
$D=(X_1+X_2+X_3+X_4)/4$	89.5%	83.3%	82.1%

values that they indeed consider completely compatible with the height that they are observing. In other words, the 1−level would contain their personal opinion and the 0−level the range that they could admit to a greater or lesser extent. In this way, fuzzy perceptions of the length as that in Fig. 5 are available.

In order to verify if the collected fuzzy information is useful for the considered classification problem, several approaches can be considered. On the one hand, a classical discriminant analysis based on the 4 variables recorded for each trapezoidal fuzzy perception T (infima and suprema of the 0− and the 1− level set), as well as an average of all of them as a defuzzifier can be applied (see Table 2).

According to the results in Tables 1 and 2 it seems that the classification results between High and Medium are better when the field valuation is considered. On the contrary, the classification results between Medium and Low are better when the DEM measures are employed.

However, the analysis in Table 2 is not taking into account the structure of fuzzy data in the classification problem. To consider the fuzzy sets as structured data, the Proximity-based Classification Criteria for Fuzzy data

Table 3 Leave-One-Out percentage of right classification for Proximity-based classification with fuzzy field valuations.

	High/Medium	Medium/Low	High/Medium/Low
heigh valuation	89.5%	90.0%	84.6%

(PCCF) in [1] can be employed. The simplified idea of PCCF is to consider the fuzzy data as observations of a fuzzy random variable X and to proceed as follows:

- the 'center' C_i of each group G_i is computed by averaging of the fuzzy data in this group.
- To classify a new fuzzy data T, the conditional probability
$$P(d(X,C_i) > d(T,C_i)/G_i)$$
is estimated. This probability is a kind of measure of the affinity of T to each one of the groups.
- T is assigned to the group G_i with highest estimated probability.

The results in Table 3 are obtained by applying PCCF to the fuzzy perceptions of the height collected in the pilot study. These results indicate that the consideration of fuzzy field valuations of the height is very valuable in comparison with the consideration of the DEM measures for High/Medium classes. Additionally, for Medium/Low and the overall classification, the results are comparable. Thus, taking into account these results, the cost of both kinds of data, and since the experts have to visit anyway the flooding plains to look for geomorphological evidences, we recommend to consider the systematic collection of fuzzy valuations.

3 Concluding Remarks

In this paper we have surveyed some of the most useful geomorphological and historical information that can be used in order to assess the flooding frequency. The different sources show various degrees of imprecision and reliability. None of them is uniformly the best to classify the flooding plains according to the expert criterion, nevertheless they supply complementary information that can be merged to build a quantitative model in the future.

The documentary sources, the historical records and the interviews can be used to determine lower bounds that may avoid critical underestimates provided by other methods.

The geomorphological evidences are, of course, very useful to classify high frequency flooding. However, these dichotomic variables are not enough to accurately determine low return periods. Thus, although the expert criterion is highly linked to these variables, they require more information to distinguish between medium and low frequency classes. One of the characteristics that the experts use for that purpose is the river bank height, as well as other quantitative variables involved in the hydrological-hydraulic models. The pilot studies have shown that considering the height is essential. Nevertheless, reliable measures are required. One of the ways of measuring this indicator is to use DEM, but there are problems in presence of lush vegetation, which is related to high frequency classes. The pilot studies have indicated that incorporating fuzzy valuations of height provided by the field researchers gives, in general, better results than using DEM measures. Additionally, it is a not expensive source of information which does not require expert knowledge on geomorphology. Nevertheless, the necessity of developing pilot studies for each new basin should be underlined, because the heights determining the cuts between classes are specific of each basin, whence the sample to train the supervised classification has to be updated.

Acknowledgements. The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grants MTM2009-09440-C02-01 and MTM2009- 09440-C02-02 and by the Principality of Asturias Projects IB09-042C1 and IB09-042C2. Their financial support is gratefully acknowledged.

References

1. Colubi, A., González-Rodríguez, G., Gil, M.A., Trutchnig, W.: Discriminant Analysis for fuzzy random variables based on nonparametric regression (submitted for publication, 2010)
2. Baker, V.R., Kochel, R.C., Patton, P.: Flood geomorphology. Wiley, New York (1988)
3. Baker, V.R.: Paleoflood hydrology: Origin, progress, prospects. Geomorphology 101, 1–13 (2008)
4. Fernández, E., Colubi, A., González-Rodríguez, G., Anadón, S.: Integrating statistical information concerning historical floods: ranking and interval return period estimation (submitted for publication, 2010)
5. González-Rodríguez, G., Colubi, A., Gil, M.A.: Fuzzy data treated as functional data. A one-way ANOVA test approach (submitted for publication, 2010)
6. Jarret, R.D.: Hydrologic and hydraulic research in mountain rivers. Hydrology of Mountainous Areas 190, 107–117 (1990)
7. Lastra, J., Fernández, E., Díez-Herrero, A., Marquínez, J.: Flood hazard delineation combining geomorphological and hydrological methods: an example in the Northern Iberian Peninsula. Natural Hazards 45, 277–293 (2008)
8. Magilligan, F.J., Phillips, J.D., James, L.A., Gómez, B.: Geomorphic and sedimentological controls on the effectiveness of an extreme Flood. J. Geology 106, 87–96 (1998)
9. Ortega, J.A., Garzón, G.: Interpretación de los depósitos de avenida como clave para establecer la dinámica de la llanura de inundación. In: Pérez Alberti, A., López Bedoya, J. (eds.) Geomorfología y territorio. Cursos e Congresos da Universidade de Santiago de Compostela, pp. 629–644 (2006)
10. Thorndycraft, V.R., Benito, G., Gregory, K.J.: Fluvial geomorphology: A perspective on current status and methods. Geomorphology 98, 2–12 (2008)

Geometric Sampling: An Approach to Uncertainty in High Dimensional Spaces

Juan Luis Fernández-Martínez, Michael Tompkins,
Tapan Mukerji, and David Alumbaugh

Abstract. Uncertainty is always present in inverse problems. The main reasons for that are noise in data and measurement error, solution non-uniqueness, data coverage and bandwidth limitations, physical assumptions and numerical approximations. In the context of nonlinear inversion, the uncertainty problem is that of quantifying the variability in the model space supported by prior information and the observed data. In this paper we outline a general nonlinear inverse uncertainty estimation method that allows for the comprehensive search of model posterior space while maintaining computational efficiencies similar to deterministic inversions. Integral to this method is the combination of model reduction techniques, a constrained mapping approach and a sparse sampling scheme. This approach allows for uncertainty quantification in inverse problems in high dimensional spaces and very costly forward evaluations. We show some results in non linear geophysical inversion (electromagnetic data).

Keywords: Inverse Problems, Geometric sampling, High Dimensional Spaces, Uncertainty.

Juan Luis Fernández-Martínez and Tapan Mukerji
Energy Resources Department, Stanford University, Palo Alto, California, USA

Juan Luis Fernández-Martínez
Department of Mathematics, University of Oviedo, Oviedo, Spain

Juan Luis Fernández-Martínez
Department of Civil and Environmental Engineering, University of California Berkeley, Berkeley, USA

Michael Tompkins and David Alumbaugh
Schlumberger-EMI Technology Center, Richmond, CA 94804, USA
e-mail: jlfm@uniovi.es,mtompkins@richmond.oilfield.slb.com,
mukerji@stanford.edu,dalumbaugh@richmond.oilfield.slb.com

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 247–254.
springerlink.com

1 Inverse Problems and Uncertainty

Inverse problems can be written in discrete form as $\mathbf{F}(\mathbf{m}) = \mathbf{d}$, where $\mathbf{m} \in \mathbf{M} \subset \mathbf{R}^n$ are the model parameters, $\mathbf{d} \in \mathbf{R}^s$ the discrete observed data, and

$$\mathbf{F}(\mathbf{m}) = (f_1(\mathbf{m}), f_2(\mathbf{m}), \ldots, f_s(\mathbf{m}))$$

is the vector field representing the forward operator and $f_j(\mathbf{m})$ is the scalar field that accounts for the j-th data. Usually $s < n$, that is, the inverse problem has an underdetermined character. Furthermore, many geophysical problems are nonlinear and poorly sampled making the inverse ill-posed, non-unique, and ill-conditioned. Ill-conditioning is an important issue when solving the inverse problem as an optimization problem, because noise in data is amplified back to the model parameters through the inverse forward operator, $\mathbf{F}^{-1}$. In addition to these difficulties, we have measurement errors, data coverage and bandwidth limitations, and numerical approximations, which all contribute to uncertainty in our inverse solutions. In the context of nonlinear inversion, the uncertainty problem is that of quantifying the variability in the model space supported by prior information, the observed data, and the errors of the method.

Global optimization algorithms can be a good alternative to deterministic solutions, because they approach the nonlinear inverse problem as a sampling problem instead of looking for the inverse operator. Also, they only need as prior information the search space of possible solutions. Typically they use as a cost (or objective) function the data prediction misfit in a certain norm p :

$$\|\mathbf{F}(\mathbf{m}) - \mathbf{d}\|_p .$$

It is possible to show analytically that the models in the neighborhood of $\mathbf{m}_0$ that fit the data within the same tolerance, tol, belongs to the following hyperquadric:

$$(\mathbf{m} - \mathbf{m}_0)^T \mathbf{JF}_{\mathbf{m}_0}^T \mathbf{JF}_{\mathbf{m}_0} (\mathbf{m} - \mathbf{m}_0) + 2\boldsymbol{\Delta}\mathbf{d}^T (\mathbf{m} - \mathbf{m}_0) + \|\boldsymbol{\Delta}\mathbf{d}\|_2^2 = tol^2.$$

$\mathbf{JF}_{\mathbf{m}_0}$ is the Jacobian matrix of the operator $\mathbf{F}$ in $\mathbf{m}_0$ and $\boldsymbol{\Delta}\mathbf{d} = \mathbf{F}(\mathbf{m}_0) - \mathbf{d}$. This means that the region of equivalent models locally in $\mathbf{m}_0$ have the direction of the vectors of the $\mathbf{V}$ base given by the singular value decomposition of $\mathbf{JF}_{\mathbf{m}_0}$ and whose axes are proportional to the inverse of the singular values λ_k in each direction. Due to the continuity of the Jacobian operator, we finally conclude that with no regularization term the misfit function has a flat and elongated valley shape. This approach assumes derivability of $\mathbf{F}$ in $\mathbf{m}_0$, which is usually the case in most inverse problems.

These types of sampling methods can be useful, but they have limitations for large spaces, since they sample each model parameter as independent variables; a property that is not necessarily true for finite resolution methods (i.e, electromagnetic imaging). In contrast, local optimization methods are not designed to approach this sampling problem, and they often fail to find a solution without regularization. These algorithms can very effectively handle inverse problems having thousands to millions of parameters. The main drawback of local methods is that they are highly dependent on the initial guess and the quality of the prior information that is built into the regularization term to achieve uniqueness and stability in the inverse solution. Furthermore they do not provide any measure of nonlinear uncertainty around the solution of the inverse problem.

An alternative to both these methods, presented by [7], is to solve the sampling problem in a bounded transformed space using optimally sparse grids. This method both accounts for the equivalence in our nonlinear inverse problem and allows for the inference of solution uncertainty by sampling the model posterior. In the following sections we review two important aspects of this technique including model parameter reduction using orthogonal transformations, and geometric sampling of the equivalent model space.

2 Model Reduction Techniques

The use of model reduction techniques act to decrease the dimension of the inverse problem. For an underdetermined linear inverse problem of the form

$$\mathbf{Gm} = \mathbf{d}$$

the method consists in expanding the solution $\mathbf{m}$ as a linear combination of a set of independent models, $\left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\right\}$:

$$\mathbf{m} \in \left\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q \right\rangle = \sum_{k=1}^{q} \alpha_k \mathbf{v}_k, \tag{1}$$

and to solve the linear system $\mathbf{B}\alpha = \mathbf{d}$ where $\mathbf{B} = \mathbf{GV}$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q]$. This methodology is related to subspace methods and can be easily generalized to nonlinear inverse problems, because once the base is determined, the search is performed on the α-space. The use of a reduced set of basis vectors that are consistent with our prior knowledge allows to regularize the inverse problem and to reduce the space of possible solutions. There are several ways of finding the base $\left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\right\}$ in order to reduce complexity. Principal Component Analysis (PCA), the Singular Value Decomposition, the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform have been presented in [2]. As stated in [2] the orthonormal base has to allow for classification of the model variability, and to be separable in order to expand this methodology to large parameterizations. Also one of these methods is covariance-based (PCA) while the other techniques are model-based, allowing for the use of different kinds of reduction techniques depending on the dimension of our inverse problem and the quality of the prior information.

3 Computing Uncertainty in High Dimensional Spaces

The methods to compute uncertainty in high dimensional spaces can be divided into two main groups depending on whether the forward problem is fast to solve or not. To the first category belong global optimization algorithms in a reduced model space [3]. Monte Carlo techniques can not address this kind of problem due to the dimensionality issue. Additionally they can be very inefficient, since they typically spend much effort sampling parts of the posterior that do not fit the observed data. Global optimization algorithms can address the non-convexity of the cost function by sampling the family of equivalent models. Nevertheless when the inverse problem has a very expensive forward problem these methods are not a good alternative. The main reason is that the tasks of sampling the posterior and the forward prediction are coupled.

3.1 The Geometric Sampling Approach

An alternative to stochastic sampling methods is geometric sampling as introduced by Tompkins and Fernández Martínez [7]. The methodology is composed of four steps: 1) parameter reduction 2) model constraint mapping by vertex enumeration, 3) sparse grid sampling, and 4) final forward evaluation. Model reduction techniques have been already presented, and we discuss sampling below.

Once we perform model reduction techniques on the model space we have at our disposal a set of orthonormal vectors $\left\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q\right\}$ allowing the linear decomposition (1). Vectors $\mathbf{v}_i$ increase their frequency with their index. Thus this linear expansion has a regularization effect on the sampling. The next step is to provide to the model $\mathbf{m}$ some lower and upper bounds, $\mathbf{l}$ and $\mathbf{u}$:

$$\mathbf{l} \leq \sum_{k=1}^{q} \alpha_k \mathbf{v}_k \leq \mathbf{u}. \tag{2}$$

At this stage we suppose a uniform prior distribution of $\mathbf{m}$ in these bounds. Condition (2) is called the vertex enumeration problem [1] and generates in the reduced model space a polytope $P \subset \mathbb{R}^p$ for the coefficients on the reduced base α_k. This idea was first suggested by Ganapathysubramanian and Zabaras [5] for heat flow problems in random media using the PCA base. Nevertheless, in their work this idea was not used to compute uncertainty associated with to non-linear inverse problems through the geometric sampling approach. Geometric Sampling consists in sampling within the polytope using an uniform distribution and sparse grids [7]. This uniform sampling distribution induces a non-uniform prior distribution in the original space $\mathbf{M}$, since:

$$P\left(m_i < c_i\right) = P\left(\sum_{k=1}^{q} \alpha_k \mathbf{v}_{ki} < c_i\right),$$

where the vectors of the base are fixed.

This complete non-uniform distribution on **M**, the posterior, is not explicitly calculated, since we only have at our disposal some independent samples whose number depend on the sparsity of the sampling scheme that has been used. However, once we have sampled α_k over our reduced space, our approximation to this posterior is determined by mapping these samples back to our original model space, **M**.

Optimizing the sampling on the reduced space is very important in our methodology, since this allows us to tailor the sampling density to the cost and to the complexity of the forward evaluations. Of particular interest are sparse sampling techniques, such as the Smolyak grids [6], that can provide for adaptive sample refinement. That is, if we wish to extend the accuracy or breadth of our sampling from some initial set to some larger set, we simply need to evaluate samples at the additional nodes over the second set, which provides a means to optimize sampling based on some criteria (in our case, convergence of statistical moments of our posterior).

In order to sample inside the polytope, defined by Equation (2), we circumscribe a hypercube in our $q-$dimensional space and perform the sampling on a Cartesian grid in the reduced model space using Smolyak nested grids [8]. Once the sampling on the hypercube is performed, models, **M**, in our original model space are reconstructed using (1). The final step in uncertainty estimation is to evaluate the posterior model samples for their likelihood (i.e., data misfit). For this, forward simulations are performed and models are accepted or rejected based on a threshold misfit. The accepted models represent the equivalence space of equiprobable models. The uncertainty of the nonlinear inverse problem then follows from either the model ensemble itself or statistical measures (e.g., mean, covariance, percentile, interquartile range) computed from it.

4 Example: A Subsurface Resistivity Image

Over the past decade, marine controlled-source electromagnetic (CSEM) surveying has emerged as a useful technique for subsurface resistivity imaging. In this method, a deep-towed electric dipole source is used to excite a low-frequency (˜0.1–10 Hz) electromagnetic signal. This signal propagates through the seawater and subsurface and is perturbed by geologic variation to depths of several kilometers. Spatially-distributed, multi-component, seafloor receivers record this electromagnetic energy at offsets up to 20km. These electromagnetic data are typically interpreted using geophysical inverse methods that attempt to reconstruct subsurface resistivities from recorded fields [4]. Inversion of this data is nonlinear, and uncertainty comes from noise in the data and assumptions about the earth model.

To demonstrate the extension of our uncertainty method to large parameter spaces, we chose a marine CSEM field dataset collected by WesternGeco in the Potiguar basin in Brazil during 2009. The electromagnetic data consisted of ∼3800 complex-valued fields at 4 frequencies (0.25, 0.50, 0.75, and 1.5Hz). For this problem, the original uniform pixel space had 33,280 parameters; however, we only considered the part of the final inverse model not occupied by air, seawater, or homogeneous resistive basement (>4500m depths). This left the inversion domain shown in Figure 1(A), which consisted of 5,461 parameters. Of particular interest, are the structures and magnitudes of the resistive features at ∼4000m depth. After subtracting a global mean from our inverse model, the SVD base was reduced to six terms that represented ∼ 81% of the variability in the residuals, following the methodology explained in [2]. We then performed optimal sampling over this six-dimensional reduced space to estimate the model posterior and solution uncertainty. The resulting posterior polytope was defined by 15,990 vertices, which, in this case, we approximated with a 6-D hypercube (i.e., 6 bases were chosen). Based on sparse sampling of this hypercube we evaluated 1942 equi-feasible models to generate the model posterior. The computational cost of evaluating these models was approximately 6 days using two 8-core workstations. With a threshold

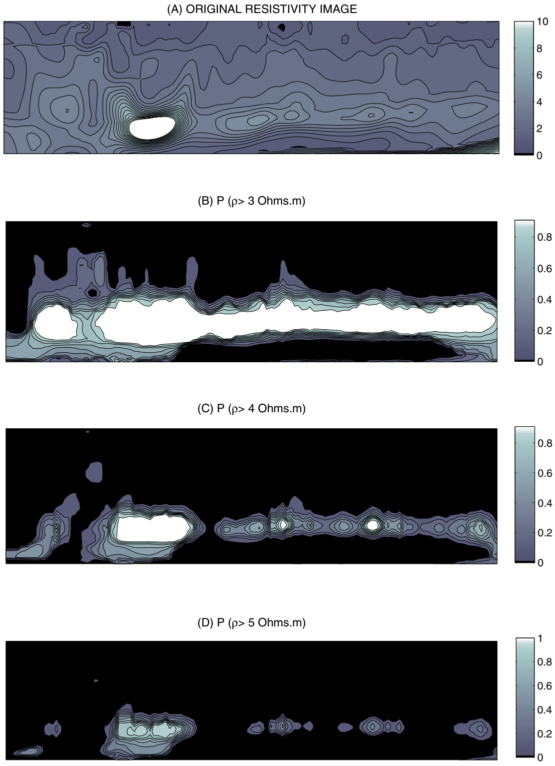


Fig. 1 Original resistivity field and probability maps for different cut-offs (3, 4 and $5\,\Omega.m$).

misfit, RMS<15%, we generated the final equivalent model set (283 models). Once we have this ensemble, we can compute statistical properties from it as well, for example, e-types, variances, or indicator probabilities. Probability (normalized frequency) maps are a useful way to visualize uncertainty. Figures 1(B)–(D) show the probability maps for different cut-offs (3, 4 and 5 Ω.m) deduced from the approximation of the model posterior over the low misfit region. These probability maps represent the probability of occurrence of a resistivity of at least 3, 4 and 5 Ω.m in our model space, and quantifies some aspects of the uncertainty in our subsurface resistivity image inverse problem. Additional measures of uncertainty are possible using any number of statistical properties of the model posterior.

5 Conclusions

The combined use of model reduction techniques, and sparse sampling allows us to approach efficiently the uncertainty problem in high dimensional spaces. This methodology can be efficiently applied to estimate nonlinear inverse model uncertainty in any kind of inverse problem. The combination of these methods can reduce the nonlinear uncertainty problem to a deterministic sampling problem in only a few dimensions, requiring only limited forward solves, and resulting in an optimally sparse representation of the posterior model space. While forward solves are required to evaluate the sampled models, our scheme optimizes sample size by iteratively increasing sampling complexity until uncertainty measures converge or a maximum number of forward solves is completed.

Acknowledgements. We acknowledge WesternGeco for allowing us to use and publish the resistivity image data set.

References

1. Avis, D., Fukuda, K.: A pivoting algorithm for convex bulls and vertex enumeration of arrangements and polyhedra. J. Discrete Comp. Geometry 8, 295–313 (1992)
2. Fernández-Martínez, J.L., Tompkins, M., Fernández-Muñiz, Z., Mukerji, T.: Inverse problems and model reduction techniques. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Soft Computing. Springer, Berlin (in this book, 2010)
3. Fernández-Martínez, J.L., Mukerji, T., García-Gonzalo, E.: Particle Swarm Optimization in high dimensional spaces. In: Proceedings of the Seventh International Conference on Swarm Intelligence, ANTS 2010, Bruxelles, Belgium (2010)

4. MacGregor, L.M., Sinha, M.C., Constable, S.: Electrical resistivity structure of the Valu Fa Ridge, Lau Basin, from marine controlled-source electromagnetic sounding. Geophys. J. Int. 146, 217–236 (2001)
5. Ganapathysubramanian, B., Zabaras, N.: Modeling diffusion in random heterogeneous media: Data-driven models, stochastic collocation and the variational multiscale method. J. Comput. Phys. 226, 326–353 (2007)
6. Smolyak, S.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Math. 4, 240–243 (1963)
7. Tompkins, M.J., Fernández-Martínez, J.L.: Scalable Solutions for Nonlinear Inverse Uncertainty Using Model Reduction, Constraint Mapping, and Sparse Sampling. In: Proceedings of the 72nd EAGE Conference & Exhibition, Barcelona, Spain (2010)
8. Xiu, D., Hesthaven, J.S.: High-order collocation methods for differential equations with random inputs. SIAM J. Sci. Comput. 27, 1118–1139 (2005)

Inverse Problems and Model Reduction Techniques

Juan Luis Fernández-Martínez, Michael Tompkins,
Zulima Fernández-Muñiz, and Tapan Mukerji

Abstract. Real problems come from engineering, industry, science and technology. Inverse problems for real applications usually have a large number of parameters to be reconstructed due to the accuracy needed to make accurate data predictions. This feature makes these problems highly underdetermined and ill-posed. Good prior information and regularization techniques are needed when using local optimization methods but only linear model appraisal (uncertainty) around the solution can be performed. The large number of parameters precludes the use of global sampling methods to approach inverse problem solution and appraisal. In this paper we show how to construct different kinds of reduced bases using Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT). The use of a reduced base helps us to regularize the inverse problem and to find the set of equivalent models that fit the data within a prescribed tolerance and are compatible with the model prior.

Keywords: Inverse Problems, Model Reduction, Orthogonal Transformations, Posterior Sampling, Uncertainty.

Juan Luis Fernández-Martínez
Department of Civil and Environmental Engineering, University of California at Berkeley, Berkeley, USA

Juan Luis Fernández-Martínez and Zulima Fernández-Muñiz
Department of Mathematics, University of Oviedo, Oviedo, Spain

Juan Luis Fernández-Martínez and Tapan Mukerji
Energy Resources Department, Stanford University, Palo Alto, California, USA

Michael Tompkins
Schlumberger-EMI Technology Center, Richmond, CA 94804, USA
e-mail: jlfm@uniovi.es,mtompkins@richmond.oilfield.slb.com,
zulima@uniovi.es,mukerji@stanford.edu

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 255–262.
springerlink.com

1 Inverse Problems and Uncertainty

Inverse problems can be written in discrete form as $\mathbf{F}(\mathbf{m}) = \mathbf{d}$, where $\mathbf{m} \in \mathbf{M} \subset \mathbb{R}^n$ are the model parameters, $\mathbf{d} \in \mathbb{R}^s$ is the discrete observed data and $\mathbf{F}(\mathbf{m}) = (f_1(\mathbf{m}), f_2(\mathbf{m}), \ldots, f_s(\mathbf{m}))$ is the vector function representing the forward operator, being $f_j(\mathbf{m})$ the scalar field that predicts the j-th data. Usually $s \ll n$, that is the inverse problem has an underdetermined character. This makes the inverse very ill posed, that is no unique solution exist and the inverse problem is very ill-conditioned. When using local techniques the prior information is built in the regularization term that is aimed at achieving uniqueness and stability in the inverse problem solution. The L-curve serves to display the trade-off between the complexity of the regularized solution and the data misfit. The results of applying this procedure is a unique model. In fact the L-curve clearly shows the nature of the equivalencies, through the family of models located along the L-curve that have similar data misfit but differ increasingly from the prior. These models are called equivalent in our approach as far as they satisfy some lower and upper bounds constraints.

2 Parameter Reduction via Orthogonal Transformations

Most inversion algorithms involve a complicated forward model with a large number of parameters needed for accomplishing accuracy on the data prediction. However the model parameterization used in the forward problem may not be the best choice for inversion since the observed data do not inform about all the components of the model. In the linear case this is of course related to the dimension of the null space of the linear operator. Model parameterization is a key concept in order to make the inverse problem less ill-conditioned. Adopting the right parameterization (i.e. basis set) also reduces the number of dimensions in which the inverse problem is going to be solved allowing performing posterior uncertainty analysis.

The use of model reduction techniques is based on the fact that the inverse model parameters are not independent. Conversely, there exist correlations between model parameters introduced by the physics of the forward problem $\mathbf{F}$ in order to fit the observed data. We propose to take advantage of this fact to reduce the number of parameters that are used to solve the identification problem. To illustrate this idea let us consider an underdetermined linear inverse problem of the form $\mathbf{G}\ \mathbf{m} = \ \mathbf{d}$ where $\mathbf{G} \in \mathscr{M}(s,n)$ is the forward linear operator and s, n stand respectively for the dimensions of the data and model spaces. The solution $\mathbf{m}$ to this linear inverse problem is expanded as a linear combination of a set of independent models $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q\}$, which are consistent with our prior knowledge:

$$\mathbf{m} \in M = \left\langle \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q \right\rangle \ \rightarrow \mathbf{m} = \sum_{k=1}^{q} \alpha_k \mathbf{v}_k, \ q \ll n$$

The problem now is equivalent to find the model $\mathbf{m}$ in a subspace of $\mathbb{R}^n$ of dimension q:

$$\mathbf{GV}\alpha = \mathbf{d} \ \rightarrow \mathbf{B}\alpha = \mathbf{d}, \qquad \mathbf{B} = \mathbf{GV} \in \mathscr{M}\left(s, q\right),$$

where $\mathbf{V} = [\mathbf{v}_1 \ \mathbf{v}_2 \ldots \mathbf{v}_q]$. This amounts to solving the linear system $\mathbf{B}\alpha = \mathbf{d}$ to find the set of weights α of the linear combination. Although this linear system might still be ill-posed, the effect of this methodology is to reduce the space of possible solutions. Additionally depending on the values of s and q the linear system $\mathbf{B}\alpha = \mathbf{d}$ might even have an over-determined character. This methodology can be easily generalized to nonlinear inverse problems, because once the base is determined, the search is performed on the α-space. The use of a reduced set of basis vectors that are consistent with our prior knowledge allows to regularize the inverse problem and to reduce the space of possible solutions. These orthonormal bases have to allow the classification of the model variability according to some criteria. Finally separability allows the generalization of our methodology to higher dimensions. In this paper we show several techniques that can be used to construct these bases, such as, the Principal Component Analysis (PCA), the Singular Value Decomposition, the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT). Some of these methods are covariance-based such as the PCA, while others techniques are model-based. Also some of these methods do not need any diagonalization (DCT, DWT) and do not require the computation of the Jacobian. This is important because both tasks might not be possible to perform in very high dimensions.

2.1 Principal Component Analysis (PCA)

Principal component analysis [7] is a well-known mathematical procedure that transforms a number of correlated variables (model parameters in this case) into a smaller number of uncorrelated variables called principal components, while maintaining their full variance and ordering the uncorrelated variables by their contributions. The resulting transformation is such that the first principal component represents the largest amount of variability, while each successive component accounts for as much of the remaining variability as possible [5]. Usually PCA is performed in the data space, but in this case it is used to reduce the dimensionality of the model space based on a priori samples obtained from conditional geostatistical realizations that have been constrained to static data. Applied to our context, PCA consists in finding an orthogonal base of the experimental covariance matrix estimated with these prior geological models, and then selecting a subset of the most important eigenvalues and associated eigenvectors that are used as a reduced model space base.

Given and ensemble of N possible models $\{\mathbf{m}_k\}_1^N \subset M \subset \mathbb{R}^n$ we compute the ensemble centered covariance

$$\mathbf{C} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{m}_k - \mu)(\mathbf{m}_k - \mu)^T$$

where $\mu = \frac{1}{N}\sum_{k=1}^{N}\mathbf{m}_k$ is the experimental mean. The covariance matrix $\mathbf{C}$ is symmetric and semi-definite positive, hence, diagonalizable with orthogonal eigenvectors $\mathbf{v}_i$, and real semi-definite positive eigenvalues λ_i such that $\mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$. Eigenvectors $\mathbf{v}_i$ are called, principal components. Eigenvalues can be ranged in decreasing order, and we can select a certain number of them to match most of the variability of the models. That is, the q first eigen-models ($q \ll N$) representing most part of the energy spectrum of the decomposition are chosen. Then, any model in the reduced space is represented as a unique linear combination of the selected eigen-models $\mathbf{m}_k = \mu + \sum_{k=1}^{q}\alpha_k\mathbf{v}_k$. The model covariance can also be a posterior covariance

$$\mathbf{C}_{pos} \propto \left(J\mathbf{F}_{\mathbf{m}_f}^T J\mathbf{F}_{\mathbf{m}_f}\right)^{-1},$$

where $J\mathbf{F}_{\mathbf{m}_f}$ is the model Jacobian matrix computed about the last iteration of the nonlinear inversion. The Jacobian matrix is rank deficient. The dimension of the null space of the Jacobian serves to account locally for the linear uncertainty analysis around the base model. Truncation (Moore Penrose pseudoinverse) and/or damping techniques can be used to invert it. In the first case a regularization is input to the inverse problem by thresholding the singular vectors that span the null space of $J\mathbf{F}_{\mathbf{m}_f}$. In the second case this inverse also gathers the influence of these singular vectors that are typically related to the high frequencies in the model. Finally, the posterior covariance can be also computed experimentally from the ensemble of the models that have been gathered in a certain region of error tolerance using global optimization algorithms.

2.2 Singular Value Decomposition (SVD)

Although the inverse problem in abstract form is written as $\mathbf{F}(\mathbf{m}) = \mathbf{d}$ we are interested in cases where the model $\mathbf{m}$ is a 2D or a 3D image. Examples of this are the conductivity or the velocity field in a 2D or 3D tomography problem.

The SVD has been applied extensively in many fields, such as, in the resolution of linear inverse problems through the Moore-Penrose pseudo inverse, in signal processing and pattern recognition, etc. The SVD allows the factorization of any rectangular matrix $\mathbf{m}_0 \in \mathscr{M}(s,n)$ in the form $\mathbf{m}_0 = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices that provide orthonormal bases for $\mathbb{R}^s, \mathbb{R}^n$ respectively. $\boldsymbol{\Sigma}$ is a s-by-n box-diagonal matrix with non negative real numbers on the diagonal, called the singular values of the matrix $\mathbf{m}_0$. These bases can be calculated as follows:

$$\mathbf{m}_0\mathbf{m}_0^T = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\mathbf{U}^T, \quad \mathbf{m}_0^T\mathbf{m}_0 = \mathbf{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\mathbf{V}^T.$$

This means that the $\mathbf{U}$ base comes from the eigenvalue decomposition (PCA) of the row correlation matrix, $\mathbf{m}_0\mathbf{m}_0^T$. Similarly the $\mathbf{V}$ base is calculated from the eigenvalue decomposition (PCA) of the column correlation matrix, $\mathbf{m}_0^T\mathbf{m}_0$. Both correlation matrices have the same rank as the original image $\mathbf{m}_0$. Once the bases $\mathbf{U}$, $\mathbf{V}$ are calculated, we project the image $\mathbf{m}_0$ onto one of the bases, as follows:

$$\mathbf{U}_{\mathbf{image}} = \mathbf{U}^T\mathbf{m}_0, \ \mathbf{V}_{\mathbf{image}} = \mathbf{m}_0\mathbf{V}.$$

The intent is to decorrelate and compress the image either to the d_u first rows of $\mathbf{U}_{\mathbf{image}}$ (or the d_v first columns of $\mathbf{V}_{\mathbf{image}}$) corresponding to most of its variability. Usually our earth models are horizontally stratified, hence $\mathbf{U}_{\mathbf{image}}$ gives a more effective compression than $\mathbf{V}_{\mathbf{image}}$. Finally (for the $\mathbf{U}_{\mathbf{image}}$ case) we define a thresholding matrix, T_i, as a zero matrix with the dimensions of $\mathbf{U}_{\mathbf{image}}$ but containing only its $i-$th row:

$$T_i(k,:) = \mathbf{U}_{\mathbf{image}}(k,:).\delta_{ik}, \ k = 1,\ldots,s, \ i = 1,\ldots,d_u,$$

where δ_{ik} stands for the Kronecker delta. The reduced base is derived by projecting the first r_d threshold matrices back onto the original canonical base as follows:

$$\mathbf{bu}_i = \mathbf{UT}_i, \qquad i = 1,\ldots,d_u.$$

A similar procedure can be done using $\mathbf{V}_{\mathbf{image}}$ if we wish to compress by our image vertically:

$$T_j(:,k) = \mathbf{V}_{\mathbf{image}}(:,k).\delta_{jk}, \ k = 1,\ldots,n, \ j = 1,\ldots,d_v,$$

$$\mathbf{bv}_j = \mathbf{T}_j\mathbf{V}^T, \qquad j = 1,\ldots,d_v.$$

2.3 The Discrete Cosine Transform (DCT)

DCT has been widely deployed by modern video coding standards. Like other transforms, the Discrete Cosine Transform (DCT) attempts to decorrelate the 2D model. The discrete cosine transform (DCT) is a discrete Fourier transform operating on real data. It expresses a signal in terms of a sum of sinusoids with different frequencies and amplitudes. This transformation is separable, that is, it can be defined in higher dimensions. For instance for an image $\mathbf{m}_0 \in \mathscr{M}(s,n)$ the DCT is defined as follows:

$$D(u,v) = c(u)\,c(v) \sum_{i=0}^{s-1}\sum_{j=0}^{n-1} m_0(i,j) \cos\left(\frac{\pi(2i+1)u}{2s}\right) \cos\left(\frac{\pi(2j+1)v}{2n}\right),$$

where $u = 0,\ldots,s-1$ and $v = 0,\ldots,n-1$, and

$$c(\alpha) = \begin{cases} \dfrac{1}{\sqrt{N}} \text{ if } \alpha = 0, \\ \sqrt{\dfrac{2}{N}} \text{ if } \alpha \neq 0. \end{cases}$$

being N either the number of rows (s) or columns (n) of the original image.

The DCT has also an inversion formula to recover the original signal from the DCT transform (see for instance [8]). DCT can be expressed in matrix form as an orthogonal transformation $D = C_{(s,s)}\mathbf{m}_0 C^T_{(n,n)}$. Thus, the method to calculate the base shown for the SVD also holds for the DCT. Although it is very easy to implement its use is very rare in geosciences maybe because in image compression the Wavelet transform achieves better results [6]. Nevertheless for our purposes it provides a very adequate model reduction.

2.4 The Discrete Wavelet Transform (DWT)

The discrete wavelet transform allows us to find two orthogonal transformations based on wavelets, named $\mathbf{U}_w$ and $\mathbf{V}_w$, such as $\mathbf{m} = \mathbf{U}_w \mathbf{m}_{LR} \mathbf{V}_w^T$. These orthogonal matrices can be constructed as follows: $\mathbf{U}_w = \mathbf{W}_L^T$ and $\mathbf{V}_w = \mathbf{W}_R^T$ where

$$\mathbf{W}_L = \left[\frac{\mathbf{H}}{\mathbf{G}}\right]_s, \ \mathbf{W}_R = \left[\frac{\mathbf{H}}{\mathbf{G}}\right]_n.$$

$\mathbf{H}$ represents a low pass or averaging portion of the matrices $\mathbf{W}$, and $\mathbf{G}$ is the high pass or differencing portion. In all of cases we have

$$\mathbf{m}_{LR} = \mathbf{W}_L \mathbf{m} \mathbf{W}_R^T = \left[\begin{array}{c|c} \mathbf{HmH}^T & \mathbf{HmG}^T \\ \hline \mathbf{GmH}^T & \mathbf{GmG}^T \end{array}\right] = \left(\begin{array}{c|c} \mathscr{B} & \mathscr{V} \\ \hline \mathscr{H} & \mathscr{D} \end{array}\right)$$

where $\mathscr{B}$ is the blur, $\mathscr{V}$ are the vertical differences, $\mathscr{H}$ are the horizontal differences and $\mathscr{D}$ are the diagonal differences. In the case of $\mathbf{U}_{image}$

$$\mathbf{U}_{image} = \mathbf{W}_L \mathbf{m} = \left[\frac{\mathbf{H}}{\mathbf{G}}\right]_s \mathbf{m} = \left[\frac{\mathbf{Hm}}{\mathbf{Gm}}\right]_{(s,n)}$$

This means that $\mathbf{U}_{image}$ has in its upper part the "partial" blur and its lower part the details. In the case of $\mathbf{V}_{image}$ the "partial" blur is on the left part and on its right the details:

$$\mathbf{V}_{image} = \mathbf{m}\mathbf{W}_R^T = \mathbf{m}\left[\mathbf{H}^T \middle| \mathbf{G}^T\right]_n = \left[\mathbf{mH}^T \middle| \mathbf{mG}^T\right]_{(s,n)}$$

Different kind of filters can be used for this purpose [2]. These families of wavelets define a discrete wavelet transform having a maximum number of vanishing moments.

2.5 Multiscale Inversion

The orthonormal character of the vectors of the reduced base allows an easy implementation of a multi-scale inversion approach adding more eigenvalues to match higher frequencies to the model $\mathbf{m}$ as needed since:

$$\mathbf{m} = (\alpha_1, \alpha_2, \ldots, \alpha_q)_{\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_q\}} = (\alpha_1, \alpha_2, \ldots, \alpha_q, 0)_{\{\mathbf{v}_1,\mathbf{v}_2,\ldots,\mathbf{v}_q,\mathbf{v}_{q+1}\}}.$$

Multiscale inversion is very easy to implement using global methods. To determine which level of detail we have to consider is an important question since all the finer scales might not be informed by the observed data. By reducing the base we are setting these finer scales of heterogeneity (high frequencies) to zero avoiding also the risk of over fitting the data. The choices of parameters q in Principal Component Analysis (PCA), d_u or d_v in the Singular Value Decomposition (SVD), and a similar parameters in the DCT and the DWT is an important question since they serve as regularization parameters. In practice these parameters are determined empirically for each problem depending on the resolution and on the value of the data misfit that we want to achieve, similarly to the L-curve procedure in non-linear least square methods [4].

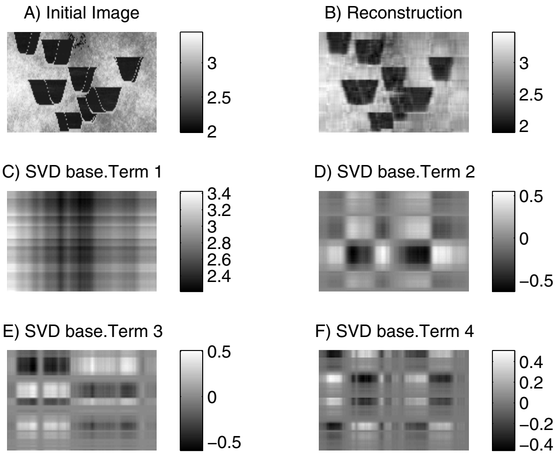


Fig. 1 Original image, reconstructed image using SVD and four terms of the SVD base.

3 Application to an Oil Reservoir Problem

To illustrate our model reduction approach we chose the P-wave velocity image from the synthetic Stanford VI reservoir [1] (Fig. 1A) consisting in 12000 parameters (80 by 150 pixels). In Fig. 1B we show its reconstruction using ten vectors of the geological base derived through the SVD. Figures 1C-1F show the four first terms of the base. This procedure allows to reduce the dimension from 12000 pixels to 10 parameters in the SVD base. Although this example is synthetic, it is important to understand that the dimensionality reduction allows to compute uncertainty around this model either using global optimization methods or the geometric sampling approach [3], [9].

References

1. Castro, S., Caers, J., Mukerji, T.: The Stanford VI Reservoir. Stanford Center for Reservoir Forecasting, SCRF (2005)
2. Daubechies, I.: Ten lectures on wavelets. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61. SIAM, Philadelphia (1992)
3. Fernández-Martínez, J.L., Tompkins, M., Mukerji, T., Alumbaugh, D.: Geometric Sampling: an Approach to Uncertainty in High Dimensional Spaces. In: Borgelt, C., González-Rodríguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis. Advances in Soft Computing. Springer, Berlin (in this book, 2010)
4. Hansen, P.C.: Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion. SIAM Monographs on Mathematical Modeling and Computation, vol. 4. SIAM, Philadelphia (1998)
5. Jolliffe, I.T.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
6. Kamoun, F., Fourati, W., Bouhlel, M.S.: Comparative survey of the DCT and wavelet transforms for image compression. RIST 14–02 (2004)
7. Pearson, K.: Principal components analysis. London, Edinburgh, Dublin Philos. Mag. J. Sci. 6(2), 566 (1901)
8. Rao, K.R., Yip, P.: Discrete Cosine Transform: Algorithms, Advantages, Applications. Academic Press, San Diego (1990)
9. Tompkins, M., Fernández-Martínez, J.L.: Scalable Solutions for Nonlinear Inverse Uncertainty Using Model Reduction, Constraint Mapping, and Sparse Sampling. In: Proceedings of the 72nd EAGE Conference & Exhibition, Barcelona, Spain (2010)

A Linearity Test for a Simple Regression Model with *LR* Fuzzy Response

Maria Brigida Ferraro, Ana Colubi, and Paolo Giordani

Abstract. A linearity test for a simple regression model with an imprecise response is investigated. The values of the imprecise response are formalized through *LR*-fuzzy numbers, and the stochastic variability through probability spaces. The linear regression model and the least squares estimators of the regression parameters are briefly recalled. The nonparametric model to be employed as reference in the testing approach is also presented. The statistic compares the variability explained by the linear regression with the one explained by the nonparametric regression, since in case of linearity, both quantities should be similar. The problem is approached by bootstrapping. A simulation study has been carried out in order to check the performance of the procedure.

Keywords: Fuzzy random variable, Fuzzy regression, Linearity test, Bootstrap approach.

1 Introduction

To formalize an imprecise value, a useful kind of fuzzy numbers is the so-called *LR* family. A linear regression model with an *LR* fuzzy response and a real explanatory variable has been introduced and analyzed in [5, 6].

Maria Brigida Ferraro and Paolo Giordani
Dipartimento di Statistica, Prob. e Stat. Applicate - SAPIENZA
Università di Roma, 00185 Roma, Italy
e-mail: `mariabrigida.ferraro@uniroma1.it,paolo.giordani@uniroma1.it`

Ana Colubi
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,
33007 Oviedo, Spain
e-mail: `colubi@indurot.uniovi.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 263–271.
springerlink.com

Among the inferential procedures in a linear regression context, it can be interesting to check the adequacy of the linear regression for modelling the relationship between the imprecise response and the explanatory variable. For this purpose, it is possible to use an expert criterion or an hypothesis test. The aim of this work is to suggest a test statistic to check the linearity of the relationship and its empirical behaviour.

The proposed linearity test takes inspiration from Azzalini & Bowman [1], who suggest to check the linearity of the relationship by comparing the residuals of the linear regression with those resulting from a nonparametric model. Here, we apply this idea in the context of the regression model with *LR* response taking into account the model in [5]. The hypothesis testing problem is approached by bootstrapping. In details, we propose a residual bootstrap test to check the linearity of the relationship.

In the next section we introduce some preliminary concepts. In Section 3 a linear regression model with *LR* fuzzy response and the estimation problem are recalled, and a nonparametric model is presented. Section 4 deals with the proposed linearity test and, in order to check its performance, simulation studies and a real-life example are carried out in Section 5. Finally, Section 6 contains some concluding remarks.

2 Preliminaries

A fuzzy set $\widetilde{A}$ is identified by the *membership function* $\mu_{\widetilde{A}} : \mathbb{R} \to [0,1]$ so that $\mu_{\widetilde{A}}(x)$ is the membership degree of x in the fuzzy set $\widetilde{A}$ [9]. A particular class of fuzzy sets very useful in practice is the *LR* family, $\mathscr{F}_{LR}$, whose members are the so-called *LR fuzzy numbers*, determined by three values: the center, the left and the right spread (see, for example, [2, 3]). Namely, a mapping $s : \mathscr{F}_{LR} \to \mathbb{R}^3$, i.e., $s(\widetilde{A}) = s_{\widetilde{A}} = (A^m, A^l, A^r)$ (where A^m, $A^l \geq 0$, $A^r \geq 0$ are, respectively, the center, the left and the right spread), is associated to each *LR* fuzzy set $\widetilde{A}$. In what follows it is indistinctly used $\widetilde{A} \in \mathscr{F}_{LR}$ or $(A^m, A^l, A^r) \in \mathbb{R}^3$. The membership function of $\widetilde{A} \in \mathscr{F}_{LR}$ 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 $\mathscr{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for intervals. In details, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathscr{F}_{LR}$ 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 $\widetilde{A} \in \mathscr{F}_{LR}$ by a positive scalar γ is

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

Yang & Ko [8] define a distance between two *LR* fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$ as follows

$$\begin{aligned} D^2_{LR}(\widetilde{A},\widetilde{B}) = (A^m - B^m)^2 &+ [(A^m - \lambda A^l) - (B^m - \lambda B^l)]^2 \\ &+ [(A^m + \rho A^r) - (B^m + \rho B^r)]^2, \end{aligned}$$

where the parameters $\lambda = \int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho = \int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ are related to the shape of the membership function. In the triangular case, $\lambda = \rho = \frac{1}{2}$ (see, for more details, [8]). In order to embed the space $\mathscr{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric a generalization of the Yang and Ko metric has been derived in [5]. Namely, given $a = (a_1, a_2, a_3)$ 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 $\lambda,\ \rho \in \mathbb{R}^+$. According to Puri & Ralescu's sense, the concept of fuzzy random variable (FRV) can be introduced. Let $(\Omega, \mathscr{A}, P)$ be a probability space, a mapping $\widetilde{X} : \Omega \to \mathscr{F}_{LR}$ 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 [7]. The expectation of an *LR* FRV $\widetilde{X}$ is the unique fuzzy set $E(\widetilde{X})$ ($\in \mathscr{F}_{LR}$) such that $(E(\widetilde{X}))_\alpha = E(X_\alpha)$ provided that $E\|\widetilde{X}\|^2_{LR} = E(X^m)^2 + E(X^m - \lambda X^l)^2 + E(X^m + \rho X^r)^2 < \infty$, 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 $X_0 = cl(\{x \in \mathbb{R} | \mu_{\widetilde{X}} \geq 0\})$. In this particular case it results $s_{E(\widetilde{X})} = (E(X^m), E(X^l), E(X^r))$. The variance of $\widetilde{X}$ can be defined as

$$\sigma^2_{\widetilde{X}} = var(\widetilde{X}) = E[D^2_{LR}(\widetilde{X}, E(\widetilde{X}))]$$

and the covariance between two *LR* FRVs $\widetilde{X}$ and $\widetilde{Y}$ is

$$\begin{gathered} \sigma_{\widetilde{X},\widetilde{Y}} = cov(\widetilde{X},\widetilde{Y}) = E\langle s_{\widetilde{X}} - s_{E(\widetilde{X})}, s_{\widetilde{Y}} - s_{E(\widetilde{Y})}\rangle_{LR} \\ = E((X^m - EX^m)(Y^m - EY^m)) \\ + E((X^m - EX^m - \lambda(X^l - EX^l))(Y^m - EY^m - \lambda(Y^l - EY^l))) \\ + E((X^m - EX^m + \rho(X^r - EX^r))(Y^m - EY^m + \rho(Y^r - EY^r))). \end{gathered}$$

3 A Linear Regression Model and a Nonparametric Model with *LR* Fuzzy Random Response and Real Explanatory Variables

Consider a random experiment in which an LR fuzzy response variable $\widetilde{Y}$ and a real explanatory variable X are observed on n statistical units, $\{\widetilde{Y}_i, X_i\}_{i=1,\ldots,n}$. Since $\widetilde{Y}$ is characterized by three real-valued random variables (Y^m, Y^l, Y^r), the regression model proposed in [5] concerns this tuple. The center Y^m can be related to the explanatory variable X through a classical regression model. Due to some difficulties entailed by the non-negativity condition of Y^l and Y^r, the authors proposed to model a transform of the left spread and a transform of the right spread of the response through simple linear regressions (on the explanatory variable X). This can be represented in the following way, letting $g : (0, +\infty) \longrightarrow \mathbb{R}$ and $h : (0, +\infty) \longrightarrow \mathbb{R}$ be invertible:

$$\begin{cases} Y^m = a_m X + b_m + \varepsilon_m, \\ g(Y^l) = a_l X + b_l + \varepsilon_l, \\ h(Y^r) = a_r X + b_r + \varepsilon_r, \end{cases} \tag{2}$$

where ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|X) = E(\varepsilon_l|X) = E(\varepsilon_r|X) = 0$. Concerning the spreads, model (2) is linear in the transformed scales represented by functions g and h.

The variance of the explanatory variable X is denoted by σ_X^2 and $\boldsymbol{\Sigma}$ stands for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, whose variances are strictly positive and finite. In the sequel we will assume the existence of all population variances and covariances involved in the developments.

In general, an LR fuzzy random variable $\widetilde{Y}$ and a (real-valued) random variable X can also be related by means of a nonparametric model. As in (2) we consider jointly three equations in which the response variables are the center Y^m and two transforms of the left and the right spreads ($g(Y^l)$ and $h(Y^r)$) of $\widetilde{Y}$, that is,

$$\begin{cases} Y^m = f_m(X) + \varepsilon_m, \\ g(Y^l) = f_l(X) + \varepsilon_l, \\ h(Y^r) = f_r(X) + \varepsilon_r. \end{cases} \tag{3}$$

To estimate model (2), a least squares (LS) approach has been employed. Let $\widetilde{Y}$ and X be two (fuzzy and real-valued) random variables satisfying model (2) observed on n statistical units, $\{\widetilde{Y}_i, X_i\}_{i=1,\ldots,n}$. It can be shown that the LS estimators for the parameters of model (2) are strongly consistent and their expressions in terms of the sample moments are (see [5])

$$\widehat{a}_m = \frac{\widehat{\sigma}_{XY^m}}{\widehat{\sigma}_X^2}, \quad \widehat{a}_l = \frac{\widehat{\sigma}_{Xg(Y^l)}}{\widehat{\sigma}_X^2}, \quad \widehat{a}_r = \frac{\widehat{\sigma}_{Xh(Y^r)}}{\widehat{\sigma}_X^2}, \quad \widehat{b}_m = \frac{\sum_{i=1}^{n} Y_i^m}{n} - \widehat{a}_m \frac{\sum_{i=1}^{n} X_i}{n},$$

$$\widehat{b}_l = \frac{\sum\limits_{i=1}^{n} g(Y_i^l)}{n} - \widehat{a}_l \frac{\sum\limits_{i=1}^{n} X_i}{n}, \quad \widehat{b}_r = \frac{\sum\limits_{i=1}^{n} h(Y_i^r)}{n} - \widehat{a}_r \frac{\sum\limits_{i=1}^{n} X_i}{n}.$$

Concerning model (3), the functions f_m, f_l and f_r can be estimated in practice by means of nonparametric smoothing. Following [1], a kernel approach can be used yielding

$$\begin{cases} \widehat{f}_m(Z) = \dfrac{\sum\limits_{i=1}^{n} Y_i^m K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \\ \widehat{f}_l(Z) = \dfrac{\sum\limits_{i=1}^{n} g(Y_i^l) K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \\ \widehat{f}_r(Z) = \dfrac{\sum\limits_{i=1}^{n} h(Y_i^r) K((Z-X_i)/w)}{\sum\limits_{i=1}^{n} K((Z-X_i)/w)}, \end{cases} \tag{4}$$

where $K\left(\frac{Z-X_i}{w}\right)$ is a kernel function and w the smoothing parameter. In this case we have used the same w for the three regression models because our aim is not to estimate such a parameter. Nonetheless, in general, three different smoothing parameters can also be considered.

For both the models, the residual sum of squares can be defined as

$$SSE = \sum_{i=1}^{n} D_{\lambda\rho}^2(\widetilde{Y}_i^T, \widehat{\widetilde{Y}^T}), \tag{5}$$

where $\widetilde{Y}_i^T = (Y_i^m, g(Y_i^l), h(Y_i^r))$ and $\widehat{\widetilde{Y}_i}^T = (\widehat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)})$, $i = 1, ..., n$.

4 A Linearity Bootstrap Test

The goal of this section is to test

$$H_0 : \begin{cases} f_m(X) = a_m X + b_m \\ f_l(X) = a_l X + b_l \\ f_r(X) = a_r X + b_r \end{cases} \tag{6}$$

against the alternative

$$H_1 : f_m(X), f_l(X), f_r(X) \text{ are smooth and non-linear functions.}$$

For testing the null hypothesis the following test statistic is used

$$T_n = \frac{SSE_0 - SSE_1}{SSE_1}, \tag{7}$$

where SSE_0 is the residual sum of squares under H_0 according to the model in (2), and SSE_1 is the residual sum of squares according to the model in (3), where $\widehat{\widetilde{Y}^T}_i = (\widehat{Y}_i^m, \widehat{g(Y_i^l)}, \widehat{h(Y_i^r)}) = (\widehat{f}_m(X), \widehat{f}_l(X), \widehat{f}_r(X))$ are the values estimated by means of kernel functions in (4).

Remark 1. We suggest to use a gaussian kernel, that is,

$$K\left(\frac{Z-X_i}{w}\right) = \frac{1}{\sqrt{2\pi}w} exp\left(-\frac{(Z-X_i)^2}{2w^2}\right).$$

In this work we propose to fix the smoothing parameter w. It has been proved that the value of w is expected not to be important since the level of the test is unaffected by this value (see, for instance, [1]). In practice, suitable values of w are from $1/n$ to $1/2$ times the range of the X-values. Nevertheless, the power of the test could be affected by the selection of the smoothing parameter.

A bootstrap approach can be used for testing the linearity. More specifically, we generate B bootstrap samples from a bootstrap population fulfilling the null hypothesis in (6), by means of a residual approach [4]. Then, a standard bootstrap algorithm can be implemented using the bootstrap statistic given by

$$T_n^* = \frac{SSE_0^* - SSE_1^*}{SSE_1^*}.$$

For the sake of convenience, the bootstrap algorithm according to the residual approach is summarized as follows:

Step 1: Compute the values $\widehat{a}_m$, $\widehat{a}_l$, $\widehat{a}_r$, $\widehat{b}_m$, $\widehat{b}_l$, $\widehat{b}_r$ and T_n.

Step 2: Compute the residuals $e_i^m = Y_i^m - \widehat{a}_m X_i - \widehat{b}_m$, $e_i^l = g(Y_i^l) - \widehat{a}_l X_i - \widehat{b}_l$, $e_i^r = h(Y_i^r) - \widehat{a}_r X_i - \widehat{b}_r$.

Step 3: Generate a bootstrap sample of the form

$$\left\{\left(X_1, Z_1^m = \widehat{Y}_1^m + e_{i_1}^m, Z_1^l = \widehat{g(X_1^l)} + e_{i_1}^l, Z_1^r = \widehat{h(X_1^r)} + e_{i_1}^r\right), \ldots, \right.$$
$$\left.\left(X_n, Z_n^m = \widehat{Y}_n^m + e_{i_n}^m, Z_n^l = \widehat{g(X_n^l)} + e_{i_n}^l, Z_n^r = \widehat{h(X_n^r)} + e_{i_n}^r\right)\right\},$$

where $\{i_1, i_2, \ldots, i_n\}$ is a random sample of the integers 1 through n, $\widehat{Y}_i^m = \widehat{a}_m X_i + \widehat{b}_m$, $\widehat{g(X_i^l)} = \widehat{a}_l X_i + \widehat{b}_l$, $\widehat{h(X_i^r)} = \widehat{a}_r X_i + \widehat{b}_r$, $i = 1, \ldots, n$, and compute the value of the bootstrap statistic T_n^*.

Step 4: Repeat Step 3 a large number B of times to get a set of B estimators, denoted by $\{T_{n1}^*, \ldots, T_{nB}^*\}$.

Step 5: Approximate the bootstrap p-value as the proportion of values in $\{T_{n1}^*, \ldots, T_{nB}^*\}$ being greater than T_n.

5 Empirical Results

A simulation experiment has been carried out in order to illustrate the empirical significance of the bootstrap test. Note that we have employed $B = 1000$ replications of the bootstrap estimator and we have considered 10000 iterations of the test at three different nominal significance levels $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$ for different sample sizes n, from 30 to 200. We have randomly generated X behaving as $Unif(-2,2)$, ε_m, ε_l, ε_r as $N(0,1)$ and $Y_m = 3X + 5 + \varepsilon_m$, $Y_2 = g(Y_l) = 1.5X + 3.4 + \varepsilon_l$, $Y_3 = h(Y_r) = 2X + 4.2 + \varepsilon_r$, and we have considered a gaussian kernel with $w = range(X)/n$. The empirical percentages of rejection under H_0 are given in Table 1. It is easy to see that also for small sample sizes n the empirical percentages of rejection are very close to the nominal level. If we consider dependent errors, namely, ε_m behaving as $N(0,1)$ and $\varepsilon_l = \varepsilon_m + \varepsilon_1$, $\varepsilon_r = \varepsilon_m + \varepsilon_2$, with ε_1 and ε_2 behaving as $N(0,0.5)$, we carry out the empirical percentages of rejection under H_0 reported in Table 2. Also in this case, we obtain satisfactory results.

We introduce a real-life example concerning the atmospheric concentration of carbon monoxide (CO) (mg/m^3) and the daily maximum temperature (T) (°C) recorded at “Villa Ada” park in Rome in April, 1-10, 1999 (see Figure 1). The first variable has been managed as a triangular *LR* fuzzy random variable where the center is the mean value of the 24 hourly observations daily recorded, the left spread is given by the deviation of the minimum value from the center and the right spread by the deviation of the maximum value from the center.

In this case we obtain a p-value equal to 0.026, that is, the null hypothesis of linearity should be rejected. Obviously, it should be noted that this result could depend on the choice of the distance, of the kernel function and the smoothing parameter.

Table 1 Empirical percentages of rejection under the hypothesis of linearity.

$n \setminus \alpha \times 100$	1	5	10
30	1.20	5.31	9.94
50	1.17	5.13	10.15
100	1.15	4.82	9.89
200	1.00	4.99	10.20

Table 2 Empirical percentages of rejection under the hypothesis of linearity (dependent errors).

$n \setminus \alpha \times 100$	1	5	10
30	1.14	4.94	9.92
50	1.08	5.12	10.24
100	1.10	5.25	9.94
200	1.12	4.60	9.44

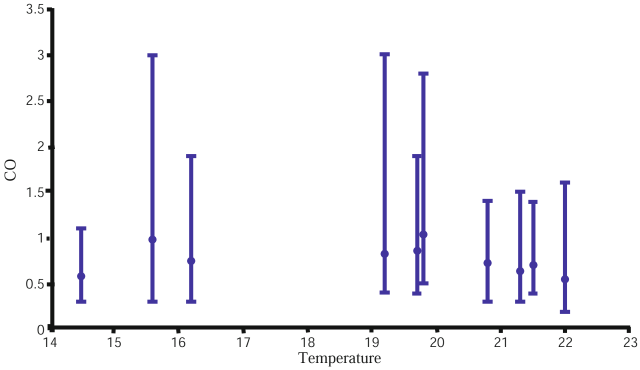


Fig. 1 The observed extreme values of the 0-level and the single-value of CO by the Temperature at "Villa Ada" park in Rome in April, 1-10, 1999

6 Conclusion

In this work we have introduced and analyzed a new linearity test to check the adequacy of a linear relationship between an LR fuzzy response and a real explanatory variable. In order to construct a test statistic, we have jointly considered three equations involving the center of the response and two transforms of the left and the right spread and we have taken into account the residual sum of squares based on a suitable distance between LR fuzzy numbers. The obtained results are as good as expected in this context. In the near future, it will be interesting to study the power of the test.

References

1. Azzalini, A., Bowman, A.: On the use of nonparametric regression for checking linear relationship. J. R. Statist. Soc. Ser. B 55, 549–557 (1993)
2. Coppi, R., D'Urso, P., Giordani, P., Santoro, A.: Least squares estimation of a linear regression model with LR fuzzy response. Comput. Statist. Data Anal. 51, 267–286 (2006)
3. Di Lascio, L., Ginolfi, L., Albunia, A., Galardi, G., Meschi, F.: A fuzzy-based methodology for the analysis of diabetic neuropathy. Fuzzy Sets Syst. 129, 203–228 (2002)
4. Efron, B., Tibshirani, R.J.: An introduction to the bootstrap. Chapman & Hall, New York (1993)
5. Ferraro, M.B., Coppi, R., González-Rodríguez, G., Colubi, A.: A linear regression model for imprecise response. Internat. J. Approx. Reason (2010), doi:10.1016/j.ijar.2010.04.003
6. Ferraro, M.B., Colubi, A., González-Rodríguez, G., Coppi, R.: A determination coefficient for a linear regression model with imprecise response. Environmetrics (accepted for publication, 2010)

7. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
8. Yang, M.S., Ko, C.H.: On a class of fuzzy *c*-numbers clustering procedures for fuzzy data. Fuzzy Sets Syst. 84, 49–60 (1996)
9. Zadeh, L.A.: Fuzzy sets. Inf. Control 8, 338–353 (1965)

Soft Methods in Robust Statistics

Peter Filzmoser

Abstract. The focus is on robust regression methods for problems where the predictor matrix has full rank and where it is rank deficient. For the first situation, various robust regression methods have been introduced, and here an overview of the most important proposals is given. For the latter case, robust partial least squares regression is discussed. The way of downweighting outlying observations is important. Using continuous weights (leading to "soft" robust methods) has advantages over 0/1 weights in terms of statistical efficiency of the estimators. This will be illustrated for both types of regression problems. Soft methods are particularly useful in high-dimensional settings.

Keywords: Robust regression, Partial least squares, High-dimensional data.

1 Introduction

The term "soft computing" was coined by Lotfi Zadeh in 1991, and it refers to the design of intelligent systems to process uncertain, imprecise and incomplete information. Since that time, many methods for soft computing have been developed, and their application offers more robust and tractable solutions than conventional techniques. The term "robust" can be seen under various aspects. In this contribution it will be treated in the light of "robust statistics" which includes statistical approaches that are less influenced by outlying observations and deviations from strict statistical model assumptions [3]. Soft computing, and hence soft methods, are also common practice in this field, and they refer to the way how data information is prepared for the statistical methodology. While classical methods give equal weight

Peter Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology,
1040 Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 273–280.
springerlink.com

to each data point, robust methods downweight atypical observations. The weights could either be chosen as 0 or 1, corresponding to rejecting the observation or not, or continuously in the interval [0,1]. The latter case can be associated with *soft methods in robust statistics.* Such methods should ideally only discard data points if they are extremely distinct from the bulk of the data. In all other cases, the information contained in the data should to some extent be taken into account. The advantage of such a procedure is usually an increase in statistical efficiency of the resulting estimator.

In this contribution we will focus on robust regression. Section 2 provides an overview of the most important proposals and explains the choice of the weight functions. Section 3 contains methods that can be used for high-dimensional problems. Here the choice of the weights is even more important. In section 4 we compare the efficiencies of the robust regression methods by a simulation study.

2 Robust Regression

In a multiple linear regression model we consider the observations $\boldsymbol{y} = (y_1,\ldots,y_n)^t$ of a response variable and an $n \times p$ matrix $\boldsymbol{X}$ of non-random predictor variables with elements x_{ij}. For a regression model with intercept the first column of $\boldsymbol{X}$ is a column of ones. The i-th observation of the predictor variables is denoted by the column vector $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ip})^t$. The linear regression model is then given by

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + e_i \quad \text{for} \quad i = 1,\ldots,n, \tag{1}$$

with $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^t$ the unknown regression coefficients, and e_i the error terms which are assumed to be i.i.d. random variables. The goal is to estimate the regression coefficients. For a given estimator $\hat{\boldsymbol{\beta}}$ the resulting i-th residual is $r_i = r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}$. The classical least squares (LS) estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 . \tag{2}$$

Under the assumptions of normally distributed errors with the same variance, and if $\boldsymbol{X}$ has full rank, this estimator is known to have excellent statistical properties. However, if the assumptions are violated, and in particular if outliers are contained either in the response, in the predictors, or in both, the performance of the LS estimator can be very poor [3].

2.1 Regression M Estimates

For this reason, the M estimator for regression was introduced as

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \tag{3}$$

where $\hat{\sigma}$ is a robust scale estimator of the residuals, which makes the regression estimator scale equivariant [2]. The function ρ controls the weighting of the scaled residuals, and it needs to be chosen carefully. It should be a bounded function such that very large residuals will have a limited influence on the estimator. A popular choice is the *bisquare* (also called *biweight*) family, with

$$\rho(r) = \begin{cases} \left(\frac{r}{k}\right)^2\left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & \text{for} \quad |r| \leq k \\ 1 & \text{otherwise} \end{cases}. \tag{4}$$

The value k is a tuning parameter, balancing efficiency and robustness. For $k \to \infty$, the corresponding estimate tends to LS and hence it becomes more efficient but at the same time less robust. Differentiation of (3) with respect to $\boldsymbol{\beta}$ gives a robustified version of the normal equations,

$$\sum_{i=1}^{n} w_i(\boldsymbol{\beta})(y_i - \boldsymbol{x}_i^t\boldsymbol{\beta})\boldsymbol{x}_i = \mathbf{0} \tag{5}$$

with the weights $w_i(\boldsymbol{\beta}) = \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) / \left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right)$ where $\psi = \rho'$. The solution can be found by the IRWLS (iteratively reweighted least squares) algorithm. However, the resulting estimator is only robust with respect to outliers in the residuals, but it is still not robust against outliers in the predictor variables. This can be seen in the definition of the weights w_i, where only outliers in the residual space are considered. The crucial point is the way how the residual scale ($\hat{\sigma}$ in Equation (3)) is estimated.

2.2 Regression S Estimates

A possibility to estimate the residual scale is to use an *M estimator of scale*, which is defined as the solution σ of the equation

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right) = \delta, \tag{6}$$

where ρ is a bounded ρ-function (e.g. the bisquare function) and δ is a fixed constant with $\delta \in (0, \rho(\infty))$. Dividing Equation (6) by $(r_i/\sigma)^2$ yields

$$\sigma^2 = \frac{1}{n\delta}\sum_{i=1}^{n} \frac{\rho\left(\frac{r_i}{\sigma}\right)}{\left(\frac{r_i}{\sigma}\right)^2} r_i^2 = \frac{1}{n\delta}\sum_{i=1}^{n} w_i r_i^2 \tag{7}$$

with weights $w_i = \rho\left(\frac{r_i}{\sigma}\right) / \left(\frac{r_i}{\sigma}\right)^2$. Given some starting value σ_0, an iterative procedure can be implemented to find the M estimator of scale $\hat{\sigma}$. Using this robust scale estimator, a robust regression estimator can be defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{S}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}\big(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})\big) \tag{8}$$

resulting in the *regression S estimator* [1]. In can be shown that regression S estimators satisfy Equation (5), which implies that they can be computed by an IRWLS algorithm. Although regression S estimators achieve highest possible robustness, the efficiency of this estimator with ρ taken as the bisquare function is only 29%, and in general it cannot exceed 33%.

2.3 Regression MM Estimates

A way to obtain the highest possible robustness with controllable efficiency is given by *regression MM estimators* [8]. The procedure for the computation is as follows [5]:

- Compute an initial estimator $\hat{\boldsymbol{\beta}}_0$; this is done by a regression S estimator (8) which is robust but inefficient.
- Compute a robust scale $\hat{\sigma}$ of the residuals $r_i(\hat{\boldsymbol{\beta}}_0)$; this is done by an M estimator of scale (6).
- Compute $\hat{\boldsymbol{\beta}}_{\mathrm{MM}}$ as a local solution of (3) using the IRWLS algorithm starting from $\hat{\boldsymbol{\beta}}_0$. The resulting MM estimator inherits its robustness from $\hat{\boldsymbol{\beta}}_0$, and the efficiency can be controlled by the parameter k from the bisquare function (4). Using $k = 3.44$ in this step yields an asymptotic efficiency of 0.85. A higher value is not recommended because this would lead to an increase of the bias [3].

2.4 Hard Rejection of Outliers for Regression

Regression MM estimators use weights for the observations from the interval $[0,1]$. The further the weights are away from 1, the less information is used from these observations. A popular regression estimator using weights of 0 and 1 for hard rejection of outliers is the LTS (least trimmed sum of squares) estimator [4]. Similar to Equation (7), this estimator minimizes a measure of scale, namely the *trimmed squares scale*

$$\sigma = \left(\frac{1}{n}\sum_{i=1}^{h} |r|^2_{(i)}\right)^{1/2}, \tag{9}$$

where $|r|_{(1)} \leq \ldots \leq |r|_{(n)}$ are the ordered absolute values of the residuals. Here, h determines the trimming proportion, and for obtaining the highest possible robustness one has to take h equal to (the integer part of) $(n+p+1)/2$.

Similar to Equation (8), the LTS estimator $\hat{\boldsymbol{\beta}}_{\mathrm{LTS}}$ is given by $\hat{\sigma}$ that results from minimizing (9). The asymptotic efficiency of the LTS estimator is only about 7%. Thus, although hard rejection of outliers results in a robust estimator, the efficiency is much lower than that of the MM estimator which uses “soft” weights corresponding to the “useful” data information.

3 Partial Robust Regression

There exist many problems where the number of the explanatory variables is much higher than the number of observations. This situation frequently occurs in chemometrics, biostatistics, in applications of marketing and econometrics, and in various other fields. Because of singularity, neither the LS estimator could be used here, nor any of the discussed robust regression methods. Partial least squares (PLS) regression, a method originally coming from chemometrics, can deal with this situation, see, e.g. [7]. The idea is to use only partial information for regression. Hence, rather than considering the regression model (1), a so-called latent variable model

$$y_i = \boldsymbol{u}_i^t \boldsymbol{\gamma} + e_i \quad \text{for} \quad i = 1, \ldots, n, \tag{10}$$

is used, where $\boldsymbol{u}_i$ are *score vectors* of length $h < p$, $\boldsymbol{\gamma}$ are the regression coefficients, and e_i the error terms. The scores $\boldsymbol{u}_i$ include only partial information contained in the original $\boldsymbol{x}_i$'s because they are of lower dimension. They are computed by $\boldsymbol{u}_i^t = \boldsymbol{x}_i^t \boldsymbol{A}$, with the so-called *loading matrix* $\boldsymbol{A}$ of dimension $p \times h$. The columns $\boldsymbol{a}_k$, $k = 1, \ldots, h$, of $\boldsymbol{A}$ are obtained sequentially by

$$\boldsymbol{a}_k = \arg\max_{\boldsymbol{a}} \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{a}) \tag{11}$$

under the constraints $\|\boldsymbol{a}\| = 1$ and $\mathrm{Cov}(\boldsymbol{X}\boldsymbol{a}, \boldsymbol{X}\boldsymbol{a}_{\boldsymbol{j}}) = 0$ for $1 \le j < k$. Once $\hat{\boldsymbol{\gamma}}$ is obtained, the final estimate for $\boldsymbol{\beta}$ for the original model (1) is directly obtained as $\hat{\boldsymbol{\beta}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}$.

The crucial point is the estimation of ‘Cov’ in Equation (11). For classical PLS regression, the sample covariance is used. For the robust case, several proposals were made, including robust covariance estimation, see [7]. Here we refer to a highly robust and efficient method called partial robust M regression [6]. The idea is to use for ‘Cov’ the sample covariance for weighted observations $w_i \boldsymbol{x}_i$ and $w_i y_i$ with weights $w_i = \sqrt{w_i^u w_i^r}$, for $i = 1, \ldots, n$. In terms of the latent variable model (10), the weights originate from

$$\hat{\boldsymbol{\gamma}}_{\mathrm{RM}} = \arg\min_{\boldsymbol{\gamma}} \sum_{i=1}^{n} w_i^u w_i^r \left(y_i - \boldsymbol{u}_i^t \boldsymbol{\gamma}\right)^2 . \tag{12}$$

‘RM’ stands for *robust M regression*, because Equation (12) corresponds to an M estimator (3) with weights w_i^r for outliers in the residuals, but has additional weights w_i^u for outliers in the scores. The latter weights make the estimator fully robust against all types of contamination. The weights can be chosen according to the so-called Fair function $f(z,c) = 1/\left(1 + |\frac{z}{c}|\right)^2$, where

$$w_i^r = f\left(\frac{r_i}{\hat{\sigma}}, c\right) \quad \text{and} \quad w_i^u = f\left(\frac{\|\boldsymbol{u}_i - \tilde{\boldsymbol{u}}\|}{\text{median}_i \|\boldsymbol{u}_i - \tilde{\boldsymbol{u}}\|}, c\right) \tag{13}$$

with $c = 4$, see [6]. Here, $r_i = r_i(\boldsymbol{\gamma}) = y_i - \boldsymbol{u}_i^t \boldsymbol{\gamma}$ are the residuals from (12), $\hat{\sigma}$ is a robust scale estimate of the residuals, and $\tilde{\boldsymbol{u}}$ denotes the robust center of the scores. Using initial robust weights, an iterative procedure can be formulated to obtain the solution $\hat{\boldsymbol{\beta}}_{\text{RM}} = \boldsymbol{A}\hat{\boldsymbol{\gamma}}_{\text{RM}}$, see [6]. The need for an iterative procedure is also the reason why this rather simple weighting scheme is recommended. An MM estimator would achieve higher efficiency, but–depending on the dimensionality of the problem–it would cause a substantial increase in computation time.

The weights in (13) are chosen from the interval $[0,1]$, and thus this is another example of "soft weighting". It is easy to modify the weights in order to get hard rejection of the outliers by replacing $f(z,c)$ in (13) by

$$\tilde{f}(z,c) = \begin{cases} 1 & \text{if } |z| \le c \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

with $c = 2.5$. The resulting estimator has the advantage that large values of $|z|$ have no effect, but the disadvantage that intermediate outliers are either completely rejected of fully included.

4 Soft versus Hard Rejection: A Simulation Study

The use of a continuous weight function, or of weights 0 and 1, will affect the efficiency of the regression estimator. It seems obvious that soft rejection, i.e. the use of "soft" weights, is able to include information that is potentially relevant to improve the statistical precision of the estimator, while hard rejection may fail to use this information. Note that with both types of weighting schemes it is possible to achieve highest possible robustness.

In the following simulation study the effects of different choices of the weights on the efficiency of the estimators will be illustrated. For the regression model (1) we generate standard normally distributed values, forming the elements of the $n \times p$ matrix $\boldsymbol{X}$. For the latent variable model (10) an $n \times h$ score matrix $\boldsymbol{U}$ and a $p \times h$ loading matrix $\boldsymbol{A}$ are generated, both filled with random standard normal numbers, and the predictor matrix is obtained by $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{A}^t$. Thus, for $h < p$ a situation with perfect collinearity is simulated. For each considered n and p, the predictor part is fixed. For both cases, the true regression parameters are denoted by $\boldsymbol{\beta}_0$, with components randomly drawn from a standard normal distribution, leading to a model

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta}_0 + e_i \quad \text{for} \quad i = 1, \ldots, n. \tag{15}$$

The error terms e_i are simulated from various distributions: standard normal, Laplace, Student t with 5 and 2 degrees of freedom, Cauchy, and Slash. The latter two are heavy-tailed distributions. From every generated sample with specific values of n, p, and h (for the latent variable model), the estimate $\hat{\boldsymbol{\beta}}^j$ is computed for $j = 1, \ldots, m$, using $m = 1000$ replications. The precision of the estimator is measured by the mean squared error (MSE), given by

$$\mathrm{MSE} = \frac{1}{m} \sum_{j=1}^{m} \|\hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}_0\|^2. \tag{16}$$

The results are shown in Figure 1 (for the regression model) and in Figure 2 (for the latent variable model). For the regression model we compare LS, LTS, S, and MM estimation. The LS estimator performs very poor under heavier-tailed distributions, while the robust regression methods are not much affected by the different error distributions. Overall, the MM estimator shows the best efficiency among the robust estimators, and it is able to compete with LS regression under normal errors.

For the latent variable model we compare in Figure 2 the results of classical estimation (PLS) and robust estimation using the weight function (13) for soft rejection (PRM) and the weights (14) for hard rejection (PRM01). Again, classical estimation dramatically fails for heavy-tailed error distributions. The efficiencies based on hard and soft weighting differ more and more with increasing dimensionality of the predictor matrix: while they differ by a factor of 1.1 to 1.8 for dimensions up to $p = 20$, the ratio increases to a value of 2.5 to 2.8 for $p = 1000$. “Intelligent” robustness–in contrast to robustness based on outlier rejection–thus becomes particularly important for high-dimensional problems, which occur frequently nowadays in practice.

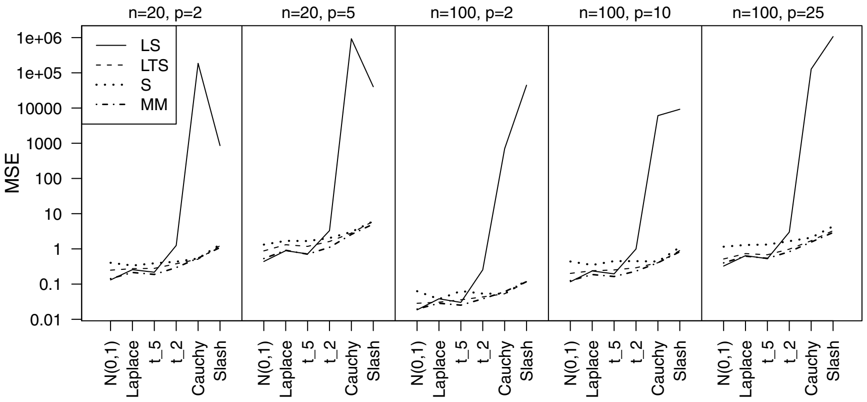


Fig. 1 Simulated MSEs for LS, LTS, S, and MM regression, using different error distributions (legend on the bottom), and different dimensions of the predictor matrix (legend on top).

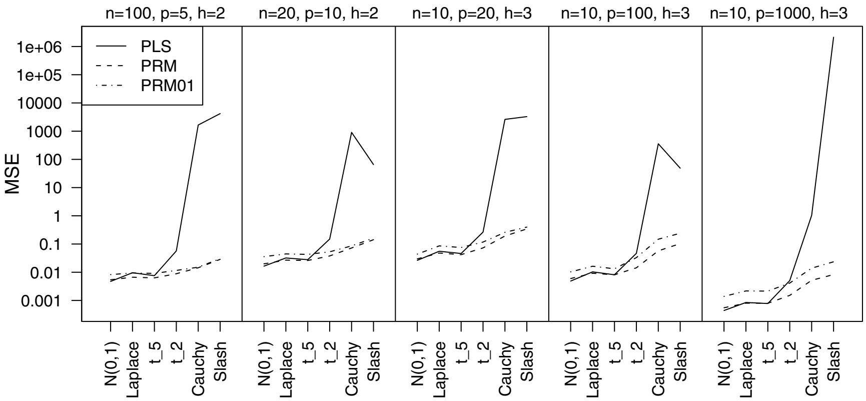


Fig. 2 Simulated MSEs for classical (PLS) and robust partial least squares regression based on soft (PRM) and hard rejection (PRM01), using different error distributions (legend on the bottom), and different dimensions and ranks of the predictor matrix (legend on top).

References

1. Davies, P.L.: Asymptotic behaviour of *S*-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. 15(3), 1269–1292 (1987)
2. Huber, P.J.: Robust statistics. John Wiley & Sons, New York (1981)
3. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust statistics: theory and methods. John Wiley & Sons Canada Ltd., Toronto (2006)
4. Rousseeuw, P.J.: Least median of squares regression. J. Amer. Statist. Assoc. 79(388), 871–880 (1984)
5. Salibian-Barrera, M., Yohai, V.J.: A fast algorithm for S-regression estimates. J. Comput. Graph. Statist. 15(2), 414–427 (2006)
6. Serneels, S., Croux, C., Filzmoser, P., Van Espen, P.J.: Partial robust M-regression. Chemometr. Intell. Lab. 79(1-2), 55–64 (2005)
7. Varmuza, K., Filzmoser, P.: Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton (2009)
8. Yohai, V.J.: High breakdown-point and high efficiency robust estimates for regression. Ann. Statist. 15(2), 642–656 (1987)

S-Statistics and Their Basic Properties

Marek Gągolewski and Przemysław Grzegorzewski

Abstract. Some statistical properties of the so-called S-statistics, which generalize the ordered weighted maximum aggregation operators, are considered. In particular, the asymptotic normality of S-statistics is proved and some possible applications in estimation problems are suggested.

Keywords: Aggregation, L-statistics, OWA, OWMax operators.

1 Introduction

The process of aggregation, i.e. combining many numerical values into a single one, plays an important role in many areas of practical human activities, such as statistics, decision making, computer science, operational research, etc. Operators projecting multidimensional state space into a single dimension are often called *aggregation functions* [5]. Among well-known examples are: the sample maximum and other quantiles, arithmetic mean, ordered weighted averaging (OWA) [11] and ordered weighted maximum (OWMax) [2] operators.

The OWA operators are a particular case of L-statistics. Their basic statistical properties were widely discussed, see e.g. [7, 10].

In this paper we consider another useful class of aggregation operators called S-statistics, which generalize OWMax. We show that S-statistics are consistent estimators of the so-called κ-index (Sect. 3). Moreover, they are asymptotically normally distributed (Sect. 4). Regarding similar constructions it seems that

Marek Gągolewski and Przemysław Grzegorzewski
Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland
e-mail: gagolews,pgrzeg@ibspan.waw.pl

Marek Gągolewski and Przemysław Grzegorzewski
Faculty of Mathematics and Information Science, Warsaw University of Technology, 00-661 Warsaw, Poland

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 281–288.
springerlink.com

S-statistics would be useful in many situations, e.g. in scientometrics to construct reliable tools for scientific activity assessment (see [3, 4, 9]), pattern matching [2] and decision making [1].

2 S-Statistics

Let $(X_1,\ldots,X_n)$ denote a sample of i.i.d. random variables, while $X_{(1)},\ldots,X_{(n)}$ are order statistics corresponding to this sample. Assume that the variance of X_i is finite and the c.d.f. F of X_i is continuous and strictly increasing in interval (a,b), where $a = \inf\{x : F(x) > 0\}$, $b = \sup\{x : F(x) < 1\}$.

Let $\kappa : [0,1] \to [a,b]$ be a strictly increasing function such that $\kappa(0) = a$ and $\kappa(1) = b$. Further on we will call such function a *control function*.

A linear combination of order statistics, called L-statistics, is a well-known tool applied especially in robust estimation or testing. Typical examples of L-statistics are trimmed and Winsorized means that are useful in situations when data follow a heavy-tailed distribution. Its subclass is known in decision making as the ordered weighted averaging (OWA) operator [11]. Below we propose another function of ordered statistics which has some interesting statistical properties.

Definition 1. *An* S-statistic *associated with a control function* κ *and a random sample* $(X_1,\ldots,X_n)$ *is a function*

$$V_{n,\kappa}(X_1,\ldots,X_n) = \bigvee_{i=1}^{n} \kappa\left(\tfrac{i}{n}\right) \wedge X_{(n-i+1)}, \tag{1}$$

where $\vee$ *and* $\wedge$ *denote the supremum (hence the name) and infimum operators, respectively.*

It can be seen that the S-statistic is a generalization of the ordered weighted maximum operator (OWMax) defined firstly in [2]. Moreover, for any control function κ, the corresponding S-statistic is a function $V_{n,\kappa} : [a,b]^n \to [a,b]$ which satisfies the following requirements:

1. $V_{n,\kappa}$ is non-decreasing in each variable, i.e. $(\forall \mathbf{x},\mathbf{y} \in [a,b]^n)\ \mathbf{x} \le \mathbf{y} \Rightarrow V_{n,\kappa}(\mathbf{x}) \le V_{n,\kappa}(\mathbf{y})$,
2. $V_{n,\kappa}$ fulfills the lower boundary condition, i.e. $\inf_{\mathbf{x}\in[a,b]^n} V_{n,\kappa}(\mathbf{x}) = a$,
3. $V_{n,\kappa}$ fulfills the upper boundary condition, i.e. $\sup_{\mathbf{x}\in[a,b]^n} V_{n,\kappa}(\mathbf{x}) = b$.

Therefore, according to the definition given e.g. in [5], $V_{n,\kappa}$ is an aggregation function. Hence, $V_{n,\kappa}$ may have (at least potentially) — like other aggregation functions — many applications in different areas. In this paper we restrict ourselves to their statistical properties related to their asymptotic distribution and estimation of a population location parameter.

Note that

$$V_{n,\kappa}(X_1,\ldots,X_n) = \kappa\left(\bigvee_{i=1}^{n} \tfrac{i}{n} \wedge \kappa^{-1}\left(X_{(n-i+1)}\right)\right). \tag{2}$$

Hence, without loss of generality, we will consider S-statistics of a form

$$V_n(Y_1,\ldots,Y_n)=\bigvee_{i=1}^{n}\tfrac{i}{n}\wedge Y_{(n-i+1)}, \tag{3}$$

where $(Y_1,\ldots,Y_n)=(\kappa^{-1}(X_1),\ldots,\kappa^{-1}(X_n))$ is a sequence of i.i.d. random variables given by the continuous c.d.f. $G:=F\circ\kappa$ defined on $[0,1]$. In other words, $V_n:=V_{n,\mathrm{id}}$, where id is the identity function.

3 κ-index

Consider the following definition.

Definition 2. *A* κ-index *of a random variable given by a c.d.f. G with respect to the control function κ is a number $\rho_\kappa\in[0,1]$ such that*

$$\rho_\kappa=1-G(\kappa(\rho_\kappa)). \tag{4}$$

If $S(x)=1-G(x)$ is a survival function then, of course, a κ-index ρ_κ satisfies

$$\rho_\kappa=S(\kappa(\rho_\kappa))=\Pr(X>\kappa(\rho_\kappa)). \tag{5}$$

Thus κ-index has an intuitive interpretation: it is such a number that the probability of assuming a value greater than $\kappa(\rho_\kappa)$ is equal to ρ_κ.

Example 1. If Y follows the Type-II Pareto distribution, i.e. $G(x)=1-1/(1+x)$ and the control function is the identity function, i.e. $\kappa(x)=x$, then $\rho_\kappa=(\sqrt{5}-1)/2=1/\varphi=\varphi-1\simeq 0.618034$, where φ is the *golden ratio.*

It appears that the S-statistic is a strongly consistent estimator of the id-index $\rho:=\rho_{\mathrm{id}}$ for any c.d.f. G defined on $[0,1]$. However, to prove it we need some lemmas given below.

Lemma 1. *For any sample $Y_1,\ldots,Y_n$ of i.i.d. random variables defined on $[0,1]$ with a continuous c.d.f. G we have*

$$\begin{aligned} V_n(Y_1,\ldots,Y_n) &= \inf\left\{x:\hat{G}_n(x)\geq 1-x\right\} && (6)\\ &= \sup\left\{x:\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i\geq x)\geq x\right\}, && (7)\end{aligned}$$

where $\hat{G}_n(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i\leq x)$ denotes the empirical distribution function and $\mathbf{1}$ is the indicator function.

Proof. Since $\sum_{i=1}^{n}\mathbf{1}(Y_i\geq x)=\max\{i:Y_{(n-i+1)}\geq x\}$ we get

$$\begin{aligned} V_n(Y_1,\ldots,Y_n) &= \max\{\tfrac{i}{n}:\tfrac{i}{n}\leq Y_{(n-i+1)}\}\vee\max\{Y_{(n-i+1)}:Y_{(n-i+1)}\leq\tfrac{i}{n}\}\\ &= \max\left\{x:\tfrac{1}{n}\max\{i:Y_{(n-i+1)}\geq x\}\geq x\right\}.\end{aligned}$$

Implication from (6) to (7) is obvious and the proof is complete. □

Recall that $(\forall x)$ we have $\hat{G}_n(x)\overset{a.s.}{\to}G(x)$ and $n\hat{G}_n(x)\sim\mathrm{Bin}(n,G(x))$.

The exact distribution of V_n is given by the next lemma.

Lemma 2. *The c.d.f. of* $V_n(Y_1,\ldots,Y_n)$ *is given by*

$$D_n(x) = 1 - \sum_{i=\lfloor xn+1\rfloor}^{n} \binom{n}{i} [1-G(x)]^i [G(x)]^{n-i} \tag{8}$$

$$= I(G(x); n-\lfloor xn\rfloor, \lfloor xn\rfloor+1) \tag{9}$$

for $x\in[0,1)$*, where* $I(p;a,b)$ *is the regularized Euler beta function and* $\lfloor y\rfloor := \max\{i\in\mathbb{N} : i\le y\}$ *is the floor function.*

Proof. The c.d.f. of the ith order statistic $Y_{i:n}$, $i=1,2,\ldots,n$, is given by

$$\begin{aligned} G_{i:n}(x) &= \Pr(Y_{i:n}\le x) \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}\int_0^{G(x)} t^{i-1}(1-t)^{n-i}\,dt \\ &= I(G(x);i,n-i+1). \end{aligned}$$

Note that V_n (by Lemma 1) is equal to the greatest number such that $\lceil nV_n\rceil = \min\{i\in\mathbb{N} : i\ge nV_n\}$ observations are not less than V_n. Hence

$$\Pr(V_n > x) = \Pr(Y_{n-\lfloor xn+1\rfloor:n} > x) = 1 - I(G(x); n-\lfloor xn\rfloor, \lfloor xn\rfloor+1),$$

and the lemma follows immediately. □

Lemma 3. *For any* $x\in(0,1)$ *we have*

$$\Pr(V_n > x) = \Pr(1-x > \hat{G}_n(x)). \tag{10}$$

Proof. Since $n\hat{G}_n(x)\sim \mathrm{Bin}(n,G(x))$ then for any $t\in(0,n)$

$$\Pr(n\hat{G}_n(x)\le t) = I(1-G(x), n-\lfloor t\rfloor, 1+\lfloor t\rfloor).$$

Now, by Lemma 2, we get for any $x\in(0,1)$

$$\begin{aligned} \Pr(V_n > x) &= 1 - I(G(x); n-\lfloor xn\rfloor, \lfloor xn\rfloor+1) \\ &= I(1-G(x); \lfloor xn\rfloor+1, n-\lfloor xn\rfloor) \\ &= I(1-G(x); n-(n-\lfloor xn\rfloor-1), 1+(n-\lfloor xn\rfloor-1)) \\ &= \Pr(n\hat{G}_n(x)\le n-(\lfloor xn\rfloor+1)) \\ &= \Pr(\hat{G}_n(x) < 1-x), \end{aligned}$$

which holds because $\lfloor xn\rfloor \le xn < \lfloor xn\rfloor+1$. Thus the proof is complete. □

The following lemma (see [6]) will be also useful.

Lemma 4 (Hoeffding's inequality). *Let* $(Z_1,\ldots,Z_n)$ *be a sequence of independent random variables with finite second moments and let* $0\le Z_i\le 1$ *for* $i=1,\ldots,n$*. Then for any* $t>0$ *the following inequality holds*

$$\Pr\left(\tfrac{1}{n}\sum_{i=1}^{n} Z_i - \tfrac{1}{n}\mathbb{E}\sum_{i=1}^{n} Z_i \ge t\right) \le e^{-2nt^2}. \tag{11}$$

The next lemma shows that the S-statistic V_n converges to ρ exponentially fast.

Lemma 5. *For any $n \in \mathbb{N}$ and $\varepsilon > 0$*

$$\Pr(|V_n - \rho| > \varepsilon) \leq 2e^{-2n\delta^2}, \tag{12}$$

where $\delta = G(\rho+\varepsilon) - (1-(\rho+\varepsilon)) \wedge 1 - (\rho-\varepsilon) - G(\rho-\varepsilon)$.

Proof. It is worth noticing that the proof of this lemma would be analogous to that of Theorem 2.3.2 [7] where a similar result on sample quantiles is discussed. For any $\varepsilon > 0$ we get (by Lemmas 3 and 4)

$$\begin{aligned}
\Pr(V_n > \rho+\varepsilon) &= \Pr(1-\rho-\varepsilon > \hat{G}_n(\rho+\varepsilon)) \\
&= \Pr\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i > \rho+\varepsilon) > \rho+\varepsilon\right) \\
&= \Pr\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i > \rho+\varepsilon) - (1-G(\rho+\varepsilon))\right. \\
&\qquad \left. > \rho+\varepsilon-(1-G(\rho+\varepsilon))\right) \\
&= \Pr\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i > \rho+\varepsilon) - \tfrac{1}{n}\mathbb{E}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i > \rho+\varepsilon)\right. \\
&\qquad \left. > G(\rho+\varepsilon)+\rho+\varepsilon-1\right) \\
&\leq \exp\left\{-2n\delta_1^2\right\}.
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
\Pr(V_n < \rho-\varepsilon) &\leq \Pr(1-\rho+\varepsilon \leq \hat{G}_n(\rho-\varepsilon)) \\
&= \Pr\left(\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbf{1}(Y_i \leq \rho+\varepsilon) - G(\rho-\varepsilon)\right. \\
&\qquad \left. \geq 1-(\rho-\varepsilon)-G(\rho-\varepsilon)\right) \\
&\leq \exp\left\{-2n\delta_2^2\right\}
\end{aligned}$$

for $\delta_2 = 1-(\rho-\varepsilon)-G(\rho-\varepsilon)$.

Hence

$$\Pr(|V_n-\rho| > \varepsilon) = \Pr(V_n > \rho+\varepsilon) + \Pr(V_n < \rho-\varepsilon) \leq 2\exp\{-2n(\min\{\delta_1,\delta_2\})^2\},$$

which completes the proof. □

Now we are ready to prove the desired result.

Theorem 1. *V_n is a strongly consistent estimator of ρ.*

Proof. $\Pr(|V_n-\rho| > \varepsilon) \to 0$ exponentially fast (by Lemma 5) w.r.t. n and therefore we get $V_n \overset{a.s.}{\to} \rho_\kappa$ (by Theorem 1.3.4 in [7]). □

4 Asymptotic Distribution of S-statistics

Unfortunately, the practical usage of the exact distribution (9) may sometimes be problematic. Therefore we are seriously interested in its approximation. In the present section we consider the asymptotic distribution of an S-statistic.

Let us also cite a well-known result that will be needed for proving the next theorem.

Lemma 6 (Berry-Esséen). *Let $Z_1, Z_2, \ldots$ denote a sequence of i.i.d. random variables with a finite expectation μ and finite variance σ^2 and such that $(\forall i)$ $\mathbb{E}\,|Z_i - \mu|^3 < \infty$. Then for all $n \in \mathbb{N}$*

$$\sup_x |H_n(x) - \Phi(x)| \le C \frac{\mathbb{E}\,|Z_1 - \mu|^3}{\sigma^3 \sqrt{n}}, \tag{13}$$

where

$$H_n(x) = \Pr\left(\frac{\sum_{i=1}^n Z_i - n\mu}{\sigma\sqrt{n}} \le x\right),$$

$\Phi(x)$ denotes the c.d.f. of the standard normal distribution and C is a positive constant independent of the distribution of Z_i.

This lemma characterizes the rate of convergence in the Lindeberg-Lévy Central Limit Theorem. Let us mention that the best currently known upper bound for C is 0,7056 (see [8]). Now we can present the asymptotic distribution of the S-statistic.

Theorem 2. *If G is a c.d.f. differentiable at ρ, then*

$$V_n \xrightarrow{D} \mathrm{N}\left(\rho, \frac{1}{1 + G'(\rho)} \sqrt{\frac{\rho(1-\rho)}{n}}\right). \tag{14}$$

Proof. Let $x \in (0,1)$ and $A > 0$ be a positive constant which will be determined later. Let

$$K_n(x) = \Pr\left(\frac{V_n - \rho}{A}\sqrt{n} \le x\right).$$

We will show that $K_n(x) \to \Phi(x)$ as $n \to \infty$.

By Lemma 3 we have

$$\begin{aligned} K_n(x) &= \Pr\left(V_n \le \rho + xAn^{-0,5}\right) \\ &= \Pr\left(1 - \rho - xAn^{-0,5} \le \hat{G}_n(\rho + xAn^{-0,5})\right). \end{aligned}$$

Assuming that $\Delta_{n,x} := \rho + xAn^{-0,5}$ and recalling that $n\hat{G}_n(\Delta_{n,x}) \sim \mathrm{Bin}(n, G(\Delta_{n,x}))$ we obtain

$$K_n(x) = \Pr\left(\frac{n\hat{G}_n(\Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}} \ge \frac{n(1 - \Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}}\right).$$

Substituting $Z^*_{n,x}$ and $\zeta_{n,x}$ given by

$$Z^*_{n,x} = \frac{n\hat{G}_n(\Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1 - G(\Delta_{n,x}))}}$$

$$\zeta_{n,x} = \frac{n(1-\Delta_{n,x}) - nG(\Delta_{n,x})}{\sqrt{nG(\Delta_{n,x})(1-G(\Delta_{n,x}))}}$$

into the previous equation we get $K_n(x) = \Pr(Z^*_{n,x} \geq \zeta_{n,x})$.

If $Z_1 \sim \mathrm{Bern}(G(\Delta_{n,x}))$, then $\mathbb{E}|Z_1 - \mathbb{E}Z_1|^3 = G(\Delta_{n,x})(1 - G(\Delta_{n,x}))((1 - G(\Delta_{n,x}))^2 + G(\Delta_{n,x})^2)$ and $\mathrm{Var}Z_1 = G(\Delta_{n,x})(1-G(\Delta_{n,x}))$.

By Lemma 6 for some $C > 0$ we obtain

$$\left|\Pr\left(Z^*_{n,x} < \zeta_{n,x}\right) - \Phi(\zeta_{n,x})\right| \leq \frac{C}{\sqrt{n}} \frac{(1-G(\Delta_{n,x}))^2 + G(\Delta_{n,x})^2}{\sqrt{G(\Delta_{n,x})(1-G(\Delta_{n,x}))}} \overset{n\to\infty}{\to} 0,$$

because $G(\Delta_{n,x})(1-G(\Delta_{n,x})) \overset{n\to\infty}{\to} (1-\rho)\rho > 0$, and since G is continuous at ρ. Finally we have

$$\begin{aligned} |\Phi(x) - K_n(x)| &= |\Pr(Z^*_n < \zeta_{n,x}) - (1-\Phi(x))| \\ &= |\Phi(x) - \Phi(-\zeta_{n,x}) + \Pr(Z^*_n < \zeta_{n,x}) - \Phi(\zeta_{n,x})| \\ &\leq |\Phi(x) - \Phi(-\zeta_{n,x})| + |Pr(Z^*_n < \zeta_{n,x}) - \Phi(\zeta_{n,x})| \\ &\to |\Phi(x) - \Phi(-\zeta_{n,x})|. \end{aligned}$$

Since our theorem will be proved when $|\Phi(x) - \Phi(-\zeta_{n,x})| \to 0$ we would determine A in such way that $-\zeta_{n,x} \to x$. It is seen that

$$\begin{aligned} -\zeta_{n,x} &= \frac{1}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}} \frac{1-\Delta_{n,x} - G(\Delta_{n,x})}{n^{-0,5}} \\ &= \frac{xA}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}} \frac{1-\rho-xAn^{-0,5} - G(\rho + xAn^{-0,5})}{xAn^{-0,5}} \\ &= -\frac{xA}{\sqrt{G(\Delta_{n,x})\,(1-G(\Delta_{n,x}))}} \frac{G(\rho + xAn^{-0,5}) - G(\rho) + xAn^{-0,5}}{xAn^{-0,5}} \\ &\overset{n\to\infty}{\to} -\frac{xA}{\sqrt{(1-\rho)\rho}} \left(G'(\rho) + 1\right) \end{aligned}$$

and hence our desired $A = \sqrt{\rho\,(1-\rho)}/(1+G'(\rho))$, QED. □

Note that Theorem 2 implies that V_n is (weakly) consistent. In practice, D_n approaches the normal distribution D^*_n quite quickly. For example if G is a beta distribution $B(0.5, 0.5)$ ($\rho = 0.5$) then for $n = 30$ we have $||D_n - D^*_n||_2 \simeq 0.013$ and $\max|D_n - D^*_n| \simeq 0.072$. and for $B(10,3)$ ($\rho \simeq 0.713494$) $||\cdot||_2 \simeq 0.009$ and $\max|\cdot| \simeq 0.071$.

5 Conclusions

L-statistics are well-known aggregation operators useful in robust statistics. S-statistics considered in this paper possess some desired statistical properties which make them useful in many areas. Asymptotic normality proved in this paper enables interval estimation and the construction of statistical tests. Of course, some questions remain open. In particular, the problem of finding well-behaving estimators of $G'(\rho)$ required for the above-mentioned constructions have to be considered in further research.

References

1. Dubois, D., Prade, H.: Weighted minimum and maximum operations in fuzzy set theory. Inform. Sci. 39, 205–210 (1986)
2. Dubois, D., Prade, H., Testemale, C.: Weighted fuzzy pattern matching. Fuzzy Sets Syst. 28, 313–331 (1988)
3. Gągolewski, M., Grzegorzewski, P.: Arity-monotonic extended aggregation operators. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) Proceedings of the 13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2010. Dortmund, Germany. Part I, CCIS, vol. 80, pp. 693–702 (2010)
4. Gągolewski, M., Grzegorzewski, P.: Possibilistic extension of some aggregation operators (submitted for publication, 2010)
5. Grabisch, M., Pap, E., Marichal, J.L., Mesiar, R.: Aggregation Functions. Cambridge University Press, New York (2009)
6. Hoeffding, W.: Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58(301), 13–30 (1963)
7. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York (1980)
8. Shevtsova, I.G.: Sharpening of the upper bound of the absolute constant in the Berry-Esseen inequality. Theor. Probab. Appl. 51(3), 549–553 (2007)
9. Torra, V., Narukawa, Y.: The *h*-index and the number of citations: two fuzzy integrals. IEEE Trans. Fuzzy Syst. 16(3), 795–797 (2008)
10. van der Vaart, A.W.: Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2000)
11. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans. Syst. Man Cybern. 18(1), 183–190 (1988)

Particle Swarm Optimization and Inverse Problems

Esperanza García-Gonzalo and Juan Luis Fernández-Martínez

Abstract. In this paper we present a powerful set of Particle Swarm optimizers for inverse modeling. Their design is based on the interpretation of the swarm dynamics as a stochastic damped mass-spring system. All the PSO optimizers have very different exploitation and exploration capabilities. Their convergence can be related to the stability of their first and second order moments of the particle trajectories. Based on these results we present their corresponding cloud algorithms where each particle in the swarm has different inertia (damping) and acceleration (rigidity) constants. These algorithms show a very good balance between exploration and exploitation and their use avoids the tuning of the PSO parameters. These algorithms have been successfully applied to environmental geophysics and petroleum reservoir engineering where the combined use of model reduction techniques allow posterior sampling in high dimensional spaces.

Keywords: Inverse Problems, PSO, PSO Family, Cloud Design.

1 Particle Swarm Optimization Applied to Inverse Problems

Particle swarm optimization is a stochastic evolutionary computation technique used in optimization, which is inspired by the social behavior of

Esperanza García-Gonzalo and Juan Luis Fernández-Martínez
Department of Mathematics, University of Oviedo, Oviedo, Spain
e-mail: espe@uniovi.es,jlfm@uniovi.es

Juan Luis Fernández-Martínez
Energy Resources Department, Stanford University, Palo Alto, California, USA

Juan Luis Fernández-Martínez
Department of Civil and Environmental Engineering,
University of California at Berkeley, Berkeley, USA

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 289–296.
springerlink.com

individuals (called particles) in nature, such as bird flocking and fish schooling [15]. Inverse problems are very important in science and technology and sometimes referred to as, parameter identification, reverse modeling, etc.

Let us consider an inverse problem of the form $\mathbf{F}(\mathbf{m}) = \mathbf{d}$, where $\mathbf{m} \in \mathbf{M} \subset \mathbf{R}^n$ are the model parameters, $\mathbf{d} \in \mathbf{R}^s$ the discrete observed data, and

$$\mathbf{F}(\mathbf{m}) = (f_1(\mathbf{m}), f_2(\mathbf{m}), \ldots, f_s(\mathbf{m}))$$

is the vector field representing the forward operator and $f_j(\mathbf{m})$ is the scalar field that accounts for the j-th data. The "classical " goal of inversion given a particular data set (often affected by noise), is to find a set of parameters $\mathbf{m}$, such the data prediction error $\|\mathbf{F}(\mathbf{m}) - \mathbf{d}\|_p$ in a certain norm p, is minimized.

The PSO algorithm to approach this inverse problem is at first glance very easy to understand and implement:

1. A prismatic space of admissible models, $\mathbf{M}$, is defined:

   $$l_j \leq m_{ij} \leq u_j, \quad 1 \leq j \leq n, \quad 1 \leq i \leq N_{\text{size}}$$

   where l_j, u_j are the lower and upper limits for the j-th coordinate of each particle in the swarm, n is the number of parameters in the optimization problem and N_{size} is the swarm size. The misfit for each particle of the swarm is calculated, $\|\mathbf{F}(\mathbf{m}) - \mathbf{d}\|_p$ and then we determine for each particle its local best position found so far (called $\mathbf{l}_i(k)$) and the minimum of all of them which is called the global best ($\mathbf{g}(k)$).
2. The algorithm updates at each iteration the positions $\mathbf{m}_i(k)$, and velocities $\mathbf{v}_i(k)$, of each model in the swarm as follows:

   $$\begin{aligned} &\mathbf{v}_i(k+1) = \omega \mathbf{v}_i(k) + \phi_1 (\mathbf{g}(k) - \mathbf{m}_i(k)) + \phi_2 (\mathbf{l}_i(k) - \mathbf{m}_i(k)), \\ &\mathbf{m}_i(k+1) = \mathbf{m}_i(k) + \mathbf{v}_i(k+1) \end{aligned}$$

   ω, a_g, a_l are the PSO parameters and are called inertia, local and global acceleration constants; $\phi_1 = r_1 a_g, \phi_2 = r_2 a_l$ are the stochastic global and local accelerations, and r_1, r_2 are vectors of random numbers uniformly distributed in (0, 1). In the classical PSO algorithm these parameters are the same for all the particles of the swarm. In an inverse problem the position are the coordinates of the model $\mathbf{m}$ on the search space and the velocities the perturbations needed to find the low misfit models.

1.1 Uncertainty or Why We Should Explore the Search Space

Inverse problems are a very special type of optimization problems that turn out to be ill-posed mainly due to several reasons:

1. There is a physical model $\mathbf{F}$ (forward operator) involved which is always a simplification of the reality. This includes physical hypothesis and numerical approximations of the forward operator. Typically the predic-

tion (forward problem) is well-posed, but not the inverse. The ill-posednes is somehow related to the kind of question we ask in the inverse problem.

2. The observed data are part of the cost function and typically are noisy and discrete in number (mainly due to economic and logistic reasons).
3. Finally, the forward problem usually involves the resolution of a partial differential, integral, or algebraic set of equations, and a very fine model discretization is used to achieve accurate data predictions. The number of model parameters is in most cases significantly greater than the number of discrete data points available.

Let us suppose that we have a model $\mathbf{m}_0$ that fulfills $\|\mathbf{F}(\mathbf{m}_0)-\mathbf{d}\|_2 < tol$. It is possible to show analytically that the models in the neighborhood of $\mathbf{m}_0$ that fit the data within the same tolerance, *tol*, belong to the following hyperquadric:

$$(\mathbf{m}-\mathbf{m}_0)^T \mathbf{JF}_{\mathbf{m}_0}^T \mathbf{JF}_{\mathbf{m}_0} (\mathbf{m}-\mathbf{m}_0) + 2\boldsymbol{\Delta}\mathbf{d}^T (\mathbf{m}-\mathbf{m}_0) + \|\boldsymbol{\Delta}\mathbf{d}\|_2^2 = tol^2$$

$\mathbf{JF}_{\mathbf{m}_0}$ is the Jacobian matrix of the operator $\mathbf{F}$ in $\mathbf{m}_0$ and $\boldsymbol{\Delta}\mathbf{d} = \mathbf{F}(\mathbf{m}_0)-\mathbf{d}$. This means that the equivalent models will have the direction of the vectors of the $\mathbf{V}$ base given by the singular value decomposition of $\mathbf{JF}_{\mathbf{m}_0}$ and whose axes are proportional to the inverse of the singular values λ_k in each direction. Due to the continuity of the Jacobian operator, we finally conclude that with no regularization term the misfit function has a flat and elongated valley shape. Also, in other kind of optimization problems (e.g., experimental fitting problems) many local minima might coexist. Thus, uncertainty in the model parameters is always important in inverse problems, forcing the modeler to explore the search space.

2 The Consistency of the PSO Family

Fernández Martínez and García Gonzalo([8], [4]), proved that the PSO algorithm can be physically interpreted as a particular discretization of a stochastic damped mass-spring system:

$$\mathbf{m}_i''(t) + (1-\omega)\mathbf{m}_i'(t) + \phi\mathbf{m}_i(t) = \phi_1\mathbf{g}(t-t_0) + \phi_2\mathbf{l}_i(t-t_0)$$

where $\phi = \phi_1 + \phi_2$. This model has been addressed as the PSO continuous model since it describes (together with the initial conditions) the continuous movement of any particle coordinate in the swarm $\mathbf{m}_i(t)$, where i stands for the particle index, and $\mathbf{g}(t)$ and $\mathbf{l}_i(t)$ are its local and global attractors. In this model the trajectories $\mathbf{m}_i(t)$ are allowed to be delayed a time t_0 with respect to the attractors, $\mathbf{g}(t-t_0)$ and $\mathbf{l}_i(t-t_0)$.

Using this physical analogy we were able to analyze the PSO particle's trajectories [8] and to explain the success in achieving convergence of some popular parameters sets found in the literature [2], [3], [16]. Also we derived a whole family of PSO algorithms [5], [14] considering different difference schemes for $\mathbf{m}_i''(t)$ and $\mathbf{m}_i'(t)$:

1. GPSO or centered-regressive PSO $(t_0 = 0)$

$$\begin{aligned} &v(t+\Delta t) = (1-(1-\omega)\Delta t)\, v(t) + \phi_1 \Delta t\, (g(t) - m(t)) + \phi_2 \Delta t\, (l(t) - m(t)),\\ &m(t+\Delta t) = m(t) + v(t+\Delta t)\Delta t. \end{aligned}$$

   The GPSO algorithm is the generalization of the PSO algorithm for any time step Δt , (PSO is the particular case for $\Delta t = 1$). These expressions for the velocity and position are obtained by employing a regressive scheme in velocity and a centered scheme in acceleration.
2. CC-PSO or centered-centered PSO $(t_0 = 0)$

$$\begin{aligned} &m(t+\Delta t) = m(t) + \left[\tfrac{2+(w-1)\Delta t}{2} v(t) + \tfrac{\Delta t}{2}\phi_1 (l(t) - m(t)) + \tfrac{\Delta t}{2}\phi_2 (g(t) - m(t))\right]\Delta t,\\ &v(t+\Delta t) = \tfrac{2+(w-1)\Delta t}{2+(1-w)\Delta t} v(t) + \tfrac{\Delta t}{2+(1-w)\Delta t} \sum_{k=0}^{1} \begin{bmatrix} \phi_1 (l(t+k\Delta t) - m(t+k\Delta t)) \\ +\phi_2 (g(t+k\Delta t) - m(t+k\Delta t)) \end{bmatrix}. \end{aligned}$$

3. CP-PSO or centered-progressive PSO $(t_0 = \Delta t)$

$$\begin{aligned} &v(t+\Delta t) = \frac{\left((1-\phi\Delta t^2) v(t) + \phi_1 \Delta t (g(t) - m(t)) + \phi_2 \Delta t (l(t) - m(t))\right)}{1 + (1-\omega)\Delta t},\\ &m(t+\Delta t) = m(t) + v(t)\Delta t. \end{aligned}$$

4. PP-PSO or progressive-progressive PSO $(t_0 = 0)$

$$\begin{aligned} &v(t+\Delta t) = (1-(1-\omega)\Delta t)\, v(t) + \phi_1 \Delta t\, (g(t) - m(t)) + \phi_2 \Delta t\, (l(t) - m(t)),\\ &m(t+\Delta t) = m(t) + v(t)\Delta t. \end{aligned}$$

5. RR-PSO or or regressive-regressive PSO $(t_0 = \Delta t)$

$$\begin{aligned} &v(t+\Delta t) = \frac{v(t) + \phi_1 \Delta t\, (g(t) - m(t)) + \phi_2 \Delta t\, (l(t) - m(t))}{1 + (1-\omega)\Delta t + \phi \Delta t^2}\\ &m(t+\Delta t) = m(t) + v(t+\Delta t)\Delta t. \end{aligned}$$

The consistency of the different PSO family members has been related to the stability of their first and second order trajectories [8], [5]. The type of mean trajectories depend on the character of the eigenvalues of the first order difference equation as a function of the inertia parameter (ω) and the total mean acceleration $(\overline{\phi} = \overline{\phi}_1 + \overline{\phi}_2 = \dfrac{a_g + a_l}{2})$. Basically there are four kind of trajectories: damped oscillatory in the complex eigenvalue region, symmetrically and asymmetrically zigzagging in the regions of negative real eigenvalues and almost monotonous decreasing character in the region of positive real eigenvalues. Maximum exploration is reached in the complex region. The second order trajectories [5] show a similar kind of behavior. The second order spectral radius controls the rate of attenuation of the second order moments of the particle trajectories (variance and temporal covariance between $m(t)$ and

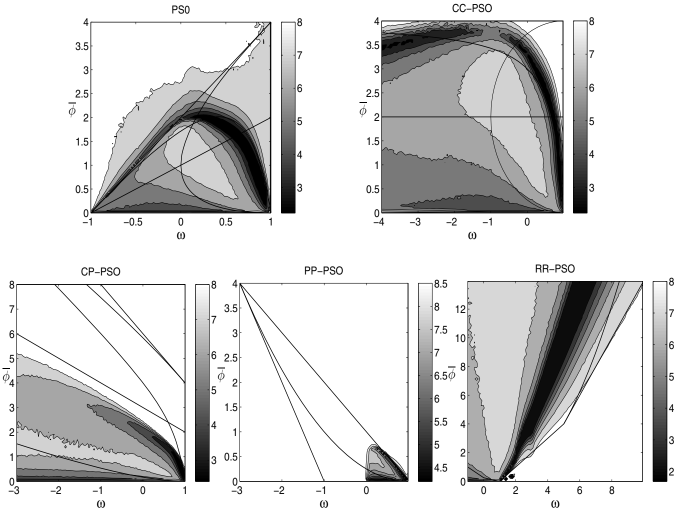


Fig. 1 Logarithmic median misfit errors for the Rosenbrock function in 50 simulations (after 300 iterations) for different family members. Similar results can be achieved for other benchmark functions.

$m(t+\Delta t)$). These results have been confirmed by numerical experiments with different benchmark functions in several dimensions. Figure 1 shows for each family member the contour plots of the misfit error (in logarithmic scale) after a certain number of iterations (300) for the Rosenbrock function. This numerical analysis is done for a lattice of $\left(\omega,\overline{\phi}\right)$ points located in the corresponding first order stability regions over 50 different simulations. For GPSO, CC-PSO and CP-PSO better parameter sets, (ω, a_g, a_l), are located on the first order complex region, close to the upper border of the second order stability region where the attraction from the particle oscillation center is lost, i.e. the variance becomes unbounded; and around the intersection to the median lines of the first stability regions where the temporal covariance between trajectories is close to zero [5]. The PP-PSO does not converge for $\omega < 0$, and the good parameter sets are in the complex region close to the limit of second order stability and $\overline{\phi} = 0$. The good parameters sets for the RR-PSO are concentrated around the line of equation $\overline{\phi} = 3(\omega - 3/2)$, mainly for inertia values greater than two. This line is located in a zone of medium attenuation and high frequency of trajectories. The CP-PSO and RR-PSO are the versions that have the greatest exploratory capabilities. Finally we performed the full stochastic analysis of the PSO continuous and discrete models [6], [7]. This analysis served to analyze the GPSO second order trajectories, to show the convergence of the discrete versions (GPSO) to the continuous PSO model as the discretization time step goes to zero, and to explain the role of the cost function on the first and second order continuous and discrete dynamical systems. Thus, PSO should not be considered heuristic.

3 How to Achieve Exploration: The Cloud Algorithms

Based on the consistency results shown above we have designed a PSO algorithm where each particle in the swarm has different inertia (damping) and local and global acceleration (rigidity) constants, being the $\left(\omega,\overline{\phi}\right)$ sets located in the low misfit regions. This idea has been implemented for the particle-cloud PSO algorithm in [13] and extended to CC-PSO and CP-PSO in [10]. Here we present the results for PP-PSO and RR-PSO. We also present the coordinates-cloud algorithm where all the same index coordinates of all the particles in the swarm will have the same (ω,a_g,a_l) constants.

The particle-cloud algorithm works as follows:

1. The misfit contours to design the clouds are based on the Rosenbrock function in 50 dimensions. The Rosenbrock function has been chosen for this purpose because in inverse problems the low misfit models are located along flat elongated valleys. Nevertheless the cloud could be designed using other benchmark functions. For each $\left(\omega,\overline{\phi}\right)$ located on the low misfit region, we generate three different (ω,a_g,a_l) points corresponding to $a_g=a_l$, $a_g=2a_l$ and $a_l=2a_g$. Particles are randomly selected depending on the iterations. The algorithm keep track of the (ω,a_g,a_l) points used to achieve the global best solution in each iteration. The criteria used to select the points belonging to the cloud is not very rigid, since points located on the low misfit region provide very good results.
2. The algorithm also uses the lime and sand modality, that is, varying Δt with iterations [4]. The first and second order stability regions increase their size when Δt goes to zero. In this case the exploration is increased around the global best solution. Conversely when Δt is greater than one the exploration is increased in the whole search space.

Table 1 shows the results obtained for different benchmark functions in 50 dimensions, using the particle cloud algorithm. The misfits are compared in to the reference values published in the literature. The RR-PSO, CC-PSO and PSO are the most performing algorithms.

The coordinate cloud algorithm gives also very good results but it is a more explorative version than the particle-cloud. Nevertheless, as pointed before, in inverse modeling it is not only important to achieve very low misfits but also to explore the space of possible solutions. When these algorithms have to be used in explorative form the cloud versions become a very interesting approach, because there is no need to tune the PSO parameters. Finally, the exploration can be also increased by introducing repulsive forces in the swarm by switching to negative the sign of the acceleration constants. This strategy has been used in [9], [11] to solve geophysical environmental inverse problems.

Table 1 Comparison between the particle-cloud modalities and the reference misfit values for Standard PSO [1] for different benchmark functions in 50 dimensions.

Median	Griewank	Rastrigin	Rosenbrock	Sphere
Standard PSO	9.8E-03	81	90	6.9E-11
PSO	9.6E-03	92	86	8.9E-19
CC-PSO	7.4E-03	99	90	1.0E-15
CP-PSO	1.8E-02	86	223	2.0E-07
PP-PSO	1.0E-01	91	251	8.4E-02
RR-PSO	1.2E-02	39	89	2.9E-25

4 Advantages and Drawbacks of Particle Swarm Optimization

Particle Swarm Optimization is a global stochastic search algorithm and it is typically used to solve optimization problems when the number of parameters is small (hundreds) and the forward problem is fast to compute. The advantage of these methods is that they address the optimization problem as a sampling problem. Thus, they do not do any regularization. In inverse problems, both a large number of parameters, and very costly forward evaluations hamper the use of global algorithms. The combined used of PSO and model reduction techniques allow us to address real world applications having thousands of parameters [12]. The use of model reduction techniques is based on the fact that the inverse model parameters are not independent. Conversely, there exist correlations between model parameters introduced by the physics of the forward problem in order to fit the observed data. Taking advantage of this fact it is possible to reduce the number of parameters that are used to solve the identification problem. Also the use of model reduction techniques helps to regularize the inverse problem, allowing to perform model appraisal by sampling the family of equivalent models that fit the observed data and are in accord with the prior information that it is at disposal. Nevertheless, their use should be evaluated from the perspective of each particular application, that is, model reduction techniques should be used with care and not just to accelerate computation.

References

1. Birge, B.: PSOt - a particle swarm optimization toolbox for use with Matlab. In: Proceedings of the 2003 IEEE Swarm Intelligence Symposium, SIS 2003, Indianapolis, Indiana, USA, pp. 182–186 (2003)
2. Carlisle, A., Dozier, G.: An off-the-shelf PSO. In: Proceedings of the Particle Swarm Optimization Workshop, Indianapolis, Indiana, USA, pp. 1–6 (2001)

3. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Comput. 6, 58–73 (2002)
4. Fernández Martínez, J.L., García-Gonzalo, E.: The generalized PSO: a new door to PSO evolution. J. Artificial Evol. Appl. ID 861275 (2008), doi:10.1155/2008/861275
5. Fernández-Martínez, J.L., García-Gonzalo, E.: The PSO family: deduction, stochastic analysis and comparison. Swarm Intell. 3, 245–273 (2009)
6. Fernández-Martínez, J.L., García-Gonzalo, E.: Stochastic stability analysis of the linear continuous and discrete PSO models. Technical Report, Department of Mathematics, University of Oviedo, Spain (2009)
7. Fernández-Martínez, J.L., García-Gonzalo, E.: What makes Particle Swarm Optimization a very interesting and powerful algorithm? In: Handbook of Swarm Intelligence – Concepts, Principles and Applications Series on Adaptation, Learning, and Optimization. Springer, Heidelberg (to appear, 2010)
8. Fernández-Martínez, J.L., García-Gonzalo, E., Fernández-Alvarez, J.P.: Theoretical analysis of particle swarm trajectories through a mechanical analogy. Int. J. Comput. Intell. Res. 4, 93–104 (2008)
9. Fernández-Martínez, J.L., García-Gonzalo, E., Fernández-Álvarez, J.P., Kuzma, H.A., Menéndez-Pérez, C.O.: PSO: A Powerful Algorithm to Solve Geophysical Inverse Problems. Application to a 1D-DC Resistivity Case. J. Appl. Geophys. (accepted for publication, 2010)
10. Fernández-Martínez, J.L., García-Gonzalo, E., Fernández-Muñiz, Z., Mukerji, T.: How to design a powerful family of Particle Swarm Optimizers for inverse modeling. New Trends on Bio-inspired Computation. Trans. Inst. Meas. Contr. (accepted for publication, 2010)
11. Fernández-Martínez, J.L., García-Gonzalo, E., Naudet, V.: Particle Swarm Optimization applied to the solving and appraisal of the Streaming Potential inverse problem. Geophys, Hydrogeophysics Special Issue (accepted for publication, 2010)
12. Fernández-Martínez, J.L., Mukerji, T., García-Gonzalo, E.: Particle Swarm Optimization in high dimensional spaces. In: Proceedings of the Seventh International Conference on Swarm Intelligence, ANTS 2010, Bruxelles, Belgium (2010)
13. García-Gonzalo, E., Fernández-Martínez, J.L.: Design of a simple and powerful Particle Swarm optimizer. In: Proceedings of the International Conference on Computational and Mathematical Methods in Science and Engineering, CMMSE 2009, Gijón, Spain (2009)
14. García-Gonzalo, E., Fernández-Martínez, J.L.: The PP-GPSO and RR-GPSO. Technical Report. Department of Mathematics. University of Oviedo, Spain (submitted for publication, 2010)
15. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, ICNN 1995, Perth, WA, Australia, pp. 1942–1948 (1995)
16. Trelea, I.C.: The particle swarm optimization algorithm: convergence analysis and parameter selection. Inf. Processing Lett. 85, 317–325 (2003)

Linear Approximations to the Power Function of Robust Tests

A. García-Pérez

Abstract. The main characteristics of robust tests, such as the power function, are computed using the asymptotic distribution of the robust test statistics because the finite sample one is unmanageable. In this paper we propose a finite sample linear approximation to the power function of a robust test obtained using the von Mises expansion of the functional Tail Probability.

Keywords: Power Function, Robust Test, Tail Area Influence Function.

1 Introduction

The finite sample distribution of robust test statistics are usually unmanageable and, for this reason, the main characteristics (such as the critical value or the power) are frequently computed using its asymptotic distribution. Hence, asymptotically equivalent tests are also equivalent from a robustness point of view. In this paper we propose an alternative procedure that consists in using the von Mises expansion of the functional tail probability as a way to compute powers using tail probabilities under the null hypothesis. We apply the results to the location M-test based on the Huber estimator, and to the fixed-carriers model based on the Huber estimator.

2 The von Mises Expansion of a Functional

Let T be a functional defined on a convex set $\mathscr{F}$ of distribution functions. If F and G are two members of $\mathscr{F}$, the (first order) von Mises expansion of T at F is ([8, 2, 4])

A. García-Pérez
Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED), 28040-Madrid, Spain
e-mail: agar-per@ccia.uned.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 297–303.
springerlink.com

$$T(F) = T(G) + \int \mathrm{IF}\,(\mathbf{x}; T, G)\, dF(\mathbf{x}) + Rem$$

where *IF* is the Hampel's Influence Function. For the functional "Tail Probability" of statistic T_n, this von Mises expansion is

$$P_F\{T_n > t\} = P_G\{T_n > t\} + \int \mathrm{TAIF}\,(\mathbf{x}; t; T_n, G)\, dF(\mathbf{x}) + Rem$$

where TAIF is the Tail Area Influence Function defined in [3] as the influence function of the tail probability

$$\mathrm{TAIF}\,(x; t; T_n, G) = \frac{\partial}{\partial \varepsilon} P_{G_{\varepsilon,x}}\{T_n(X_1, .., X_n) > t\}\bigg|_{\varepsilon=0}$$

where $G_{\varepsilon,x} = (1-\varepsilon)G + \varepsilon\delta_x$. In [5], an extension of this definition to the multivariate situation and an exact expression for the TAIF are obtained.

In [6] we extend these definitions and results to non-identically distributed random vectors, defining the *i-th Partial Tail Area Influence Functions* of T_n at $\mathbf{G} = (G_1, ..., G_n)$ with relation to G_i, $i = 1, ..., n$, as

$$\mathrm{TAIF}_i(\mathbf{x}; t; T_n, \mathbf{G}) = \frac{\partial}{\partial \varepsilon} P_{G_i^{\varepsilon,\mathbf{x}}}\{T_n(\mathbf{X}_1, ..., \mathbf{X}_n) > t\}\bigg|_{\varepsilon=0}$$

in those $\mathbf{x} \in \mathscr{X}$ where each limit exists. In the computation of TAIF_i only G_i is contaminated and the other distributions remain fixed, being the von Mises expansion

$$P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) > t\} = P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) > t\}$$

$$+ \sum_{i=1}^{n} \int_{\mathscr{X}} \mathrm{TAIF}_i\,(\mathbf{x}; t; T_n, \mathbf{G})\, dF_i(\mathbf{x}) + Rem$$

where the remainder term

$$Rem = \frac{1}{2} \int\int T^{(2)}_{\mathbf{G_F}}(\mathbf{x}_1, \mathbf{x}_2)\, d[\mathbf{F}(\mathbf{x}_1) - \mathbf{G}(\mathbf{x}_1)]\, d[\mathbf{F}(\mathbf{x}_2) - \mathbf{G}(\mathbf{x}_2)]$$

is small if distributions $\mathbf{F}$ and $\mathbf{G}$ are close. ($T^{(2)}_{\mathbf{G_F}}$ is the *second derivative* of the functional tail probability at the mixture distribution $\mathbf{G_F} = (1-\lambda)\mathbf{G} + \lambda\mathbf{F}$, for some $\lambda \in [0,1]$.)

Hence, if $\mathbf{F}$ and $\mathbf{G}$ are close, we can write (using the results obtained in [5] and [6])

$$P_{\mathbf{F}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) > t\} = P_{F_1,...,F_n}\{T_n(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) > t\}$$

$$\simeq P_{\mathbf{G}}\{T_n(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n) > t\} + \sum_{i=1}^{n} \int_{\mathscr{X}} \mathrm{TAIF}_i\,(\mathbf{x}; t; T_n, \mathbf{G})\, dF_i(\mathbf{x})$$

$$= (1-n)P_{\mathbf{G}}\{T_n(\mathbf{X}_1,\mathbf{X}_2,...,\mathbf{X}_n) > t\} + \int_{\mathscr{X}} P_{G_2,...,G_n}\{T_n(\mathbf{x},\mathbf{X}_2,...,\mathbf{X}_n) > t\}\, dF_1(\mathbf{x})$$

$$+ \int_{\mathscr{X}} P_{G_1,G_3,...,G_n}\{T_n(\mathbf{X}_1,\mathbf{x},...,\mathbf{X}_n) > t\}\, dF_2(\mathbf{x}) + \cdots$$

$$+ \int_{\mathscr{X}} P_{G_1,...G_{n-1}}\{T_n(\mathbf{X}_1,...,\mathbf{X}_{n-1},\mathbf{x}) > t\}\, dF_n(\mathbf{x})$$

that allows an approximation to the tail probability $P_{\mathbf{F}}\{T_n > t\}$ under models $(F_1,...,F_n)$, knowing the value of this tail probability under near models $(G_1,...,G_n)$.

3 A Linear Approximation to the Power Function of a Test

Let us consider a level α-test of the null hypothesis $H_0 : \theta \in \Theta_0$ against an alternative $H_1 : \theta \in \Theta_1$. We shall suppose that it rejects H_0 for large values of T_n and that the critical value is k_n^α. The previous von Mises approximation can be used to obtain a linear approximation to the power function, considering as distributions $(F_1,...,F_n)$ the models under an alternative $\theta \in \Theta_1$ and, as distributions $(G_1,...,G_n)$ the models under the null. In this case, for location families, we obtain

$$Power(\theta) \simeq (1-n)\alpha + P_{F_{1;\theta_0},...,F_{n;\theta_0}}\{T_n(\mathbf{X}_1 + (\theta-\theta_0),\mathbf{X}_2,...,\mathbf{X}_n) > k_n^\alpha\}$$

$$+ P_{F_{1;\theta_0},...,F_{n;\theta_0}}\{T_n(\mathbf{X}_1,\mathbf{X}_2 + (\theta-\theta_0),...,\mathbf{X}_n) > k_n^\alpha\} + \ldots$$

$$+ P_{F_{1;\theta_0},...,F_{n;\theta_0}}\{T_n(\mathbf{X}_1,...,\mathbf{X}_{n-1},\mathbf{X}_n + (\theta-\theta_0)) > k_n^\alpha\}.$$

This linear approximation is very accurate if the alternative θ is close to the null θ_0. If this is not the case, we can extend its application using an iterative procedure that consists in considering intermediate distributions $\mathbf{F}_j = (F_{1;\theta_j},...,F_{n;\theta_j}) \equiv (F_{1j},...,F_{nj})$, with location parameter $\theta_j = \theta_0 + j(\theta-\theta_0)/(k+1)$, $j = 1,...,k+1$, between the model distribution under the null hypothesis θ_0, $\mathbf{F}_0 \equiv \mathbf{F}_{\theta_0} = (F_{1;\theta_0},...,F_{n;\theta_0})$, and the model distribution under the alternative θ, $\mathbf{F}_{k+1} \equiv \mathbf{F}_\theta = (F_{1;\theta},...,F_{n;\theta})$, obtaining the approximation

$$Power(\theta) \simeq \alpha + \sum_{j=1}^{k+1} [P_{H_0}\left\{T_n(\mathbf{X}_1 + c_{2j},\mathbf{X}_2 + c_{1j},...,\mathbf{X}_n + c_{1j}) > k_n^\alpha\right\}$$

$$+ P_{H_0}\left\{T_n(\mathbf{X}_1 + c_{1j},\mathbf{X}_2 + c_{2j},\mathbf{X}_3 + c_{1j} + ...,\mathbf{X}_n + c_{1j}) > k_n^\alpha\right\} + ...$$

$$+P_{H_0}\left\{T_n(\mathbf{X}_1+c_{1j},...,\mathbf{X}_{n-1}+c_{1j},\mathbf{X}_n+c_{2j})>k_n^{\alpha}\right\}$$

$$-nP_{H_0}\left\{T_n(\mathbf{X}_1+c_{1j},...,\mathbf{X}_n+c_{1j})>k_n^{\alpha}\right\}]$$

where $c_{1j}=(j-1)(\theta-\theta_0)/(k+1)$ and $c_{2j}=j(\theta-\theta_0)/(k+1)$.

In this approximation we must express the tail probabilities in a different way for any different problem. But, with the approximation, we transfer the computation under the alternative hypothesis to computations under the null.

In the rest of the paper we shall consider tests based on the Huber estimator where the usual expression that connects the tail probability of an M-estimator with its score function is used. Nevertheless, the method exposed in the paper can be used not only for these problems but also to other more complex ones; see for instance the approximations for saddlepoint tests in [6].

4 Huber Location Test

The previous linear approximation to the power function of the level α-test of $H_0:\theta=\theta_0$ against $H_1:\theta>\theta_0$ considering as test statistics the Huber statistics with score function ψ_b (iid situation, $F_1=...=F_n=F$) is

$$Power(\theta)\simeq\alpha+n\sum_{j=1}^{k+1}\left[P_{H_0}\left\{{}^1T_n(\mathbf{X}_1,...,\mathbf{X}_n)>k_n^{\alpha}-c_{1j}\right\}\right. \tag{1}$$

$$\left.-P_{H_0}\left\{T_n(\mathbf{X}_1,...,\mathbf{X}_n)>k_n^{\alpha}-c_{1j}\right\}\right]\cdot\left[F_{\theta_0}(k_n^{\alpha}-c_{1j}+b)-F_{\theta_0}(k_n^{\alpha}-c_{1j}-b)\right]$$

where ${}^1T_n$ is the M-estimator with score function ${}^1\psi=\psi_b+(\theta-\theta_0)/((k+1)n)$. We can see from this expression, for instance, that as we increase the tuning constant b, the power increases obtaining the maximum power test when $b\rightarrow\infty$, i.e., for the sample mean.

In this simple case of the sample mean test, if $F_\theta\equiv N(\theta,1)$, the exact power function is $Power(\theta)=1-\Phi_s(z_{1-\alpha}-\theta\sqrt{n})$ and approximation (1)

$$Power(\theta)\simeq\alpha+n\sum_{j=1}^{k+1}\left[\Phi_s((k_n^{\alpha}-c_{1j})\sqrt{n})-\Phi_s((k_n^{\alpha}-c_{1j}-\theta/(n(k+1)))\sqrt{n})\right]$$

where Φ_s is the standard normal cumulative distribution function. Figure 1 shows the exact power of this test (solid line) and the approximation (dashed line) for $n=3$ and $\alpha=0.025$. In the left side we used only $k=6$ iterations and $k=15$ in the right one.

For other b values, it is possible to use saddlepoint approximations given, for instance, in [1] (or just the asymptotic distribution of an M-estimator), to compute the tail probabilities in (1), always under the null hypothesis; i.e., with the linear approximations that we propose in the paper, we can use known approximations to the p-value to compute power functions.

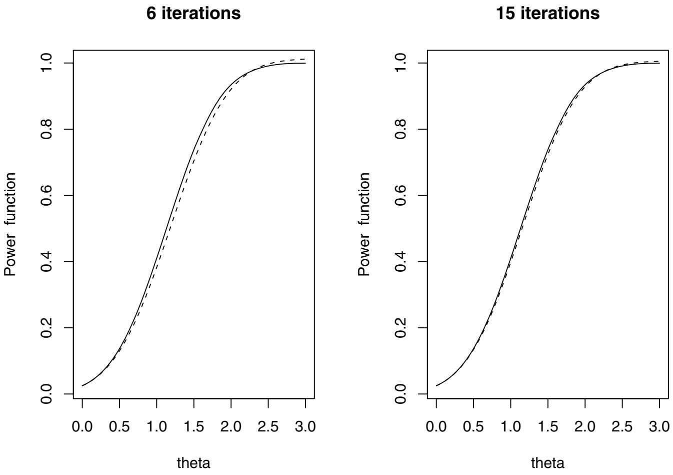


Fig. 1 Exact (solid line) and approximated (dashed line) power functions for the mean test

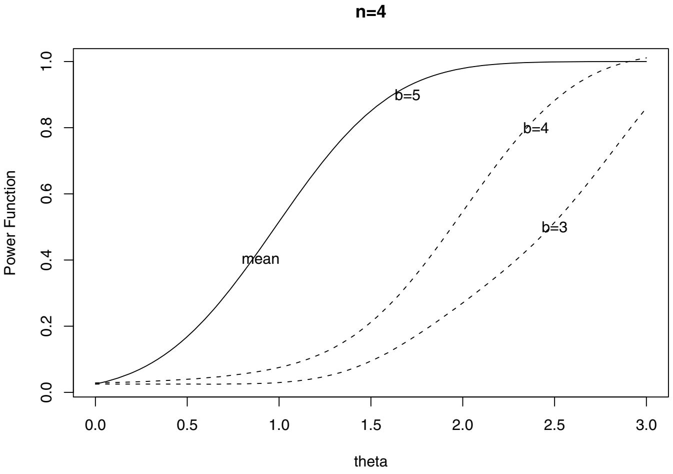


Fig. 2 Approximate power function of the Huber test for different tuning constants *b*

The linear approximations of the Huber test for different tuning constants appear in Figure 2.

5 Fixed-Carriers Model Based on the Huber Estimator

Let us consider the Simple Linear Regression Model with fixed-carriers. In this model, we suppose that

$$Y_i = a + \beta x_i + e_i \qquad\qquad i = 1, ..., n$$

where the assumptions are that Y_i ($Y_i|x_i$ for the not fixed-carriers model) follows a model distribution F_i with mean $\mu_i = E[Y_i] = a + \beta x_i$ and variance σ^2, being e_i iid variables with a common distribution W.

Let us also suppose that we wish to test the usual null hypothesis $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ using as test statistic T_n, the Huber estimator for β with score function ψ_b.

Because the test is bilateral and we suppose with equal tails, if k_n^α here is such that $P_{H_0}\{T_n > k_n^\alpha\} = \alpha/2$, assuming that, under H_0, the distribution of T_n is symmetric,

$$Power(\beta) = P_\beta\{T_n < -k_n^\alpha\} + P_\beta\{T_n > k_n^\alpha\} = 2P_\beta\{T_n > k_n^\alpha\}.$$

The linear approximation (with no iterations) is now

$$Power(\beta) \simeq \alpha + 2\sum_{i=1}^{n}\left(P_{H_0}\left\{{}^iT_n(Y_1,...,Y_n) > k_n^\alpha\right\} - \frac{\alpha}{2}\right)$$
$$\cdot\,(W(k_n^\alpha x_i + b\sigma) - W(k_n^\alpha x_i - b\sigma))$$

where ${}^iT_n(Y_1,...,Y_i,...,Y_n) = T_n(Y_1,...,Y_i + \beta x_i,...,Y_n)$ is an M-estimator (the i-th shifted version of $T_n(Y_1,...,Y_i,...,Y_n)$) with score function

$${}^i\psi = \psi_b + \frac{\beta x_i}{n\sigma}.$$

Although it would be possible to use a saddlepoint approximation to compute the tail probabilities under the null hypothesis, we shall consider here the approximation to this distribution given by [7], obtaining finally,

$$Power(\beta) \simeq \alpha + 2\sum_{i=1}^{n}\left[1 - \frac{\alpha}{2} - \Phi\left(\frac{c_n^*(k_n^\alpha - \beta)\gamma}{{}^i\sigma}\right)\right]$$
$$\cdot\,[W(k_n^\alpha x_i + b\sigma) - W(k_n^\alpha x_i - b\sigma)]$$

With the linear approximations proposed in the paper, we always obtain analytic expressions for the power function. With them we can obtain general conclusions. For instance, in these examples we see that, as we increase the tuning constant b the power increases.

Moreover, since these power functions can easily be computed (we have used R), we can, for example, to determine the tuning constant b for a given power, just moving the argument b in the R function until the power is obtained.

Acknowledgements. This research is supported by the Spanish Science and Innovation Department (MTM2009-10072).

References

1. Daniels, H.E.: Saddlepoint approximations for estimating equations. Biometrika 70(1), 89–96 (1983)
2. Fernholz, L.T.: von Mises calculus for statistical functionals. Lecture Notes in Statistics, vol. 19. Springer, New York (1983)
3. Field, C.A., Ronchetti, E.M.: A tail area influence function and its application to testing. Commun. Stat. 4(1-2), 19–41 (1985)
4. García-Pérez, A.: von Mises approximation of the critical value of a test. Test 12(2), 385–411 (2003)
5. García-Pérez, A.: Another look at the Tail Area Influence Function. Metrika (in press, 2010)
6. García-Pérez, A.: A linear approximation to the power function of a test (submitted for publication, 2010)
7. Yohai, V.J., Maronna, R.A.: Asymptotic behavior of *M*-estimators for the linear model. Ann. Statist. 7, 258–268 (1979)
8. Withers, C.S.: Expansions for the distribution and quantiles of a regular functional of the empirical distribution with applications to nonparametric confidence intervals. Ann. Statist. 11, 577–587 (1983)

Decision Support for Evolving Clustering

Olga Georgieva and Sergey Nedev

Abstract. An evolving clustering algorithm applying the adaptive-distance measure is developed. An incorporated fuzzy decision support procedure classifies the current income. The decision support increases the algorithm robustness. As it discovers on-line clusters with different shape and orientation it is applicable to a wide range of practical tasks as diagnostics and prognostics, system identification, real time classification.

Keywords: Dynamic Data Mining, Evolving Clustering, Real Time Classification, Prognostics.

1 Introduction

The advantages of the clustering methods are effectively explored as one of the challenging theoretical and practical problems in data mining. The commonly used Fuzzy C-Means (FCM) clustering method [4] discovers spherical clusters with equal volumes and density. Several clustering algorithms extend the original FCM method to the case of clusters with a general shape [3], [7], [13]. Among them the Gustafson-Kessel (GK) clustering algorithm [10] is widely used as a powerful clustering technique with numerous applications in various domains including image processing, classification and system identification. Its main feature is the local adaptation of the distance metric to the shape of the cluster.

Most of the clustering methods are based on the concept of batch clustering, i.e. the data set is assumed to be available before the clustering analysis

Olga Georgieva
Department of Software Engineering, Sofia University, Bulgaria
e-mail: o.georgieva@fmi.uni-sofia.bg

Sergey Nedev
Department of Computer Systems, Technical University - Sofia, Bulgaria

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 305–312.
springerlink.com

is carried out. In a wide range of applications, however, the data is presented to the clustering algorithm in real time. A growing number of methods try to cope the problem of evolving data streams clustering [6], [11], [12]. Large amount of these methods are based on the single pass clustering [8], [9], [15] comprising techniques that are in contrast to the iterative strategies like K-means and FCM based clustering.

Successful solution of the real time modeling and classification task [3], [14] have been proposed by incorporating the Mountain clustering algorithm [14] and its modification - the Subtractive clustering algorithm [5]. The on-line extension of the Subtractive clustering utilizes recursive and noniterative techniques for calculating the potential of the new data point in order to update the existing clusters or to discover new ones [1], [2].

In this paper we propose a new evolving clustering algorithm that builds upon the advantages of the GK algorithm enable to identify clusters with a generic shape and orientation. In order to deal with the vagueness of the classification task a fuzzy decision support algorithm was incorporated in the evolving clustering procedure. The algorithm presents robust properties according to variance of the algorithm parameters.

2 GK Clustering

Objective function-based clustering aims minimization of a criterion J that represents the fitting error of the clusters with respect to the data. The underlying objective function for most of the clustering algorithms is [4]:

$$J(V,U,F) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^{m} d_{ik}^{2}, \tag{1}$$

where N is the number of data points, c is the number of clusters; u_{ik} and d_{ik} denote correspondingly the membership degree and distance between the k-th data point $x_k = [x_{k1}, x_{k2}, ..., x_{kn}]$ and i-th cluster prototype v_i; $U = \{u_{ik}\}$, $i = 1, ..., c$, $k = 1, ..., N$ is a partition matrix and n is the number of features describing each data point. $V = [v_1, v_2, ..., v_c]^T$ is a cluster prototypes matrix and $v_i, i = 1, ..., c$ is the prototype vector $v_i = [v_{i1}, v_{i2}, ..., v_{in}]$ of the i^{-th} cluster. In the simplest case, the cluster prototype is a single point called cluster centre. The fuzzifier $m \in [1, \infty)$ is the weighted exponent coefficient which determines how much clusters may overlap.

In case of GK clustering the distance d_{ik} is a squared inner-product distance norm that depends on a positive definite symmetric matrix A_i:

$$d_{ik}^2 = \|x_k - v_i\|_{A_i}^2 = (x_k - v_i)A_i(x_k - v_i)^T. \tag{2}$$

The matrix A_i determines the shape and orientation of the selected cluster. Commonly oblong clusters with different orientation in the space are presented in the data set. In order to cover such clusters the algorithm should employ an adaptive distance norm unique for every cluster. For this the norm inducing matrix A_i is calculated according to the data covariance:

$$A_i = [\rho_i det(F_i)]^{1/n} F_i^{-1}, \tag{3}$$

where ρ_i is the cluster volume and F_i is the fuzzy covariance matrix of the i-th cluster:

$$F_i = \frac{\sum_{k=1}^{N} u_{ik}^m (x_k - v_i)^T (x_k - v_i)}{\sum_{k=1}^{N} u_{ik}^m}. \tag{4}$$

Without any prior knowledge the cluster volume $\rho_i = 1$, $i = 1,...,c$ is simply fixed at one for each cluster.

The parameters that minimise the criterion (1) are the membership degrees

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}^2}{d_{jk}^2})^{\frac{1}{m-1}}} \tag{5}$$

and the cluster parameters namely cluster centers

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m x_k}{\sum_{k=1}^{N} u_{ik}^m}, \tag{6}$$

that finally determine the distance d_{ik}. The optimization scheme that finds a proper partition alternatively considers one of the parameter sets, either the membership degrees or the cluster centers as fixed, while the other parameter set is optimized, until the algorithm finally converges.

3 Evolving Clustering Based on GK Algorithm

The objective function clustering algorithms are not able to deal with data streams. They process a fixed data set assuming that the number of clusters is known in advance by that applying iterative optimization scheme. The new incoming data could not be processed by the original algorithm scheme as the stream of incoming data changes the data structure over time. Evolving algorithm variants are need in this case.

Originally, we assume that the GK has been applied to identify an initial set of c clusters of the previously collected data. Each of those clusters is defined by its center v_i and fuzzy covariance matrix F_i. The assumption is realistic as in almost all evolving applications some initial knowledge about the data being processed is available.

We assume that the boundary of each cluster is defined by the cluster radius. We determine the radius r_i of the i-th cluster equal to the maximal distance between the cluster center v_i and the points belonging to this cluster with membership degree larger or equal to a given threshold membership degree u_h:

$$r_i = max\left\|v_i - x_j\right\|_{A_i} \text{for} \forall x_j \in i^{th} \text{ cluster and } u_{ij} \geq u_h \tag{7}$$

where $\|.\|_{A_i}$ is the GK distance norm determined according to equation (2) for which the data x_j belongs to the i-th cluster with membership degree u_{ij} such that $u_{ij} \geq u_h$. Three possibilities should be evaluated if a new data point x_k incomes currently. 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 it may define a new cluster. Alternatively, x_k could be an outlier, which does not affect the data structure. Bellow those three possibilities are considered in detail.

The similarity between the new data point x_k and each of the existing c clusters is evaluated by checking the GK distances defined by eqs. (2) and (3):

$$d_{ik} = \sqrt{(x_k - v_i)[det(F_i)]^{1/n} F_i^{-1}(x_k - v_i)^T}, \tag{8}$$

where each cluster volume is fixed at one. The minimal distance d_{pk} of the k-th data determines the closest cluster p as

$$p = argmin_{i=1,\ldots,c}(d_{ik}). \tag{9}$$

The data point x_k is assigned to the cluster p if the distance d_{pk} is less or equal to the radius r_p, i.e. if the condition

$$d_{pk} \leq r_p \tag{10}$$

is satisfied. If this is the case we recalculate the p-th cluster parameters-cluster center and covariance matrix according to Eqs. (4)-(6).

If condition (10) fails a new potential cluster is investigated. The credibility of the estimated cluster is assessed by the number of points belonging to this cluster with a certain degree of membership. Its lower bound is estimated from the minimal number of data points necessary to learn the parameters of the covariance matrix. Apparently, large amount of data not only guarantees the validity of the covariance matrices but improves the robustness of the algorithm with respect to outliers. Thus, we suggest a larger threshold P_{tol} that corresponds to the desired minimal amount of points falling within the r_i boundary of each cluster. The threshold value is context determined due to the specificity of the considered data set.

If the threshold P_{tol} is satisfied the number of clusters is incremented

$$c := c + 1. \tag{11}$$

Then, the incoming data x_k is accepted as a center of the new cluster v_{new} with a covariance matrix F_{new} initialized by the covariance matrix of the closest cluster

$$v_{new} = x_k, \quad F_{new} = F_p. \tag{12}$$

In opposite case, if the threshold P_{tol} is not passed the data x_k is treated as an outlier, which does not affect the data structure.

The choice of the membership threshold u_h depends on the density of the data set and the level of cluster overlapping. The default value of u_h is 0,5 but being stricter in the identification of proper clusters the prescribed threshold membership degree should be chosen larger. For a more tolerant identification it should be chosen smaller.

4 Decission Support Procedure

Decision of the data assignment is a multivariable task depending on different parameters. During the data income the cluster radius and prototype's coordinates are changed in a real time mode. On the other hand the number of points that form a valuable cluster should not be fixed but rather varied in a certain diapason. The uncertainty of the clustering decision is dictated by the necessity to take into account these vague parameters. A possible solution could be found through a fuzzy decision support technique.

The distance $\mathbf{D}$ between the current income x_k and the closest cluster as well as the amount of the points $\mathbf{P_{inc}}$ surrounding this income are considered as linguistic variables. In the simplest case two fuzzy values - *small* and *large* constructed by respectively left and right opened Γ membership functions grasp each variable universe. The respective membership functions over the universe of $\mathbf{D}$ should have a projection of their intersection point equal to the radius r_p of the closest cluster. The universe of $\mathbf{P_{inc}}$ is characterized by *low* and *high* membership functions that are left and right opened Γ functions having an intersection point projected over the most acceptable number of data P_{tol} that can form a cluster.

The rule base covers the possible combinations of the fuzzy values:

R1: If $\mathbf{D}$ is *small* then classify as *Assign*

R2: If $\mathbf{D}$ is *large* & $\mathbf{P_{inc}}$ is *low* then classify as *Outlier*

R3: If $\mathbf{D}$ is *large* & $\mathbf{P_{inc}}$ is *high* then classify as *New*

Singleton values determine the consequent of the three rules that correspond to the three classification cases: a) *Assign* - assigning the income data to the closest cluster; b) *Outlier* - the income is an outlier which does not change the data structure and c) *New* - the income is surrounded with large number of data that form a new cluster.

The intersection operation *and* in the antecedent of the rules R2 and R3 is accomplished by Larsen product. It involves both input membership degrees in calculation the rule degree of fulfillment. Every rule is solved by applying the Mamdani implication mechanism. The output is obtained as a value within the interval $[0, 1]$. The decision for the final classification is settled to the maximal value among the three outputs calculated for the current income. In the extreme case of equal maximal output degrees we give a preference to keeping the data structure.

In practice, the membership functions of the fuzzy values govern the evolving classification process. Different strategies could be implemented in order to define their parameters. For instance, if the data set presents clusters having equal size then the membership functions could be constantly determined. If the available clusters are much different the membership functions of the decision system should be changed according to the radius r_p and P_{tol} in a real time mode.

The decision procedure was incorporated in the evolving clustering algorithm by the following step description procedure:

1. **Initialization:** Calculate the initial number of clusters c and the corresponding matrices V, U and $F = [F_1, ..., F_c]$ by off-line GK algorithm. Choose in advance: u_h; also r_p and P_{tol} that are needed to construct the membership functions of the fuzzy rule base.
   Repeat for every new data point x_k
2. Calculate d_{ik} by eq. (8).
3. Determine the closest cluster p by eq. (9).
4. Calculate the radius r_p of the closest cluster by eq. (7).
5. Calculate the outputs of the three fuzzy rules.
6. Classify the new data point x_k according to the maximal output value:

   If *Assign* is the largest output then keep the structure. Recalculate V, $F = [F_1, ..., F_c]$ and U;
   If *Outlier* is the largest output then keep the structure;
   If *New* is the largest output then create a new cluster: $c = c+1, v_{c+1} = x_k; F_{c+1} = F_p$
   end

5 Data Set Example and Discussion

A two dimensional artificial data set of 500 data having clusters with different shape and orientation among outliers was clustered to explore the algorithm properties. The two fuzzy values for each linguistic variable of the fuzzy rule base have been defined as Γ membership functions. The left opened functions are set for the small values and the right opened ones for the large values. Each corresponding pair of left and right membership function has a fixed intersection. The two projections are set to $r_p = 4$ for the distance universe and $P_{tol} = 12$ for surrounding points universe.

Initial clusters (Fig. 1a) for the first 200 data points were identified by batch GK procedure. The third cluster (Fig. 1b) was determined for the next income that has been surrounded by enough data. The forth cluster is recognized at 377 data income (Fig. 1c). The next added outliers do not change the clustering result (Fig. 1d). The partition was obtained for the default threshold $u_h = 0,5$. The incorporated fuzzy decision support not only reflects the existing uncertainty but provides robust properties. Thus, by reducing the cluster credibility parameter $P_{tol} = 10$ the algorithm recognizes strongly overlapping clusters (Fig. 2).

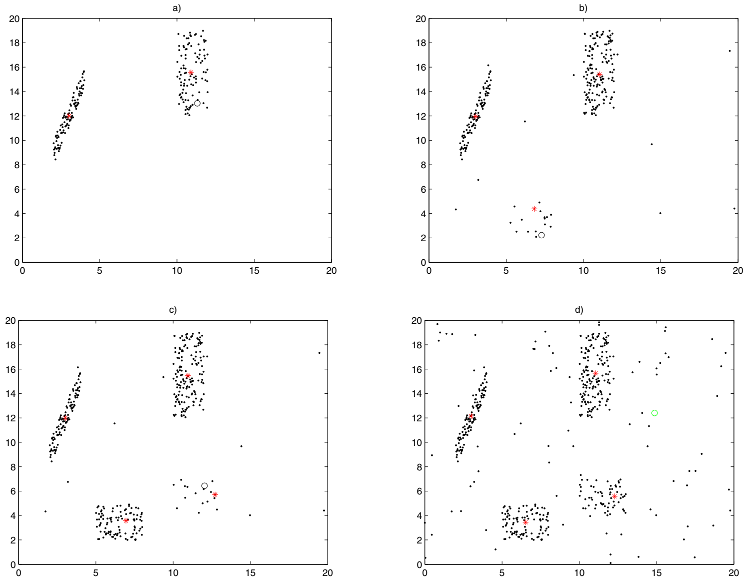


Fig. 1 Data are given by dots, cluster centers - by stars and current income - by circle.

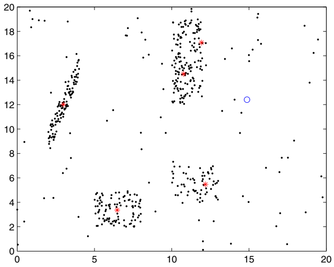


Fig. 2 Five clusters have been discovered by more tolerant clustering

6 Conclusion

The proposed algorithm is applicable in real time clustering tasks. It is based on GK distance metrics and grasps well the data structure even it presents clusters with different shape and orientation. Due to the incorporated fuzzy decision support the algorithm poses robustness characteristics according to the variation of the clustering parameters - the threshold membership degree that affect the cluster radius and number of data that can form a new reliable cluster. By extending the antecedent input vector of the decision support rule base additional clustering parameters could be easily included in order to increase the classification properties.

Acknowledgements. The research was realized within the support of "**SISTER**: Strengthening the IST Research Capacity of Sofia University" project funded from 7 FP of the EU.

References

1. Angelov, P.: Evolving Rule-Based Models: A tool for design of flexible adaptive systems. Springer, Heidelberg (2002)
2. Angelov, P., Filev, D.: An approach to online identification of Takagi-Sugeno fuzzy models. IEEE Trans. Syst. Man Cybern., Part B: Cybern. 34(1), 484–498 (2004)
3. Babuska, R.: Fuzzy modeling for control. Kluwer Academic Publishers, Boston (1998)
4. Bezdek, J.C.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
5. Chiu, S.L.: Fuzzy model identification based on cluster estimation. J. Intell. Fuzzy Syst. 2, 267–278 (1994)
6. Crespo, F., Weber, R.: A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets Syst. 150, 267–284 (2005)
7. Gath, I., Geva, A.B.: Unsupervised optimal fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 7, 773–781 (1989)
8. Georgieva, O., Klawonn, F.: Dynamic data assigning assessment clustering of streaming data. Appl. Soft Computing 8(4), 1305–1313 (2008)
9. Gupta, C., Grossman, R.L.: A single pass generalized incremental algorithm for clustering. In: Proceedings of the Fourth SIAM International Conference on Data Mining, SDM 2004, Lake Buena Vista, FL, USA, pp. 137–153. SIAM, Philadelphia (2004)
10. Gustafson, D.E., Kessel, W.C.: Fuzzy clustering with a fuzzy covariance matrix. In: Proceeding ot the IEEE Conference on Decision and Control, San Diego, pp. 761–766 (1979)
11. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, KDD 2001, San Francisco, California, pp. 97–106. ACM Press, New York (2001)
12. Kasabov, N., Song, Q.: DENFIS: Dynamic evolving neuro-fuzzy inference system and its application for time-series prediction. IEEE Trans. Fuzzy Syst. 10, 144–154 (2002)
13. Keller, A., Klawonn, F.: Adaptation of cluster sizes in objective function based fuzzy clustering. In: Leondes, C.T. (ed.) Intelligent Systems: Technology and Applications. Database and Learning Systems, vol. IV, pp. 181–199. CRC Press, Boca Raton (2003)
14. Yager, R., Filev, D.: Essentials of Fuzzy Modeling and Control. John Wiley & Sons, New York (1994)
15. Yang, J.: Dynamic clustering of evolving streams with a single pass. In: Proceedings of the 19th International Conference on Data Engineering, ICDE 2003, Bangalore, India, pp. 695–697 (2003)

On Jaffray's Decision Model for Belief Functions

Phan H. Giang

Abstract. In this paper, two decision models for Dempster-Shafer belief functions proposed by Jaffray and Giang-Shenoy respectively are compared. Jaffray's model is applicable for general belief function while Giang-Shenoy's model works for the partially consonant class (pcb). Pcb has been shown by Walley as the only class that is consistent with the likelihood principle of statistics. While both models share many nice properties such as tractability, the separation of risk attitude, ambiguity attitude from ambiguity belief, they differ on important aspects. The comparison is made possible by application of both models to pcb. It is shown that due to a Hurwicz-type condition imposed on decision under ignorance, Jaffray's approach violates the consequentialism property (analogous to the law of iterated expectation in probability theory) that is satisfied by Giang-Shenoy approach.

Keywords: Decision theory, Partially consonant, Belief function, Ambiguity attitude.

1 Introduction

Twenty years ago, Jaffray [7] proposed a model of decision making for Dempster-Shafer belief functions that has some remarkable properties. Wakker [12] notes recently that "[Jaffray's] models, developed 20 years ago, achieve a tractability and a separation between risk attitudes, ambiguity attitudes, and ambiguity beliefs that have not yet been obtained in other models popular today". From an axiomatic perspective, Jaffray and Wakker [9] show that the model is a result of weakening Savage's sure-thing principle for

Phan H. Giang
Department of Health Administration and Policy, George Mason University,
Fairfax, Virginia 22030, USA
e-mail: pgiang@gmu.edu

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 313–320.
springerlink.com

unambiguous events. Belief function theory grew out of Dempster's effort in 1960s to generalize Bayesian statistics [1] and later Shafer's proposal of evidential reasoning [11]. Since then a rich body of literature on belief function theory has been accumulated. One of the unsettled issues is decision making with belief function. The problem is that going beyond the realm of probability also means leaving behind the established Bayesian decision theory with many of its nice properties such as the dynamic consistency, consequentialism, immunity from Dutch book argument.

This note is the result of comparison of Jaffray's model with the decision model with consonant belief functions proposed in [4]. Several observations motivate a comparative analysis. Jaffray and Wakker's axiomatization [9] in Savage's style for general belief functions while [4] is in von Neumann-Morgenstern style for a subclass of partially consonant belief function (pcb). The most puzzling fact is that [9] and [4] behave differently when applied for pcb. Given close links between decision and statistical reasoning tasks such as hypothesis testing and estimation, it is important to understand the source and consequences of such a difference.

2 Jaffray-Wakker Derivation

We begin with the framework and notations in which the approaches in [9] and [4] can be described and compared. Ω is a finite set of *states* and $\mathscr{U} = [0,1]$ is a set of *outcomes*. An outcome value is measured in risk-adjusted utility rather than monetary unit. This assumption allows a focus on ambiguity. States can be thought in terms of variables which are denoted by the capital letters to the end of the alphabet e.g., X,Y,Z. Variable instances are denoted by lower case letters. Events are subsets of *states* and are denoted by capital letters to the start of the alphabet e.g., A,B,C. A state is a tuple of instances of all variables. An act is a mapping $d : \Omega \rightarrow \mathscr{U}$. An act is *resolved* on a variable if the knowledge of the variable value is enough to determine the outcome of act. The uncertainty is described by belief functions over the states or, as result of act, over the outcomes. Belief functions are denoted by lower case letters in the middle of the alphabet f,g,h,m etc. In section 3 a concept of utility that has two components is used. Greek letter λ is reserved for the *left* and ρ for the *right* component. Finally, for the rest of this paper slash (/) does not denote arithmetic division, but to separate act's outcome from associating uncertainty degree.

For the sake of self-containedness we repeat basic definitions and well known facts about DS belief functions. A probability *mass* function m is a mapping from the power set of Ω to the unit interval such that sum of masses is 1. The subsets with positive mass are called *foci.* $m : 2^{\Omega} \rightarrow [0,1]$ such that $\sum_{A \subseteq \Omega} m(A) = 1$. Two other forms of a belief function are *belief* (Bel) and *plausibility* (Pl) defined from m as follows: $Bel(B) = \sum_{A \subseteq B} m(A)$ and $Pl(B) = \sum_{A \cap B \neq \emptyset} m(A)$ for any $B \subseteq \Omega$. The most basic fact about DS belief functions is that three forms m, Bel and Pl are equivalent in the sense that given any form the others can be computed. The computation technique is based on the Möbius transform and its inverse. After learning an event B, a belief function is *conditioned* on that set. In terms of plausibility function, conditional belief assume the familiar form of conditional probability: $Pl(A|B) = \frac{Pl(A \cap B)}{Pl(B)}$ for $A, B \subseteq \Omega$.

As a mapping from Ω to $\mathscr{U}$, an act would "carry" uncertainty from the states to the outcomes in the sense that the mass assigned to A would be carried to $d(A) \overset{\text{def}}{=} \cup_{s \in A} d(s)$. So an act is identified with the belief function it induces on $\mathscr{U}$. A *neutrality principle* [9] requires that two acts that induce the same belief function on $\mathscr{U}$ are indifferent. This assumption allows focus on a preference relation $\succeq$ over set of belief functions on $\mathscr{U}$ denoted by $\mathscr{F}$. As usual, $\succeq$ is assumed to be a *weak order* (complete and transitive) with asymmetric and symmetric parts denoted by $\succ$ and $\sim$.

An event (subset) A is called *ambiguous* if there exists a focus that intersects with both A and its complement $\bar{A}$. An ambiguous event A is characterized by a strict inequality $Pl(A) + Pl(\bar{A}) > 1$. This is the case because (*i*) there exists a focus B whose strictly positive mass $m(B)$ is counted twice in both $Pl(A)$ and $Pl(\bar{A})$ and (*ii*) the mass of any other focus is counted at least in either $Pl(A)$ or $Pl(\bar{A})$. Conversely, a *unambiguous* event is characterized by equality $Pl(A) + Pl(\bar{A}) = 1$. Because of symmetry, (un)ambiguity is a property of both an event and its complement. Intuitively, a unambiguous event and its complement separate the foci into two non-overlapping groups.

Jaffray and Wakker show that [9] the utility of act f is the weighted (by foci masses) average of the utilities of elementary acts e_B induced by its foci if if relation $\succeq$ satisfies mixture continuity and and weak sure-thing principle. This is a consequence of Herstein and Milnor's theorem [6].

$$v(f) = \sum_{B \subseteq \mathscr{U}} m_f(B) v(e_B) \tag{1}$$

The *weak* sure-thing principle requires that two pairs of acts (d_1, d_2) and (d'_1, d'_2) such that (*i*) $d_1(s) = d_2(s) = c$ and $d'_1(s) = d'_2(s) = c'$ for all $s \in \bar{A}$; and (*ii*) $d_1(v) = d'_1(v)$ and $d_2(v) = d'_2(v)$ for all $v \in A$ where A is a unambiguous event should have the same preference direction i.e., $d_1 \succeq d_2$ iff $d'_1 \succeq d'_2$. This is weaker than the original Savagian sure-thing principle by adding a condition on the unambiguity of the common-outcome set.

Belief functions are closed under linear mixture. For belief functions $f, g \in \mathscr{F}$ and $0 \leq \mu \leq 1$ is a real number, a *linear mixture* of f and g: $h \equiv \mu f + (1-\mu) g$ defined as $m_h(A) \overset{\text{def}}{=} \mu m_f(A) + (1-\mu) m_g(A)$ for any $A \subseteq \mathscr{U}$ is also a belief function (m_f is the mass function of f). The corresponding equalities for Bel and Pl also hold. The *mixture continuity* condition [6] requires that for a triple $f, g, h \in \mathscr{F}$ satisfying $f \succeq g \succeq h$ there is a real number μ such that $\mu f + (1-\mu) h \sim g$. The idea is that you can fine-tune mixture weight μ so that mixture spans whole range from h to g.

So the key problem is how to determine the utility of elementary belief functions $v(e_B)$. Jaffray argued for the adoption of the principle of "total ignorance" because e_B represents situation where the only information available is that outcome will be in B. In particular,

$$v(e_B) = \alpha(\bot_B, \top_B)\bot_B + (1 - \alpha(\bot_B, \top_B))\top_B \tag{2}$$

where $\bot_B$ ($\top_B$) is the minimal (maximal) element in B and α is two-place real function. Interestingly, in $v(e_B)$ all intermediate outcomes are ignored and only the top and the bottom elements of B matter. Thus, $v(f) =$

$$\sum_{B \subseteq \mathscr{U}} m_f(B)\,(\alpha(\bot_B, \top_B)\bot_B + (1 - \alpha(\bot_B, \top_B))\top_B) \tag{3}$$

This utility expression separates uncertain information represented by f and the ambiguity attitude represented by $\alpha(\cdot,\cdot)$.[1] It is tractable because it is only necessary to determine α value for every pair of elements of $\mathscr{U}$ rather than for each subset.

3 A Decision Model with Partially Consonant Belief Functions

In [4] Giang and Shenoy proposed a decision model for partially consonant belief functions (pcb). First studied by Walley [13], pcb is a class of belief functions with foci partitioned into non-overlapping groups and within each group, they are nested. For example $\{A_{10} \supset A_{11} \ldots \supset A_{1n_1}\}$, $\{A_{20} \supset A_{21} \ldots \supset A_{2n_2}\} \ldots \{A_{m0} \supset A_{m1} \ldots \supset A_{mn_m}\}$ and $A_{i0} \cap A_{j0} = \emptyset$. We can assume without loss of information that $\cup_i A_{i0} = \Omega$. This class includes both probability and possibility functions as special cases. The fundamental importance of pcb is due to Walley's result [13] that pcb is the only class of DS belief functions that is consistent with the likelihood principle of statistics.

Consider algebra $\mathscr{A}$ formed from $\{A_{10}, A_{20} \ldots A_{m0}\}$, an event A is unambiguous iff $A \in \mathscr{A}$. A pcb can be decomposed into a probability function on $\mathscr{A}$: $Pl(A_{i0}) \overset{\text{def}}{=} Pl(A_{i0})$ for $1 \leq i \leq m$ and m conditional possibility functions π_i on A_{i0}: $\pi_i(C) \overset{\text{def}}{=} Pl(C|A_{i0})$ for $C \subseteq A_{i0}$.

An act $d : \Omega \rightarrow \mathscr{U}$ can be rewritten $d = [E_1/w_1, E_2/w_2 \ldots, E_k/w_k]$ where $E_i = d^{-1}(w_i)$. A key observation is that act d under pcb can be viewed as two-stage act $[A_{i0}/[B_{i1}/w_1, \ldots B_{ik}/w_k]]_{i=1}^{m}$ where $B_{ij} = A_{i0} \cap E_j$ for $i = 1, m$ and $j = 1, k$. Thus, the first stage is a probabilistic act and the second stage is a set of possibilistic acts. This is amenable to evaluation in a folding back manner. First, the second-stage possibilistic lotteries are evaluated and their certainty equivalences are plugged into the probabilistic lottery which in its turn is evaluated by standard expected utility. This fold-back procedure implies *consequentialism* property [10].

[1] The risk attitude is ignored in this review for the sake of clarity.

In [5], possibilistic lotteries are evaluated by a two-component or "binary" utility function $t : \mathscr{F}_\pi \to \Psi$ where $\mathscr{F}_\pi$ is the set of possibilistic functions on $\mathscr{U}$ and $\Psi \overset{\text{def}}{=} \{\langle\lambda,\rho\rangle \,|\, \lambda,\rho \in [0,1]$ and $\max(\lambda,\rho) = 1\}$. An order $\geqq$ on Ψ, a component-wise operation *max* on pairs and product of a scalar and a pair can be defined as follows

$$\langle\lambda,\rho\rangle \geqq \langle\lambda',\rho'\rangle \quad \text{iff} \quad \lambda \geq \lambda' \ \text{ and } \ \rho \leq \rho' \tag{4}$$

$$\max(\langle\lambda,\rho\rangle,\langle\lambda',\rho'\rangle) \overset{\text{def}}{=} \langle\max(\lambda,\lambda'),\max(\rho,\rho')\rangle \tag{5}$$

$$\pi\langle\lambda,\rho\rangle \overset{\text{def}}{=} \langle\pi\lambda,\pi\rho\rangle \tag{6}$$

$$t([\pi_{ij}/w_j]_{j=1}^{k}) = \max\{\pi_{ij}t(w_j)\} = \left\langle \max_j(\pi_{ij}\lambda_j), \max_j(\pi_{ij}\rho_j) \right\rangle \tag{7}$$

For a continuous t, one can define $t^{-1} : \Psi \to [0,1]$ as follows: for $w \in [0,1]$ $t^{-1}(\langle\lambda,\rho\rangle) \mapsto w$ if $t(w) = \langle\lambda,\rho\rangle$. This definition of t^{-1} justifies familiar cancellations: $t(t^{-1}(\langle\lambda,\rho\rangle)) = \langle\lambda,\rho\rangle$ and $t^{-1}(t(w)) = w$ for $w \in [0,1]$. The utility function for pcb lotteries [4] has the following form

$$u([p_i/[\pi_{i1}/w_1,\pi_{i2}/w_2\ldots\pi_{ik}/w_k]_{i=1}^{m}) = \sum_{i=1}^{m} p_i t^{-1}(\max_j\{\pi_{ij}t(w_j)\}) \tag{8}$$

4 A Comparison

We have seen that an act under uncertainty described by pcb can be evaluated in two different ways according Jaffray's model (J-utility function v) vs the model in [4] (GS-utility function u). Our inquiry is to answer two questions (a) is Jaffray's model equivalent to GS model when the belief function is pcb and (b) if it is not then why and what properties of GS model that are not held by Jaffray's model and vice versa.

Let us consider the simplest case of elementary belief functions e_B with the only focus B i.e., mass $m_{e_B}(B) = 1$. J-utility $v(e_B) = \alpha(\bot_B,\top_B)\bot_B + (1-\alpha(\bot_B,\top_B))\top_B$. This is a linear combination of utilities of the bottom and top elements with the weight equal to $\alpha(\bot_B,\top_B)$.

As for GS-utility, $u(e_B) = t^{-1}(\max\{t(b)|b \in B\})$. Suppose $t(\bot_B) = \langle\lambda_\bot^B,\rho_\bot^B\rangle$, $t(\top_B) = \langle\lambda_\top^B,\rho_\top^B\rangle$ and $t(b) = \langle\lambda_b,\rho_b\rangle$. Because $\top_B \geqq b$ for any $b \in B$ by definition of $\geqq$, for the left component $\lambda_\top^B \geq \lambda_b$. Similarly, because $b \geqq \bot_B$ for the right component $\rho_\bot^B \geq \rho_b$ for any $b \in B$. Therefore $u(e_B) = t^{-1}(\max\{t(b)|b \in B\}) = t^{-1}\left(\langle\lambda_\top^B,\rho_\bot^B\rangle\right)$. $u(e_B)$ depends on $\bot_B$ and $\top_B$ only. Thus, in both J-utility and GS-utility, for decision under ignorance only extreme outcome matter, the intermediate outcomes are ignored.

However, there is difference in the behavior of u and v. Consider the case when the bottom element of B is still preferable to the fair gamble $\bot_B \succeq [1/1,1/0]$. It follows that $\lambda_\bot^B = 1$. So $u(e_B) = t^{-1}(\langle 1,\rho_\bot^B\rangle) = t^{-1}(\langle\lambda_\bot^B,\rho_\bot^B\rangle) = t^{-1}(t(\bot_B)) = \bot_B$. Thus, GS-utility equalizes e_B with its bottom element $\bot_B$.

Analogously, it can be shown that if the fair gamble is preferable to the top element of B, then $u(e_B) = \top_B$. This particular behavior is consistent with the "optimistic" and "pessimistic" modes of possibilistic decision described in [3]. J-utility, being a linear combination of top and bottom elements, can not have this behavior unless $\alpha(\perp_B, \top_B) = 1$ or $\alpha(\perp_B, \top_B) = 0$.

Let us consider an example to clarify properties of GS-utility that are not satisfied by J-utility. There are two variables X, Y with domains $\mathscr{X} = \{x_1, x_2\}$ and $\mathscr{Y} = \{y_1, y_2\}$. $\Omega = \{x_1y_1, x_1y_2, x_2y_1, x_2y_2\}$. Suppose a belief function with three nested foci is given as follows. $Pl(x_1) = 1$, $Pl(x_2) = 1$, $Pl(y_1|x_1) = 1$, $Pl(y_2|x_1) = 0.3$, $Pl(y_1|x_2) = 1$ and $Pl(y_2|x_2) = 0.5$.

focus	Ω	x_1y_1, x_2y_1, x_2y_2	x_1y_1, x_2y_1
mass	0.3	0.2	0.5

Consider act d with the first stage resolved on X. If $X = x_1$ then L_1; if $X = x_2$ then L_2. The second stage is resolved on Y. If $Y = y_1$ then both L_1 and L_2 deliver 1; if $Y = y_2$ then both L_1 and L_2 deliver 0. In two-stage view $L = [x_1/L_1, x_2/L_2]$, $L_1 = [y_1/1, y_2/0]$ and $L_2 = [y_1/1, y_2/0]$. This act will be evaluated by J-utility and GS-utility in a single-stage and a two-stage views.

GS-utility calculates two-stage view of d as follows. Since $u(L_1) = \langle 1, 0.3\rangle$ and $u(L_2) = \langle 1, 0.5\rangle$, hence $u(L) = \langle 1, 0.5\rangle$. In one-stage view of d, $L' = [\{x_1y_1, x_2y_1\}/1,\ \{x_1y_2, x_2y_2\}/0]$ i.e., if $\{x_1y_1, x_2y_1\}$ occurs, the outcome is 1; otherwise it is 0. Because $Pl(\{x_1y_1, x_2y_1\}) = 1$ and $Pl(\{x_1y_2, x_2y_2\}) = 0.5$ $u(L') = \langle 1, 0.5\rangle$. Thus, two views of an act are equal under GS-utility. In general, GS-utility satisfies the consequentialism for possibilistic acts.

J-utility calculates one-stage view. $d(\Omega) = \{0, 1\}$, $d(\{x_1y_1, x_2y_1, x_2y_2\}) = \{0, 1\}$ and $d(\{x_1y_1, x_2y_1\}) = \{1\}$. The belief function f on $\mathscr{U}$ induced by d and m has 2 foci $m_f(\{0, 1\}) = 0.5$ and $m_f(\{1\}) = 0.5$. J-utility is $v(f) = 0.5v(e_{\{0,1\}}) + 0.5v(e_{\{1\}})$. Because $v(e_{\{0,1\}}) = 1 - \alpha(0, 1)$, $v(e_{\{1\}}) = 1$

$$v(f) = 1 - 0.5\alpha(0, 1). \tag{9}$$

In two stage view, to compute utility for second stage lotteries it is necessary to compute belief functions obtained from m by conditioning on $X = x_1$ and $X = x_2$ as shown in the following table.

focus	Ω	x_1y_1, x_2y_1, x_2y_2	x_1y_1, x_2y_1
mass	0.3	0.2	0.5
conditioning event	x_1y_1, x_1y_2		
conditional foci	x_1y_1, x_1y_2	x_1y_1	x_1y_1
conditioning event	x_2y_1, x_2y_2		
conditional foci	x_2y_1, x_2y_2	x_2y_1, x_2y_2	x_2y_1

Denote the conditionals on $X = x_i$ by m_i for $i = 1, 2$. $m_1(\{x_1y_1, x_1y_2\}) = 0.3$ and $m_1(\{x_1y_1\}) = 0.7$. $m_2(\{x_2y_1, x_2y_2\}) = 0.5$ and $m_2(\{x_2y_1\}) = 0.5$. The belief function on $\mathscr{U}$ induced by m_i is f_i. $m_{f_1}(\{0, 1\}) = 0.3$, $m_{f_1}(\{1\}) = 0.7$, $m_{f_2}(\{0, 1\}) = 0.5$ and $m_{f_2}(\{1\}) = 0.5$. J-utility calculation yields

$$v(f_1) = 0.3(1-\alpha(0,1)) + 0.7 = 1 - 0.3\alpha(0,1) \tag{10}$$
$$v(f_2) = 0.5(1-\alpha(0,1)) + 0.5 = 1 - 0.5\alpha(0,1) \tag{11}$$

The first stage is that if $X = x_1$ then the outcome is $w_1 = 1 - 0.3\alpha(0,1)$ and if $X = x_2$ then it is $w_2 = 1 - 0.5\alpha(0,1)$. The induced belief function on $\mathscr{U}$ is $e_{\{w_1,w_2\}}$ (i.e., the masses of the all original foci are transferred to $\{w_1,w_2\}$). So the J-utility is

$$\begin{aligned} v(e_{\{w_1,w_2\}}) = & \quad \alpha(w_1,w_2)w_1 + (1-\alpha(w_1,w_2))w_2 & (12)\\ & = 1 - 0.5\alpha(0,1) + 0.2\alpha(0,1)\alpha(w_1,w_2) & (13) \end{aligned}$$

The difference between (9) and (13) indicates that J-utility values of one-stage and two-stage views of an act differently except for trivial weights. In other words, the consequentialism does not hold for J-utility and hence the folding-back procedure is not applicable.

The consequentialism failure of J-utility requires a careful examination not just because folding-back is a convenient evaluation method but also it embodies a *normative* requirement that the *rational* value of a choice should not depend on the way it is *presented* to the decision maker (although in real life such manipulations do have effect on choice). Shifting between one-stage and two-stage views of an act does not add or lose any new information, therefore, the utilities should not change. This is the idea behind the law of iterated expectation in probability $E[X] = E[E[X|Y]]$. If this property is violated, a DM can be subjected to a Dutch-book type trap in which she is led to make a collection of choices that in toto costs her a sure loss. To understand the reason behind this failure, we note that (1) set of pcb lotteries is closed under mixture and (2) GS-utility satisfies both the mixture continuity and independence conditions. By Herstein and Milnor's theorem [6], u is consistent with the form of (1). This observation points to Hurwicz's condition (2), that distinguishes J-utility from GS-utility, as responsible for the failure. Jaffray and Jeleva [8] observe the problem of Hurwicz's criterion in decision context not involving belief function.

5 Conclusion

This note analyzes and compares two models of decision making with DS belief function using J-utility and GS-utility. Jaffray's model is applicable for general belief function while Giang-Shenoy's model is applicable only for partially consonant belief function. Several nice properties such as tractability, separation of risk attitude, ambiguity attitude from ambiguity belief and the satisfaction of mixture continuity and independence are satisfied by both models. The difference, however, is that while consequentialism holds for GS-utility model when the secondary lotteries are possibilistic, J-utility loses this property because Hurwicz's condition is used for decision under ignorance. The loss could have undesired consequences that the users of J-utility should keep in mind. For example, a folding-back algorithm can not be used and J-utility of an act may depend on the way the act is presented.

An obvious question arisen at this point is about the consequence of restriction on pcb class in GS model. An answer is contained in Walley's result [13] that pcb is the only class that is consistent with the likelihood principle, hence, pcb is not a restriction at all the statistical reasoning context. We also note Dempster's view [2] that a belief function is the image of a probability distribution on a space S through a multi-valued mapping from S to Ω. We can extend the focus $A_i \subseteq \Omega$ which is the image of s_i to s_iA_i on $S \times \Omega$. The foci on the extended space do not intersect. So, any belief function on Ω can be viewed as a pcb on the extended space $S \times \Omega$.

References

1. Dempster, A.: A generalization of Bayesian inference (with discussion). J. Roy. Statist. Soc. Ser. B 30, 205–247 (1968)
2. Dempster, A.: Upper and lower probability induced by multivalued mapping. Ann. Probab. 38, 325–339 (1967)
3. Dubois, D., Godo, L., Prade, H., Zapico, A.: On the possibilistic decision model: from decision under uncertainty to case-based decision. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 7(6), 631–670 (1999)
4. Giang, P.H., Shenoy, P.P.: Decision making with partially consonant belief functions. In: Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI 2003). Morgan Kaufmann, San Francisco (2003)
5. Giang, P.H., Shenoy, P.P.: Two axiomatic approaches to decision making using possibility theory. European J. Oper. Res. 162(2), 450–467 (2005)
6. Herstein, I., Milnor, J.: An axiomatic approach to measurable utility. Econometrica 21, 291–297 (1953)
7. Jaffray, J.-Y.: Linear utility theory for belief functions. Oper. Res. Lett. 8, 107–112 (1989)
8. Jaffray, J.-Y., Jeleva, M.: Information processing under imprecise risk with an insurance demand illustration. Internat. J. Approx. Reason. 49(1), 117–129 (2008)
9. Jaffray, J.-Y., Wakker, P.: Decision making with belief functions: Compatibility and incompatibility with the sure-thing principle. J. Risk Uncertain. 8(3), 255–271 (1994)
10. Machina, M.: Dynamic consistency and non-expected utility models of choice under uncertainty. J. Econ. Liter. 27, 1622–1668 (1989)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
12. Wakker, P.: Jaffray's ideas on ambiguity. Working paper. Econometric Institute, Erasmus University, Rotterdam, the Netherlands, `http://people.few.eur.nl/wakker/pdf/jaffray.pdf`
13. Walley, P.: Belief function representation of statistical evidence. Ann. Statist. 15(4), 1439–1465 (1987)

Quasi Conjunction and p-Entailment in Nonmonotonic Reasoning

A. Gilio and G. Sanfilippo

Abstract. We study, in the setting of coherence, the extension of a probability assessment defined on n conditional events to their quasi conjunction. We consider, in particular, two special cases of logical dependencies; moreover, we examine the relationship between the notion of p-entailment of Adams and the inclusion relation of Goodman and Nguyen. We also study the probabilistic semantics of the QAND rule of Dubois and Prade; then, we give a theoretical result on p-entailment.

Keywords: Coherence, Lower/upper probability bounds, Quasi conjunction, QAND rule, p-entailment.

1 Introduction

In classical (monotonic) logic, if a conclusion C follows from some premises, then C also follows when the set of premises is enlarged; that is, adding premises never invalidates any conclusions. Differently, in (nonmonotonic) commonsense reasoning typically we are in a situation of partial knowledge and a conclusion reached from a set of premises may be retracted, when some premises are added. Nonmonotonic reasoning is a relevant topic in the field of artificial intelligence and has been studied in literature by many, symbolic or numerical, formalisms (see, e.g. [2, 3, 8]). A remarkable theory, related with

A. Gilio
Dipartimento di Metodi e Modelli Matematici,
University of Rome "La Sapienza", Italy
e-mail: gilio@dmmm.uniroma1.it

G. Sanfilippo
Dipartimento di Scienze Statistiche e Matematiche "S. Vianelli",
University of Palermo, Italy
e-mail: sanfilippo@unipa.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 321–328.
springerlink.com

nonmonotonic reasoning, has been given by Adams in his probabilistic logic of conditionals ([1]). We recall that the approach of Adams can be developed with full generality by exploiting a coherence-based probabilistic reasoning, which allows a direct assignment of conditional probabilities, without assuming that conditioning events have a positive probability ([4]). A basic notion in the work of Adams is the quasi conjunction of conditionals, which also plays a relevant role in the work of Dubois and Prade on conditional objects, where a suitable QAND rule is introduced to characterize entailment from a knowledge base. In our paper we deepen some probabilistic aspects related with QAND rule and with the conditional probabilistic logic of Adams.

The paper is organized as follows: In Section 2 we recall the p-consistency and p-entailment notions in the setting of coherence; in Sections 3 and 4 we study the lower and upper probability bounds for the quasi conjunction of conditional events, by relating them to Lukasiewicz t-norm and Hamacher t-conorm, respectively; we also examine two special cases of logical dependencies related with the inclusion relation of Goodman and Nguyen and with the compound probability theorem; in Section 5 we deepen the analysis on the lower and upper probability bounds for the quasi conjunction, by examining further aspects; in Section 6 we examine the relation between the notion of p-entailment and the inclusion relation of Goodman and Nguyen; then, we study the probabilistic semantics of *QAND* rule, by proving the p-entailment from any given family of conditional events $\mathscr{F}$ to the quasi conjunction $\mathscr{C}(\mathscr{F})$; finally, we prove the equivalence between p-entailment from $\mathscr{F}$ and p-entailment from the $\mathscr{C}(\mathscr{S})$, for some non-empty subset $\mathscr{S}$ of $\mathscr{F}$; in Section 7 we give some conclusions.

Due to the lack of space (almost) all proofs of our results are omitted.

2 Some Preliminary Notions

In this section we recall, in the setting of coherence ([4, 5]), the notions of p-consistency and p-entailment of Adams ([1]). Given a conditional knowledge base $\mathscr{K}_n = \{H_i \mid\!\sim E_i\,,\, i = 1,\dots,n\}$, we denote by $\mathscr{F}_n = \{E_i|H_i\,,\, i = 1,\dots,n\}$ the associated family of conditional events.

Definition 1. *The knowledge base* $\mathscr{K}_n = \{H_i \mid\!\sim E_i\,,\, i = 1,\dots,n\}$ *is* p-consistent *iff, for every set of lower bounds* $\{\alpha_i, i = 1,\dots,n\}$*, with* $\alpha_i \in [0,1)$*, there exists a coherent probability assessment* $\{p_i, i = 1,\dots,n\}$ *on* $\mathscr{F}_n$*, with* $p_i = P(E_i|H_i)$*, such that* $p_i \geq \alpha_i, i = 1,\dots,n$*.*

We say that $\mathscr{F}_n$ is p-consistent when it is p-consistent the associated knowledge base $\mathscr{K}_n$; then, we point out that the property of p-consistency for $\mathscr{F}_n$ is equivalent to the coherence of the assessment $(p_1, p_2,\dots,p_n) = (1,1,\dots,1)$ on $\mathscr{F}_n$ (strict p-consistency, [4]).

Definition 2. *A p-consistent knowledge base* $\mathscr{K}_n = \{H_i \mid\!\sim E_i\,,\, i = 1,\dots,n\}$ p-entails *the conditional* $A \mid\!\sim B$*, denoted* $\mathscr{K}_n \Rightarrow_p A \mid\!\sim B$*, iff there exists a* *non-empty subset* $\Gamma \subseteq \{1,\ldots,n\}$ *such that, for every* $\alpha \in [0,1)$, *there exists a set of lower bounds* $\{\alpha_i, i \in \Gamma\}$, *with* $\alpha_i \in [0,1)$, *such that for all coherent probability assessments* $\{z, p_i, i \in \Gamma\}$ *defined on* $\{B|A, E_i|H_i,\, i \in \Gamma\}$, *with* $z = P(B|A)$ *and* $p_i = P(E_i|H_i)$, *if* $p_i \geq \alpha_i$ *for every* $i \in \Gamma$, *then* $z \geq \alpha$.

Remark 1. We say that a family of conditional events $\mathscr{F}_n$ p-entails a conditional event $B|A$ when the associated knowledge base $\mathscr{K}_n$ p-entails the conditional $A \mid\!\sim B$. Therefore, p-entailment of $B|A$ from $\mathscr{F}_n$ amounts to the existence of a non-empty subset $\mathscr{S} = \{E_i|H_i,\, i \in \Gamma\}$ of $\mathscr{F}_n$ such that, defining $P(E_i|H_i) = p_i, P(B|A) = z$, for every $\alpha \in [0,1)$, there exist lower bounds $\alpha_i, i \in \Gamma$, with $\alpha_i \in [0,1)$, such that $p_i \geq \alpha_i$, $i \in \Gamma$, implies $z \geq \alpha$.

3 Lower and Upper Bounds for Quasi Conjunction

Let A,H,B,K be logically independent events, with $H \neq \emptyset, K \neq \emptyset$. The quasi conjunction of two conditional events $A|H$ and $B|K$, as defined in ([1]), is given by $\mathscr{C}(A|H,B|K) = (AH \vee H^c) \wedge (BK \vee K^c)|(H \vee K)$. We recall that quasi conjunction plays a key role in the logic of conditional objects ([3]).

It can be easily verified that, for every pair (x,y), with $x \in [0,1], y \in [0,1]$, the probability assessment (x,y) on $\{A|H,B|K\}$ is coherent. Then, it can be verified (see [5]) that, for each given assessment (x,y) on $\{A|H,B|K\}$, the probability assessment $\mathscr{P} = (x,y,z)$ on $\mathscr{F} = \{A|H,B|K,\mathscr{C}(A|H,B|K)\}$, with $z = P[\mathscr{C}(A|H,B|K)]$, is a coherent extension of (x,y) if and only if

$$\max(x+y-1,0) = l \leq z \leq u = \begin{cases} \frac{x+y-2xy}{1-xy}, & (x,y) \neq (1,1), \\ 1, & (x,y) = (1,1). \end{cases}$$

We observe that the lower bound l coincides with the Lukasiewicz t-norm $T_L(x,y)$, while the upper bound u coincides with the Hamacher t-conorm $S_0^H(x,y)$, with parameter $\lambda = 0$ (see [7]).

Remark 2. Notice that, if the events A,B,H,K were not logically independent, then some constituents C_h's (at least one) would become impossible and the lower bound l could increase, while the upper bound u could decrease. To examine this aspect we will consider two special cases of logical dependencies.

3.1 The Case $A|H \subseteq B|K$

We recall the Goodman & Nguyen relation of inclusion for conditional events ([6]). Given two conditional events $A|H$ and $B|K$, we say that $A|H$ implies $B|K$, denoted by $A|H \subseteq B|K$, if and only if $AH \subseteq BK$ and $B^cK \subseteq A^cH$. Given any conditional events $A|H, B|K$, we denote by Π_x the set of coherent probability assessment x on $A|H$, by Π_y the set of coherent probability assessment y on $B|K$ and by Π the set of coherent probability assessment (x,y) on $\{A|H,B|K\}$; moreover we indicate by $T_{x \leq y}$ the triangle $\{(x,y) \in [0,1]^2 : x \leq y\}$. In the next result, to avoid the specific analysis of some trivial cases, we assume $\Pi_x = \Pi_y = [0,1]$. We have

Theorem 1. *Let $A|H,B|K$ be two conditional events, with $\Pi_x = \Pi_y = [0,1]$. Then: $A|H \subseteq B|K \Longleftrightarrow \Pi \subseteq T_{x\leq y}$.*

Actually, concerning Theorem 1, the implication $\Longrightarrow$ also holds in trivial cases where $\Pi_x \subset [0,1]$, or $\Pi_y \subset [0,1]$.

Remark 3. We observe that, under the hypothesis $A|H \subseteq B|K$, we have $\mathscr{C}(A|H, B|K) = (AH \vee H^cBK)\,|\,(H \vee K)$ and, as we can verify, it is

$$A|H \subseteq \mathscr{C}(A|H,B|K) \subseteq B|K\,. \tag{1}$$

Moreover, if we do not assume further logical relations, then $\Pi = T_{x\leq y}$ and, for each coherent assessment (x,y) on $\{A|H,B|K\}$, the extension $z = P[\mathscr{C}(A|H,B|K)]$ is coherent if and only if $l \leq z \leq u$, where

$$l = x = \min(x,y)\,,\; u = y = \max(x,y)\,.$$

We remark that the values l,u may change if we add further logical relations; in particular, if $H = K$, it is $\mathscr{C}(A|H,B|H) = A|H$, in which case $l = u = x$.

Finally, in agreement with Remark 2, we observe that

$$T_L(x,y) \leq \min(x,y) \leq \max(x,y) \leq S_0^H(x,y).$$

3.2 Compound Probability Theorem

We now examine the quasi conjunction of $A|H$ and $B|AH$, with A,B,H logically independent events. As it can be easily verified, we have $\mathscr{C}(A|H,B|AH) = AB|H$; moreover, the probability assessment (x,y) on $\{A|H,B|AH\}$ is coherent, for every $(x,y) \in [0,1]^2$. Hence, by the compound probability theorem, if the assessment $\mathscr{P} = (x,y,z)$ on $\mathscr{F} = \{A|H,B|AH,AB|H\}$ is coherent, then $z = xy$; that is, $l = u = xy$. In agreement with Remark 2, we observe that $T_L(x,y) \leq xy \leq S_0^H(x,y)$. More in general, given a family $\mathscr{F} = \{A_1|H,A_2|A_1H,\ldots,A_n|A_1\cdots A_{n-1}H\}$, by iteratively exploiting the associative property, we have

$$\mathscr{C}(\mathscr{F}) = \mathscr{C}(\mathscr{C}(A_1|H,A_2|A_1H),A_3|A_2A_1H,\ldots,A_n|A_1\cdots A_{n-1}H) =$$

$$= \mathscr{C}(A_1A_2|H,A_3|A_2A_1H,\ldots,A_n|A_1\cdots A_{n-1}H) = \cdots = A_1A_2\cdots A_n|H\,;$$

thus, by the compound probability theorem, if the assessment $\mathscr{P} = (p_1,\ldots, p_n,z)$ on $\mathscr{F} \cup \{\mathscr{C}(\mathscr{F})\}$ is coherent, then $z = l = u = p_1 \cdot p_2 \cdots p_n$.

4 Lower and Upper Bounds for the Quasi Conjunction of n Conditional Events

Given the family $\mathscr{F}_n = \{E_1|H_1, \ldots, E_n|H_n\}$, we denote by $\mathscr{C}(\mathscr{F}_n)$ the quasi conjunction of the conditional events in $\mathscr{F}_n$. By the associative property of quasi conjunction, defining $\mathscr{F}_k = \{E_1|H_1, \ldots, E_k|H_k\}$, for each $k = 2, \ldots, n$, it is $\mathscr{C}(\mathscr{F}_k) = \mathscr{C}(\mathscr{C}(\mathscr{F}_{k-1}), E_k|H_k)$. Then, we have

Theorem 2. *Given a probability assessment $\mathscr{P}_n = (p_1, p_2, \ldots, p_n)$ on $\mathscr{F}_n = \{E_1|H_1, \ldots, E_n|H_n\}$, let $[l_k, u_k]$ be the interval of coherent extensions of the assessment $\mathscr{P}_k = (p_1, p_2, \ldots, p_k)$ on the quasi conjunction $\mathscr{C}(\mathscr{F}_k)$, where $\mathscr{F}_k = \{E_1|H_1, \ldots, E_k|H_k\}$. Then, assuming $E_1, H_1, \ldots, E_n, H_n$ logically independent, for each $k = 2, \ldots, n$, we have*

$$l_k = T_L(p_1, p_2, \ldots, p_k), \quad u_k = S_0^H(p_1, p_2, \ldots, p_k),$$

where T_L is the Lukasiewicz t-norm and S_0^H is the Hamacher t-conorm, with parameter $\lambda = 0$.

4.1 The Case $E_1|H_1 \subseteq E_2|H_2 \subseteq \ldots \subseteq E_n|H_n$

In this subsection we give a result on quasi conjunctions when $E_i|H_i \subseteq E_{i+1}|H_{i+1}, i = 1, \ldots, n-1$. We have

Theorem 3. *Given a family $\mathscr{F}_n = \{E_1|H_1, \ldots, E_n|H_n\}$ of conditional events such that $E_1|H_1 \subseteq E_2|H_2 \subseteq \ldots \subseteq E_n|H_n$, and a coherent probability assessment $\mathscr{P}_n = (p_1, p_2, \ldots, p_n)$ on $\mathscr{F}_n$, let $\mathscr{C}(\mathscr{F}_k)$ be the quasi conjunction of $\mathscr{F}_k = \{E_i|H_i, i = 1, \ldots, k\}$, $k = 2, \ldots, n$. Moreover, let $[l_k, u_k]$ be the interval of coherent extensions on $\mathscr{C}(\mathscr{F}_k)$ of the assessment $(p_1, p_2, \ldots, p_k)$ on $\mathscr{F}_k$. We have: (i) $E_1|H_1 \subseteq \mathscr{C}(\mathscr{F}_2) \subseteq \ldots \subseteq \mathscr{C}(\mathscr{F}_n) \subseteq E_n|H_n$; (ii) by assuming no further logical relations, any probability assessment $(z_2, \ldots, z_k)$ on $\{\mathscr{C}(\mathscr{F}_2), \ldots, \mathscr{C}(\mathscr{F}_k)\}$ is a coherent extension of the assessment $(p_1, p_2, \ldots, p_k)$ on $\mathscr{F}_k$ if and only if $p_1 \leq z_2 \leq \cdots \leq z_k \leq p_k$, $k = 2, \ldots, n$; moreover*

$$l_k = \min(p_1, \ldots, p_k) = p_1, \; u_k = \max(p_1, \ldots, p_k) = p_k, \; k = 2, \ldots, n.$$

Proof. (i) By iteratively applying (1) and by the associative property of quasi conjunction, we have $\mathscr{C}(\mathscr{F}_{k-1}) \subseteq \mathscr{C}(\mathscr{F}_k) \subseteq E_k|H_k$, $k = 2, \ldots, n$;
(ii) by exploiting the logical relations in point (i), the assertions immediately follow by applying a reasoning similar to that in Remark 3. □

5 Further Aspects on the Lower and Upper Bounds

Now, given any coherent assessment (x, y) on $\{A|H, B|K\}$, we examine further probabilistic aspects on the lower and upper bounds, l and u, for the coherent extensions $z = P[\mathscr{C}(A|H, B|K)]$. More precisely, given any number $\gamma \in [0, 1]$, we are interested in finding:

(i) the set $\mathscr{L}_\gamma$ of the coherent assessments (x,y) on $\{A|H,B|K\}$ such that, for each $(x,y) \in \mathscr{L}_\gamma$, one has $l \geq \gamma$;
(ii) the set $\mathscr{U}_\gamma$ of the coherent assessments (x,y) on $\{A|H,B|K\}$ such that, for each $(x,y) \in \mathscr{U}_\gamma$, one has $u \leq \gamma$.
Case (i). Of course, $\mathscr{L}_0 = [0,1]^2$; hence we can assume $\gamma > 0$. It must be $l = \max\{x+y-1,0\} \geq \gamma$, i.e., $x+y \geq 1+\gamma$ (as $\gamma > 0$); hence $\mathscr{L}_\gamma$ coincides with the triangle having the vertices $(1,1),(1,\gamma),(\gamma,1)$; that is

$$\mathscr{L}_\gamma = \{(x,y) : \gamma \leq x \leq 1, 1+\gamma-x \leq y \leq 1\}.$$

Notice that $\mathscr{L}_1 = \{(1,1)\}$; moreover, for $\gamma \in (0,1)$, $(\gamma,\gamma) \notin \mathscr{L}_\gamma$.
Case (ii). Of course, $\mathscr{U}_1 = [0,1]^2$; hence we can assume $\gamma < 1$. We recall that $u = \frac{x+y-2xy}{1-xy}$; hence

$$u-x = \frac{y(1-x)^2}{1-xy} \geq 0\,,\;\; u-y = \frac{x(1-y)^2}{1-xy} \geq 0\,; \tag{2}$$

then, from $u \leq \gamma$ it follows $x \leq \gamma, y \leq \gamma$; hence $\mathscr{U}_\gamma \subseteq [0,\gamma]^2$. Then, taking into account that $x \leq \gamma$ and hence $1-(2-\gamma)x > 0$, we have

$$\frac{x+y-2xy}{1-xy} \leq \gamma \iff y \leq \frac{\gamma-x}{1-(2-\gamma)x}\,; \tag{3}$$

therefore

$$\mathscr{U}_\gamma = \left\{(x,y) :\; 0 \leq x \leq \gamma,\;\; y \leq \frac{\gamma-x}{1-(2-\gamma)x}\right\}.$$

Notice that $\mathscr{U}_0 = \{(0,0)\}$; moreover, for $x=y=\gamma \in (0,1)$, it is $u = \frac{2\gamma}{1+\gamma} > \gamma$; hence, for $\gamma \in (0,1)$, $\mathscr{U}_\gamma$ is a strict subset of $[0,\gamma]^2$.
Of course, for every $(x,y) \notin \mathscr{L}_\gamma \cup \mathscr{U}_\gamma$, it is $l < \gamma < u$.
In the next result we determine in general the sets $\mathscr{L}_\gamma, \mathscr{U}_\gamma$.

Theorem 4. *Given a coherent assessment $(p_1,p_2,\ldots,p_n)$ on the family $\{E_1|H_1,\ldots,E_n|H_n\}$, where $E_1,H_1,\ldots,E_n,H_n$ are logically independent, we have*

$$\mathscr{L}_\gamma = \{(p_1,\ldots,p_n) \in [0,1]^n : p_1+\cdots+p_n \geq \gamma+n-1\}\,,\; \gamma > 0\,,$$

$$\mathscr{U}_\gamma = \{(p_1,\ldots,p_n) \in [0,1]^n : 0 \leq p_1 \leq \gamma,\;\; p_{k+1} \leq r_k\,,\; k=1,\ldots,n-1\}\,,\; \gamma < 1\,, \tag{4}$$

where $r_k = \frac{\gamma-u_k}{1-(2-\gamma)u_k}$, $u_k = S_0^H(p_1,\ldots,p_k)$, with $\mathscr{L}_0 = \mathscr{U}_1 = [0,1]^n$.

6 QAND Rule and Probabilistic Entailment

We recall that in [3], based on a three-valued calculus of conditional objects, a logic for nonmonotonic reasoning has been proposed. Conditional objects can be seen as the counterpart of the conditional assertions considered in [8] and, for what concerns logical operations, we can look at them as conditional events. Given a set of conditional objects $\mathscr{K}$, we denote by $\mathscr{C}(\mathscr{K})$ the quasi conjunction of the conditional objects in $\mathscr{K}$. In [3] the following inference rule, named QAND, derivable by applying the inference rules of System $\boldsymbol{P}$ (see [8]), has been introduced

$$\text{(QAND)} \qquad\qquad \mathscr{K} \Rightarrow \mathscr{C}(\mathscr{K})\,.$$

As shown in Section 2, the notions of p-consistency and p-entailment of Adams can be suitable defined in the setting of coherence (see [4, 5]). In the next theorem, to avoid a specific analysis of trivial cases, we assume $\Pi_x = \Pi_y = [0,1]$. We have

Theorem 5. *Given two conditional events $A|H, B|K$, with $\Pi_x = \Pi_y = [0,1]$, we have*

$$A|H \Rightarrow_p B|K \iff A|H \subseteq B|K\,.$$

The next result, related to the approach of Adams, deepens in the framework of coherence the probabilistic semantics of the QAND rule.

Theorem 6. *Given a p-consistent family $\mathscr{F}_n = \{E_i|H_i, i = 1,\ldots,n\}$ and denoting by $\mathscr{C}(\mathscr{F}_n)$ the associated quasi conjunction, for every $\varepsilon \in (0,1]$ there exist $\delta_1 \in (0,1],\ldots,\delta_n \in (0,1]$ such that, for every coherent assessment $(p_1,\ldots,p_n,z)$ on $\mathscr{F}_n \cup \{\mathscr{C}(\mathscr{F}_n)\}$, where $p_i = P(E_i|H_i)$, $z = P(\mathscr{C}(\mathscr{F}_n))$, if $p_1 \geq 1-\delta_1,\ldots,p_n \geq 1-\delta_n$, then $z \geq 1-\varepsilon$. Hence, we have $\mathscr{F}_n \Rightarrow_p \mathscr{C}(\mathscr{F}_n)$.*

Recalling Remark 1, in the next result we show that p-entailment of a conditional event $B|A$ from a family $\mathscr{F}_n$ is equivalent to the existence of a non-empty subset $\mathscr{S}$ of $\mathscr{F}_n$ such that $\mathscr{C}(\mathscr{S})$ p-entails $B|A$.

Theorem 7. *A p-consistent family of conditional events $\mathscr{F}_n$ p-entails a conditional event $B|A$ if and only if there exists a non-empty subset $\mathscr{S}$ of $\mathscr{F}_n$ such that $\mathscr{C}(\mathscr{S})$ p-entails $B|A$.*

An example. We illustrate Theorem 7 by using the well known inference rules Cautious Monotonicity (CM), Or, and Cut, as shown below.
(CM) If $\{C|A, B|A\} \subseteq \mathscr{F}_n$, then $\mathscr{F}_n \Rightarrow_p C|AB$. The assertion follows by observing that, defining $\mathscr{S} = \{C|A, B|A\}$, it is $\mathscr{C}(\mathscr{S}) = BC|A \subseteq C|AB$, so that $\mathscr{C}(\mathscr{S}) \Rightarrow_p C|AB$.
(Or) If $\{C|A, C|B\} \subseteq \mathscr{F}_n$, then $\mathscr{F}_n \Rightarrow_p C|(A \vee B)$. The assertion follows by observing that, defining $\mathscr{S} = \{C|A, C|B\}$, it is $\mathscr{C}(\mathscr{S}) = C|(A \vee B)$, so that, trivially, $\mathscr{C}(\mathscr{S}) \Rightarrow_p C|(A \vee B)$.
(Cut) If $\{C|AB, B|A\} \subseteq \mathscr{F}_n$, then $\mathscr{F}_n \Rightarrow_p C|A$. The assertion follows by observing that, defining $\mathscr{S} = \{C|AB, B|A\}$, it is $\mathscr{C}(\mathscr{S}) = BC|A \subseteq C|A$, so that $\mathscr{C}(\mathscr{S}) \Rightarrow_p C|A$.
Of course, in the previous inference rules, the entailment of the conclusion from $\mathscr{F}_n$ also follows by directly applying Definition 2, as made in [4].

7 Conclusions

We have studied, in a coherence-based setting, the extensions of a given probability assessment on n conditional events to their quasi conjunction, by also considering two cases of logical dependency. We have analyzed further probabilistic aspects on quasi conjunction, by also examining the relation between the notion of p-entailment and the inclusion relation of Goodman and Nguyen. Then, we have shown that each p-consistent family $\mathscr{F}$ p-entails the quasi conjunction $\mathscr{C}(\mathscr{F})$. Finally, we have given a result on the equivalence between p-entailment from $\mathscr{F}$ and p-entailment from $\mathscr{C}(\mathscr{S})$, where $\mathscr{S}$ is some non-empty subset of $\mathscr{F}$.

Acknowledgements. We thank the anonymous referees for their useful comments and suggestions.

References

1. Adams, E.W.: The Logic of Conditionals. D. Reidel Publishing Company, Dordrecht (1975)
2. Biazzo, V., Gilio, A., Lukasiewicz, T., Sanfilippo, G.: Probabilistic logic under coherence: complexity and algorithms. Ann. Math. Artif. Intell. 45(1-2), 35–81 (2005)
3. Dubois, D., Prade, H.: Conditional Objects as Nonmonotonic Consequence relationships. IEEE Trans. Syst. Man Cybern. 24, 1724–1740 (1994)
4. Gilio, A.: Probabilistic reasoning under coherence in System P. Ann. Math. Artif. Intell. 34(1-3), 5–34 (2002)
5. Gilio, A.: On Császár's condition in nonmonotonic reasoning. In: Proceedings of the 10th International Workshop on Non-monotonic Reasoning. Special Session: Uncertainty Frameworks in Non-Monotonic Reasoning, NMR 2004, Whistler BC, Canada (2004), `http://events.pims.math.ca/science/2004/NMR/uf.html`
6. Goodman, I.R., Nguyen, H.T.: Conditional objects and the modeling of uncertainties. In: Gupta, M.M., Yamakawa, T. (eds.) Fuzzy Computing: Theory, Hardware, and Applications, pp. 119–138. North-Holland / Elsevier Science Publishers B.V., Amsterdam (1988)
7. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Academic Publishers, Dordrecht (2000)
8. Kraus, S., Lehmann, D.J., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artif. Intell. 44(1-2), 167–207 (1990)

Elements of Robust Regression for Data with Absolute and Relative Information

Karel Hron and Peter Filzmoser

Abstract. Robust regression methods have advantages over classical least-squares (LS) regression if the strict model assumptions used for LS regression are violated. We briefly review LMS and LTS regression as robust alternatives to LS regression, and illustrate their advantages. Furthermore, it is demonstrated how robust regression can be used if the response variable contains relative rather than absolute information.

Keywords: Multiple linear regression, Robustness, Relative and absolute information, Compositional data.

1 Introduction

In multiple linear regression we consider a linear combination of several explanatory variables, and use this aggregated information to predict a response variable. It results in estimations of parameters of a linear functional that reveal how the response depends on the set of explanatory variables. The least-squares method that is commonly used to obtain the estimations, leads to the best statistical efficiency if certain model assumptions are fulfilled. On the other hand, this method is also very sensitive to outlying observations that could completely destroy the results and thus make any interpretation meaningless. For this reason, many robust counterparts were proposed in the literature. They are usually less efficient than the classical approach, but they

Karel Hron
Palacký University, 77146 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

Peter Filzmoser
Department of Statistics and Probability Theory, Vienna University of Technology, 1040 Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 329–335.
springerlink.com

are in general substantially more resistant to outliers or other deviations from the underlying regression model assumptions. The robust methods thus represent a practical and meaningful alternative to the classical approach, as far as both the response variable and the covariates carry absolute information. However, in many areas data occur which include only relative information (known nowadays under the term *compositional data*) where all the relevant information is contained in the ratios rather then in the absolute values as in the usual case. As these data induce another sample space, they need to be transformed before regression analysis is carried out.

This contribution is organized as follows. In Section 2 a brief review of the classical and robust regression estimators is provided. In Section 3 the basic concepts of compositional data are presented. The final section shows how the relative information can be used in (robust) regression analysis using a real data example.

2 Classical and Robust Linear Regression

Multiple regression analysis forms a tool for prediction of values of a quantity, the response variable, using known (independent) variables. The main task is to find a functional relationship (here assumed to be a linear one) between the response and covariates, i.e. to estimate parameters of the regression function [4]. Let $x_1,\dots,x_q$ be the q variables that we use for prediction of the response variable y. Under the standard regression assumptions, y is a random variable and $x_1,\dots,x_q$ are assumed to be non-random. Having n observations of both y and the explanatory quantities, the linear multiple regression model is

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_qx_{iq}+\varepsilon_i,\quad \text{for } i=1,\dots,n, \tag{1}$$

or in matrix form

$$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}, \tag{2}$$

with the n-dimensional vector $\boldsymbol{y}$ containing the observations of the response variable, the random vector of errors $\boldsymbol{\varepsilon}$ (are assumed to have mean zero), and the $n\times(q+1)$ dimensional design matrix $\mathbf{X}$ with full column rank. Under the assumption of uncorrelated components ε_i, with variance $\mathrm{var}(\varepsilon_i)=\sigma^2$, the vector of unknown parameters can be estimated using the least-squares (LS) method as

$$\widehat{\boldsymbol{\beta}}_{LS}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3}$$

Obviously, the estimate $\widehat{\boldsymbol{\beta}}_{LS}$ minimizes the term

$$\sum_{i=1}^{n}\varepsilon_i^2(\boldsymbol{\beta})=(\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta})=\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}. \tag{4}$$

It is easy to verify that $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, and under the additional assumption of normality of $\boldsymbol{\varepsilon}$ it is also the maximum-likelihood estimator of $\boldsymbol{\beta}$. Consequently, it can be used to obtain the *predicted values* $\widehat{\boldsymbol{y}}_{LS}$ of $\boldsymbol{y}$ as

$$\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the *hat matrix*. The *estimated residuals* are

$$\widehat{\boldsymbol{\varepsilon}}_{LS} = \sqrt{\sum_{i=1}^{n} \varepsilon_i^2(\widehat{\boldsymbol{\beta}}_{LS})} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_{LS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \tag{6}$$

where $\boldsymbol{I}$ stands for identity matrix of order n.

LS-estimation may fail if the model assumptions are violated. Data points deviating from the linear trend can have a strong influence on the estimation because LS regression is based on the squares of the residuals, which then can become very large. We now illustrate this effect in linear regression with one predictor variable.

Figure 1 (left) shows five points that approximately follow a linear trend. Moving one observation in y-direction has a strong influence on the LS parameters, because also the regression line follows this movement in y-direction (right). Also the robust regression method LTS regression (see below) has been applied here, and the movement of the point has no effect on this estimate: the dashed line representing the resulting LTS line coincides with the LS-line of the original data (dotted).

An even worse behaviour is shown in Figure 2, where in the left picture a similar design is presented as in Figure 1 (left). When now an observation is moved in x-direction, the LS regression line is completely changed (right). For

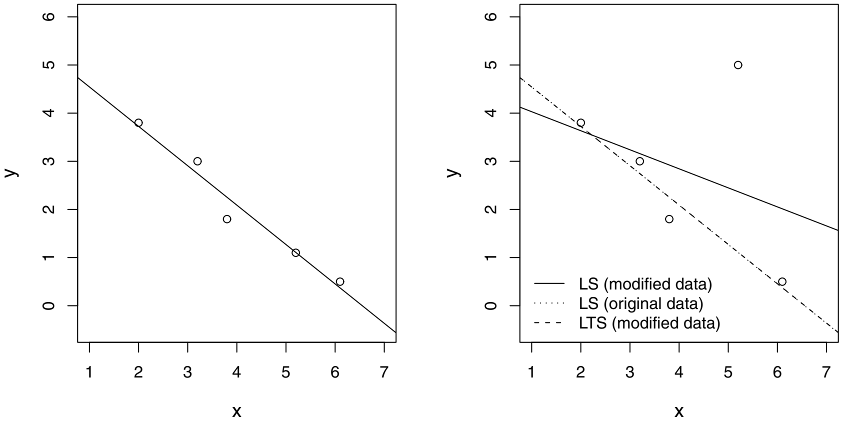


Fig. 1 Influence of an outlier in y-direction on classical LS and robust LTS regression.

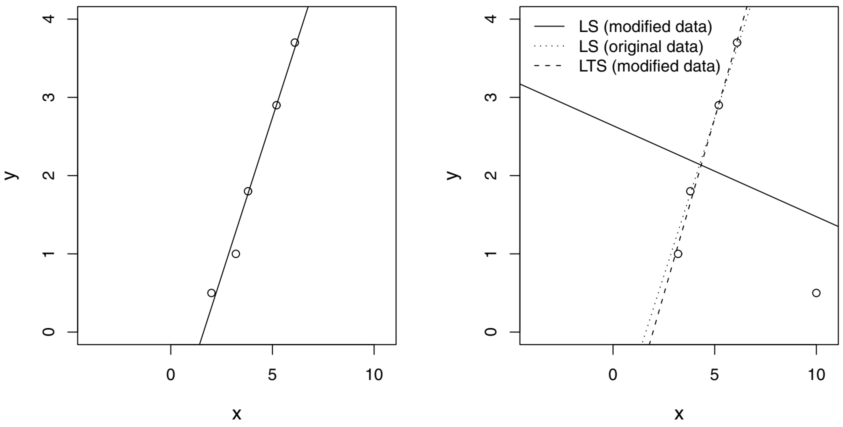


Fig. 2 Influence of an outlier in x-direction on classical LS and robust LTS regression.

this reason, x-outliers are also called *leverage points* because they can "lever" the LS regression line. This undesirable behaviour of LS regression can be avoided by robust regression. The solution of LTS regression for the modified data is almost the same as that for LS regression for the original data.

The basic principle of robust regression is to fit the model to the data majority that follows the linear trend [5]. Accordingly, for *Least Median of Squares* (LMS) regression the function

$$\text{median}_i\, \varepsilon_i^2(\boldsymbol{\beta}) \tag{7}$$

is minimized. Here, the sum from (4) is simply replaced by a median. However, any explicit solution for the regression coefficients as for LS regression is not available, it has to be found using approximative algorithms. For LMS regression it turns out that up to 50% of the data points can be moved arbitrarily without any substantial change of the regression coefficients. This behaviour is expressed by the *breakdown point* which equals 0.5.

Another very robust regression method is *Least Trimmed Sum of Squares* (LTS) regression, where the term

$$\sum_{i=1}^{h} (\varepsilon_i^2(\boldsymbol{\beta}))_{(i)} \tag{8}$$

is minimized, again using a numerical procedure. Here $(\varepsilon_i^2(\boldsymbol{\beta}))_{(1)} \leq \cdots \leq (\varepsilon_i^2(\boldsymbol{\beta}))_{(n)}$ are the sorted squared residuals. By taking $h \approx n/2$, the method has a breakdown point of about 0.5, for larger h it moves to $(n-h)/n$.

3 Relative Information and Compositional Data

As far as the response variable y carries absolute information, the preceding considerations can be used directly. However, in many practical situations the information is not absolute but relative, often expressed in proportions or percentages. Examples of relative information are the unemployment rate in selected countries, proportions of people working in agriculture, percentages of inhabitants with tertiary education, or proportions of the household budget spent on foodstuff. Here the usual model assumptions fail because the values of the response variable are bounded in a certain interval, e.g. in $(0,100)$ in case of percentages, and the assumption of normal distribution is thus not meaningful. However, the problem is in fact a conceptual one and it is inherent to the nature of the data. Namely, here the idea of the relative scale is quite an intuitive concept of differences for them. While the difference between 5% and 10% is the same as between 45% and 50%, the proportions show a quite contrasting relation, because 5% is half of 10%, while 45% is 0.9 of 50%. Thinking in terms of differences in ratios is natural for this kind of data, called in general compositional data (or compositions for short) [1], where only the relative information is of interest. They induce the simplex as the sample space with an own geometry, called nowadays the Aitchison geometry. Thus, compositional data need to be moved from the simplex to the usual Euclidean space isometrically before any statistical analysis can be carried out. This causes in fact that the relative information is transformed into absolute information. The best transformation for this purpose seems to be the isometric logratio (ilr) transformation [2], for both theoretical and practical reasons.

Here we consider a situation where only the response variable includes relative information, but not the explanatory variables. Thus we deal with the problem of an univariate analysis of compositional data [3]. In this case, the ilr transformation of the response variable y simplifies to a new variable (that reminds to the well-known logit transformation)

$$z = \frac{1}{\sqrt{2}} \ln \frac{y}{c-y} \ , \tag{9}$$

where c corresponds to the total value of the whole (1, 100%, total amount of inhabitants working in agriculture, total household budget in Euro) for each observation. After ilr transformation, the values can already be used for regression analysis in the sense of the previous section. After regression analysis and a corresponding prediction for z, the results can be back-transformed to obtain an interpretation in the sense of the original variable y.

4 Use of Robust Regression for Compositional Data

To demonstrate the theoretical considerations numerically, we apply regression analysis to an example where the relation between the percentage of employees in the tertiary sector and the value of the Gross Domestic Product (GDP) per capita in the member states of the European Union is investigated. The considered data are from the year 2009. The tertiary sector is also called "service" sector, where service provision is defined as an economic activity that does not result in ownership, and this is in contrast to providing physical goods. The GDP is a basic measure of a country's overall economic output. It is the market value of all final goods and services made within the borders of a country in a year. The data were obtained from public sources of the internet encyclopedia Wikipedia. Figure 3 (left) shows the data without Luxembourg, where the response variable is already ilr-transformed. Thus both variables contain absolute information and the regression analysis in sense of the previous section can be applied. In the lower right corner of the plot an outlier is clearly visible: Ireland, with a GDP of 30.900 Euro per capita, but with only 49% of employees in tertiary sector. This outlier can be considered as *y*-outlier, because it is still not exceptional in *x*-direction. Still, a strong effect on LS estimation (solid line) is visible, LTS regression is not affected by the outlier, and when excluding Ireland from the analysis, LS would practically coincide with the LTS line.

Figure 3 (right) shows the original data, together with the regression lines from the left picture back-transformed to the original space. Note that the back-transformation is unique, because from Equation (9) we obtain y by

$$y = \frac{c \cdot \exp(\sqrt{2}z)}{1+\exp(\sqrt{2}z)} \ . \tag{10}$$

Due to the different geometry of the simplex, the back-transformed regression lines are no longer linear. To make the effect of the ilr transformation visible,

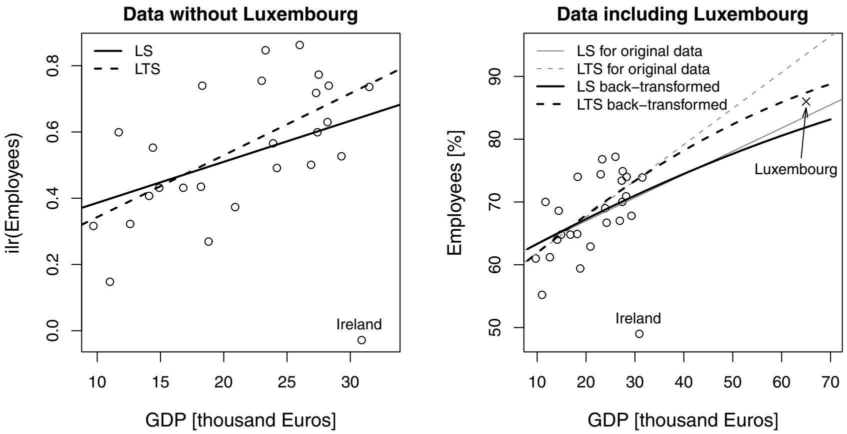


Fig. 3 Regression analysis for the percentage of employees in the tertiary sector after ilr transformation (response variable) and the GDP per capita (explanatory variable) in the member states of the European Union.

classical and robust regression is also applied in the wrong geometry using the original *y*-variable. The results are shown by gray lines. Now Luxembourg is projected into the plot. The GDP of Luxembourg is exceptionally high with 65.009 Euro per capita, and 86% of the employees are in the tertiary sector. The prediction from LTS regression in the ilr-space is closest to the true value, while LS regression, as well as regression analysis (classical and robust) applied in the wrong geometry differ substantially. The reasonability of the robust approach applied in the ilr-space is confirmed by the fact that the resulting regression line is almost unchanged if the outlier Luxembourg is included already at the beginning of the analysis.

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall, London (1986)
2. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric logratio transformations for compositional data analysis. Math. Geol. 35(3), 279–300 (2003)
3. Filzmoser, P., Hron, K., Reimann, C.: Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Sci. Total Environ. 407, 6100–6108 (2009)
4. Johnson, R., Wichern, D.: Applied multivariate statistical analysis, 6th edn. Prentice-Hall, London (2007)
5. Maronna, R., Martin, R.D., Yohai, V.J.: Robust statistics: theory and methods. Wiley, New York (2006)

On Testing Fuzzy Independence Application in Quality Control

Olgierd Hryniewicz

Abstract. In many practical applications of statistics it is assumed that the observed realizations of measurements are mutually independent. This assumption is usually made in order to ease necessary computations. However, for real data sets, especially large ones, the application of statistical tests of independence very often leads to the rejection of the hypothesis of independence even if actual dependence is very weak, and does not have any practical impact on phenomena of interest. Therefore, there is a practical need to define a concept of “near-independence”. In the paper we analyze the possibility of the usage of Kendall’s τ for this purpose. It has been shown, using an example from statistical quality control (Shewhart control charts) that the requirement on τ (or any other similar coefficient) is not sufficient for the construction of statistical tests for testing fuzzy “near-independence”.

Keywords: Fuzzy independence, Control chart, Fuzzy data, Kendall τ.

1 Introduction

Statistical analysis of dependencies in data sets is one of the most applicable areas of statistics. Statisticians usually are interested in finding dependencies in observed data in order to use this knowledge for solving many practical problems. There exist, however, applications of statistical methods where we are rather interested in confirmation that the considered variables or, in general, phenomena are actually independent.The practical reasons for these interests stem from the fact that probabilistic models of complex phenomena are much simpler when variables which are used for their description are statistically independent. For example, in the reliability analysis of complex

Olgierd Hryniewicz
Systems Research Institute, 01-447 Warsaw, Poland
e-mail: hryniewi@ibspan.waw.pl

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 337–344.
springerlink.com

systems the mathematical models used for the computations of probabilities of failures may be very complicated even when failures of system's components are statistically independent. In case of existing dependence between such failures the necessary computations can be performed only in very few, rather simple, cases. Another example of the need of independence comes from expert systems. When an expert system is built we cannot assume independence between considered features and phenomena, as the rationale of expert systems is based on the assumption of such dependencies. However, in practice we usually assume the existence of conditional independence between considered variables. In all these problems, in contrast to the problems mentioned previously, we are interested in *not* rejecting the hypothesis of statistical independence.

Statistical test of independence have been proposed by many authors during the last one hundred (or even more) years. The most popular of them, such as tests based on the Pearson coefficient of correlation ρ and non-parametric rank tests based on the Spearman rank correlation statistic ρ_S are known for more than one hundred years. Many other tests of independence, both parametric and non-parametric, have been proposed by numerous authors, and this research area is still considered interesting among statisticians who try to built more efficient, and in certain applications optimal, procedures.

When we perform statistical tests of independence we either could accept the tested hypothesis or we should reject it at a given significance level. When we are interested in having statistical independence the acceptance of the hypothesis of independence does not create any problems. The problem begins when our data do not let us to assume that the considered random variables are independent. This situation may happen when we analyse large data sets. When we treat these data sets as large samples taken from hypothetical infinite populations even a very small departure form independence will cause the rejection of the hypothesis of independence. The practical question arises then if such rejection indicates that the models built on the assumption of independence cannot be used in practice. Therefore, there is often a practical need to soften the independence requirements by defining the state of "near-independence". The question arises then, how to evaluate this state using statistical data.

The vague concept of "near-independence" has been introduced in Hryniewicz [4]. In contrast to the case of independence, that is very precisely defined in terms of the theory of probability, the concept of "near-independence" is a vague one. Hryniewicz [4] proposed that existing different measures of the strength of dependence may be used - depending on the context - for the evaluation of the state of "near - independence". He claims that these measures might be used for the analysis of dependence when the state of independence is defined, using Zadeh's terminology, "to a degree". Let α be a certain measure of the strength of dependence which in the case of independence adopts the value α_0 (usually equal to zero). For this particular value the independence is definitely to a degree one. However, if we know that $0 < |\alpha - \alpha_0| \leq \varepsilon, \varepsilon > 0$ we can talk about the independence to a degree μ_ε depending on the value of ε, and a given practical context. In this paper we present a practical case, taken from statistical quality control, which shows that this simple approach is insufficient in practice. We show that for the same values of a certain measure of the strength of dependence, such as Kendall's τ coefficient of association, practical consequences of the departure from independence may be different for different structures of dependence described in terms of copulas.

In Section 2 we present a general mathematical framework for dealing with the problem of "near-independence". We focus our attention on using copulas for the description of dependence between random variables. When we use copulas for the description of dependent data, Kendall's τ seems to be the most useful measure for the measurement of the strength of dependence. Therefore, in this section we also present some general results that can be useful for the analysis of "near-independence" using Kendall's τ. In Section 3 we present results of Monte Carlo experiments which show how certain characteristics of a basic tool of statistical quality control, known as a Shewhart control chart, depend not only on the strength of dependence measured using Kendall's τ, but on the type of dependence as well. From the analysis of this Monte Carlo experiment we derive recommendation for the construction of a fuzzy test of independence that can be useful for testing "near-independence" in the context of control charts.

2 Mathematical Modelling of Statistical Dependence

Mathematical models used for the description of dependent random variables are well known for many years. In the simplest two-dimensional case we are interested in the description of dependence between two random variables X and Y having marginal distributions described by cumulative probability functions $F(x)$ and $G(y)$, respectively. In his fundamental work Sklar [10] showed that for a two-dimensional distribution function $H(X,Y)$ with marginal distribution functions $F(X)$ and $G(Y)$ there exists a copula C such that $H(x,y) = C(F(x), G(y))$. For more information about copulas see e.g the book by Nelsen [8].

All well known multivariate probability distributions, the multivariate normal distribution included, can be generated by parametric families C_α of copulas, where real- or vector-valued parameter α describes the strength of dependence between the components of the random vector. The number of papers devoted to the theory and applications of copulas is still growing rapidly. For more recent results the reader should consult already mentioned book by Nelsen [8]. In this paper we focus our attention on three types of copulas. First is the normal copula, which in the two-dimensional case is defined as follows:

$$C(u_1, u_2; \rho) = \Phi_N(\phi^{-1}(u_1), \phi^{-1}(u_2); \rho) \tag{1}$$

where $\Phi_N(u_1,u_2)$ is the cumulative probability distribution function of the bivariate normal distribution, $\phi^{-1}(u)$ is the inverse of the cumulative probability function of the univariate normal distribution (the quantile function), and ρ is the well known coefficient of correlation.

Another copula, considered in this paper is the Farlie-Gumbel-Morgenstern (FGM) copula frequently used for modelling weak dependencies. This copula is defined by the following formula:

$$C(u_1,u_2;\theta) = u_1u_2 + \theta u_1u_2(1-u_1)(1-u_2), |\theta| \leq 1 \tag{2}$$

The remaining three copulas considered in this paper belong to a general class of symmetric copulas, named the Archimedean copulas. They are generated using a class Φ of functions $\phi : [0,1] \rightarrow [0,\infty]$, named generators. It can be proved that in the two-dimensional case a generator ϕ induces a copula if and only if it is convex. In case of more dimensions similar conditions have not been clarified yet. For more information see, e.g., [7]. Every member of the class of Archimedean copulas generates the following multivariate distribution function for the random vector $(X_1,\ldots,X_p)$:

$$C(u_1,\ldots,u_p) = Pr(F_1(X_1) \leq u_1,\ldots,F_p(X_p) \leq u_p) = \phi^{-1}[\phi(u_1) + \cdots + \phi(u_p)] \tag{3}$$

The two-dimensional Archimedean copulas that are investigated in this paper are defined by the following formulae (copulas and their respective generators):

- Clayton's

$$C(u,v) = \max\left(\left[u^{-\alpha} + v^{-\alpha} - 1\right]^{-1/\alpha}, 0\right), \alpha \in [-1,\infty) \setminus 0 \tag{4}$$

$$\phi(t) = (t^{-\alpha} - 1)/\alpha, \alpha \in [-1,\infty) \setminus 0 \tag{5}$$

- Frank's

$$C(u,v) = -\frac{1}{\alpha} \ln\left(1 + \frac{(e^{-\alpha u} - 1)(e^{-\alpha v} - 1)}{e^{-\alpha} - 1}\right), \alpha \in (-\infty,\infty) \setminus 0 \tag{6}$$

$$\phi(t) = \ln\left(\frac{1 - e^{-\alpha}}{1 - e^{-\alpha t}}\right), \alpha \in (-\infty,\infty) \setminus 0 \tag{7}$$

- Gumbel's

$$C(u,v) = \exp\left(-\left[(-\ln u)^{1+\alpha} + (-\ln v)^{1+\alpha}\right]^{\frac{1}{1+\alpha}}\right), \alpha \in (0,\infty) \tag{8}$$

$$\phi(t) = (-\ln(t))^{\alpha+1}, \alpha \in (0,\infty) \tag{9}$$

In case of independence the dependence parameter α_{ind} adopts the value of 0 (in Clayton's and Frank's copulas as an appropriate limit).

Genest and MacKay [2] considered the population version of the well known Kendall's association coefficient τ. This characteristic can be used for the description of the strength of dependence in copulas, and its importance in characterizations of copulas has been shown e.g. in the paper by Nelsen et. al. [9]. Statistical properties of this statistic for the considered in this paper serially correlated data are presented in the paper by Ferguson et al. [1]. Let $K(t)$ be the cumulative probability function of the random variable $T = C(U_1, U_2)$, where U_1 and U_2 are random variables uniformly distributed on $[0,1]$. The following relation links a copula with Kendall's τ:

$$\tau = 3 - 4\int_0^1 K(t)dt \tag{10}$$

Estimation of $K(t)$ for the case of two-dimensional copulas, and thus the estimation of τ was considered by Genest and Rivest [3].

Closed formulae for Kendall's τ are available only for some copulas. In the case of the normal copula we have the following expression

$$\tau_{Norm} = arcsin(\rho)/(\pi/2). \tag{11}$$

For the FGM copula we can compute Kendall's τ from a very simple formula

$$\tau_{FGM} = 2\theta/9 \tag{12}$$

For the family of Archimedean copulas there exists the following general formula that links Kendall's τ with the generator function ϕ:

$$\tau_{Arch} = 1 + 4\int_0^1 \frac{\phi(v)}{\phi'(v)}dv \tag{13}$$

For specific cases of the considered in this paper Archimedean copulas we have

- Clayton's copula
$$\tau = \frac{\alpha}{\alpha+2} \tag{14}$$
- Frank's copula
$$\tau = 1 + 4\left(\frac{1}{\alpha}\int_0^{\alpha} \frac{t}{e^t - 1}dt - 1\right)/\alpha \tag{15}$$
- Gumbel's copula
$$\tau = \frac{\alpha}{\alpha+1} \tag{16}$$

In the context of the analysis of "near-independence" we are interested in cases when the values of the dependence parameter are close to the independence value equal to zero. Using elementary technique of the expansion in the Taylor series around zero (Maclaurin series) we have very simple relations: $\tau_{Norm} \approx 2\rho/\pi$ for the normal copula, $\tau_{FGM} = 2\theta/9$ for the FGM copula,

$\tau_{Gumb} \approx \alpha$ for Gumbel's copula, and $\tau_{Clay} \approx \alpha/2$ for Clayton's copula. One can also apply this extension technique for the general class of Archimedean copulas. The expansion of (13) yields the following general formula

$$\tau_{Arch} \approx 4\alpha \int_0^1 [-xg(x) + x^2 g'(x) ln(x)] dx \tag{17}$$

where

$$g(x) = \lim_{\alpha \to 0} \frac{d}{d\alpha} \phi_\alpha(x), \tag{18}$$

and $\phi_\alpha(x)$ is the generator of the copula. The general formula (17) let us find, after some straightforward but tedious computations, that for Frank's copula we have $\tau_{Frank} \approx \alpha/9$. The accuracy of the estimators of τ for different copulas is generally unknown. However, simulation experiments show that for the values of τ close to zero (i.e. in the case of "near-independence") the variance of the estimator is close to that obtained in the case of independence (see [1]).

3 The Concept of "Near-Independence" in Statistical Quality Control

One of the most frequently applied procedure of statistical quality control is a control chart introduced by W.Shewhart in the 1920s. The aim of the control chart is to assist a process operator in keeping this process under control. The process under consideration is sampled, and the results of measurements, summarized in a form of certain statistics like sample mean, sample standard deviation or sample range, are plotted against time. On each control chart there are also plotted control lines. The area between the control lines represents the set of those values of the results of process' inspection which indicate that is very probable (probability larger than 0.99 in usual applications) that the process is under control. When the observed value of a plotted statistic falls beyond the control lines an alarm is triggered, as it seems to be very likely that the process went out of control.

The most important characteristic of a control chart is the *Average Run Length* (*ARL*) defined as the average number of inspected samples till the moment of an alarm. If the control limits are too tight there is significant probability of false alarm. On the other hand, when the area between the control lines is too wide the probability of triggering a necessary alarm is too low. Therefore, it is necessary to design the chart very carefully, taking into account all pertaining probabilities. In everyday practice all these probabilities are calculated under the assumption of *statistical independence* between the consecutive measurements. It has been shown by many researchers (see Hryniewicz and Szediw [6] for references) that the existence of the dependence between observations may change dramatically the performance of the chart. When we want to take all these dependencies into account the design and maintenance of a control chart may become very difficult, and practically impossible for the majority of users. Therefore, it is important to know that the existing dependency does not influence very much the assumed performance of the chart. Thus, we are confronted with the problem of the description of "near-independence" in the context of statistical quality control.

In order to investigate the influence of the type of dependence (represented by 5 different copulas) on the performance of control charts we have performed Monte Carlo experiments with the aim to evaluate the value of *ARL*. Consecutive observations were generated in such a way that the joint probability distribution of two consecutive observations was described by a copula characterized by a predefined value of Kendall's τ. Note that in this experiment the actual value of Kendall's τ, due to the serial correlation - even in the case of independence - of the consecutive pairs of observations, is different that the value used for the design of the experiment. A part of obtained results in case of the so called "3-σ" decision rule, is presented in Table 1. Similar results in case of the "6 in a row increasing (decreasing)" decision rule are presented in Table 2.

The results of simulation experiments show undoubtedly that the concept of "near-independence" strongly depends upon the context. It depends upon a general type of dependence (positive or negative) and the type of applicable copula. For example, in the case described in Table 1 departures from independence strongly depend upon the value of a coefficient c_τ in the approximate formula $\tau \approx c_\tau \alpha$ when the dependence is positive. On the other hand, when the dependence is negative, "near-independence" can be described

Table 1 ARLs ("3-σ" rule) for different types of dependence in data

Kendall's τ	Normal	FGM	Clayton	Frank	Gumbel
0,1	371,1	369,6	384,4	370,5	456,4
0,05	370,5	369,3	374,7	372,1	443,3
0	**370,5**	**370,5**	**370,5**	**370,5**	**370,5**
-0,05	371,3	370,8	370,0	371,7	x
-0,1	371,3	368,9	370,6	371,2	x

Table 2 ARLs ("6 in a row" rule) for different types of dependence in data

Kendall's τ	Normal	FGM	Clayton	Frank	Gumbel
0,1	97,1	147,3	95,3	96,1	92,4
0,05	119,2	146,4	117,1	118,5	115,0
0	**147,1**	**147,1**	**147,1**	**147,1**	**147,1**
-0,05	183,2	147,1	182,6	186,2	x
-0,1	225,7	146,6	225,8	236,7	x

by a fuzzy requirement "τ is near zero" without taking into account possible differences between considered copulas. Thus, the fuzzy requirement for the value of τ representing "near-independence" should be asymmetric around zero with the membership function of a rectangular shape for the negative values of τ, and triangular shape for the positive ones. In the case presented in Table 2 the departures from independence are equally important both for positive and negative dependence, except for the FGM copula which is known as the one that well describes small departures from independence. Therefore, "near-independence" can be modelled by a rectangular membership function of the fuzzy τ which is symmetric around zero.

The conclusions from this experiment are, in a certain sense, *negative.* It seems to be impossible to test a fuzzy concept of "near-independence" independently on a particular context, as it has been proposed in [4]. The construction of the membership function of this fuzzy hypothesis should be context dependent. In testing the fuzzy hypothesis of "near-independence" (see [5] for the information about this type of statistical tests) we must take into account not only particular application, but also additional information about the type of observed dependence.

References

1. Ferguson, T.S., Genest, C., Hallin, M.: Kendall's tau for serial dependence. Canad. J. Statist. 28, 587–604 (2000)
2. Genest, C., McKay, R.J.: The joy of copulas: Bivariate distributions with uniform marginals. Amer. Statist. 88, 1034–1043 (1986)
3. Genest, C., Rivest, L.-P.: Statistical Inference Procedures for Bivariate Archimedean Copulas. J. Amer. Statist. Assoc. 88, 1034–1043 (1993)
4. Hryniewicz, O.: On testing fuzzy independence. In: Lawry, J., Miranda, E., Bugarin, A., Li, S., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Integrated Uncertainty Modeling. Advances in Intelligent and Soft Computing, vol. 37. Springer, Heidelberg (2006)
5. Hryniewicz, O.: Possibilistic decisions and fuzzy statistical tests. Fuzzy Sets Syst. 157, 2665–2673 (2006)
6. Hryniewicz, O., Szediw, A.: Sequential Signals on a Control Chart Based on Nonparametric Statistical Tests. In: Lenz, H.-J., Wilrich, P.-T., Schmid, W. (eds.) Frontiers in Statistical Quality Control, vol. 9. Physica-Verlag, Heidelberg (2010)
7. McNeil, A.J., Nešlehová, J.: Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. Ann. Statist. 37, 3059–3097 (2009)
8. Nelsen, R.B.: An Introduction to Copulas. Springer, New York (2006)
9. Nelsen, R.B., Quesada-Molina, J.J., Rodríguez-Lallena, J.A., Úbeda-Flores, M.: Kendall distribution functions. Statist. Probab. Lett. 65, 263–268 (2003)
10. Sklar, A.: Fonctions de répartition á n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)

The Fisher's Linear Discriminant

Iuliana F. Iatan

Abstract. In this paper we have chosen a proper vector b in order to prove that the MSE discriminant function a^tY is directly related to Fisher's linear discriminant. The Fisher's criterion is in the range of techniques for performing linear discrimination in the two class case. The Fisher's linear discriminant is a criterion function that involves all of the samples, while the perceptron criterion function is focussed on the misclassified samples.

1 Introduction

Through the mapping from d- dimensional input space X to $d+1$- dimensional space Y

$$Y^t = (1, x_1, \ldots, x_d) = (1, X^t), \;\; a^t = (a_0, \ldots, a_d) = (\omega_0, w^t),$$

the decision rule

$$g(X) = \begin{cases} w^tX + \omega_0 > 0 \implies X \in \omega_1, \\ w^tX + \omega_0 < 0 \implies X \in \omega_2. \end{cases}$$

becomes

$$g(Y) = \begin{cases} a^tY > 0 \implies Y \in \omega_1, \\ a^tY < 0 \implies Y \in \omega_2. \end{cases} \tag{1}$$

Suppose now that we have a set of N samples $\{Y_1, Y_2, \ldots, Y_N\} \subseteq \Re^{d+1}$, some labelled ω_1 and some labelled ω_2. We want to use these samples to determine the weights a in a linear discriminant function $g(Y) = a^tY$.

Iuliana F. Iatan
Department of Mathematics and Computer Science, Technical University of Civil Engineering of Bucharest, Romania
e-mail: iuliafi@yahoo.com

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 345–352.
springerlink.com

If such a weight vector exists, the samples are said to be *linearly separable.*

A sample Y_i is classified correctly if

$$\begin{cases} a^t Y_i > 0 \ for \ Y_i \in \omega_1, \\ a^t Y_i < 0 \ for \ Y_i \in \omega_2, \end{cases}$$

namely

$$\begin{cases} a^t Y_i > 0 \ for \ d_i = 1, \\ a^t Y_i < 0 \ for \ d_i = -1, \end{cases}$$

where $d_i \in \{-1,1\}, i = \overline{1,N}$ is the label of the vector i.

This suggests a normalization that simplifies the treatment of the two category case, which consists in the replacement of all samples labelled ω_2 by their negatives. With this normalization we can forget the labels and look for a weight vector a such that $a^t Z_i > 0$ for all of the samples, where

$$Z_i = \begin{cases} Y_i > 0 \ for \ d_i = 1, \\ -Y_i < 0 \ for \ d_i = -1. \end{cases}$$

Such a weight vector is called a separating vector or more generally a solution vector.

The Perceptron and relaxation procedures help us for finding a separating vector when the samples are linearly separable. All of these methods are called error correcting procedures, because they call for a modification of the weight vector when and only when an error is encountered. Therefore, the criterion functions considered in the case of the previous methods have focussed their attention on the misclassified samples.

We shall now consider a criterion function that involves all of the samples. Where previously we have sought a weight vector a making all of the inner products $a^t Y_i$ positive, now we shall try to make $a^t Y_i = b_i$, where the b_i are some arbitrarily specified positive constants. Thus, we have replaced the problem of finding the solution to a set of linear inequalities with the more stringent but better understood problem of finding the solution to a set of linear equations. The treatment of simultaneous linear equations is simplified by introducing matrix notation. Let Y be $n \times \hat{d}$, matrix $\hat{d} = d + 1$ whose ith row is the vector Y_i^t and let b the column vector $b = (b_1, \ldots, b_n)^t$, $b_i > 0, i = \overline{1, \ n}$, n being the number of vectors. Our problem is to find a weight vector a satisfying

$$\begin{pmatrix} Y_{10} \ Y_{11} \ \ldots \ Y_{1d} \\ Y_{20} \ Y_{21} \ \ldots \ Y_{2d} \\ \ldots \ \ldots \ \ldots \ \ldots \\ Y_{n0} \ Y_{n1} \ \ldots \ Y_{nd} \end{pmatrix} (a_0, a_1, \ldots, a_d)^t = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{2}$$

or

$$Ya = b. \tag{3}$$

If Y were nonsingular, we could write $a = Y^{-1}b$ and obtain a formal solution at once. Since Y is rectangular (usually with more rows than columns) no solution exists for the system (3). However, we can seek a weight vector a that minimizes some function of the error between Ya and b. If we define the error vector e by $e = Ya - b$, the one approach is to try to minimize the squared length of the error vector. This is equivalent to minimizing the sum of squared error criterion function

$$J_s(a) = \|Ya - b\|^2 = \sum_{i=1}^{n} \left(a^t Y_i - b_i\right)^2. \tag{4}$$

From the condition $\nabla J_s(a) = 0$, where

$$\nabla J_s(a) = 2\sum_{i=1}^{n} \left(a^t Y_i - b_i\right) Y_i = 2Y^t \left(Ya - b\right)$$

we shall deduce

$$Y^t Ya = Y^t b. \tag{5}$$

In this way we have converted the problem of solving $Ya = b$ to that of solving $Y^t Ya = Y^t b$, which has the great advantage that the $\hat{d}$-by-$\hat{d}$ matrix $Y^t Y$ is square and often nonsingular. The system (5) has the unique solution (is a MSE=Minimum Squared Error solution)

$$a = \left(Y^t Y\right)^{-1} Y^t b = Y^* b, \tag{6}$$

where $Y^* = (Y^t Y)^{-1} Y^t$ is a $\hat{d}$-by-n matrix called the *pseudoinverse* of Y.

The MSE solution depends on the margin vector b and we shall see that different choices for b give the solution different properties.

2 The Relation Between MSE Solution and the Fisher's Linear Discriminant

In this section we shall prove that with the proper choice of the vector b, the MSE discriminant function $a^t Y$ is directly related to Fisher's linear discriminant.

Theorem 1. *We assume that we have a set of n d- dimensional samples $\{X_1^{(1)}, \ldots, X_{n_1}^{(1)}, X_1^{(2)}, \ldots, X_{n_2}^{(2)}\}$, where n_i samples are labelled ω_i, $i = \overline{1,2}$, $n_1 + n_2 = n$. The matrix* $\mathbf{Y}$, *the vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *can be partitioned as follows:*

$$\mathbf{Y} = \begin{bmatrix} 1_1 & X^{(1)} \\ -1_2 & -X^{(2)} \end{bmatrix}, \ \mathbf{a} = \begin{bmatrix} \omega_0 \\ w \end{bmatrix}, \ \mathbf{b} = \begin{bmatrix} \frac{n}{n_1} \cdot 1_1 \\ \frac{n}{n_2} \cdot 1_2 \end{bmatrix} \tag{7}$$

where

- 1_i *is a column vector of* n_i *ones,* $i=\overline{1,2}$,
- $X^{(i)}$ *is an* n_i*-by-d matrix whose rows are the samples labelled* ω_i, $i=\overline{1,2}$.

The special choice for **b** *links the MSE solution to the Fisher's linear discriminant.*

Proof. We shall write (5) as

$$\begin{bmatrix} 1_1^t & -1_2^t \\ X^{(1)^t} & -X^{(2)^t} \end{bmatrix} \cdot \begin{bmatrix} 1_1 & X^{(1)} \\ -1_2 & -X^{(2)} \end{bmatrix} \cdot \begin{bmatrix} \omega_0 \\ w \end{bmatrix} = \begin{bmatrix} 1_1^t & -1_2^t \\ X^{(1)^t} & -X^{(2)^t} \end{bmatrix} \cdot \begin{bmatrix} \frac{n}{n_1}\cdot 1_1 \\ \frac{n}{n_2}\cdot 1_2 \end{bmatrix} \qquad (8)$$

We define

$$m_i = \frac{1}{n}\sum_{X\in\omega_i} X$$

the mean vector of ω_i, $i=\overline{1,2}$ and

$$S_w = \sum_{i=1}^{2}\sum_{X\in\omega_i}(X-m_i)(X-m_i)^t$$

the pooled sample scatter matrix.

We shall have

$$S_w = \begin{pmatrix} X_{11}^{(1)}-m_{11} \\ \vdots \\ X_{1d}^{(1)}-m_{1d} \end{pmatrix} \left(X_{11}^{(1)}-m_{11},\ldots,X_{1d}^{(1)}-m_{1d}\right)+\ldots$$

$$+\begin{pmatrix} X_{n_11}^{(1)}-m_{11} \\ \vdots \\ X_{n_1d}^{(1)}-m_{1d} \end{pmatrix} \left(X_{n_11}^{(1)}-m_{11},\ldots,X_{n_1d}^{(1)}-m_{1d}\right)$$

$$+\begin{pmatrix} X_{11}^{(2)}-m_{21} \\ \vdots \\ X_{1d}^{(2)}-m_{2d} \end{pmatrix} \left(X_{11}^{(2)}-m_{21},\ldots,X_{1d}^{(2)}-m_{2d}\right)+$$

$$+\ldots+\begin{pmatrix} X_{n_21}^{(2)}-m_{21} \\ \vdots \\ X_{n_2d}^{(2)}-m_{2d} \end{pmatrix} \left(X_{n_21}^{(2)}-m_{21},\ldots,X_{n_2d}^{(2)}-m_{2d}\right) = \begin{pmatrix} s_{11} & \ldots & s_{1d} \\ \ldots & \ldots & \ldots \\ s_{d1} & \ldots & s_{dd} \end{pmatrix}.$$

where

$$\begin{aligned}
s_{11} =&(X_{11}^{(1)}-m_{11})^2+\ldots+(X_{n_1 1}^{(1)}-m_{11})^2+(X_{11}^{(2)}-m_{21})^2+\ldots+(X_{n_2 1}^{(2)}-m_{21})^2,\\
s_{1d} =&(X_{11}^{(1)}-m_{11})(X_{1d}^{(1)}-m_{1d})+\ldots+(X_{n_1 1}^{(1)}-m_{11})(X_{n_1,d}^{(1)}-m_{1d})\\
&+(X_{11}^{(2)}-m_{21})(X_{1d}^{(2)}-m_{2d})+\ldots+(X_{n_2 1}^{(2)}-m_{21})(X_{n_2 d}^{(2)}-m_{2d}),\\
s_{d1} =&s_{1d},\\
s_{dd} =&(X_{1d}^{(1)}-m_{1d})^2+\ldots+(X_{n_1 d}^{(1)}-m_{1d})^2+(X_{1d}^{(2)}-m_{2d})^2+(X_{n_2,d}^{(2)}-m_{2d})^2. \qquad (9)
\end{aligned}$$

Since

$$m_1=\begin{pmatrix} m_{11}\\ \vdots\\ m_{1d}\end{pmatrix}=\frac{1}{n_1}\begin{pmatrix} X_{11}^{(1)}+\ldots+X_{n_1 1}^{(1)}\\ \vdots\\ X_{1d}^{(1)}+\ldots+X_{n_1 d}^{(1)}\end{pmatrix}\Longrightarrow\begin{cases} X_{11}^{(1)}+\ldots+X_{n_1 1}^{(1)}=n_1m_{11}\\ \vdots\\ X_{1d}^{(1)}+\ldots+X_{n_1 d}^{(1)}=n_1m_{1d}\end{cases}$$

and

$$m_2=\begin{pmatrix} m_{21}\\ \vdots\\ m_{2d}\end{pmatrix}=\frac{1}{n_2}\begin{pmatrix} X_{11}^{(2)}+\ldots+X_{n_2 1}^{(2)}\\ \vdots\\ X_{1d}^{(2)}+\ldots+X_{n_2 d}^{(2)}\end{pmatrix}\Longrightarrow\begin{cases} X_{11}^{(2)}+\ldots+X_{n_2 1}^{(2)}=n_2m_{21}\\ \vdots\\ X_{1d}^{(2)}+\ldots+X_{n_2 d}^{(2)}=n_2m_{2d}\end{cases}$$

we shall obtain

$$\sum_{i=1}^{n_1}(X_{i1}^{(1)}-m_{11})^2+\sum_{i=1}^{n_2}(X_{i1}^{(2)}-m_{21})^2=\sum_{i=1}^{n_1}{X_{i1}^{(1)}}^2-2m_{11}n_1m_{11}+n_1m_{11}^2$$

$$+\sum_{i=1}^{n_2}{X_{i1}^{(2)}}^2-2m_{21}n_2m_{21}+n_2m_{21}^2=\sum_{i=1}^{n_1}{X_{i1}^{(1)}}^2-n_1m_{11}^2+\sum_{i=1}^{n_2}{X_{i1}^{(2)}}^2-n_2m_{21}^2;$$

The relation (9) becomes

$$\begin{aligned}
s_{11} =&\sum_{i=1}^{n_1}{X_{i1}^{(1)}}^2-n_1m_{11}^2+\sum_{i=1}^{n_2}{X_{i1}^{(2)}}^2-n_2m_{21}^2,\\
s_{1d} =&(X_{11}^{(1)}-m_{11})(X_{1d}^{(1)}-m_{1d})+\ldots+(X_{n_1 1}^{(1)}-m_{11})(X_{n_1,d}^{(1)}-m_{1d})\\
&+(X_{11}^{(2)}-m_{21})(X_{1d}^{(2)}-m_{2d})+\ldots+(X_{n_2 1}^{(2)}-m_{21})(X_{n_2 d}^{(2)}-m_{2d}),\\
s_{d1} =&s_{1d},\\
s_{dd} =&\sum_{i=1}^{n_1}{X_{id}^{(1)}}^2-n_1m_{1d}^2+\sum_{i=1}^{n_2}{X_{id}^{(2)}}^2-n_2m_{2d}^2. \qquad (10)
\end{aligned}$$

We can note that

$$U=\begin{bmatrix}1_1^t & -1_2^t\\ X^{(1)^t} & -X^{(2)^t}\end{bmatrix}\cdot\begin{bmatrix}1_1 & X^{(1)}\\ -1_2 & -X^{(2)}\end{bmatrix},$$

namely

$$U=\begin{pmatrix}n_1+n_2 & n_1m_{11}+n_2m_{21} & \dots & n_1m_{1d}+n_2m_{2d}\\ n_1m_{11}+n_2m_{21} & A & \dots & B\\ \dots & \dots & \dots & \dots\\ n_1m_{1d}+n_2m_{2d} & B & \dots & A\end{pmatrix}, \tag{11}$$

where

$$A=\sum_{i=1}^{n_1}X_{i1}^{(1)^2}+\sum_{i=1}^{n_2}X_{i1}^{(2)^2},$$

$$B=X_{11}^{(1)}X_{1d}^{(1)}+\ldots+X_{n_11}^{(1)}X_{n_1d}^{(1)}+X_{11}^{(2)}X_{1d}^{(2)}+\ldots+X_{n_21}^{(2)}X_{n_2d}^{(2)}.$$

From (10) we shall calculate

$$s_{1d}=X_{11}^{(1)}X_{1d}^{(1)}+\ldots+X_{n_11}^{(1)}X_{n_1d}^{(1)}-m_{1d}n_1m_{11}-m_{11}n_1m_{1d}+n_1m_{11}m_{1d}$$

$$+X_{11}^{(2)}X_{1d}^{(2)}+\ldots+X_{n_21}^{(2)}X_{n_2d}^{(2)}-m_{2d}n_2m_{21}-m_{21}n_2m_{2d}+n_2m_{21}m_{2d}.$$

Thus, finally we obtain

$$s_{1d}=X_{11}^{(1)}X_{1d}^{(1)}+\ldots+X_{n_11}^{(1)}X_{n_1d}^{(1)}-n_1m_{11}m_{1d}+X_{11}^{(2)}X_{1d}^{(2)}+\ldots+X_{n_21}^{(2)}X_{n_2d}^{(2)}-n_2m_{21}m_{2d}. \tag{12}$$

From (11) and (12) it results

$$U=\begin{bmatrix}n & (n_1m_1+n_2m_2)^t\\ n_1m_1+n_2m_2 & S_w+n_1m_1m_1{}^t+n_2m_2m_2{}^t\end{bmatrix}. \tag{13}$$

Let's evaluate now

$$V=\begin{bmatrix}1_1^t & -1_2^t\\ X^{(1)^t} & -X^{(2)^t}\end{bmatrix}\cdot\begin{bmatrix}\frac{n}{n_1}\cdot 1_1\\ \frac{n}{n_2}\cdot 1_2\end{bmatrix}=\begin{pmatrix}0\\ n(m_{11}-m_{21})\\ \vdots\\ n(m_{1d}-m_{2d})\end{pmatrix}.$$

We can write

$$V=\begin{bmatrix}0\\ n(m_1-m_2)\end{bmatrix}. \tag{14}$$

Introducing (13) and (14) in (8) we shall obtain

$$\begin{bmatrix}n & (n_1m_1+n_2m_2)^t\\ n_1m_1+n_2m_2 & S_w+n_1m_1m_1{}^t+n_2m_2m_2{}^t\end{bmatrix}\cdot\begin{bmatrix}\omega_0\\ w\end{bmatrix}=\begin{bmatrix}0\\ n(m_1-m_2)\end{bmatrix}, \tag{15}$$

namely

$$n\omega_0 + (n_1m_1 + n_2m_2)^t w = 0 \tag{16}$$

and

$$(n_1m_1 + n_2m_2)\omega_0 + (S_w + n_1m_1m_1{}^t + n_2m_2m_2{}^t)w = n(m_1 - m_2). \tag{17}$$

We denote

$$m = \frac{1}{n}\left[\sum_{i=1}^{n_1} X_i^{(1)} + \sum_{i=1}^{n_2} X_i^{(2)}\right] = \frac{1}{n}\begin{bmatrix} X_{11}^{(1)} + \ldots + X_{n_11}^{(1)} + X_{11}^{(2)} + \ldots + X_{n_21}^{(2)} \\ \vdots \\ X_{1d}^{(1)} + \ldots + X_{n_1d}^{(1)} + X_{1d}^{(2)} + \ldots + X_{n_2d}^{(2)} \end{bmatrix}$$

$$= \frac{1}{n}\begin{bmatrix} n_1m_{11} + n_2m_{21} \\ \vdots \\ n_1m_{1d} + n_2m_{2d} \end{bmatrix};$$

therefore

$$m = \frac{1}{n}(n_1m_1 + n_2m_2). \tag{18}$$

From (16) and (18) it results $n\omega_0 + nm^t w = 0$ and further,

$$\omega_0 = -m^t w. \tag{19}$$

Introducing (19) in (17) we shall have

$$nm(-m^t w) + (S_w + n_1m_1m_1{}^t + n_2m_2m_2{}^t)w = n(m_1 - m_2) \tag{20}$$

or

$$\left(\frac{1}{n}S_w + Q\right)w = m_1 - m_2, \tag{21}$$

where

$$Q = -\frac{1}{n^2}\left(n_1^2m_1m_1{}^t + n_1n_2m_1m_2{}^t + n_1n_2m_2m_1{}^t + n_2^2m_2m_2{}^t\right) + \frac{n_1}{n}m_1m_1{}^t + \frac{n_2}{n}m_2m_2{}^t.$$

We shall write Q as $Q = \frac{n_1n_2}{n^2}(m_1 - m_2)(m_1 - m_2)^t$. Thus (21) becomes

$$\left[\frac{1}{n}S_w + \frac{n_1n_2}{n^2}(m_1 - m_2)(m_1 - m_2)^t\right]w = m_1 - m_2. \tag{22}$$

From (22) we shall have

$$w = n\left(I + \frac{n_1n_2}{n^2}S_w{}^{-1}(m_1 - m_2)(m_1 - m_2)^t\right)^{-1}S_w{}^{-1}(m_1 - m_2)$$

therefore $w = \alpha n S_w^{-1}(m_1 - m_2)$, where $\alpha = \left(I + \frac{n_1n_2}{n^2}S_w{}^{-1}(m_1 - m_2)(m_1 - m_2)^t\right)^{-1}$. □

From the decision rule (1) we deduce $a^tY > 0 \Longleftrightarrow \begin{bmatrix} \omega_0^t & w^t \end{bmatrix} \cdot \begin{bmatrix} 1 \\ X \end{bmatrix} > 0$ and taking into account (19) we have $\begin{bmatrix} -w^tm & w^t \end{bmatrix} \cdot \begin{bmatrix} 1 \\ X \end{bmatrix} > 0 \Longleftrightarrow w^t(X-m) > 0$.

We shall obtain the Fisher's decision rule: $\begin{cases} w^t(X-m) > 0 \implies X \in \omega_1, \\ w^t(X-m) < 0 \implies X \in \omega_2. \end{cases}$

3 Conclusions

In this paper we shall prove that with the proper choice of the vector b, the MSE discriminant function a^tY is directly related to Fisher's linear discriminant. The equation $w = \alpha n S_w^{-1}(m_1 - m_2)$, except for an unimportant scale factor, is identical to the solution for Fisher's linear discriminant. The Fisher's linear discriminant is a criterion function that involves all of the samples, while the perceptron criterion function is focussed on the misclassified samples.

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. In: Information Science and Statistics. Springer, New York (2006)
2. Brunelli, R.: Template Matching Techniques in Computer Vision: Theory and Practice. Wiley, New York (2009)
3. Hilbe, J.M.: Logistic Regression Models. Texts in Statistical Science Series. Chapman & Hall/CRC Press, Boca Raton (2009)
4. Iatan, I.: Statistical Methods for Pattern Recognition. Lambert Academic Publishing AG& Co., KG (2010)
5. Liberati, C., Howe, J.A., Bozdogan, H.: Data Adaptive Simultaneous Parameter and Kernel Selection in Kernel Discriminant Analysis Using Information Complexity. J. Pattern Recogn. Res. 4(1), 119–132 (2009)
6. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Elsevier, London (2009)
7. Webb, A.: Statistical Pattern Recognition, 2nd edn. John Wiley and Sons, New York (2002)

Testing Archimedeanity

Piotr Jaworski

Abstract. The aim of this paper is to provide a simple asymptotic test for Archimedeanity. The main idea is to test the associativity, which is a property that distinguish Archimedean among other two-dimensional copulas.

Keywords: Archimedean Copulas, Asymptotic Tests, Empirical Copula Processes.

1 Introduction

A copula $C:[0,1]^2 \longrightarrow [0,1]$ is called *Archimedean* if it can be expressed in the form

$$C(x,y) = \psi(\varphi(x)+\varphi(y)), \tag{1}$$

where $\psi:[0,+\infty] \longrightarrow [0,1]$ and $\varphi:[0,1] \longrightarrow [0,+\infty]$ are continuous convex nonincreasing functions such that

$$\psi \circ \varphi = id \quad \text{and} \quad \varphi(1) = 0.$$

Depending on the source, either ψ ([5, 2]) or φ ([6]) is called the generator. Furthermore, it is known that if ψ and φ are as above then $\psi(\sum_{i=1}^{2}\varphi(u_i))$ is a copula ([6], Theorem 4.1.4).

Archimedean copulas are frequently used in modeling bivariate dependence, because they are easy to handle. They are given by closed analytical formulas and are easy to simulate (see [1], Chapter 6).

Therefore there is a need to have a tool which will help to decide whether some empirical data come from the distribution governed by an Archimedean copula. To provide such a tool we will construct an asymptotic test based on the associativity.

Piotr Jaworski
Institute of Mathematics, University of Warsaw, Poland
e-mail: P.Jaworski@mimuw.edu.pl

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 353–360.
springerlink.com

We recall that a copula may be considered as a binary operation on the unit interval. We say that the copula C is associative if the corresponding binary operation is, i.e. if

$$\forall x,y,z \in [0,1] \quad C(x,C(y,z)) = C(C(x,y),z). \tag{2}$$

This property is distinguishing the Archimedean copulas. Namely (see [6] Theorems 4.1.5 and 4.1.6 and [4]):

Theorem 1. *A copula C is associative if and only if it is Archimedean or it is equal to the upper Fréchet-Hoeffding bound $M(x,y) = \min(x,y)$ or it is an ordinal sum of Archimedean copulas and M's.*

Corollary 1. *If a copula C is associative and cannot be represented as an ordinal sum of two copulas then it is Archimedean.*

There is no copula C differentiable at points (t,t) such that $0 \neq t \neq 1$ and $C(t,t) = t$. Therefore we get the following criterion:

Corollary 2. *If a copula C is associative and is differentiable at every point (t,t), $t \in (0,1)$, then it is Archimedean.*

Let (X,Y) be a pair of random variables with copula C and continuous marginals F and G. Basing on independent copies $(X_1,Y_1),\dots,(X_n,Y_n)$, we construct the empirical copula function. Namely, let $(x_1,y_1),\dots,(x_n,y_n)$ be the realizations. To each of them we associate a pair of ranks (u_i,v_i)

$$u_i = \frac{\#\{j : x_j \le x_i\}}{n}, \quad v_i = \frac{\#\{j : y_j \le y_i\}}{n}.$$

The empirical copula function

$$C_n : [0,1]^2 \longrightarrow [0,1]$$

is given by the formula

$$C_n(u,v) = \frac{1}{n} \cdot \#\{k : u_k \le u, v_k \le v\}. \tag{3}$$

Note that although C_n is not a copula, its restriction to the lattice $L = \{(\frac{i}{n},\frac{j}{n}) : i,j = 0,\dots,n\}$ is a subcopula.

Next we construct a two-parameter family of test statistics

$$\mathscr{T}_n(x,y) = \sqrt{n}(C_n(x,C_n(y,y)) - C_n(C_n(x,y),y)), \quad x,y \in (0,1). \tag{4}$$

Basing on $\mathscr{T}_n$ we will provide a test for associativity, hence also for Archimedeanity.

2 Asymptotic Results

Let $(X_k,Y_k)_{k=1}^{+\infty}$ be a data generating process consisting of independent copies of (X,Y). For the sample of size n we construct the statistics $\mathscr{T}_n$. The first observation is that under slight assumptions on differentiability of C these statistics are asymptotically normal.

Theorem 2. *Let $0<x<y<1$ and $x+y>1$. If C has continuous derivatives on a square $[C(x,y),y]^2$ then as n grows to $+\infty$*

$$\mathscr{T}_n(x,y) - \sqrt{n}(C(x,C(y,y)) - C(C(x,y),y)) \xrightarrow{d} N(0,\sigma^2), \tag{5}$$

where σ does not exceed 2.

The above estimate on the asymptotic standard deviation can be improved when C is associative.

Theorem 3. *Let $0<x<y<1$ and $x+y>1$. If C is associative and has continuous derivatives on a square $[C(x,y),y]^2$ then as n grows to $+\infty$*

$$\mathscr{T}_n(x,y) \xrightarrow{d} N(0,\sigma^2), \tag{6}$$

where σ does not exceed $\sqrt{2(1-y)}$.

Note that we do not require the differentiability of C at the points $(0,0)$ or $(1,1)$ which would be very restrictive. As a matter of fact if we choose x and y such that $0<x<y<1$ and $x+y>1$ then it is enough to assume that C is C^1 on the square $[x+y-1,y]^2$. Therefore if C is continuously differentiable on the open square $(0,1)^2$ then we can choose any x and y subject to the condition $0<x<y<1$ and $x+y>1$.

The proofs of the above theorems are provided in the Appendix.

3 Test

We consider the following null hypothesis:

- H_0: The copula C is Archimedean[1] and is continuously differentiable on the rectangle $(0,1)^2$;

and an alternative hypothesis:

- H_1: The copula C is not associative.

We fix the significance level α and apply the following decision rule:

1. Select a point (x,y), such that $0<x<y<1$ and $x+y>1$.
2. Count the value t of the test statistic $\mathscr{T}_n(x,y)$.

[1] The word Archimedean may be replaced by associative - see Corollary 2

3. Determine the critical value $t^* = q(1-\alpha/2)\sqrt{2(1-y)}$, where q is a quantile from the standard normal distribution.
4. If the absolute value of t exceeds t^* then reject H_0.

Alternatively we may instead of point 3 and 4 apply:

3'. Determine the p-value, $p = 2F(-|t|(2(1-y))^{-0.5})$, where F is the distribution function of $N(0,1)$.

4'. If p is smaller then α, then reject H_0.

Remark 1. 1. Theorem 3 implies that the asymptotic size of the above test does not exceed α.

2. The test can be repeated for different values x and y.
3. Theorem 2 gives some guidance how big should be the sample so that the test will be able to reject with given power a nonassociative copula C^* with known value of $\delta = C^*(x, C^*(y,y)) - C^*(C^*(x,y),y)$. Generally one should take $n >> \delta^{-2}$ i.e. at least about 10 000.

4 Example - Real Data

As an illustration of our test we will study four copulas associated with exchange rates of British Pound (GBP) and Japan Yen (JPY) against American Dollar (USD).

Let X and Y denote the one day logarithmic returns of prices of 1 GPB in USD and of 1 JPY in USD. We assume that the pairs of daily returns are independent and identically distributed. Let C_{00}, C_{10}, C_{01} and C_{11} be the copulas describing the dependencies between respectively X and Y, $-X$ and Y, X and $-Y$, $-X$ and $-Y$. Basing of the data for the period 4th January 1971 - 26th February 2010 ($n = 9844$ observations)[2] we perform the tests for four copulas and two points $(x,y) = (0.4, 0.75)$ and $(x,y) = (0.45, 0.8)$. The results are gathered in Table 1. The Archimedeanity of the copulas C_{10} and C_{01} is rejected on both levels, the Archimedeanity of the copula C_{00} is rejected only on the level $\alpha = 0.1$. For the last copula C_{11} the null hypothesis is not rejected. The results of the test suggest that only the survival copula C_{11} ought to be modelled as an Archimedean one.

5 Example - Artificial Data

To get some insight on the power with which the test is rejecting the non Archimedean copulas we generated random variates from the Student two dimensional distribution with two degrees of freedom and $\rho \in \{-0.5, 0, 0.5\}$. We put $n = 10000$ and repeated the draw 100 times for each copula. In Table 2 we present the obtained results. In subsequent columns there are values of the generalized correlation ρ, x, y, $\delta = C_\rho(x, C_\rho(y,y)) - C_\rho(C_\rho(x,y),y)$, the

[2] Source: PACIFIC Exchange Rate Service `http://fx.sauder.ubc.ca/`

Table 1 Tests for Archimedeanity

copula	x	y	test	p-value	$t^*(0.05)$	$t^*(0.1)$
C_{00}	0,4	0,75	1,119	0,114	1,386	1,163
C_{00}	0,45	0,8	1,149	0,069	1,240	1,040
C_{10}	0,4	0,75	1,512	0,033	1,386	1,163
C_{10}	0,45	0,8	1,038	0,101	1,240	1,040
C_{01}	0,4	0,75	1,411	0,046	1,386	1,163
C_{01}	0,45	0,8	1,411	0,026	1,240	1,040
C_{11}	0,4	0,75	0,554	0,433	1,386	1,163
C_{11}	0,45	0,8	0,655	0,300	1,240	1,040

Table 2 Testing Student copulas

ρ	x	y	δ	mean	σ	$t^*(0.05)$	$t^*(0.1)$	$N(0.05)$	$N(0.1)$
-0,5	0,4	0,75	0,02097	2,104	0,25	1,386	1,163	99	100
0	0,4	0,75	0,01651	1,668	0,25	1,386	1,163	85	98
0,5	0,4	0,75	0,01162	1,171	0,21	1,386	1,163	17	47

mean value of the sample statistics $\mathscr{T}_n(x,y)$ and the standard deviation – σ, critical values for $\alpha = 0.05$ and $\alpha = 0.1$, the number of times (out of 100) when the test rejected the Archimedeanity for $\alpha = 0.05$ and $\alpha = 0.1$.

6 Appendix: Proofs

The proofs of Theorems 2 and 3 follow from the fact that in points of differentiability the empirical copula process $\sqrt{n}(C_n - C)$ converges weakly to the Gaussian process $\mathbb{G}_C$ - compare [3] and [7]. Namely if C is continuously differentiable on a square $[a,b]^2$ then

$$\sqrt{n}(C_n - C)(u,v) \xrightarrow{d} \mathbb{G}_C(u,v) \quad \text{in} \quad l^\infty([a,b]^2). \tag{7}$$

The limiting Gaussian process can be written as

$$\mathbb{G}_C(u,v) = \mathbb{B}_C(u,v) - \partial_1 C(u,v)\mathbb{B}_C(u,1) - \partial_2 C(u,v)\mathbb{B}_C(1,v), \tag{8}$$

where $\mathbb{B}_C$ is a Brownian bridge on $[0,1]^2$ with covariance function

$$\mathbb{E}(\mathbb{B}_C(u,v) \cdot \mathbb{B}_C(u',v')) = C(\min(u,u'), \min(v,v')) - C(u,v)C(u',v'). \tag{9}$$

It is not difficult to check the variance of $\mathbb{G}_C$ is uniformly bounded. Indeed:

Lemma 1. *For every copula C and every point of its differentiability*

$$\mathbb{D}^2(\mathbb{G}_C(u,v)) \leq \frac{1}{4}. \tag{10}$$

Proof

$$\mathbb{D}^2(\mathbb{G}_C(u,v)) = \mathbb{E}(\mathbb{B}_C(u,v) - \partial_1 C(u,v)\mathbb{B}_C(u,1) - \partial_2 C(u,v)\mathbb{B}_C(1,v))^2$$

$$= C(u,v)(1-C(u,v)) - 2\partial_1 C(u,v)(C(u,v)(1-u)) - 2\partial_2 C(u,v)(C(u,v)(1-v))$$
$$+\partial_1 C(u,v)^2 u(1-u) + \partial_2 C(u,v)^2 v(1-v) + 2\partial_1 C(u,v)\partial_2 C(u,v)(C(u,v)-uv).$$

Since C is nondecreasing in both variables and Lipschitz with constant 1, its partial derivatives belong to the interval $[0,1]$ ([6] Theorem 2.2.7). Therefore the variance of $\mathbb{G}_C$ is bounded by the value of the quadratic function

$$\Psi(\xi_1,\xi_2) = C(u,v)(1-C(u,v)) - 2\xi_1 C(u,v)(1-u) - 2\xi_2 C(u,v)(1-v)$$
$$+\xi_1^2 u(1-u) + \xi_2^2 v(1-v) + 2\xi_1\xi_2(C(u,v)-uv)$$

at the vertices of the unit square.

Since $C(u,v) \in [0,1]$, we get

$$\Psi(0,0) = C(u,v)(1-C(u,v)) \leq \frac{1}{4}.$$

Since $0 \leq u - C(u,v) \leq 1$, we get

$$\Psi(1,0) = C(u,v)(1-C(u,v)) - 2C(u,v)(1-u) + u(1-u)$$
$$= (C(u,v)-u+1)(u-C(u,v)) \leq \frac{1}{4}.$$

In the same way we get

$$\Psi(0,1) \leq \frac{1}{4}.$$

Finally, since $0 \leq u+v-C(u,v) \leq 1$, we get

$$\Psi(1,1) = C(u,v)(1-C(u,v)) - 2C(u,v)(1-u) - 2C(u,v)(1-v)$$
$$+u(1-u) + v(1-v) + 2(C(u,v)-uv)$$
$$= (C(u,v)-u-v)(-1+u+v-C(u,v)) \leq \frac{1}{4}.$$

□

From formula 7 and the delta method ([7] §3.9) we get:

Lemma 2. *Under the assumptions of Theorem 2*

$$\sqrt{n}((C_n(x,C_n(y,y)) - C_n(C_n(x,y),y)) - (C(x,C(y,y)) - C(C(x,y),y))) \\ \xrightarrow{d} \mathbb{H}_C(x,y), \tag{11}$$

where $\mathbb{H}_C$ is a Gaussian process, which can written as

$$\begin{aligned} \mathbb{H}_C(x,y) &= \mathbb{G}_C(x,C(y,y)) - \mathbb{G}_C(C(x,y),y) \\ &\quad -\partial_1 C(C(x,y),y)\mathbb{G}_C(x,y) + \partial_2 C(x,C(y,y))\mathbb{G}_C(y,y). \end{aligned} \tag{12}$$

Theorem 2 is a direct corollary of above lemmas. Indeed, since $\mathbb{H}_C$ and $\mathbb{G}_C$ are Gaussian, we get

$$\begin{aligned} \sigma(\mathbb{H}_C(x,y)) &\leq \sigma(\mathbb{G}_C(x,C(y,y)) + \sigma(\mathbb{G}_C(C(x,y),y)) \\ &\quad +\partial_1 C(C(x,y),y)\sigma(\mathbb{G}_C(x,y)) + \partial_2 C(x,C(y,y))\sigma(\mathbb{G}_C(y,y)) \\ &\leq 4 \cdot \frac{1}{2} = 2 \end{aligned} \tag{13}$$

The proof of Theorem 3 requires more detailed study of $\mathbb{H}_C$. Due to associativity we have the following relations:

$$C(x,C(y,z)) = C(C(x,y),z), \tag{14}$$

$$\partial_1 C(x,C(y,z)) = \partial_1 C(C(x,y),z)\partial_1 C(x,y), \tag{15}$$

$$\partial_2 C(x,C(y,z))\partial_2 C(y,z) = \partial_2 C(C(x,y),z), \tag{16}$$

$$\partial_2 C(x,C(y,z))\partial_1 C(y,z) = \partial_1 C(C(x,y),z)\partial_2 C(x,y), \tag{17}$$

which, after substitution $z = y$, imply:

Lemma 3. *If the copula C is associative then*

$$\begin{aligned} \mathbb{H}_C(x,y) &= \mathbb{B}_C(x,C(y,y)) - \mathbb{B}_C(C(x,y),y) \\ &\quad +\partial_1 C(C(x,y),y)(\mathbb{B}_C(C(x,y),1) - \mathbb{B}_C(x,y)) \\ &\quad +\partial_2 C(x,C(y,y))(-\mathbb{B}_C(1,C(y,y)) + \mathbb{B}_C(y,y)) \\ &\quad +\partial_1 C(C(x,y),y)\partial_2 C(x,y)(\mathbb{B}_C(1,y) - \mathbb{B}_C(y,1)). \end{aligned} \tag{18}$$

Now we are in the position to prove Theorem 3.

Since partial derivatives of C belong to the interval $[0,1]$, the variance of $\mathbb{H}_C$ is bounded by the maximum value of the quadratic function $\Psi(\xi_1,\xi_2,\xi_3)$, where $\xi_1,\xi_2,\xi_3 \in [0,1]$ and $\xi_3 \leq \min(\xi_1,\xi_2)$.

$$\Psi(\xi_1,\xi_2,\xi_3) = 2[C(x,C(y,y)) - C(C(x,y),C(y,y))$$

$$+(\xi_1+\xi_2)(C(C(x,y),C(y,y)) - C(x,C(y,y)))$$

$$+\xi_1^2(C(x,y)-C(C(x,y),y))+\xi_2^2(C(y,y)-C(y,C(y,y)))$$
$$+\xi_3^2(y-C(y,y))+\xi_1\xi_2(-C(C(x,y),C(y,y))+2C(x,C(y,y))-C(x,y))$$
$$+\xi_1\xi_3(C(C(x,y),y)-C(x,y))+\xi_2\xi_3(C(y,C(y,y))-C(y,y))]$$

Since the domain is a convex polyhedron, Ψ is attaining the maximum at one of its vertices. So, it is enough to check the value of Ψ at the points (0,0,0), (1,0,0), (0,1,0), (1,1,0), (1,1,1). Basing on the fact that C is Lipschitz we get

$$\frac{1}{2}\Psi(0,0,0) = C(x,C(y,y)) - C(C(x,y),C(y,y)) \le x - C(x,y) \le 1-y,$$
$$\frac{1}{2}\Psi(1,0,0) = C(x,y) - C(x,C(y,y)) \le y - C(y,y) \le 1-y,$$
$$\frac{1}{2}\Psi(0,1,0) = C(y,y) - C(y,C(y,y)) \le y - C(y,y) \le 1-y,$$
$$\frac{1}{2}\Psi(1,1,0) = C(y,y) - C(y,C(y,y)) \le y - C(y,y) \le 1-y,$$
$$\frac{1}{2}\Psi(1,1,1) = C(x,C(y,y)) - C(x,y) - C(y,y) + y \le y - C(y,y) \le 1-y,$$

which finishes the proof.

Acknowledgements. Partially supported by Polish Ministry of Science and Higher Education, Grant N N201 547838.

References

1. Cherubini, U., Luciano, E., Vecchiato, W.: Copula Methods in Finance. John Wiley & Sons, New York (2004)
2. Durante, F., Sempi, C.: Copula Theory: an Introduction. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds.) Copula Theory and Its Applications. Proceedings of the Workshop Held in Warsaw, September 25-26, 2009. Lecture Notes in Statistics - Proceedings, vol. 198. Springer, Heidelberg (2010)
3. Fermanian, J.-D., Radulović, D., Wegkamp, M.: Weak convergence of empirical copula processes. Bernoulli 10(5), 847–860 (2004)
4. Klement, E.P., Mesiar, R., Pap, E.: Triangular norms. Trends in Logic—Studia Logica Library, vol. 8. Kluwer Academic Publishers, Dordrecht (2000)
5. McNeil, A.J., Nešlehová, J.: Multivariate Archimedean copulas, d-monotone functions and ℓ_1-norm symmetric distributions. Ann. Statist. 37(5B), 3059–3097 (2009)
6. Nelsen, R.B.: An Introduction to Copulas. Springer, New York (2006)
7. van der Vaart, A., Wellner, J.A.: Weak convergence and empirical processes: With Applications to Statistics, 2nd edn. Springer, New York (2000)

An Attempt to Define Graphical Models in Dempster-Shafer Theory of Evidence

Radim Jiroušek

Abstract. The goal of this paper is to introduce graphical models in Dempster-Shafer theory of evidence. The way the models are defined is a natural and straightforward generalization of the approach from probability theory. The models possess the same "Global Markov Properties", which holds for probabilistic graphical models. Nevertheless, the last statement is true only under the assumption that one accepts a new definition of conditional independence in Dempster-Shafer theory, which was introduced in Jiroušek and Vejnarová (2010). Therefore, one can consider this paper as an additional reason supporting this new type of definition.

Keywords: Graphical Markov models, Conditional independence, Factorization, Multidimensional basic assignment.

1 Introduction

Graphical Markov models [8] developed to their variety and proficiency in the last two decades of the 20th century, have become a benchmark with which models from other theories of uncertainty are often compared. Here we have in mind Bayesian networks (perhaps the most popular member of graphical Markov models), decomposable models (indisputably the most efficient from the computational point of view) and also "classical" graphical models. The last models were originally studied within the class of log-linear models as distributions whose interactions can be described with the help of simple graphs.

Radim Jiroušek
Faculty of Management, University of Economics, Jindřichuv Hradec, and Institute of Information Theory and Automation, Academy of Sciences,
Prage, Czech Republic
e-mail: radim@utia.cas.cz

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 361–368.
springerlink.com

In this paper we want to show that the idea upon which graphical models were founded can be (almost straightforwardly) exploit also within Dempster-Shafer Theory of evidence. In fact, the only new idea of the approach is that not all subsets of the considered space of discernment may be focal elements.

2 Basic Concepts and Notation

In the following text we will need just basic concepts od Dempster-Shafer theory of evidence. However, to make the explanation more lucid we will explain our motivation originated in probability theory. Naturally, when speaking about graphical models we cannot avoid a couple of notions from graph theory. All these concepts will be briefly introduced in this section.

All our considerations will concern finite multidimensional space

$$\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n. \tag{1}$$

The reader can interpret it either as a space of possible combinations of values of n (random) variables, or as an n-dimensional space on which the respective measures will be defined. Subsets of $N = \{1,2,\ldots,n\}$ will be denoted by K,L,M with possible indices. So, $\mathbf{X}_K$ will denote a Cartesian product of those $\mathbf{X}_i$, for which $i \in K$:

$$\mathbf{X}_K = \times_{i\in K}\mathbf{X}_i.$$

A *projection* of $x = (x_1,x_2,\ldots,x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e. for $K = \{i_1,i_2,\ldots,i_\ell\}$

$$x^{\downarrow K} = (x_{i_1},x_{i_2},\ldots,x_{i_\ell}) \in \mathbf{X}_K.$$

Analogously, for $K \subset L \subseteq N$ and $A \subset \mathbf{X}_L$, $A^{\downarrow K}$ will denote a *projection* of A into $\mathbf{X}_K$:

$$A^{\downarrow K} = \{y \in \mathbf{X}_K : \exists x \in A \quad (y = x^{\downarrow K})\}.$$

Let us remark that we do not exclude situations when $K = \emptyset$: $A^{\downarrow\emptyset} = \emptyset$.

One of the most important notions of this text will be a *join* of two subsets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$, which is defined

$$A \otimes B = \{x \in \mathbf{X}_{K\cup L} : x^{\downarrow K} \in A \quad \& \quad x^{\downarrow L} \in B\}. \tag{2}$$

Notice that if K and L are disjoint then the join of the corresponding sets is just their Cartesian product $A \otimes B = A \times B$. For $K = L$, $A \otimes B = A \cap B$. If $K \cap L \neq \emptyset$ and $A^{\downarrow K\cap L} \cap B^{\downarrow K\cap L} = \emptyset$ then also $A \otimes B = \emptyset$.

In view of this paper it is important to realize that if $x \in C \subseteq \mathbf{X}_{K\cup L}$, then $x^{\downarrow K} \in C^{\downarrow K}$ and $x^{\downarrow L} \in C^{\downarrow L}$, which means that always $C \subseteq C^{\downarrow K} \otimes C^{\downarrow L}$. However, it does not mean that $C = C^{\downarrow K} \otimes C^{\downarrow L}$. For example, considering only a 2-dimensional frame of discernment $\mathbf{X}_{\{1,2\}}$ with $\mathbf{X}_i = \{a_i,\bar{a}_i\}$ for both $i = 1,2$, and $C = \{a_1a_2,\bar{a}_1a_2,a_1\bar{a}_2\}$ one gets

$$C^{\downarrow\{1\}} \otimes C^{\downarrow\{2\}} = \{a_1,\bar{a}_1\} \otimes \{a_2,\bar{a}_2\} = \{a_1a_2,\bar{a}_1a_2,a_1\bar{a}_2,\bar{a}_1\bar{a}_2\} \neq C.$$

2.1 Graph Notions

In the paper we will exclusively consider simple graphs $G=(N,E)$ with a set of nodes N corresponding to the previously introduced index set. It means that the considered graphs contain neither oriented nor multiple edges and also no loops.

An important notion is that of a *clique*, which denotes a maximal subset of N inducing a complete subgraph (i.e. all pairs of nodes of a clique are connected by an edge and adding an additional node to the clique violates this property). The graph in Figure 1(a) has three cliques: $\{1,2,3,4\},\{3,4,5\},\{6\}$, the graph in Figure 1(b) has five cliques: $\{1,2,3\},\{1,4\},\{3,6\},\{4,5\},\{5,6\}$.

A graph is *decomposable* if its cliques $K_1,K_2,\ldots,K_r$ can be ordered in the way that the sequence meets the so called *running intersection property* (RIP):

$$\forall i=2,\ldots,r \quad \exists j(1\leq j<i): \quad K_i\cap(K_1\cup\ldots\cup K_{i-1})\subseteq K_j. \tag{3}$$

Notice that this property is met by any ordering of the cliques of the graph in Figure 1(a), and that the cliques of the graph in Figure 1(b) cannot be ordered to meet this property. It means that from the mentioned two graphs only the former is decomposable. The graph in Figure 1(c) is also decomposable, because the ordering of its cliques $\{1,2,4\},\{2,3,4\},\{4,6\},\{3,4,5\}$ meets RIP (in spite of the fact that, for example, $\{3,4,5\},\{1,2,4\},\{2,3,4\},\{4,6\}$ does not meet this property).

The last notions we will need are notions of *separation* and a *separating set.* We say that two different nodes $i,j\in N$ are *separated by a set* $K\subseteq N\setminus\{i,j\}$ if we cannot go along the graph edges from i to j without going through a node from K. So, if there is no path from i to j (as, for example there is no path from 1 to 6 in the graph in Figure 1(a)) then even the empty set may be a separating set. A set K is a *minimal separating set* if there exists a pair of nodes i and j, which is separated by K but no proper subset of K separates i and j. Notice that in the graph in Figure 1(c) both $\{2,4\}$ and $\{4\}$ are minimal separating sets; the former is a minimal separating set for 1 and 3, whereas the latter is a minimal separating set for 1 and 6.

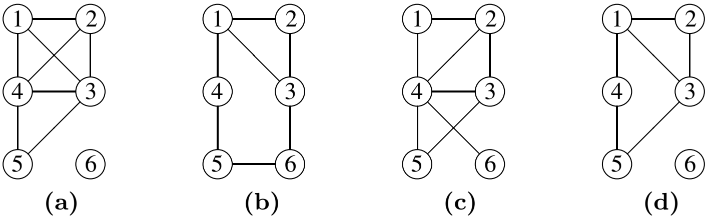


Fig. 1 Graphs with 6 nodes

If graph $G = (N,E)$ is not complete then it is always possible to find a couple of subsets $L,M \subset N$ (usually there are lot of such couples; the exception is a graph consisting of only two cliques, for which this couple is unique) such that

- $L \cup M = N$;
- $L \cap M$ is a minimal separating set;
- each pair of nodes $i \in L \setminus M$, $j \in M \setminus L$ is separated by $L \cap M$.

The set of all these couples will be denoted by symbol $\mathscr{S}(G)$ - for examples concerning all the graphs in Figure 1 see Table 1. Now, we are ready to introduce a class of subsets of $\mathbf{X}_N$ whose structures *comply with graph* G (these sets will be used in the definition of graphical models in Section 4):

$$\mathscr{R}(G) = \{A \subseteq \mathbf{X}_N : \forall (L,M) \in \mathscr{S}(G) \ \ (A = A^{\downarrow L} \otimes A^{\downarrow M})\}. \qquad (4)$$

Table 1 $\mathscr{S}(G)$ for graphs in Figure 1

Graph G	Couples (L,M) from $\mathscr{S}(G)$
(a)	$(\{1,2,3,4\},\{3,4,5,6\})$ $(\{1,2,3,4,6\},\{3,4,5\})$ $(\{1,2,3,4,5\},\{6\})$
(d)	$(\{1,2,3,4,5\},\{6\})$ $(\{1,2,3\},\{1,3,4,5,6\})$ $(\{1,2,3,6\},\{1,3,4,5\})$ $(\{1,2,3,5,6\},\{1,4,5\})$ $(\{1,2,3,5\},\{1,4,5,6\})$ $(\{1,2,3,4\},\{3,4,5,6\})$ $(\{1,2,3,4,6\},\{3,4,5\})$
(b)	$(\{1,2,4\},\{2,3,4,5,6\})$ $(\{1,2,3,4\},\{3,4,5,6\})$ $(\{1,2,3,4,5\},\{3,5,6\})$ $(\{1,2,3,4,6\},\{4,5,6\})$
(c)	$(\{1,2,4\},\{2,3,4,5,6\})$ $(\{1,2,4,6\},\{2,3,4,5\})$ $(\{1,2,3,4\},\{3,4,5,6\})$ $(\{1,2,3,4,6\},\{3,4,5\})$ $(\{1,2,3,4,5\},\{4,6\})$

2.2 Probabilistic Factorization

Consider a probability measure π on $\mathbf{X}_N$ and $L,M \subseteq N$ such that $L \cup M = N$. We say that π factorizes with respect to a couple (L,M) if the exist functions

$$\phi : \mathbf{X}_L \longrightarrow [0,+\infty), \qquad \psi : \mathbf{X}_M \longrightarrow [0,+\infty),$$

such that for all $x \in \mathbf{X}_N$

$$\pi(x) = \phi(x^{\downarrow L}) \cdot \psi(x^{\downarrow M}).$$

It is well known that π factorizes with respect to (L,M) if and only if for all $x \in \mathbf{X}_N$

$$\pi(x) \cdot \pi^{\downarrow L\cap M}(x^{\downarrow L\cap M}) = \pi^{\downarrow L}(x^{\downarrow L}) \cdot \pi^{\downarrow M}(x^{\downarrow M}),$$

which corresponds to the conditional independence $L \setminus M \perp\!\!\!\perp M \setminus L | L \cap M\,[\pi]$.

This notion forms a basis for a more general notion of a *graphical model*, which is a probability distribution factorizing with respect to a graph $G = (N, E)$ [8].

Consider a graph $G = (N, E)$ with r cliques $K_1, K_2, \dots, K_r$. We say that a probability distribution π *factorizes with respect to graph* G if there exist r functions $\phi_1, \phi_2, \dots, \phi_r$,

$$\phi_i : \mathbf{X}_{K_i} \longrightarrow [0, +\infty),$$

such that for all $x \in \mathbf{X}_N$

$$\pi(x) = \prod_{i=1}^{r} \phi_i(x^{\downarrow K_i}).$$

What is the advantage of graphical models? Naturally, first of all we can represent such a distribution with the help of (in the binary case) $\prod_{i=1}^{r} 2^{|K_i|}$ parameters (factors), which is usually much less than 2^n, the number of probabilities necessary to define a general n-dimensional distribution. Moreover, graphical models have their "semantics" expressible with the help of their conditional independence structure: If distribution π factorizes with respect to $G = (N, E)$ and $K \subset N$ separates in G nodes $i, j \in N$, then $i \perp\!\!\!\perp j \,|\, K\, [\pi]$.

2.3 Basic Assignment Notation

The role of a probability distribution from a probability theory is in Dempster-Shafer theory played by any of the set functions: belief function, plausibility function, commonality function or basic (*probability or belief*) assignment [4, 9]. In this text we will exclusively use normalized basic assignments for the purpose. Such a *basic assignment* m on $\mathbf{X}_K$ $(K \subseteq N)$ is a function

$$m : \mathscr{P}(\mathbf{X}_K) \longrightarrow [0, 1],$$

for which $m(\emptyset) = 0$, and $\sum_{A \subseteq \mathbf{X}_K} m(A) = 1$. All the sets A for which $m(A)$ is positive are called *focal elements* of m.

Having a basic assignment m on $\mathbf{X}_K$ we will consider its *marginal assignment* on $\mathbf{X}_L$ (for $L \subseteq K$), which is defined (for each $\emptyset \neq B \subseteq \mathbf{X}_L$):

$$m^{\downarrow L}(B) = \sum_{A \subseteq \mathbf{X}_K : A^{\downarrow L} = B} m(A).$$

3 Factorization and Independence

Unconditional (*marginal*) *independence* has been introduced in Dempster-Shafer theory in several equivalent ways; mostly as an application of *conjunctive combination rule* (non-normalized *Dempster's rule* of combination) [1, 3, 7, 10], or with the help of *commonality functions* [12, 11]. Here, we will use another (and as it was showed in [6] still equivalent) definition.

Let $K,L \subset N$ be disjoint. For a basic assignment m the independence $K \perp\!\!\!\perp L[m]$ holds if for all $A \subseteq \mathbf{X}_{K\cup L}$

$$m^{\downarrow K\cup L} = \begin{cases} m^{\downarrow K} \cdot m^{\downarrow L} & \text{if } A = A^{\downarrow K} \otimes A^{\downarrow L} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Having a probability measure π defined on a 2-dimensional space $\mathbf{X}_1 \times \mathbf{X}_2$ and factorizing with respect to $(\{1\},\{2\})$ we know that there exist functions ϕ and ψ such that for each $x \in \mathbf{X}_1 \times \mathbf{X}_2$

$$\pi(x) = \phi(x^{\downarrow\{1\}}) \cdot \psi(x^{\downarrow\{2\}}). \tag{6}$$

It means that $|\mathbf{X}_1| \cdot |\mathbf{X}_2|$ probabilities of measure π is defined with the help of $|\mathbf{X}_1|$ and $|\mathbf{X}_2|$ values of the factor functions ϕ and ψ. This fits the product rule expressed by formula (6).

Is it possible to transfer this simple idea directly into Dempster-Shafer theory? Basic assignment m on $\mathbf{X}_1 \times \mathbf{X}_2$ is defined with the help of $2^{|\mathbf{X}_1|\cdot|\mathbf{X}_2|}$ values, whereas factor functions

$$\mu : \mathscr{P}(\mathbf{X}_1) \longrightarrow [0,+\infty), \qquad \nu : \mathscr{P}(\mathbf{X}_2) \longrightarrow [0,+\infty),$$

are defined with the help of $2^{|\mathbf{X}_1|}$ and $2^{|\mathbf{X}_2|}$ values, respectively. Thus using an analogy to a product rule we can get only $2^{|\mathbf{X}_1|+|\mathbf{X}_2|}$ different values. However noticing that factorization (in this simple 2-dimensional situation) should yield the independence $\{1\} \perp\!\!\!\perp \{2\}$, and looking at the definition formula (5), we see that we do not need to define values of m for all subset $A \subseteq \mathbf{X}_1 \times \mathbf{X}_2$, but only for those A for which $A = A^{\downarrow|1|} \otimes A^{\downarrow|2|}$.

Generalizing the above consideration to a more complex, overlapping factorization we proposed the following definition of factorization in [5].

Definition 1. ***Simple Factorization.*** Consider two nonempty sets $K \cup L = N$. We say that basic assignment *m factorizes with respect to* (K,L) if there exist two nonnegative set functions

$$\mu : \mathscr{P}(\mathbf{X}_K) \longrightarrow [0,+\infty), \quad \nu : \mathscr{P}(\mathbf{X}_L) \longrightarrow [0,+\infty),$$

such that for all $A \subseteq \mathbf{X}_{K\cup L}$

$$m(A) = \begin{cases} \phi(A^{\downarrow K}) \cdot \psi(A^{\downarrow L}) & \text{if } A = A^{\downarrow K} \otimes A^{\downarrow L} \\ 0 & \text{otherwise.} \end{cases}$$

It is almost obvious that for this notion the following simplified version of Factorization Lemma is valid [13].

Lemma 1. *Let $K,L \subseteq N$ be disjoint and nonempty, $K \cup L = N$. m factorizes with respect to (K,L) if and only if $K \setminus L \perp\!\!\!\perp L \setminus K | K \cap L$ $[m]$.*

4 Graphical Models

Definition 2. *Let $G = (N, E)$ be a graph with r cliques $K_1, K_2, \ldots, K_r$. We say that basic assignment m* factorizes with respect to graph *G if there exist r functions $\mu_1, \mu_2, \ldots, \mu_r$, ($\mu_i : \mathscr{P}(\mathbf{X}_{K_i}) \longrightarrow [0, +\infty)$), such that for all $A \subseteq \mathbf{X}_N$*

$$m(A) = \begin{cases} \prod_{i=1}^{r} \mu_i(A^{\downarrow K_i}), & \textit{if } A \in \mathscr{R}(G), \\ 0 & \textit{otherwise.} \end{cases}$$

Example 1. Consider a 6-dimensional basic assignment factorizing with respect to the graph in Figure 1(d). If all $\mathbf{X}_i$ are binary, then general basic assignment may have up to $2^{64} - 1$ focal elements. Nevertheless, since the considered graph consists of 5 cliques: $\{1,2,3\}$, $\{1,4\}$, $\{3,5\}$, $\{4,5\}$ and $\{6\}$, all the necessary factor functions are defined with by $2^8 + 3 \cdot 2^4 + 2^2 = 308$ numbers.

We believe that the above presented example sufficiently illustrates an efficiency with which graphical models can be represented in Dempster-Shafer theory. What remains to be showed that it possesses also the second advantageous property of probabilistic graphical models, i.e. that the dependence structure of the distribution is somehow encoded in the graph. We do not have enough space to formalize the property in a form of a theorem and to prove it but an analogy of the probabilistic statement presented at the end of Section 2.2 holds: If basic assignment m factorizes with respect to $G = (N, E)$ and $K \subset N$ separates nodes i, j in G, then $i \perp\!\!\!\perp j \,|\, K \,[m]$.

For this, however, we have to say what we understand by conditional independence in Dempster-Shafer theory. Namely, we cannot apply the definition used by most of the other authors (e.g. [2, 10, 12]) but the following definition introduced in [6].

Definition 3. ***Conditional Independence.*** Let $K, L, M \subset N$ be disjoint, K, L nonempty. We say that for a basic assignment m conditional independence $K \perp\!\!\!\perp L | M \; [m]$ holds if for any $A \subseteq \mathbf{X}_{K \cup L \cup M}$ such that $A = A^{\downarrow K \cup M} \otimes A^{\downarrow L \cup M}$ the equality

$$m^{\downarrow K \cup L \cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K \cup M}(A^{\downarrow K \cup M}) \cdot m^{\downarrow L \cup M}(A^{\downarrow L \cup M})$$

holds, and $m^{\downarrow K \cup L \cup M}(A) = 0$ for all the remaining $A \subseteq \mathbf{X}_{K \cup L \cup M}$, for which $A \neq A^{\downarrow K \cup M} \otimes A^{\downarrow L \cup M}$.

5 Conclusions

We have introduced graphical models in Dempster-Shafer theory as a simple and natural generalization of probabilistic graphical models. Analogously to probabilistic case, also for Dempster-Shafer graphical models one can show that they can be efficiently represented with a reasonable number of parameters and that some conditional independence relations can be read from the respective graphs. This holds, however, only when a new definition of conditional independence in Dempster-Shafer theory (see Definition 3) is accepted. Thus the paper brings an additional reason supporting this new definition. Recall that the new concept of conditional independence does not suffer from *inconsistency with marginalization* (for details and a Studený's example see [2]), for Bayesian basic assignments coincides with probabilistic conditional independence, and meets all the semigraphoid axioms.

Acknowledgements. This work was supported by National Science Foundation of the Czech Republic under the grants no. ICC/08/E010, and 201/09/1891, and by Ministry of Education under the grants 1M0572 and 2C06019.

References

1. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief Function Independence: I. The Marginal Case. Internat. J. Approx. Reason. 29, 47–70 (2002)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief Function Independence: II. The Conditional Case. Internat. J. Approx. Reason. 31, 31–75 (2002)
3. Couso, I., Moral, S., Walley, P.: Examples of independence for imprecise probabilities. In: Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications, ISIPTA 1999, Ghent, Belgium, pp. 121–130 (1999)
4. Dempster, A.: Upper and lower probabilities induced by a multi-valued mapping. Ann. Math. Statist. 38, 325–339 (1967)
5. Jiroušek, R.: Factorization and Decomposable Models in Dempster-Shafer Theory of Evidence. In: Workshop on the Theory of Belief Functions, Brest, France (2010)
6. Jiroušek, R., Vejnarová, J.: Compositional models and conditional independence in evidence theory. Internat. J. Approx. Reason (2010), doi:10.1016/j.ijar.2010.02.005
7. Klir, G.J.: Uncertainty and Information. In: Foundations of Generalized Information Theory. Wiley, Hoboken (2006)
8. Lauritzen, S.L.: Graphical models. Oxford University Press, Oxford (1996)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, New Jersey (1976)
10. Shenoy, P.P.: Conditional independence in valuation-based systems. Internat. J. Approx. Reason. 10, 203–234 (1994)
11. Studený, M.: Formal properties of conditional independence in different calculi of AI. In: Moral, S., Kruse, R., Clarke, E. (eds.) ECSQARU 1993. LNCS, vol. 747, pp. 341–351. Springer, Heidelberg (1993)
12. Studený, M.: On stochastic conditional independence: the problems of characterization and description. Ann. Math. Artif. Intell. 35, 323–341 (2002)
13. Vejnarová, J.: On conditional independence in evidence theory. In: Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK, pp. 431–440 (2009)

Comparison of Time Series via Classic and Temporal Protoforms of Linguistic Summaries: An Application to Mutual Funds and Their Benchmarks

Janusz Kacprzyk and Anna Wilbik

Abstract. We present a new approach to the evaluation of similarity of time series that are characterized by linguistic summaries. We consider so-called temporal data summaries, i.e. novel linguistic summaries that explicitly include a temporal aspect. We consider the case of a mutual (investment) fund and its underlying benchmark(s), and the new comparison method is based not on the comparison of the consecutive values or segments of the fund and its benchmark but on the comparison of classic and temporal linguistic summaries (i.e. based on a classic and temporal protoform) best describing their past behavior.

Keywords: Time series comparison, Linguistic data summarization, Temporal protoform, Fuzzy logic, Computing with words.

1 Introduction

As in our previous works cited in the literature, we consider the following setting: a decision maker has to decide on how much and in which mutual fund (or a financial instrument) to invest. The decision maker has information on some objective aspects of the past behavior of the mutual fund quotations, exemplified by results of statistical analyses, macroeconomic data, exchange rates, etc. and also has some additional information and knowledge, resulting from experience, informal analyses, personal sources of information, intuition, etc. which are of a *tacit knowledge* type, difficult to codify or share.

We follow the decision support philosophy, that is, the decision maker is to make his/her investment decision autonomously, and additional information, insight into the data, etc. may be of help.

Janusz Kacprzyk and Anna Wilbik
Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl,wilbik@ibspan.waw.pl

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 369–377.
springerlink.com

One of most interesting and relevant questions of a professional analyst or a customer may be: how similar was the temporal evolution of quotations of the particular investment fund and its benchmark(s)? This has clearly to do with the measuring of similarity of time series. The problem of similarity of time series or finding similar subsequences of time series is an important issue, e.g., in indexing [21, 20], clustering [6] or motif discovery [3, 23]. Many approaches to the similarity measures of time series were proposed, cf. those based on the Euclidean distance [5], possibly with scaling or shifting [3], the Dynamic Time Warping (DTW) [20], and Longest Common Subsequence (LCS) [5], using the Discrete Fourier Transform (DFT) [1], wavelets [2], etc.

We also dealt with this problem (cf. [11]) and computed a degree of similarity of two time series by comparing the membership values of linguistic descriptions of the consecutive extracted segments (trends) of the two time series, and then by aggregating the results obtained.

In this work we further develop the approach proposed in our previous paper cf. [11]) by using a new type of protoforms of linguistic summaries of time series data concerning the (past) quotations of a mutual fund under consideration. Basically, the new protoforms – called *temporal protoforms* – involve explicitly a time aspect with which a particular summary deals. For instance, the traditional protoforms used in our previous papers may be "Among all trends, *most* are *short*" or "Among all *short* trends, *almost all* are *increasing*". The new temporal protoforms proposed by us may :. for instance, "*recently*, among all trends, *most* are of *high variability*" or "*In the peak period of the economic crisis*, among all *decreasing* trends, *most* are *long*.

It is easy to see that the form of the new protoforms of linguistic summaries proposed in more sophisticated and a comparison of linguistic summaries whose structure follows such protoforms will be considerably more difficult. The main problem is the additional time dimension that is present.

We follow the main idea proposed in Kacprzyk and Wilbik(cf. [11]) that a very human consistent method of comparison of two time series (here: the past quotations of an mutual fund and its benchmark) may be performed a linguistic quantifier driven aggregation of partial comparison results. Then, that idea was extended by Kacprzyk and Wilbik [12] to the comparison of linguistic summaries of time series, not their numerical series of values.

The degree of similarity of two time series is the degree to which, for instance, for a "majority" of distinctive time periods "most" valid summaries of the fund have the truth values similar with a majority of similar summaries of the benchmark". This is an extension of our previous proposal [12] and has very much to do with a soft definition of consensus meant as a degree to which, for instance, "*most* of the *important* individuals agree as to *almost all* of the *relevant* options". This was introduced by Kacprzyk and Fedrizzi [7, 8, 9], Kacprzyk, Fedrizzi and Nurmi [10], and recently by Kacprzyk and Zadrożny [16]. It is clearly to see that the use of temporal protoforms makes the definitions similarity based on consensus difficult because of the temporal character of the protoform of the summaries assumed.

2 Linguistic Summaries of Time Series

A linguistic summary of data (database) is a (usually short) sentence (or a few sentences) that captures the very essence of the data; the date is assumed here numeric, its amount is large and not comprehensible for human users. We use here Yager's [24] basic approach, and a linguistic summary includes: a summarizer P, a quantity in agreement Q, i.e. a linguistic quantifier, truth (validity) $\mathscr{T}$ of the summary and optionally, a qualifier R; an extended, implementable approach was shown in Kacprzyk and Yager [14], and Kacprzyk, Yager and Zadrożny [18].

Thus, the core of a linguistic summary is a linguistically quantified proposition in the sense of Zadeh [25] which may be written, respectively, as

$$Qy's \text{ are } P \qquad\qquad QRy's \text{ are } P \tag{1}$$

which may be exemplified, respectively by: "*Most* of employees earn *low* salary", $\mathscr{T}$=0.7, or "*Most* of *young* employees earn *low* salary", $\mathscr{T}$=0.82.

We focus on trends, linear segments extracted from the time series, obtained via using a piecewise linear segmentation method (cf. [22]). We consider three features of (global) trends in time series: (1) dynamics of change, (2) duration, and (3) variability. All of them are treated as linguistic variables, and for for all we use a fuzzy granulation to represent the values by a small set of linguistic labels as, e.g.: increasing, constant, decreasing, equated with fuzzy sets.

Classic protoforms

For clarity and convenience we employ Zadeh's protoforms for dealing with linguistic summaries – cf. Kacprzyk and Zadrożny [15]. A *protoform* is a more or less abstract prototype (template) of a linguistically quantified proposition. We have two types of protoforms of linguistic summaries of trends:

– a simple (short) form, e.g., "Among all segments, *most* are *increasing*":

$$\text{Among all segments, } Q \text{ are } P \tag{2}$$

– an extended form, e.g., "Among all *short* segments, *most* are *slowly increasing*":

$$\text{Among all } R \text{ segments, } Q \text{ are } P \tag{3}$$

The basic quality criterion is the truth value (degree) [24]. Using Zadeh's calculus of linguistically quantified propositions [25] it is calculated as:

$$\mathscr{T}(\text{Among all } y\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_P(y_i)\right) \tag{4}$$

$$\mathscr{T}(\text{Among all } Ry\text{'s, } Q \text{ are } P) = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_R(y_i)\wedge\mu_P(y_i)}{\sum_{i=1}^{n}\mu_R(y_i)}\right) \tag{5}$$

where Q is a fuzzy set representing the linguistic quantifier in the sense of Zadeh [25], i.e. regular, nondecreasing and monotone and $\wedge$ is the minimum (or a t-norm, cf. Kacprzyk, Wilbik and Zadrożny [19]).

Temporal protoforms
We extended (cf. [13]) our protoforms given in (2) and (3) by adding a temporal expression E_T like: "recently", "in the very beginning" or "in May 2010", "initially", etc., and the *temporal protoforms* are:
– a simple form as, e.g., "*Recently* among all segments, *most* are *short*":

$$E_T \text{ among all segments, } Q \text{ are } P \tag{6}$$

– an extended form as, e.g., "*Initially* among all *short* segments, *most* are *slowly increasing*":

$$E_T \text{ among all } R \text{ segments, } Q \text{ are } P \tag{7}$$

The truth values of the temporal protoforms (6) and (7) are, respectively:

$$\mathcal{T}(E_T \text{ among all } y\text{'s, } Q \text{ are } P) = \mu_Q \left(\frac{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^{n} \mu_{E_T}(y_i)} \right) \tag{8}$$

$$\mathcal{T}(E_T \text{ among all } Ry\text{'s, } Q \text{ are } P) = \mu_Q \left(\frac{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_R(y_i) \wedge \mu_P(y_i)}{\sum_{i=1}^{n} \mu_{E_T}(y_i) \wedge \mu_R(y_i)} \right) \tag{9}$$

where $\mu_{E_T}(y_i)$ is degree to which a trend (segment) occurs during the time span described by E_T.

3 Evaluation of the Similarity of Two Time Series Based on a Set of Their Best Linguistic Summaries

We assume that if two time series are described by similar linguistic summaries, then they may be considered as similar. So the *degree of similarity* of two time series is defined as the degree to which, e.g., "'most" valid summaries of the fund have similar truth values as a majority of similar summaries describing the benchmark', i.e. we compare two groups of the most valid linguistic summaries, with the truth value higher than some value. Notice that the similarity as meant in this paper is of a somehow compound character because in involves the similarity of fuzzy sets representing the particular linguistic terms in the summaries, the similarity of temporal expressions, as well as the similarity of the very structure of the summary. We follow a more intuitive approach to the definition of similarity, expressed by the above cited linguistically quantified proposition that represents a "softy" concept of consensus. Clearly, the particular elements of similarity, the above mentioned "subsimilarities" have a clear meaning and formal definitions, for instance as discussed in Cross and Sudkamp [4].

To get a deeper insight, we can generate the linguistic summaries with the temporal expressions like "initially", "in the the middle of considered time span", "recently", etc. – cf. [13]). Then we can compare the best of those summaries of the time series of the same temporal expression. So the *degree of similarity* of two time series is now the degree to which, e.g., 'for a "majority" of temporal expressions, "most" valid temporal summaries of the fund have the truth values similar with a majority of similar temporal summaries of the benchmark'.

In case of the comparison based on a classic protoform, assume that we wish to compare two time series A and B, A described by k linguistic summaries s_{A_j}, $j = 1,\dots,k$ and B described by l summaries, s_{B_j}, $j = 1,\dots,l$, with their respective truth values denoted as $\mathscr{T}_{A_j}$, $j = 1,\dots,k$ and $\mathscr{T}_{B_j}$, $j = 1,\dots,l$. First, we calculate the degree of similarity between each summary describing A and each summary describing B, $\mathrm{sim}(s_{A_i}, s_{B_j})$ – cf. Kacprzyk and Wilbik [12]. Then, the degree to which a summary from time series A is similar to the most valid summaries of B is:

$$\mathrm{sim}(s_{A_i}, B) = \mu_{\mathrm{some}}\left(\frac{\sum_{j=1}^{l}(1 - |\mathscr{T}_{A_i} - \mathscr{T}_{B_j}|)\ \mathrm{sim}(s_{A_i}, s_{B_j})}{\sum_{j=1}^{l}\mathrm{sim}(s_{A_i}, s_{B_j})}\right) \tag{10}$$

where $1 - |\mathscr{T}_{A_i} - \mathscr{T}_{B_j}|$ is an evaluation of similarity of truth values.

Then, to aggregate the above similarity values, we use a linguistic quantifier driven aggregation:

$$\mathrm{sim}(A, B) = \mu_{\mathrm{most}}\left(\frac{1}{k}\sum_{i=1}^{k}\mathrm{sim}(s_{A_i}, B)\right) \tag{11}$$

In the latter case, we first compare the summaries with the same temporal expression, E_{T_p} as described above. Then, we aggregate the similarity values for each temporal expression using a linguistic quantifier driven aggregation. If $\mathrm{sim}(A, B, E_{T_p})$ is the similarity value for temporal expression E_{T_p}, and we have t such expressions, then

$$\mathrm{sim}(A, B) = \mu_{\mathrm{most}}\left(\frac{1}{k}\sum_{p=1}^{t}\mathrm{sim}(A, B, E_{T_p})\right) \tag{12}$$

4 Numerical Results: The Similarity of Time Series which Are Represented as Linguistic Summaries

We will present now some numerical results of similarity evaluation of some linguistic summaries representing the time series in question. The linguistic summaries considered are based on novel temporal protoforms, and we will use a new method of comparison for such a type of protoforms. We will only deal with the comparison of linguistic summaries of time series, and not with the comparison of the original numerical time series which is the domain of conventional methods of the comparison of time series. That is why a comparison between the results obtained by our approach and by conventional methods may be not meaningful as it would concern different entities and aspects.

The method proposed in this paper was tested on data on quotations of an investment (mutual) fund that invests at least 50% of assets in shares listed at the Warsaw Stock Exchange (WSE). We have used two benchmarks, the WIG index, the benchmark for the fund given in its prospectus, and WIG20, the index of 20 biggest and most liquid companies; cf. (`http://www.gpw.pl`).

The data from from January 2002 to December 2009 contain a stable, growth, and fall periods – cf. Fig. 1, with the single share value of PLN (Polish zloty) 12.06 in the beginning to PLN 35.82 at the end, and with PLN 9.35 as minimal and PLN 57.85 as maximal. The biggest daily increase was PLN 2.32, while the biggest daily decrease was PLN 3.46. The values for the benchmark (WIG and WIG20), cf. Fig. 1, measured in different units, were: 13995.24 and 1217.32 at the beginning, 39985.99 and 2388.72 at the end, with the minimal of 12582.38 and 1039.20, and the maximal of 67568.51 and 3917.87, respectively.

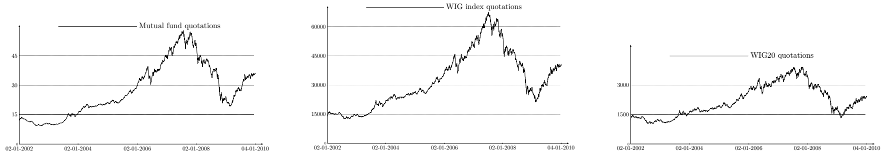


Fig. 1 Daily quotations of an investment fund in question and WIG and WIG20 indices

Comparison based on a classic protoform

We considered the linguistic summaries the mutual fund with the least truth value of 0.8. Basically, all 15 summaries describing the mutual fund were also the 15 summaries describing the WIG. The WIG20 index was described by 14 summaries only, and all of them were also describing the fund. Next we calculated the similarities between the summaries and aggregated those values obtaining the degree of similarity of those time series. The degree of similarity of the fund and WIG was 0.9426, and the degree of similarity of the fund and WIG20 was 0.9386. The WIG and WIG20 time series are very similar.

Comparison based on a temporal protoform

We used the same threshold values, i.e. 0.8 for the truth value and 0.2 for the degree of focus, and only 3 temporal expressions: “initially” (first 3 years), “in the middle of the considered time span” and “recently” (from the beginning of the financial crisis).

Table 1 Degrees of similarities for given temporal expressions

temporal expression	fund–WIG similarity	fund–WIG20 similarity
initially	0.9694	0.8675
in the middle	0.9027	0.9133
recently	0.9680	0.9667

We obtained that the degree of similarity of the fund and WIG is 0.9467. and the degree of similarity of the fund and WIG20 is 0.9158 so that the difference is even bigger than for the comparison based on a clasic protoform. We also compared the similarities on some time spans as shown in the Table 1.

The first results are encouraging and promising. It seems that by combining a comparison of times series based on their global characteristics by (a set of) linguistic summaries and temporal linguistic summaries of segments (trends) may be very human consistent and intuitively appealing, hence best serving the purpose of decision support.

5 Concluding Remarks

We presented a new approach to the evaluation of the similarity of time series that are characterized by linguistic summaries of their past performance. We considered an investment (mutual) fund and its underlying benchmark(s), and the new comparison method is based not on the comparison of the consecutive values or segments of the fund and its benchmark but on the comparison of classic and temporal linguistic summaries (i.e. based on a classic and temporal protoform) best describing their past behavior. The degree of similarity of two time series was basically based on the soft degree of consensus and was the degree to which, for instance, for a "majority" of distinctive time periods "most" valid summaries of the fund have the truth values similar with a majority of similar summaries of the benchmark". The results obtained indicated that the new method for the evaluation of similarity gives very human consistent and intuitively appealing results.

References

1. Agrawal, R., Faloutsos, C., Swami, A.: Efficient similarity search in sequence databases. In: Lomet, D.B. (ed.) FODO 1993. LNCS, vol. 730, pp. 69–84. Springer, Heidelberg (1993)
2. Chan, K.P., Fu, W.C.: Efficient time series matching by wavelets. In: Proceedings of the 15th International Conference on Data Engineering, ICDE 1999, Sydney, Austrialia, p. 126. IEEE Computer Society, Los Alamitos (1999)
3. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proceedings of the the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 493–498 (2003)

4. Cross, V., Sudkamp, T.: Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications. Springer, Heidelberg (2002)
5. Das, G., Gunopulos, D., Mannila, H.: Finding similar time series. In: Komorowski, J., Żytkow, J.M. (eds.) PKDD 1997. LNCS, vol. 1263, pp. 88–100. Springer, Heidelberg (1997)
6. Geurts, P.: Pattern extraction for time series classification. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, p. 115. Springer, Heidelberg (2001)
7. Kacprzyk, J., Fedrizzi, M.: 'Soft' consensus measures for monitoring real consensus reaching processes under fuzzy preferences. Control Cybernet 15, 309–323 (1986)
8. Kacprzyk, J., Fedrizzi, M.: A 'soft' measure of consensus in the setting of partial (fuzzy) preferences. European J. Oper. Res. 34, 315–325 (1988)
9. Kacprzyk, J., Fedrizzi, M.: A 'human-consistent' degree of consensus based on fuzzy logic with linguistic quantifiers. Math. Social Sci. 18, 275–290 (1989)
10. Kacprzyk, J., Fedrizzi, M., Nurmi, H.: Group decision making and consensus under fuzzy preferences and fuzzy majority. Fuzzy Sets Syst. 49, 21–31 (1992)
11. Kacprzyk, J., Wilbik, A.: Using fuzzy linguistic summaries for the comparison of time series: an application to the analysis of investment fund quotations. In: Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA/EUSFLAT 2009, Lisbon, Portugal, pp. 1321–1326 (2009)
12. Kacprzyk, J., Wilbik, A.: A comprehensive comparison of time series described by linguistic summaries and its application to the analysis of performance of a mutual fund and its benchmark. In: Proceedings of the 2010 World Conference on Computational Intelligence, WCCI 2010, Barcelona, Spain (in press, 2010)
13. Kacprzyk, J., Wilbik, A.: Temporal linguistic summaries of time series using fuzzy logic. In: Proceedings of International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2010, Dortmund, Germany (in press, 2010)
14. Kacprzyk, J., Yager, R.R.: Linguistic summaries of data using fuzzy logic. Int. J. Gen. Syst. 30, 33–154 (2001)
15. Kacprzyk, J., Zadrożny, S.: Linguistic database summaries and their protoforms: toward natural language based knowledge discovery tools. Inform. Sci. 173, 281–304 (2005)
16. Kacprzyk, J., Zadrożny, S.: Towards a general and unified characterization of individual and collective choice functions under fuzzy and nonfuzzy preferences and majority via the ordered weighted average operators. Int. J. Intell. Syst. 24(1), 4–26 (2009)
17. Kacprzyk, J., Zadrożny, S.: Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries and natural language generation. IEEE Trans. Fuzzy Syst. (to appear, 2010)
18. Kacprzyk, J., Yager, R.R., Zadrożny, S.: A fuzzy logic based approach to linguistic summaries of databases. Int. J. Appl. Math. Comput. Sci. 10, 813–834 (2000)
19. Kacprzyk, J., Wilbik, A., Zadrożny, S.: Linguistic summarization of time series under different granulation of describing features. In: Kryszkiewicz, M., Peters, J.F., Rybiński, H., Skowron, A. (eds.) RSEISP 2007. LNCS (LNAI), vol. 4585, pp. 230–240. Springer, Heidelberg (2007)

20. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowl. Inform. Syst. 7(3), 358–386 (2005)
21. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Locally adaptive dimensionality reduction for indexing large time series databases. In: Proceedings of ACM SIGMOD Conference on Management of Data, Santa Barbara, CA, pp. 151–162 (2001)
22. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: Data Mining in Time Series Databases, World Scientific Publishing, Singapore (2004)
23. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2009, Sparks, Nevada, USA, pp. 473–484 (2009)
24. Yager, R.R.: A new approach to the summarization of data. Inform. Sci. 28, 69–86 (1982)
25. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst. 9(2), 111–127 (1983)

Mining Gradual Dependencies Based on Fuzzy Rank Correlation

Hyung-Won Koh and Eyke Hüllermeier

Abstract. We propose a novel framework and an algorithm for mining gradual dependencies between attributes in a data set. Our approach is based on the use of fuzzy rank correlation for measuring the strength of a dependency. It can be seen as a unification of previous approaches to evaluating gradual dependencies and captures both, qualitative and quantitative measures of association as special cases.

1 Introduction

In association analysis, a widely applied data mining technique, the goal is to find "interesting" associations in a data set, that is, dependencies between so-called itemsets (binary attributes) $\mathscr{A}$ and $\mathscr{B}$ expressed in terms of rules of the form "IF $\mathscr{A}$ THEN $\mathscr{B}$". The intended meaning of a rule of that kind is that, if $\mathscr{A}$ is present in a transaction, then $\mathscr{B}$ is likely to be present, too. Association rule mining has also been extended to the fuzzy case, in which the presence of an item in a transaction is a matter of degree [7].

Another type of association rule, called *gradual dependency*, has been introduced in [10] and was further studied in [2, 11]. As explained in Section 2, the idea is to express dependencies, not between the presence or absence of attributes, but between the *change* of the presence of fuzzy items in a transaction. The contribution of this paper is a novel framework for mining gradual dependencies that is based on the use of fuzzy rank correlation as a measure of confidence (Section 3). This framework can be seen as a unification of previous approaches and captures both, qualitative and quantitative measures of association (Section 4). We also propose an algorithm for mining gradual dependencies and illustrate the method on a wine quality data set.

Hyung-Won Koh and Eyke Hüllermeier
Department of Mathematics and Computer Science,
University of Marburg, Germany
e-mail: koh,eyke@mathematik.uni-marburg.de

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 379–386.
springerlink.com

2 Gradual Dependencies

We adopt a feature-based representation of transactions (data records) and denote by $\mathbb{A}$ the (finite) set of underlying fuzzy attributes. Thus, each transaction is represented in terms of a feature vector $\boldsymbol{u}$, and for each $A \in \mathbb{A}$, $A(\boldsymbol{u}) \in [0,1]$ indicates the degree to which $\boldsymbol{u}$ has feature A or, say, to which A is present in $\boldsymbol{u}$. Correspondingly, the degree of presence of a feature subset $\mathscr{A} = \{A_1,\ldots,A_m\} \subset \mathbb{A}$, considered as a conjunction of primitive features $A_1,\ldots,A_m$, is given by $\mathscr{A}(\boldsymbol{u}) = \top(A_1(\boldsymbol{u}),A_2(\boldsymbol{u}),\ldots,A_m(\boldsymbol{u}))$, where $\top$ is a triangular norm (t-norm) serving as a generalized conjunction.

Given a data set consisting of N transactions $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_N$, a standard problem in (fuzzy) association analysis is to find all rules $\mathscr{A} \rightharpoonup \mathscr{B}$ whose support and confidence, defined as

$$\mathrm{supp} = \sum_{i=1}^{N} \top(\mathscr{A}(\boldsymbol{u}_i),\mathscr{B}(\boldsymbol{u}_i)), \quad \mathrm{conf} = \frac{\sum_{i=1}^{N} \top(\mathscr{A}(\boldsymbol{u}_i),\mathscr{B}(\boldsymbol{u}_i))}{\sum_{i=1}^{N} \mathscr{A}(\boldsymbol{u}_i)}, \tag{1}$$

exceed user-defined thresholds. A rule of such kind indicates the frequent occurrence of $\mathscr{B}$ given $\mathscr{A}$ (confidence), confirmed by sufficiently many examples (support). On a logical level, the meaning of a standard association rule $\mathscr{A} \rightharpoonup \mathscr{B}$ is captured by the material conditional. On a natural language level, such a rule is understood as an IF–THEN construct: If the antecedent $\mathscr{A}$ holds true, so does the consequent $\mathscr{B}$.

As mentioned above, another type of pattern, called *gradual dependency*, was introduced in [10]. Here, the idea is to express dependencies between the *direction of change* of attribute values. This idea is closely connected to so-called *gradual rules* in fuzzy logic. On a logical level, such rules are modeled in terms of residuated implication operators. Semantically, a rule $\mathscr{A} \rightharpoonup \mathscr{B}$ is often understood as "THE MORE the antecedent $\mathscr{A}$ is true, THE MORE the consequent $\mathscr{B}$ is true", for example "The larger an object, the heavier it is" [8]. This interpretation is arguable, however. In fact, to satisfy a gradual fuzzy rule in a logical sense, it is enough that $\mathscr{A}(\boldsymbol{u}) \leq \mathscr{B}(\boldsymbol{u})$; thus, there is actually no consideration of the change of an attribute value and, therefore, no examination of a tendency.

2.1 Evaluating Gradual Dependencies

Instead of pursing a logical approach using implication operators to evaluate a rule $\mathscr{A} \rightharpoonup \mathscr{B}$, it was proposed in [10] to take the so-called *contingency diagram* as a point of departure. A contingency diagram is a two-dimensional diagram in which every transaction $\boldsymbol{u}$ defines a point $(x,y) = (\mathscr{A}(\boldsymbol{u}),\mathscr{B}(\boldsymbol{u})) \in [0,1]^2$. Thus, for every transaction $\boldsymbol{u}$, the values on the abscissa and ordinate are given, respectively, by the degrees $x = \mathscr{A}(\boldsymbol{u})$ and $y = \mathscr{B}(\boldsymbol{u})$ to which it satisfies the antecedent and the consequent part of a candidate rule.

Informally speaking, a gradual dependency is then reflected by the relationship between the points in the contingency diagram. In particular, a "THE MORE ... THE MORE" relationship manifests itself in an increasing trend, i.e., an approximate functional dependency between the x- and y-values: the higher x, the higher y tends to be. In [10], it was therefore suggested to analyze contingency diagrams by means of techniques from statistical regression analysis. For example, if a linear regression line with a significantly positive slope can be fit to the data, this suggests that indeed a higher $x = \mathscr{A}(\boldsymbol{u})$ tends to come along with a higher $y = \mathscr{B}(\boldsymbol{u})$.

A qualitative, non-parametric alternative to this numerical approach was proposed in [2]. Roughly speaking, to evaluate a candidate rule $\mathscr{A} \rightharpoonup \mathscr{B}$, the authors count the number of pairs of points (x,y) and (x',y') in the contingency diagram for which $x < x'$ and $y < y'$. As an advantage of this approach, note that it is more flexible in the sense of not making any assumption about the type of functional dependency; as opposed to this, the regression approach implicitly assumes a linear dependency. On the other hand, since the actual distances between the points are ignored, there is also a disadvantage, namely a loss of information about the strength of a relationship.

The two above approaches, the numerical and the qualitative one, essentially come down to looking for two types of correlation between the x- and y-values, namely the standard Pearson correlation and the rank correlation. The goal of this paper is to combine the advantages of both approaches. To this end, we propose to measure the strength of a dependency in terms of a *fuzzy rank correlation* measure that combines properties of both types of correlation. As will be seen, this measure is able to capture the strength of a tendency while remaining flexible and free of specific model assumptions. Our proposal is related to the approach presented in [12] but additionally offers a sound theoretical justification.

3 Fuzzy Rank Correlation

Consider $n \geq 2$ paired observations $\{(x_i,y_i)\}_{i=1}^n \subset (\mathbb{X} \times \mathbb{Y})^n$ of two variables X and Y, where $\mathbb{X}$ and $\mathbb{Y}$ are two linearly ordered domains; we denote $\boldsymbol{x} = (x_1,x_2,\ldots,x_n)$ and $\boldsymbol{y} = (y_1,y_2,\ldots,y_n)$. The goal of a rank correlation measure is to measure the dependence between the two variables in terms of their tendency to increase and decrease in the same or the opposite direction. If an increase in X tends to come along with an increase in Y, then the (rank) correlation is positive. The other way around, the correlation is negative if an increase in X tends to come along with a decrease in Y. If there is no dependency of either kind, the correlation is (close to) 0.

Several rank correlation measures are defined in terms of the number C of *concordant*, the number D of *discordant*, and the number N of *tied* data points. For a given index pair $(i,j) \in \{1,\ldots,n\}^2$, we say that (i,j) is concordant, discordant or tied depending on whether $(x_i - x_j)(y_i - y_j)$ is positive, negative or 0, respectively. A well-known example is Goodman and Kruskal's *gamma rank correlation* [9], which is defined as $\gamma = (C-D)/(C+D)$.

3.1 Fuzzy Equivalence and Order Relations

Bodenhofer and Klawonn [5] propose a fuzzy extension of the gamma coefficient based on concepts of fuzzy orderings and $\top$-equivalence relations, where $\top$ denotes a t-norm [3].

A fuzzy relation $E : \mathbb{X} \times \mathbb{X} \to [0,1]$ is called *fuzzy equivalence* with respect to a t-norm $\top$, for brevity $\top$-equivalence, if it is reflexive ($E(x,x) = 1$), symmetric ($E(x,y) = E(y,x)$), and $\top$-transitive ($\top(E(x,y), E(y,z)) \leq E(x,z)$). Moreover, a fuzzy relation $L : \mathbb{X} \times \mathbb{X} \to [0,1]$ is called *fuzzy ordering* with respect to a t-norm $\top$ and a $\top$-equivalence E, for brevity $\top$-*E-ordering*, if it is E-reflexive ($E(x,y) \leq L(x,y)$), $\top$-E-antisymmetric ($\top(L(x,y), L(y,x)) \leq E(x,y)$), and $\top$-transitive($\top(L(x,y), L(y,z)) \leq L(x,z)$). We call a $\top$-E-ordering L *strongly complete* if $\max(L(x,y), L(y,x)) = 1$ for all $x, y \in \mathbb{X}$. Finally, let R denote a strict fuzzy ordering associated with a strongly complete $\top$-E-ordering L; in the case of the well-known Łukasiewicz t-norm, defined by $\top(x,y) = \max(0, x+y-1)$, this relation can simply be taken as $R(x,y) = 1 - L(x,y)$ [4].

3.2 The Fuzzy Gamma Rank Correlation

Consider a set of paired data points $\{(x_i, y_i)\}_{i=1}^n \subset (\mathbb{X} \times \mathbb{Y})^n$ and assume to be given two $\top$-equivalences $E_{\mathbb{X}}$ and $E_{\mathbb{Y}}$ and two strict fuzzy order relations $R_{\mathbb{X}}$ and $R_{\mathbb{Y}}$. Using these relations, the concepts of concordance and discordance of data points can be generalized as follows: Given an index pair (i,j), the degree to which this pair is concordant, discordant, and tied is defined, respectively, as

$$\tilde{C}(i,j) = \top(R_{\mathbb{X}}(x_i, x_j), R_{\mathbb{Y}}(y_i, y_j)), \tag{2}$$

$$\tilde{D}(i,j) = \top(R_{\mathbb{X}}(x_i, x_j), R_{\mathbb{Y}}(y_j, y_i)), \tag{3}$$

$$\tilde{T}(i,j) = \bot(E_{\mathbb{X}}(x_i, x_j), E_{\mathbb{Y}}(y_i, y_j)), \tag{4}$$

where $\top$ is a t-norm and $\bot$ is the dual t-conorm of $\top$ (i.e. $\bot(x,y) = 1 - \top(1-x, 1-y)$). The following equality holds for all index pairs (i,j):

$$\tilde{C}(i,j) + \tilde{C}(j,i) + \tilde{D}(i,j) + \tilde{D}(j,i) + \tilde{T}(i,j) = 1.$$

Adopting the simple sigma-count principle to measure the cardinality of a fuzzy set, the number of concordant and discordant pairs can be computed, respectively, as

$$\tilde{C} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{C}(i,j), \qquad \tilde{D} = \sum_{i=1}^{n} \sum_{j \neq i} \tilde{D}(i,j).$$

The *fuzzy ordering-based* gamma rank correlation measure $\tilde{\gamma}$, or simply "fuzzy gamma", is then defined as

$$\tilde{\gamma} = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}} \ . \tag{5}$$

From the definition of $\tilde{\gamma}$, it is clear that the basic idea is to decrease the influence of "close-to-tie" pairs (x_i, y_i) and (x_j, y_j). Such pairs, whether concordant or discordant, are turned into a partial tie, and hence are ignored to some extent. Or, stated differently, there is a smooth transition between being concordant (discordant) and being tied; see Fig. 1.

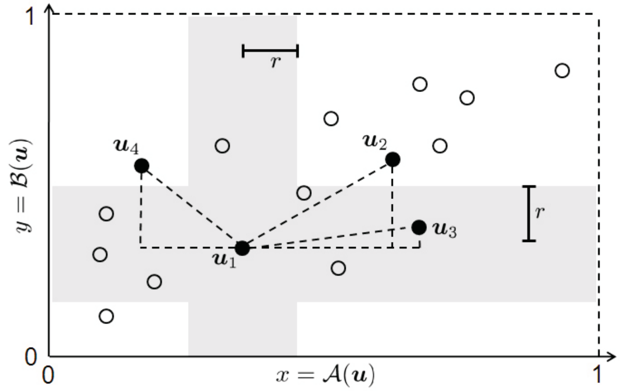


Fig. 1 Example of a contingency diagram. The pair $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ is concordant, while $(\boldsymbol{u}_1, \boldsymbol{u}_4)$ is discordant. Points with a distance $< r$ from $\boldsymbol{u}_1$ in one of the dimensions (gray region) are considered as partially tied with $\boldsymbol{u}_1$. For example, the pair $(\boldsymbol{u}_1, \boldsymbol{u}_3)$ is concordant to a degree < 1.

4 Mining Gradual Dependencies

Our idea is to evaluate a gradual dependency $\mathscr{A} \rightharpoonup \mathscr{B}$ in terms of two measures, namely the number of concordant pairs, $\tilde{C}$, and the rank correlation $\tilde{\gamma}$ as defined in (5). Comparing this approach with the classical setting of association analysis, $\tilde{C}$ plays the role of the support of a rule, while $\tilde{\gamma}$ corresponds to the confidence. These measures can also be nicely interpreted within the formal framework proposed in [7], in which every observation (in our case a pair of points $(\mathscr{A}(\boldsymbol{u}), \mathscr{B}(\boldsymbol{u}))$ and $(\mathscr{A}(\boldsymbol{v}), \mathscr{B}(\boldsymbol{v})))$ is considered, to a certain degree, as an *example* of a pattern, as a *counterexample*, or as being *irrelevant* for the evaluation of the pattern. In our case, these degrees are given, respectively, by the degree of concordance, the degree of discordance, and the degree to which the pair is a tie.

4.1 Evaluation of Candidate Rules

More formally, we define the support and confidence of a gradual dependency $\mathscr{A} \rightharpoonup \mathscr{B}$ as follows:

$$\operatorname{supp}(\mathscr{A} \rightharpoonup \mathscr{B}) = \tilde{C}, \quad \operatorname{conf}(\mathscr{A} \rightharpoonup \mathscr{B}) = \frac{\tilde{C} - \tilde{D}}{\tilde{C} + \tilde{D}},$$

where

$$\tilde{C} = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \tilde{C}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \top \left(R\left(\mathscr{A}(\boldsymbol{u}_i), \mathscr{A}(\boldsymbol{u}_j)\right), R\left(\mathscr{B}(\boldsymbol{u}_i), \mathscr{B}(\boldsymbol{u}_j)\right)\right),$$
$$\tilde{D} = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \tilde{D}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \sum_{\boldsymbol{u}_i} \sum_{\boldsymbol{u}_j} \top \left(R\left(\mathscr{A}(\boldsymbol{u}_i), \mathscr{A}(\boldsymbol{u}_j)\right), R\left(\mathscr{B}(\boldsymbol{u}_j), \mathscr{B}(\boldsymbol{u}_i)\right)\right).$$

Considering the special case of the Łukasiewicz t-norm, it can be verified that $E(x,y) = [1 - |x-y|/r]_0^1$ is a $\top$-equivalence on $\mathbb{R}$ and $R(x,y) = [(x-y)/r]_0^1$ is a strict fuzzy ordering, where $[\cdot]_0^1$ denotes the mapping $a \mapsto \min(1, \max(0,a))$. Note that these relations are parameterized by the value $r \in (0,1]$. For $r \rightarrow 0$, the confidence measure converges toward the classical (non-fuzzy) rank correlation, whereas for $r = 1$, we obtain $R(x,y) = x - y$ if $x \geq y$ and $= 0$ otherwise. The degree of concordance (discordance) is then proportional to the Euclidean distances, which means that this case is very close to the numerical evaluation in terms of Pearson correlation.

4.2 Rule Mining and Algorithmic Issues

Due to the associativity of a t-norm, the support of a rule $\mathscr{A} \rightharpoonup \mathscr{B}$ just corresponds to the support of the itemset $\mathscr{I} = \mathscr{A} \cup \mathscr{B}$. In other words, to compute a degree of concordance, there is no need to separate an itemset into an antecedent and a consequent part of a rule. Moreover, it is easy to see that the support measure is anti-monotone, i.e., $\operatorname{supp}(\mathscr{I}) \leq \operatorname{supp}(\mathscr{J})$ for $\mathscr{J} \subset \mathscr{I}$. Consequently, the candidate generation and pruning techniques of the standard Apriori framework can be used to find all frequent itsemsets, i.e., all itemsets whose support exceeds a user-defined threshold [1].

To compute the support of an itemset, we adopt some ideas that were presented in [11] for the binary case and can easily be extended to the fuzzy case. Suppose that, for a given itemset $\mathscr{I}$, the concordance degrees $\tilde{C}(\boldsymbol{u}_i, \boldsymbol{u}_j)$ are stored in an $|N| \times |N|$ matrix. From this matrix, $\operatorname{supp}(\mathscr{I})$ can easily be computed by summing all entries. Moreover, given the matrices for two itemsets $\mathscr{I}$ and $\mathscr{J}$, the matrix for the union $\mathscr{I} \cup \mathscr{J}$ is obtained by a simple position-wise t-norm combination. This approach is appealing for programming languages specifically tailored to matrix computations. In general, however, the storage requirements will be too high, especially noting that the matrices are normally quite sparse. More efficient implementations should hence exploit dedicated techniques for handling sparse matrices that, amongst others, avoid the storage of zero entries.

For each itemset $\mathscr{I}$ exceeding the given support threshold, a set of candidate rules $\mathscr{A} \rightharpoonup \mathscr{B}$ is derived by splitting $\mathscr{I}$ into antecedent part $\mathscr{A}$ and consequent part $\mathscr{B}$. For reasons of comprehensibility, we restrict ourselves to the case $|\mathscr{B}| = 1$, i.e., to consequents with a single attribute. A candidate rule of that kind is presented to the user if it exceeds the confidence threshold. While the concordance of the rule, $\tilde{C}$, is already known, this decision requires the additional computation of the discordance $\tilde{D}$.

4.3 Illustration

To illustrate our method (a thorough empirical evaluation is precluded due to space restrictions), we applied it to the Wine Quality data set from the UCI repository, in which each data record corresponds to a red wine described in terms of 11 numerical attributes and a quality degree between 0 and 10. Each attribute was replaced by two fuzzy attributes `small` and `large` with membership degrees 1 (0) and 0 (1) for the smallest and largest value, respectively, and linearly interpolating in-between. Using $r = 0.1$, we found the following rules exceeding a confidence threshold of 0.6:

- The more fixed acids and the more alcohol, the better the quality.
- The more volatile acids and sulfur dioxides, the lower the quality.
- The more volatile acids and the less alcohol, the lower the quality.
- The more sulfur dioxides and the less sulfates, the lower the quality.
- The more sulfur dioxides and the less alcohol, the lower the quality.
- The more sulfates and alcohol, the better the quality.

Roughly, one can observe that the amounts of volatile acids, sulfates and alcohol seem to have the strongest influence on the quality of the wine, with the former in a negative and the latter two in a positive manner. These results seem to agree quite nicely with oenological theory [6].

5 Concluding Remarks

We have presented a unified framework for mining fuzzy gradual dependencies, in which the strength of association between itemsets is measured in terms of a fuzzy rank correlation coefficient. As explained above, this framework generalizes previous proposals and allows for a seamless transition from a purely qualitative to a quantitative assessment.

An important aspect to be addressed in future work concerns more efficient algorithms and implementations for mining gradual dependencies. Due to the need to compare *pairs* of observations, the inherent problem complexity increases from linear to quadratic in the size of the data set. Thus, in order to guarantee scalability, efficient pruning techniques are needed that avoid unnecessary comparisons. Since the concordance relation in rank correlation is in direct correspondence to Pareto-dominance in preference modeling, it might be interesting to exploit algorithms that have recently been developed for the computation of so-called *skylines* (Pareto sets) of a database [13].

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD, Washington, D.C., USA, pp. 207–216 (1993)
2. Berzal, F., Cubero, J.C., Sanchez, D., Vila, M.A., Serrano, J.M.: An alternative approach to discover gradual dependencies. Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 15(5), 559–570 (2007)
3. Bodenhofer, U.: Representations and constructions of similarity-based fuzzy orderings. Fuzzy Sets Syst. 137, 113–136 (2003)
4. Bodenhofer, U., Demirci, M.: Strict fuzzy orderings with a given context of similarity. E Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 16(2), 147–178 (2008)
5. Bodenhofer, U., Klawonn, F.: Robust rank correlation coefficients on the basis of fuzzy orderings: Initial steps. Mathware Soft Comput. 15, 5–20 (2008)
6. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. 47(4), 547–553 (2009)
7. Dubois, D., Hüllermeier, E., Prade, H.: A systematic approach to the assessment of fuzzy association rules. Data Min. Knowl. Discov. 13(2), 167–192 (2006)
8. Dubois, D., Prade, H.: Gradual inference rules in approximate reasoning. Inform. Sci. 61(1,2), 103–122 (1992)
9. Goodman, L.A., Kruskal, W.H.: Measures of Association for Cross Classifications. Springer, New York (1979)
10. Hüllermeier, E.: Association rules for expressing gradual dependencies. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 200–211. Springer, Heidelberg (2002)
11. Laurent, A., Lesot, M.J., Rifqi, M.: GRAANK: Exploiting rank correlations for extracting gradual itemsets. In: Proceedings of FQAS 2009, Roskilde, Denmark, pp. 382–393 (2009)
12. Molina, C., Serrano, J.M., Sanchez, D., Vila, M.: Measuring variation strength in gradual dependencies. In: Proceedings of EUSFLAT 2007, Ostrava, Czech Republic, pp. 337–344 (2007)
13. Papadias, D., Tao, Y., Fu, G., Seeger, B.: Progressive skyline computation in database systems. ACM Trans. Database Syst. 30(1), 41–82 (2005)

From Probabilities to Belief Functions on MV-Algebras

Tomáš Kroupa

Abstract. In this contribution we generalize belief functions to many-valued events represented by elements of the finite product of standard MV-algebras. Our definition is based on the mass assignment approach from Dempster-Shafer theory of evidence. The generalized belief function is totally monotone and it has Choquet integral representation w.r.t. a classical belief function.

Keywords: Belief function, State, MV-algebra, Algebra of fuzzy sets.

1 Introduction

A main aim of this paper is to study belief functions in the more general setting than Boolean algebras of events. This effort is in the line with a growing interest in the generalization of classical probability towards "many-valued" events, such as those resulting from formulas in Łukasiewicz infinite-valued logic. An algebra of such many-valued events is called an MV-algebra (Definition 1). The counterpart of a probability on a Boolean algebra is a so-called state on an MV-algebra—see [10, 14, 11] for a detailed discussion of probability on MV-algebras including its interpretation in terms of bookmaking over many-valued events. The recent articles [7, 6, 9] focus on more general functionals on MV-algebras: namely, upper (lower) probabilities and possibility (necessity) measures. The presented paper is thus an attempt to fill the gap in the classification of uncertainty measures on MV-algebras.

Section 2 contains basic definitions related to MV-algebras and totally monotone functions. In Section 3 we will recall the notion of state and the integral representation of states (Theorem 1). The states on an MV-algebra

Tomáš Kroupa
Institute of Information Theory and Automation of the ASCR,
182 08 Prague, Czech Republic
e-mail: kroupa@utia.cas.cz

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 387–394.
springerlink.com

of certain set functions will be of particular interest (Example 4). Section 4 is devoted to belief functions. We will restrict their discussion only to the MV-algebra of all $[0,1]$-valued functions on a finite set. The case of belief functions on a general MV-algebra is left aside for future investigations since it involves intricate mathematical tools such as topologies on spaces of closed sets. Hence we develop a many-valued generalization of the usual notion of belief function in the finite setting [15]. In particular, Definition 3 of a belief function in the many-valued framework is based on a natural generalization of the notion of mass assignment. The properties of such belief functions are explored through Choquet integral representation (Proposition 1). This representation implies total monotonicity of belief functions and some other properties (Proposition 2). Finally, we give a complete description of the convex set of all belief functions by finding its extreme points (Proposition 3). Some proofs are omitted due to the lack of space.

2 Preliminaries

If X is any set, then $\mathscr{P}(X)$ denotes the set of all subsets of X. Put $\mathscr{P}_\emptyset(X) = \mathscr{P}(X) \setminus \{\emptyset\}$. For any $A \in \mathscr{P}(X)$, the *characteristic function* of the set A is given by $1_A(x) = 1$, if $x \in A$, and $1_A(x) = 0$, otherwise, for every $x \in X$.

A *fuzzy set* on X is a function $X \to [0,1]$. A set of fuzzy sets on X can be endowed with an algebraic structure to introduce the union, intersection, and other possible operations with fuzzy sets. This stream of research is based on so-called tribes of fuzzy sets parameterized by a t-norm [1]—see [12] for the latest exposition. Another motivation for investigating algebras of fuzzy sets stems from mathematical fuzzy logics. In this contribution we will tacitly confine to Łukasiewicz infinite-valued logic whose associated algebras of truth values are so-called MV-algebras—see [3] for their in-depth study. MV-algebras play the same role in Łukasiewicz logic as Boolean algebras in the classical two-valued logic.

Definition 1. *An* MV-algebra *is an algebra $\langle M,\oplus,\neg,0\rangle$ with a binary operation $\oplus$, a unary operation $\neg$ and a constant 0 such that $\langle M,\oplus,0\rangle$ is an abelian monoid and the following equations hold true for every $f,g \in M$: $\neg\neg f = f$, $f \oplus \neg 0 = \neg 0$, $\neg(\neg f \oplus g) \oplus g = \neg(\neg g \oplus f) \oplus f$.*

On every MV-algebra M we define $1 = \neg 0$, $f \odot g = \neg(\neg f \oplus \neg g)$. For any two elements $f,g \in M$ we write $f \leq g$ if $\neg f \oplus g = 1$. The relation $\leq$ is in fact a partial order. Further, the operations $\vee,\wedge$ defined by $f \vee g = \neg(\neg f \oplus g) \oplus g$ and $f \wedge g = \neg(\neg f \vee \neg g)$, respectively, make the algebraic structure $\langle M,\wedge,\vee,0,1\rangle$ into a distributive lattice with bottom element 0 and top element 1.

Example 1. The most important example of an MV-algebra is the *standard MV-algebra,* which is the real unit interval $[0,1]$ equipped with operations $f \oplus g = \min(1, f+g)$ and $\neg f = 1-f$. Note that we have $f \odot g = \max(0, f+g-1)$. The operations $\odot,\oplus$ are also known under the names

Łukasiewicz t-norm and *Łukasiewicz t-conorm,* respectively. The partial order $\leq$ on the MV-algebra $[0,1]$ coincides with the usual order of reals.

Example 2. More generally, the set $[0,1]^X$ of all fuzzy sets on a set X becomes an MV-algebra if the operations $\oplus$ and $\neg$ and the element 0 are defined pointwise. The corresponding lattice operations $\vee, \wedge$ are then the pointwise maximum and the pointwise minimum of two real functions, respectively.

Example 3. MV-algebras generalize Boolean algebras in the following sense. Every (Boolean) algebra of sets is an MV-algebra in which $\oplus$ coincides with $\vee$ and $\odot$ coincides with $\wedge$, where $\vee$ and $\wedge$ is the union and the intersection of two sets, respectively. The operation $\neg$ becomes the complement of a set.

We say that an MV-algebra is *semisimple* if it is (isomorphic to) an MV-algebra of continuous functions $[0,1]$ defined on some compact Hausdorff space. In particular, all the MV-algebras from Examples 1–3 are semisimple. Semisimple MV-algebras can be viewed as many-valued counterparts of algebras of sets.

Throughout the remainder we deal with real functions on an MV-algebra whose successive differences of all orders are nonnegative. This property (so-called total monotonicity) of real functions was studied already by Choquet in his foundational work about capacities [2]. Total monotonicity is the common property of belief functions studied in different settings such as a finite algebra of sets [15], any algebra of sets [16] or Borel σ-algebra of the real line [4]. We consider the difference operator with respect to the lattice operations of an MV-algebra M. This leads to the following definition. Let $b : M \to \mathbb{R}$ and put $\Delta_g b(f) = b(f) - b(f \wedge g)$, for every $f,g \in M$.

Definition 2. *A function $b : M \to \mathbb{R}$ is* totally monotone *if*

$$\Delta_{g_n} \cdots \Delta_{g_1} b(f) \geq 0, \quad \textit{for every } n \geq 1 \textit{ and every } f, g_1, \ldots, g_n \in M.$$

It is possible to show that b is totally monotone if and only if

(i) $b(f) \leq b(g)$ whenever $f \leq g$, for every $f,g \in M$,
(ii) for each $n \geq 2$ and every $f_1, \ldots, f_n \in M$:

$$b\left(\bigvee_{i=1}^{n} f_i\right) \geq \sum_{\substack{I \subseteq \{1,\ldots n\} \\ I \neq \emptyset}} (-1)^{|I|+1}\, b\left(\bigwedge_{i \in I} f_i\right).$$

3 Probabilities on MV-Algebras

A *state* on an MV-algebra M is a mapping $s : M \to [0,1]$ such that $s(1) = 1$ and $s(f \oplus g) = s(f) + s(g)$, for every $f,g \in M$ with $f \odot g = 0$. In case that M is an algebra of sets, then the notion of state agrees with that of finitely additive probability measure. Properties of states are best analyzed through their correspondence to Borel probability measures: it turns out that every state on a semisimple MV-algebra is integral—see [8] or [13].

Theorem 1. *If s is a state on a semisimple MV-algebra M, then there exists a uniquely determined Borel probability measure μ on the compact Hausdorff space X such that $s(f) = \int f \,\mathrm{d}\mu$, for each $f \in M$.*

It is possible to show by using linearity of Lebesgue integral that s is a totally monotone function on M.

The following example is crucial for the investigation of belief functions in the next section. We will introduce an MV-algebra whose elements are set functions and single out a particular class of states for later use.

Example 4. Let X be a finite nonempty set. Consider the MV-algebra $[0,1]^{\mathscr{P}(X)}$ of all functions $\mathscr{P}(X) \to [0,1]$. We will deal only with those states s on $[0,1]^{\mathscr{P}(X)}$ for which $s(1_{\{\emptyset\}}) = 0$. Theorem 1 says that each such state s corresponds to a unique finitely additive probability μ on $\mathscr{P}(\mathscr{P}(X))$ satisfying $s(q) = \sum_{A \in \mathscr{P}(X)} q(A)\mu(\{A\})$ and $\mu(\{\emptyset\}) = 0$, for every $q \in [0,1]^{\mathscr{P}(X)}$. The set $\mathbf{S}$ of all states s on $[0,1]^{\mathscr{P}(X)}$ with $s(1_{\{\emptyset\}}) = 0$ can be identified with a convex subset of the $(2^{|X|} - 1)$-dimensional Euclidean space. Since the correspondence between $\mathbf{S}$ and the set of all probabilities μ on $\mathscr{P}(\mathscr{P}(X))$ with $\mu(\{\emptyset\}) = 0$ is a one-to-one affine mapping, the convex set $\mathbf{S}$ is in fact a $(2^{|X|} - 2)$-simplex. The extreme points of $\mathbf{S}$ are in one-to-one correspondence with the nonempty subsets of X: every state s_A, $A \in \mathscr{P}_{\emptyset}(X)$ such that $s_A(q) = q(A)$, for each $q \in [0,1]^{\mathscr{P}(X)}$, is an extreme point of $\mathbf{S}$. This characterization of state space and its extreme points is a consequence of the description of state space of any MV-algebra—see [10] or [8].

4 BFs on Finite Product of Standard MV-Algebras

The domain of belief functions introduced in this section is limited to those MV-algebras $[0,1]^X$ with X finite. Each such MV-algebra is in algebraic terms just a *finite product* of standard MV-algebras.

We will repeat basic definitions of Dempster-Shafer theory of belief functions [15]. Let X be a finite nonempty set. We say that a function $\beta : \mathscr{P}(X) \to [0,1]$ is a *belief function on* $\mathscr{P}(X)$ if there is a mapping $m : \mathscr{P}(X) \to [0,1]$ with $m(\emptyset) = 0$ and $\sum_{A \in \mathscr{P}(X)} m(A) = 1$ such that $\beta(A) = \sum_{B \subseteq A} m(B)$, for every $A \in \mathscr{P}(X)$. The function m is usually called a *basic assignment.* Observe that an equivalent description of a belief function β is possible by a finitely additive probability $\mu : \mathscr{P}(\mathscr{P}(X)) \to [0,1]$ with $\mu(\{\emptyset\}) = 0$ and such that

$$\beta(A) = \mu(\{B \in \mathscr{P}(X) \mid B \subseteq A\}), \quad \text{for every } A \in \mathscr{P}(X). \tag{1}$$

Every belief function β on $\mathscr{P}(X)$ is totally monotone on the lattice $\mathscr{P}(X)$.

A point of departure for the generalization of the notion of belief function to an MV-algebra $[0,1]^X$ of all $[0,1]$-valued functions from the finite set X is the introduction of the following operator. Let the operator $\rho : [0,1]^X \to [0,1]^{\mathscr{P}(X)}$ be defined for every $f \in [0,1]^X$ as

$$\rho(f)(B) = \begin{cases} \min\{f(x) \mid x \in B\}, & B \in \mathscr{P}_\emptyset(X), \\ 1, & B = \emptyset. \end{cases}$$

Given $A, B \in \mathscr{P}(X)$, observe that $\rho(1_A)(B) = 1$ if and only if $B \subseteq A$. This means that $\rho(1_A)$ is the characteristic function of $\{B \in \mathscr{P}(X) \mid B \subseteq A\}$. Thus, we can rewrite (1) with a slight abuse of notation as

$$\beta(A) = \mu(\rho(1_A)), \quad \text{for every } A \in \mathscr{P}(X). \tag{2}$$

The preceding considerations lead naturally to the following definition of belief function.

Definition 3. *Let X be a finite nonempty set. A mapping $b : [0,1]^X \to [0,1]$ is called a* belief function on $[0,1]^X$ *if there is a state on the MV-algebra $[0,1]^{\mathscr{P}(X)}$ such that $s(1_{\{\emptyset\}}) = 0$ and $b(f) = s(\rho(f))$, for every $f \in [0,1]^X$. The state s is called a* state assignment.

We are going to generalize the integral representation theorem for states (Theorem 1) to belief functions. This requires introduction of Choquet integral [5]. Although we are integrating only the functions defined on the finite set X, we keep the integral notation to emphasize the analogy with Theorem 1 in this setting.

If f is a function $X \to [0,1]$ and β is a set function $\mathscr{P}(X) \to [0,1]$ with $\beta(\emptyset) = 0$, then *Choquet integral* of f with respect to β is defined as $\mathscr{C}\!\!\int f \,\mathrm{d}\beta = \int_0^1 \beta\big(f^{-1}([t,1])\big)\,\mathrm{d}t$. Since X is finite, the Choquet integral $\mathscr{C}\!\!\int f\,\mathrm{d}\beta$ exists and takes the form of a finite sum. Indeed, assume the set X has n elements $x_1, \ldots, x_n$ indexed in such a way that the numbers $y_i = f(x_i), i = 1, \ldots, n$ satisfy $y_1 \geq \cdots \geq y_n$. Put $y_{n+1} = 0$ and $S_i = \{x_1, \ldots, x_i\}, i = 1, \ldots, n$. Then $\mathscr{C}\!\!\int f\,\mathrm{d}\beta = \sum_{i=1}^n (y_i - y_{i+1})\beta(S_i)$.

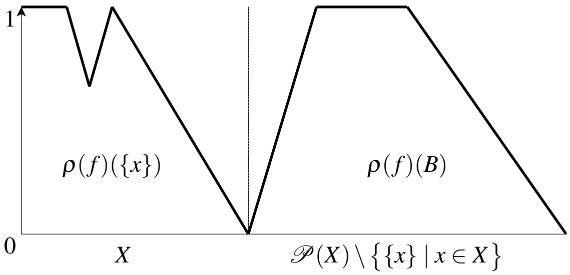


Fig. 1 Continuation of an element of the MV-algebra $[0,1]^X$ to $[0,1]^{\mathscr{P}(X)}$

Proposition 1. *For every belief function b on $[0,1]^X$ there exists a unique belief function β on $\mathscr{P}(X)$ such that $b(f) = \mathcal{C}\!\!\!\!\int f \,\mathrm{d}\beta$, $f \in [0,1]^X$.*

Proof. Let s be the state assignment on $[0,1]^{\mathscr{P}(X)}$ corresponding to b. According to Example 4 there is a unique probability μ on $\mathscr{P}(\mathscr{P}(X))$ such that $s(q) = \sum_{A \in \mathscr{P}(X)} q(A)\mu(\{A\})$ and $\mu(\{\emptyset\}) = 0$, for every $q \in [0,1]^{\mathscr{P}(X)}$. This means that b can be expressed as

$$b(f) = s(\rho(f)) = \sum_{A \in \mathscr{P}(X)} \rho(f)(A)\mu(\{A\}). \tag{3}$$

For every $A \in \mathscr{P}_{\emptyset}(X)$ and $B \in \mathscr{P}(X)$, let $\varepsilon_A(B) = 1$, whenever $A \subseteq B$, and $\varepsilon_A(B) = 0$, otherwise. Then $\rho(f)(A) = \min\{f(x) \mid x \in A\} = \mathcal{C}\!\!\!\!\int f \,\mathrm{d}\varepsilon_A$. The equality (3) together with linearity of Choquet integral with respect to the integrating set functions ε_A yield

$$b(f) = \sum_{A \in \mathscr{P}_{\emptyset}(X)} \mu(\{A\}) \mathcal{C}\!\!\!\!\int f \,\mathrm{d}\varepsilon_A = \mathcal{C}\!\!\!\!\int f \,\mathrm{d}\left(\sum_{A \in \mathscr{P}_{\emptyset}(X)} \mu(\{A\})\varepsilon_A\right).$$

It suffices to show that the function $\beta = \sum_{A \in \mathscr{P}_{\emptyset}(X)} \mu(\{A\})\varepsilon_A$ is a belief function on $\mathscr{P}(X)$. For each $B \in \mathscr{P}(X)$,

$$\beta(B) = \sum_{A \in \mathscr{P}_{\emptyset}(X)} \mu(\{A\})\varepsilon_A(B) = \sum_{A \subseteq B} \mu(\{A\}) = \mu(\{A \in \mathscr{P}(X) \mid A \subseteq B\}). \qquad \square$$

$$\begin{tikzcd}
\text{BF } b \text{ on } [0,1]^X \arrow[rr, leftrightarrow, "\mathcal{C}\!\!\!\!\int"] \arrow[d, leftrightarrow, "\rho"'] & & \text{BF } \beta \text{ on } \mathscr{P}(X) \arrow[d, leftrightarrow, "\rho"] \\
\text{State } s \text{ on } [0,1]^{\mathscr{P}(X)} \arrow[rr, leftrightarrow, "\int"] & & \text{Probability } \mu \text{ on } \mathscr{P}(\mathscr{P}(X))
\end{tikzcd}$$

Fig. 2 The relation between belief functions (BF), states, and probabilities

The derived Choquet integral representation coincides with the definition of a belief function on "formulas" of Łukasiewicz logic proposed in [9]. Due to Proposition 1 the properties of belief functions on $[0,1]^X$ are completely determined by the properties of Choquet integral. These are the most important among them—see [5].

Proposition 2. *Let b be a belief function on $[0,1]^X$. Then b is totally monotone and for every $f, g \in [0,1]^X$:*

(i) $b(0) = 0, b(1) = 1$
(ii) if $f \odot g = 0$, *then* $b(f \oplus g) \geq b(f) + b(g)$
(iii) $b(f) + b(\neg f) \leq 1$

(iv) b *is a state if the state assignment* s *satisfies* $s(q) = 0$ *for each* $q \in [0,1]^{\mathscr{P}(X)}$ *such that* $q(A) > 0$ *for some* $A \in \mathscr{P}(X)$ *with* $|A| > 1$
(v) $b(f) = \min\{s(f) \mid s \text{ state on } [0,1]^X \text{ with } s \geq b\}$

The property (*ii*) is so-called *superadditivity*. The condition (*iv*) is a generalization of the analogous fact about belief functions on $\mathscr{P}(X)$: a belief function β on $\mathscr{P}$ is a probability iff the corresponding basic assignment satisfies $m(A) = 0$ for each $A \in \mathscr{P}(X)$ with $|A| > 1$. The last property (*v*) means that b is a lower probability in the sense of [6, Definition 4.1], which enables interpreting the belief function b in the game-theoretical framework based on a notion of coherence.

The geometrical structure of the set of all belief functions on $[0,1]^X$ is fully determined by the associated simplex of state assignments on $[0,1]^{\mathscr{P}(X)}$. For each $A \in \mathscr{P}_\emptyset(X)$, a belief function $b_A(f) = \min\{f(x) \mid x \in A\}, f \in [0,1]^X$ corresponds to the state assignment s_A (see Example 4). Consequently, we obtain the following characterization of the set of all belief functions.

Proposition 3. *The set of all belief functions on* $[0,1]^X$ *is a* $(2^{|X|}-2)$*-simplex whose set of extreme points is* $\{b_A \mid A \in \mathscr{P}_\emptyset(X)\}$.

Observe that every b_A preserves finite minima since for every $f, g \in [0,1]^X$ we have $b_A(f \wedge g) = b_A(f) \wedge b_A(g)$. In general, it can be shown that each minimum-preserving function $b : [0,1]^X \to [0,1]$ with $b(0) = 0, b(1) = 1$ is a belief function. Such functions are termed *necessity measures* and they were recently investigated on formulas of finitely-valued Łukasiewicz logic in [7].

Acknowledgements. The work of the author was supported by the grant GA ČR 201/09/1891 and by the grant No.1M0572 of the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Butnariu, D., Klement, E.P.: Triangular Norm Based Measures and Games with Fuzzy Coalitions. Kluwer Academic Publishers, Dordrecht (1993)
2. Choquet, G.: Theory of capacities. Ann. Inst. Fourier, Grenoble 5, 131–295 (1953–1954) (1955)
3. Cignoli, R.L.O., D'Ottaviano, I.M.L., Mundici, D.: Algebraic foundations of many-valued reasoning. Trends in Logic—Studia Logica Library, vol. 7. Kluwer Academic Publishers, Dordrecht (2000)
4. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Statist. 38, 325–339 (1967)
5. Denneberg, D.: Non-additive measure and integral, Theory and Decision Library. Series B: Mathematical and Statistical Methods, vol. 27. Kluwer Academic Publishers, Dordrecht (1994)
6. Fedel, M., Keimel, K., Montagna, F., Roth, W.: Imprecise probabilities, bets and functional analytic methods in Łukasiewicz logic (submitted for publication, 2010)

7. Flaminio, T., Godo, L., Marchioni, E.: On the logical formalization of possibilistic counterparts of states over n-valued Łukasiewicz events. J. Logic Comput. (2010), doi:10.1093/logcom/exp012
8. Kroupa, T.: Every state on semisimple MV-algebra is integral. Fuzzy Sets Syst. 157(20), 2771–2782 (2006)
9. Kroupa, T.: Belief functions on formulas in Łukasiewicz logic. In: Kroupa, T., Vejnarova, J. (eds.) Proceedings of the 8th Workshop on Uncertainty Processing, WUPES 2009, Liblice, Czech Republic (2009)
10. Mundici, D.: Averaging the truth-value in Łukasiewicz logic. Studia Logica 55(1), 113–127 (1995)
11. Mundici, D.: Bookmaking over infinite-valued events. Internat. J. Approx. Reason. 43(3), 223–240 (2006)
12. Navara, M.: Triangular norms and measures of fuzzy sets. In: Klement, E., Mesiar, R. (eds.) Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms, pp. 345–390. Elsevier, Amsterdam (2005)
13. Panti, G.: Invariant measures in free MV-algebras. Comm. Algebra 36(8), 2849–2861 (2008)
14. Riečan, B., Mundici, D.: Probability on MV-algebras. In: Handbook of Measure Theory, vol. I, II, pp. 869–909. North-Holland, Amsterdam (2002)
15. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
16. Shafer, G.: Allocations of probability. Ann. Probab. 7(5), 827–839 (1979)

Soft Methods in Trend Detection

Piotr Ładyżyński and Przemysław Grzegorzewski

Abstract. The problem of trend detection in fuzzy time series is considered. Two fuzzy tests are suggested and their basic properties are examined. Moreover, the general problem of fuzziness in statistical data which might be either inherent or imposed is discussed.

Keywords: Fuzzy numbers, Fuzzy sets, Granular computing, Time series, Trend.

1 Introduction

Trend detection is a crucial problem in many real-life problems. It appears whenever a time series, i.e. a sequence of observations measured at successive time moments, is considered. Investors, e.g., try to predict trend in prices recorded on a stock market since information on its strength and direction (i.e. whether it is increasing or decreasing) is fundamental for their financial decisions. Trend detection is important, of course, not only in finance but for any forecasting in economy, industry, social sciences, medicine, etc.

It seems that the notion of trend is quite obvious but in fact there is no unique definition of that concept. Moreover, available data may be imprecise or vague, especially if they come from humans and are expressed in a natural language. However, even if the data are just real numbers (like prices) their adequate interpretation leading to reliable forecasts may be neither trivial not straightforward. Thus, although broad statistical literature devoted to

Piotr Ładyżyński and Przemysław Grzegorzewski
Faculty of Mathematics and Information Science, Warsaw University of Technology,
00-661 Warsaw, Poland

Przemysław Grzegorzewski
Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland
e-mail: pgrzeg@ibspan.waw.pl

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 395–402.
springerlink.com

time series analysis is easily accessible, classical methods for trend detection are not satisfactory in numerous practical situations. This is the reason why soft methods applied in the decision support based on time series seem to be desirable.

The paper is organized as follows: In Sect. 2 we try do define trend and discuss some classical method for trend detection. In Sect. 3 we consider the problem of fuzziness which either appears in time series or is intentionally introduced to receive granule information. Then in Sect. 4 we propose some trend detection tests for fuzzy data and discuss their properties.

2 How to Define and Detect Trend?

Nearly everybody is intuitively familiar with the concept of trend. However, mere intuition might be insufficient for obtaining a proper time series analysis. Unfortunately, there is no unique definition of trend. In the Encyclopedia of Statistical Sciences [2] we can find such definitions: "The trend corresponds to sustained and systematic variations over a long period of time" or "The trend is generally thought of as a smooth and slow movement over a long term". The same Encyclopedia states that: "the identification of trend has always posed a serious statistical problem. The problem is not one of mathematical or analytical complexity but of conceptual complexity" (see [2]).

Despite general problems with the precise meaning of trend and its identification several statistical procedures for trend detection have been proposed (e.g. see [1]). Unfortunately, they often require strong assumptions. Tests based on linear regression can identify only linear trends; t-tests require i.i.d. observations from the normal distribution, etc. Such situation is typical e.g. in Statistical Process Control where we usually assume that the process is normally distributed (see, e.g., [5, 9]). To avoid specific assumptions on the underlying distribution nonparametric tests are often used. Many procedures developed for detecting trends are based on runs like the test based on the length of the longest run, test based on the shorter of the longest run, test based on the longer of the longest run, test based on the number of runs or test based on the total number of runs of signs (see [3, 4]). Other popular distribution-free tests for detecting trend are the Spearman rank correlation test and the Mann-Kendall test (see [8]).

Another approach to trend detection based on the averaging methods usually used for time series smoothing and prediction was proposed in [7]. Let $(X_i)_{i=1}^{n+1}$ denote a sequence of observations and let

$$m_i = \frac{X_{i+1} - X_i}{X_i}, \tag{1}$$

denote the rate of returns $(t = 1, \ldots, n)$. A test statistic of the first test is given by

$$T^{(arith)} = \frac{\sum_{i=1}^{n} m_i - nm}{\sigma\sqrt{n}}, \tag{2}$$

where $m = Em_i$, $i = 2,\ldots,n$. If the rate of returns m_i are i.i.d. random variables (which is a common assumption for the Black-Scholes market) and there is no trend in data then $T^{(aryt)}$ is asymptotically normal.

We can also use the exponential smoothing for trend detection, i.e.

$$m_{i+1}^{(exp)} = \lambda m_{i+1} + (1-\lambda)m_i^{(exp)}, \tag{3}$$

where $\lambda \in [0,1]$ is a smoothing parameter. The greater values of λ the lesser weights of the observations from the past and greater weights of the recent observations. A test statistic of the second test is then defined as follows:

$$T^{(exp)} = m_n^{(exp)}. \tag{4}$$

The distribution of statistics (4) under null hypothesis of no trend is calculated numerically using the Monte-Carlo method.

3 Time Series and Fuzziness

All statistical methods mentioned above were constructed for the real time series, i.e. observed variables that assume precise real values. However, in many real-life situations observations are not precise. There are many reasons for vagueness in data. Suppose we would like to know whether given star becomes more active, or whether the pollution in given area increases, or the water level of the river raises significantly. In all such situations imprecise measuring instruments and environmental interferences causes that measurements also become imprecise. If we consider, e.g., the possible green mass growth at given afforested territory then our observations are imprecise by definition since we do not have any precise instrument and all we make very rough estimates. Such situation is also typical when the data come directly from respondents/users, i.e. from persons that express their opinions or perceptions using natural language which abounds in vagueness. In all such cases fuzzy set theory might be useful both for modelling and processing imprecise data.

However, sometimes too precise data may also be, paradoxically, not very convenient to grasp the heart of the matter. Consider a following example.

Example 1. We have to program an automatic trading system which would buy and sell shares on a stock market without any human supervision. Suppose that signals generated by the system are connected with trends on the market: we would like to buy in the presence of the increasing trend and conversely, we would like to sell our shares when the trend of the prices becomes decreasing. Of course, the operation rules of a true system are much more complicated and such crude requirement may be considered only as one of many criteria for opening and closing positions.

The majority of the long and middle term trading methods utilize the end of the day price for calculations. One can use these very prices to identify possible trend in data but then he will loose the structure and the essential behavior of prices of analyzed share during the whole day. Suppose that we register a share price every 15 minutes during the trading session. Often some moments in the session may be more important for predicting the future price than the others. For example, the prices on the Warsaw Stock Exchange can drastically change after 3:30 p.m. when stock exchanges in US open and polish investors are affected by the situation on US market. Let $(Y_i)_{i=1}^{31}$ represent prices of a share collected during day session from 9:00 a.m. to 4:30 p.m. every 15 minutes. Suppose that experts claim that prices collected after 3:30 p.m. are three times more valuable for price prediction than observations recorded in other moments. However, experts suggested that the next day price also depends on time structure of prices from the last day. For example, if one day between 9:00 a.m. and 12:00 a.m. our share costed 100 USD, between 12:00 a.m. and 1:00 p.m. - 107 USD, while between 1:00 p.m. and 4:30 p.m. - 100 USD, the next day price will be rather 100 USD than 107 USD. Suppose we want to combine all the prices Y_i recorded during this day into a single fuzzy set which might be considered as a kind of information granule containing not only observational data but also some expert knowledge and other a priori data. To perform this transformation we may utilize the following algorithm:

1. Duplicate two times last five observations $(Y_i)_{i=27}^{31}$ - the prices registered from 3:30 to 4:30 p.m. Now we have 41 observations, where $(Y_i)_{i=1}^{41}$. $Y_{27} = Y_{32} = Y_{37}$, $Y_{28} = Y_{33} = Y_{38}$,..., $Y_{21} = Y_{36} = Y_{41}$.
2. A membership function $\mu : \mathbb{R} \to [0,1]$ of a fuzzy granule corresponding to a single day is given by:

$$\mu(y) = \frac{1}{\sup_{y \in \mathbb{R}} \hat{f}(y)} \hat{f}(y), \tag{5}$$

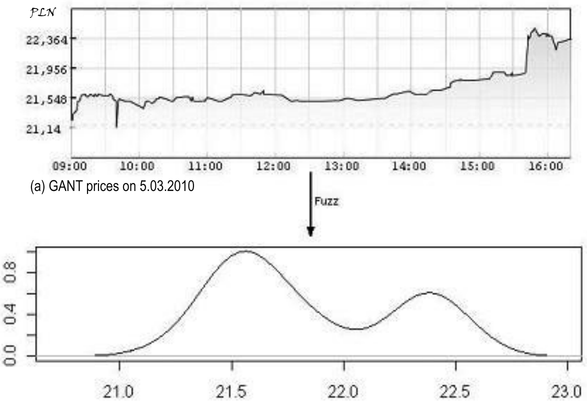
   where $\hat{f}(y)$ is estimated from $(Y_i)_{i=1}^{n}$, $n = 41$,

$$\hat{f}(y) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_n}\right), \tag{6}$$

   and where K is a positive kernel (e.g. the gaussian density function).

The suggested transformation of daily prices of GANT (Warsaw Stock Exchange) share into fuzzy price is shown in Fig. 1. Of course this algorithm is only an example and everybody can construct his own algorithm according to his experiences and knowledge in stock trading. Moreover, if we need not arbitrary fuzzy numbers but fuzzy objects with some regular membership function, e.g. fuzzy numbers or particular type of fuzzy numbers, we may enrich the algorithm adding all necessary requirements.

Whatever is the source of vagueness in data we may use fuzzy sets for modelling imprecise time series. And despite of the nature of imprecision we need statistical tools for time series analysis and especially for trend detection. In the next section we suggest such statistical procedures for detecting trend in fuzzy data and discuss some basic properties of these utilities.



(a) GANT prices on 5.03.2010

(b) Membership function of fuzzy price

Fig. 1 Transformation of the daily prices into a fuzzy granule.

4 Testing Trend with Fuzzy Data

Suppose that a random experiment is described as usual by a probability space $(\boldsymbol{\Omega}, \mathbb{A}, \boldsymbol{P})$, where $\boldsymbol{\Omega}$ is a set of all possible outcomes of the experiment, $\mathbb{A}$ is a $\sigma-$algebra of subsets of $\boldsymbol{\Omega}$ (the set of all possible events) and $\boldsymbol{P}$ is a probability measure. A fuzzy random variable (f.r.v.) is a mapping $\mathscr{X}$: $\boldsymbol{\Omega} \to \mathbb{FN}(\mathbb{R})$, where $\mathbb{FN}(\mathbb{R})$ is a family of fuzzy numbers, which satisfies the following properties:

i) $\{\mathscr{X}(\alpha, \omega) : \alpha \in (0,1]\}$ is a set representation of $\mathscr{X}(\omega)$ for all $\omega \in \boldsymbol{\Omega}$,
ii) for each $\alpha \in (0,1]$ both $\mathscr{X}_\alpha^L = \mathscr{X}_\alpha^L(\omega) = \inf \mathscr{X}_\alpha(\omega)$ and $\mathscr{X}_\alpha^U = \mathscr{X}_\alpha^U(\omega) = \sup \mathscr{X}_\alpha(\omega)$, are usual real-valued random variables on $(\boldsymbol{\Omega}, \mathbb{A}, \boldsymbol{P})$.

Thus a f.r.v. $\mathscr{X}$ is considered as a perception of an unknown usual random variable $X : \boldsymbol{\Omega} \to \mathbb{R}$, called an *original* of $\mathscr{X}$ (if only vague data are available it is impossible to show which of the possible originals is the true one). Similarly n-dimensional fuzzy random sample $\mathscr{X}_1, \ldots, \mathscr{X}_n$ may be treated as a fuzzy perception of the usual random sample $X_1, \ldots, X_n$ (see [6]).

Let U and V denote two fuzzy numbers with α-cuts: $U_\alpha = \{x \in \mathbb{R} : \mu_U(x) \geq \alpha\} = \left[U_\alpha^L, U_\alpha^U\right]$, $V_\alpha = \{x \in \mathbb{R} : \mu_V(x) \geq \alpha\} = \left[V_\alpha^L, V_\alpha^U\right]$, respectively. Below we use a following distance between fuzzy numbers U and V:

$$D_K^2(U,V) = \int_0^1 \left(U_\alpha^U - V_\alpha^U\right)^2 d\alpha + \int_0^1 \left(U_\alpha^L - V_\alpha^L\right)^2 d\alpha. \qquad (7)$$

Suppose we consider a fuzzy sample $(\mathscr{X}_i)_{i=1}^{n+1}$ representing a fuzzy time series under study. To generalize methods for a trend detection described above into a fuzzy context let us firstly compute a fuzzy rate of return

$$\widetilde{m}_i = \frac{\mathscr{X}_{i+1} - \mathscr{X}_i}{\mathscr{X}_i}, \tag{8}$$

where $i = 1, \ldots n$, and division is performed according to fuzzy arithmetic. Let $\overline{m}$ denote a fuzzy average, i.e. $\widetilde{M} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{m}_i$. Then we estimate a variance of these fuzzy rates of return by

$$S^2_{\widetilde{m}} = \tfrac{1}{n-1}\sum_{i=1}^{n} D^2_K(\widetilde{m}_i, \widetilde{M}). \tag{9}$$

By $\tilde{0}$ we denote a fuzzy zero, i.e. a fuzzy number with the following α-cuts:

$$\widetilde{0}_\alpha = [0 - S_{\widetilde{m}}(1-\alpha), 0 + S_{\widetilde{m}}(1-\alpha)]. \tag{10}$$

If the arithmetic mean of the fuzzy rate of return is close to zero, trend does not occur. Thus we compare obtained fuzzy arithmetic mean with a fuzzy zero using the appropriate metric. Therefore, we will test the null hypothesis H_0 : *no trend*, against alternative H_1 : *trend exists*. Then the test statistics is given by:

$$T^{(arith)}_{fuzz} = D_K\left(\widetilde{M}, \widetilde{0}\right). \tag{11}$$

Assuming the significance level δ we reject H_0 if statistics (11) belongs into a critical region

$$\mathscr{K}_\alpha = (-\infty, c_{\frac{\delta}{2}}] \cup [c_{1-\frac{\delta}{2}}, +\infty), \tag{12}$$

where $c_{\frac{\delta}{2}}$ and $c_{1-\frac{\delta}{2}}$ are the quantiles of order $\frac{\delta}{2}$ and $1 - \frac{\delta}{2}$, respectively, of statistic (11) under H_0, which are calculated by the Monte-Carlo method.

In a similar way we may generalize the test based on exponential smoothing for fuzzy data. Now, assuming the smoothing parameter $\lambda \in [0,1]$ we obtain a following test statistic:

$$T^{(exp)}_{fuzz} = D_K\left(\widetilde{m}^{(exp)}_n, \widetilde{0}\right), \tag{13}$$

where

$$\widetilde{m}^{(exp)}_n = \lambda \widetilde{m}_n + (1-\lambda)\widetilde{m}^{(exp)}_{n-1}. \tag{14}$$

Here again the critical region is obtained by the Monte-Carlo method. Please note, that now to get a fuzzy zero (10) we substitute a fuzzy mean $\widetilde{M}$ in the fuzzy variance (9) by (14).

To examine the power of the two fuzzy test proposed above we performed a simulation study. Assuming the Black-Scholes model we generated classical trajectories $X_t = X_0 exp\left(\left(r - \frac{\sigma^2}{2}\right)t + \sigma W_t\right)$, where $r \in \mathbb{R}, \sigma > 0$ which were then fuzzified. Situation with $r = 0$ corresponds to trajectory without trend. Hence changing a value of r we may model trajectories with arbitrary trend. Some empirical results showing the comparison of the power for our two tests are given in Fig. 2.

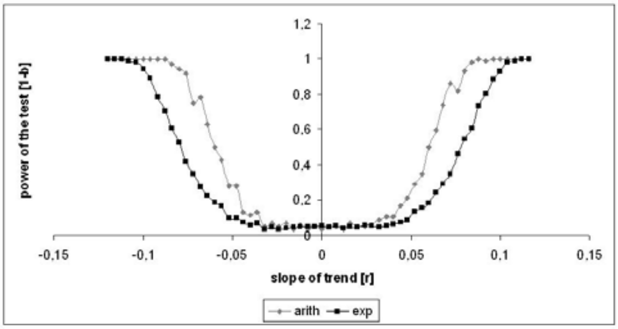


Fig. 2 The power of the fuzzy arithmetic mean test vs. fuzzy exponential test.

One may conclude that the power of the fuzzy arithmetic mean test dominates the fuzzy exponential test. This result is not surprising since the same situation happens for crisp data (see [7]).

During our simulation experiment we have also examined other distances than (7). In particular, we have considered some weighted metrics like

$$D^2_{K'}(U,V) = \int_0^1 \alpha^3 \left(U^U_\alpha - V^U_\alpha\right)^2 d\alpha + \int_0^1 \alpha^3 \left(U^L_\alpha - V^L_\alpha\right)^2 d\alpha. \tag{15}$$

For this very metric we have observed an interesting result. Namely, the comparison of the power of the fuzzy arithmetic mean test with the classical arithmetic mean test performed on the defuzzified data (by the popular center of gravity method) showed that sometimes it is much better to process fuzzy data than to defuzzify them too early (see Fig.3).

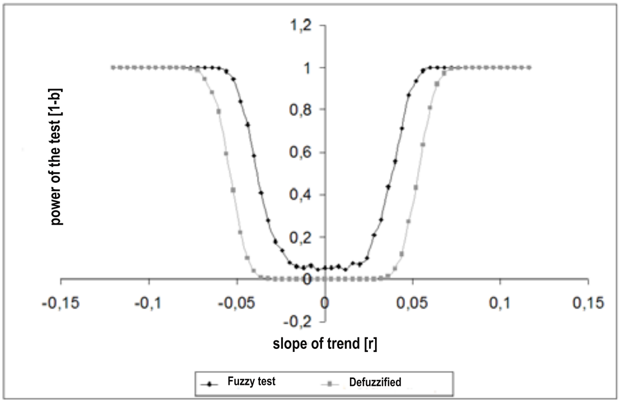


Fig. 3 The comparison of fuzzy test and appropriate classical test applied on defuzzified data.

5 Conclusions

In this paper we have proposed two trend detection tests for fuzzy data. Simulation study shows that the so called fuzzy arithmetic mean test is more powerful than the second one. Of course, the suggested solution is just the example how we could handle with imprecise time series but it does not determine the hole problem and many questions still remain open. Although trend detection is a crucial point of any time series analysis some other problems, like analysis of the seasonal components and reliable forecasting based on fuzzy data is also of interest.

References

1. Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco (1970)
2. Dagum, C., Dagum, E.B.: Trend. In: Kotz, S., Johnson, N. (eds.) Encyclopedia of Statistical Sciences. John Wiley & Sons, New York (1985)
3. Domański, C.: Theoretical foundations of nonparametric tests and their application to economic and social sciences. Acta Universitatis Lodziensis (1986)
4. Gibbons, J.D.: Nonparametric Statistical Inference. McGraw-Hill, New York (1971)
5. Grzegorzewski, P.: Shewhart control charts and the problem of the detection of trend. In: von Collani, E., Göb, R., Kiesmüller, G. (eds.) Proceedings of the 4th Würzburg-Umea Conference in Statistics, Würzburg, Germany, pp. 175–185 (1996)
6. Kruse, R., Meyer, K.D.: Statistics with Vague Data. Riedel Publishing Company, Dordrecht (1987)
7. Ładyżyński, P.: Trend detection tests. Masters Thesis, Warsaw University of Technology (2009)
8. Mann, H.: Non-parametric tests against trend. Econometrica 13, 245–259 (1945)
9. Western Electric Statistical Quality Control Handbook. Western Electric Corporation, Indianapolis, Ind. (1956)

Linguistic Decision Trees for Fusing Tidal Surge Forecasting Models

Jonathan Lawry and Hongmei He

Abstract. The use of linguistic decision trees as represented within the label semantics framework, is proposed for the fusion of multiple forecasting models. The learning algorithm LID3 is applied to infer a decision tree with branches representing a set of rules each identifying a probability distribution on the available models and where the constraints in each rule are generated from fuzzy labels describing the relevant input attributes. The resulting aggregated forecast for a given vector of input attributes $\mathbf{x}$, is then taken to be the mean value of the forecasts from each model relative to a probability distribution on models conditional on $\mathbf{x}$ as determined from the linguistic decision tree. The potential of this approach is then investigated through its application to the fusion of tidal surge forecasting models for the east coast of the UK.

1 Introduction

In environmental modelling there are often multiple forecasting models available for a given problem. Typically the performance of these models will vary as environmental conditions change, resulting in different models giving best predictive accuracy for different input values. A number of different model fusion strategies have been studied in the literature. For example, Winkler and Makridakis [9] proposed a weighted average combination of different time series forecasting models where the weights were based on the relative predictive performance of each model on a training data set. While this approach can be effective, the weights are not dependent on input values and consequently cannot capture those cases where different models perform best given

Jonathan Lawry and Hongmei He
Department of Engineering Mathematics, University of Bristol,
Bristol, BS8 1TR, United Kingdom
e-mail: j.lawry@bris.ac.uk

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 403–410.
springerlink.com

values in different regions of the input space. Abrahart and See [1] carried out a study of techniques for model fusion, including neural networks and fuzzy methods, applied to river level and discharge forecasting on the River Ouse and Upper River Wye catchments. Inputs to the fusion algorithms include current discharge or flow and current model accuracy for each model. Results suggested that a combined probabilistic fuzzy model performed well and indeed gave the best fusion results on the River Ouse data.

In this paper we propose the use of linguistic decision trees based on label semantics [7] for model fusion. These tree structured rules integrate fuzzy description labels and probabilistic uncertainty in a single coherent framework (i.e label semantics [4], [5]). As such they naturally generate a weighted probabilistic aggregation of forecasting models, while also providing a transparent rule-base according to which decisions can be traced and explained. The potential of this approach will be explored by application to the fusion of tidal surge forecasting models for the east coast of the UK. Accurate forecasting of surges in this area is vital in order to provide advanced warning of high water levels in the Thames, so that the Thames Barrier can be closed to protect London from flooding.

2 Linguistic Decision Trees

Linguistic decision trees (LDT) [7] are a tree-structured classification model based on label semantics. The information heuristics used for building the tree are modified from Quinlan's ID3 [8] in accordance with the label semantics framework. The nodes of a LDT are linguistic descriptions of variables and leaves are sets of appropriate labels. In such decision trees, the probability estimates for branches across the whole tree is used for classification, instead of the majority class of the single branch into which the examples fall. Linguistic expressions such as small, medium and large are used to learn from data and build a linguistic decision tree guided by information based heuristics. For each branch, instead of labeling it with a certain class (such as positive or negative in binary classification) the probability of members of this branch belonging to a particular class is evaluated from a given training dataset. Unlabeled data is then classified by using probability estimation of classes across the whole decision tree.

2.1 Brief Overview of Label Semantics

Label semantics [4] is a methodology for using linguistic expressions or fuzzy labels to describe (typically numerical) values. Label semantics proposes two fundamental and inter-related measures of the appropriateness of labels as descriptions of an object or value. We begin by identifying a finite set of basic labels $LA = \{L_1, \ldots, L_n\}$ for describing elements from the underlying universe Ω. These are building blocks for more complex compound expressions which can then also be used as descriptors as follows. A countably infinite set of expressions LE can be generated through recursive applications of logical connectives to the basic labels in LA. The measure of appropriateness of an expression $\theta \in LE$ as a description of instance x is denoted by $\mu_\theta(x)$ and quantifies an agent's subjective probability that θ can be appropriately used to describe x. From an alternative perspective, when faced with describing instance x, an agent may consider each label in LA and attempt to identify the subset of labels that are appropriate to use. Let this complete set of appropriate labels for x be denote by $\mathscr{D}_x$. Uncertainty concerning $\mathscr{D}_x$ is then represented by a mass function m_x defined on sets of labels.

Definition 1. *Mass Function on Labels*
$\forall x \in \Omega$ *a mass function on labels is a function* $m_x : 2^{LA} \rightarrow [0,1]$ *such that* $\sum_{S \subseteq LA} m_x(S) = 1$

Note that there is no requirement for the mass associated with the empty set to be zero. Instead, $m_x(\emptyset)$ quantifies the agent's belief that none of the labels are appropriate to describe x.

Appropriateness measures for labels are then related to mass functions according to the rule that $\mu_{L_i}(x)$, denoting the appropriateness of L_i to describe x, corresponds to the sum of m_x over those subsets of labels containing L_i.

Definition 2. *Appropriateness of Labels*

$$\forall x \in \Omega, \forall L_i \in LA,\ \mu_{L_i}(x) = \sum_{F \subseteq LA : L_i \in F} m_x(F)$$

In many cases we assume that for any $x \in \Omega$ the subsets of labels for which m_x is non-zero forms a nested sequence. This is referred to as the *consonance assumption* and is particularly justifiable in cases where the appropriateness of labels is judged based on a single shared criterion. See [4] or [5] for a more detailed justification of this assumption. Making the consonance assumption means that m_x can be determined directly from the values for $\mu_{L_i}(x)$ if these are known for the basic labels $L_i \in LA$. Specifically, given appropriateness measures $\mu_{L_1}(x), \ldots, \mu_{L_n}(x)$ ordered such that $\mu_{L_i}(x) \geq \mu_{L_{i+1}}(x)$ for $i = 1, \ldots, n-1$ then assuming consonance the mass function m_x is given by:

$$\begin{gathered} m_x(\{L_1, \ldots, L_n\}) = \mu_{L_n}(x) \\ m_x(\{L_1, \ldots, L_i\}) = \mu_{L_i}(x) - \mu_{L_{i+1}}(x) :\ i = 1, \ldots, n-1 \\ \text{and } m_x(\emptyset) = 1 - \mu_{L_1}(x) \end{gathered}$$

There are clear links between label semantics and other uncertainty theories including probability theory, Dempster-Shafer theory, and Possibility Theory. More details of these connections can be found in [5].

2.2 The LID3 Algorithm

Consider a classification problem where examples are to be classified on the basis of attributes $\mathbf{x} = \langle x_1, \ldots, x_m \rangle$ as one of k classes $C_1, \ldots, C_k$. Here we assume that $x_i \in \Omega_i = [a_i, b_i]$ where $a_i < b_i \in \mathbb{R}$ and that $LA_i = \{L_{i,1}, \ldots, L_{i,n_i}\}$ is a predefined set of labels for describing the elements of Ω_i. Furthermore, we assume that $L_{i,j}$ is defined by an appropriateness measure $\mu_{L_{i,j}} : \Omega_i \to [0,1]$. In this context a linguistic decision tree is a probabilistic tree structured classifier with nodes corresponding to the description sets $\mathscr{D}_{x_i}$ for attributes $x_1, \ldots, x_m$. The branches of the tree are then generated by the possible values of $\mathscr{D}_{x_i}$ corresponding to those subsets of LA_i which have non-zero mass function value for some element of Ω_i. More formally, the possible values of node $\mathscr{D}_{x_i}$ are the elements of the set $\mathscr{F}_i$ defined by:

$$\mathscr{F}_i = \{F \subseteq LA_i : \exists x \in \Omega_i,\ m_x(F) > 0\}$$

Consequently a branch B of a linguistic decision tree is a conjunction of the form:

$$(\mathscr{D}_{x_{i_1}} = F_{i_1}) \wedge (\mathscr{D}_{x_{i_2}} = F_{i_2}) \wedge \ldots \wedge (\mathscr{D}_{x_{i_d}} = F_{i_d})$$

where $1 \leq d \leq m$ $i_j \neq i_r$ for $j \neq r$ and $F_{i_j} \in \mathscr{F}_{i_j}$ for $j = 1, \ldots, d$. Associated with each branch B there is a conditional probability distribution on the classes $P(C_1|B), \ldots, P(C_k|B)$. Then given an instantiation of the attribute vector $\mathbf{x}$ Jeffrey's rule is applied across the branches B of the decision tree, to obtain a probability distribution on classes conditional on $\mathbf{x}$ as follows:

$$P(C_l|\mathbf{x}) = \sum_B P(C_l|B) P(B|\mathbf{x}) \text{ where } P(B|\mathbf{x}) = \prod_{j=1}^{d} m_{x_{i_j}}(F_{i_j})$$

The LID3 algorithm is an extension of the well-known ID3 algorithm introduced by Quinlan [8]. LID3 infers a linguistic decision tree from a training database DB of examples each corresponding to a vector of attribute values together with their associated class:

$$DB = \{\left\langle \mathbf{x}^{(r)}, C^{(r)} \right\rangle : r = 1, \ldots, N\}$$

Using this database LID3 applies the standard ID3 entropy search heuristic to identify the most informative attributes but where the relevant branch and class probabilities are determined by:

$$P(B) = \frac{1}{N} \sum_{r=1}^{N} P(B|\mathbf{x}^{(r)}) \text{ and } P(C_l|B) = \frac{\sum_{r: C^{(r)} = C_l} P(B|\mathbf{x}^{(r)})}{\sum_{r=1}^{N} P(B|\mathbf{x}^{(r)})}$$

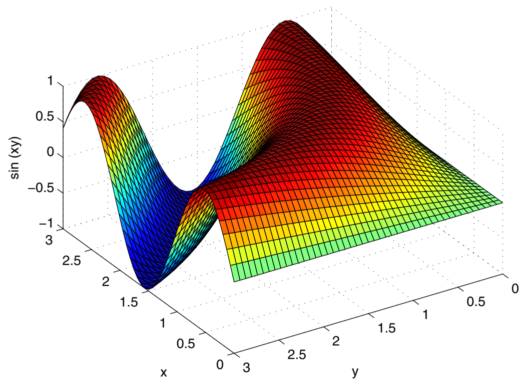


Fig. 1 Surface plot of the function $z = sin(x \times y)$ across a regular grid

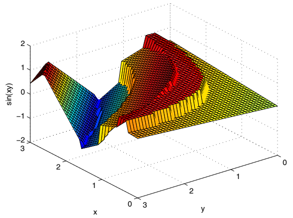


Fig. 2 Piecewise combination of $f_1, \cdots, f_6$

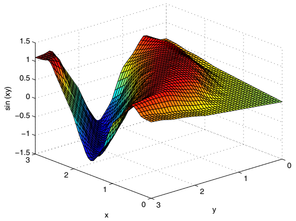


Fig. 3 Fused function obtained from the linguistic decision tree

3 Model Fusion

Consider a regression problem where the aim is to identify a functional mapping from $\Omega_1 \times \ldots \times \Omega_m$ into Ω_{m+1}[1] which is consistent with data relating to an underlying functional mapping $g : \Omega_1 \times \ldots \times \Omega_m \rightarrow \Omega_{m+1}$. Let DBR denote a regression training database of the form:

$$DBR = \{(\mathbf{x}^{(r)}, x^{(r)}_{m+1}) : r = 1, \ldots, N\} \text{ where } x^{(r)}_{m+1} = g(\mathbf{x}^{(r)})$$

Now suppose we have k functions $f_l : \Omega_1 \times \ldots \times \Omega_m \rightarrow \Omega_{m+1} : l = 1, \ldots k$ approximating g. Here we investigate the possibility of using linguistic decision trees to fuse (or aggregate) these functions. To do so we generate a classification database DBC from DBR so that:

$$DBC = \{(\mathbf{x}^{(r)}, f^{(r)}) : r = 1, \ldots, N\} \text{ where}$$
$$f^{(r)} = \arg\min\{|f(\mathbf{x}^{(r)}) - x^{(r)}_{m+1}| : f \in \{f_1, \ldots, f_k\}\}$$

[1] Here we assume $\Omega_i = [a_i, b_i]$ where $a_i < b_i \in \mathbb{R}$ for $i = 1, \ldots, m+1$.

Applying LID3 to *DBC* then generates a linguistic decision tree from which we can determine a conditional probability $P(f_l|\mathbf{x})$ for $l = 1, \ldots, k$ and for any instantiation of the attribute vector $\mathbf{x}$. We can then generate an aggregated approximation of g by taking the mean value of $f_l(\mathbf{x})$ with respect to this distribution so that:

$$\hat{f}(\mathbf{x}) = \sum_{l=1}^{k} f_l(\mathbf{x}) P(f_l|\mathbf{x})$$

Example 1. In this example a database of 529 points was generated describing a surface defined according to the equation $z = \sin(x \times y)$ where $x, y \in [0, 3]$ as shown in figure 1. This database was then partitioned into 6 clusters $C_1, \ldots, C_6$ according to $0 \leq xy < \frac{\pi}{4}$, $\frac{\pi}{4} \leq xy < \frac{\pi}{2}$, $\frac{\pi}{2} \leq xy < \pi$, $\pi \leq xy < \frac{3\pi}{2}$, $\frac{3\pi}{2} \leq xy < \frac{5\pi}{2}$ and $\frac{5\pi}{2} \leq xy < \frac{7\pi}{2}$ respectively. Least squares linear regression was then applied to C_i to generate linear function f_i for $i = 1, \ldots, 6$. Using these different models a classification database was generated as above and LID3 was applied to infer a linguistic decision tree. Figure 1 shows the different functions f_i applied in a piecewise manner to each set C_i, while figure 2 shows the resulting fused approximation obtained from the linguistic decision tree.

4 Tidal Surge Forecasting

Decisions regarding the closure of the Thames Barrier in London are based, in part, on forecasts of sea level in the Thames Estuary. Sea level is measured by water height at some fixed point and is decomposed into two main components:

- The astronomical tide, classified as the periodic movements of the ocean with a coherent amplitude, that is dependent on the astronomical effects of the sun, moon, and their position with respect to the earth.
- The meteorological effects produced by weather conditions.

In our study we used data from tidal gauges for the years 1997-2001. These gauges form part of the UK National Tidal Gauge Network which records sea level at 44 different locations around the UK coast. The data has been preprocessed by the British Oceanographic Data Centre (BODC) in order to, for example, remove null or improbable values. Figure 4 shows a time series plot for the Sheerness gauge between 2000 and 2001 with data recorded at 30 minute intervals. In addition, we have generated residual data by subtracting the tidal component, which can be accurately estimated using harmonic tidal prediction methods [3]. The data was then divided into a training set (years 1997 to 1999) and a test set (years 2000 to 2001). In this paper we focus on the problem of forecasting the residual at Sheerness 8 hours ahead, denoted y_{t+8} based on current and previous residual values further up the coast at Whitby, denoted x_t and $x_{t-0.5}$. Consequently we simplifying the forecasting problem by assuming an underlying function g according to which $y_t = g(x_t, x_{t-0.5})$.

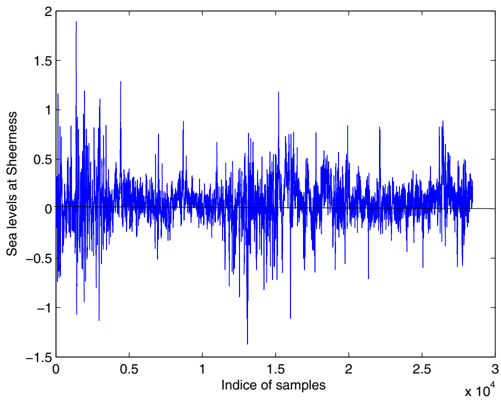
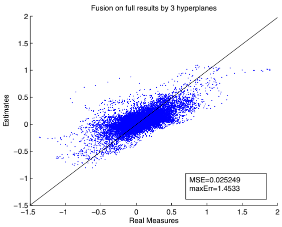


Fig. 4 Sea level data from the tidal gauge at Sheerness between 2000 and 2001



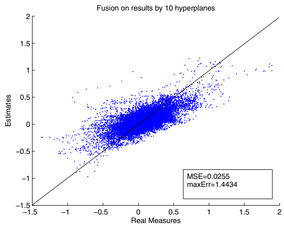


Fig. 5 Scatter plot of predicted against actual residual values for the test data, where forecasts are based on decision tree fusion of f_1, f_2 and f_3

Fig. 6 Scatter plot of predicted against actual residual values for the test data, where forecasts are based on decision tree fusion of 10 k-plane hyperplanes

Model fusion is applied to this forecasting problem by generating a number of different linear approximations to g and then aggregating these by applying LID3 to learn a linguistic decision tree as outlined in section 3. In the first instance the training data was divided into three subsets based on the residual value at Sheerness with C_1, C_2 and C_3 corresponding to those triples $\langle x_{t-0.5}, x_t, y_{t+8} \rangle$ where $y_{t+8} \leq -0.39$, $-0.39 < y_{t+8} \leq 0.52$ and $0.52 < y_{t+8}$ respectively. Least squares linear regression was then employed to infer functions f_1, f_2 and f_3 from C_1, C_2 and C_3 respectively. The intuitive idea is that these three functions provide different linear approximations of g for different levels of the residual at Sheerness. A classification database was then generated form the training data as described in section three and LID3 was applied to infer a linguistic decision tree to fuse these three function. On the test set the fused model gave a Mean Squared Error (MSE) of $0.025249m^2$ and a maximal error value of $1.4533m$. Figure 5 provides a graphical representation of the accuracy of this fused model in the form of a scatter plot of actual against forecast residual values for the test data. Here perfect accuracy is represented by the $y = x$ line on the plot. This result is comparable with, and indeed a small improvement on, a previous study involving this data in which a fuzzy Bayesian approach gave a MSE of $0.026m^2$ [6].

In a second experiment we applied *k-plane clustering* [2] as an unsupervised learning algorithm to learn 10 clusters with associated linear functions. Again by generating a classification database for the training data and applying LID3 we learnt a linguistic decision tree to fuse these functions. This resulted in a MSE of $0.0255m^2$ and a maximal error of $1.443m$. Figure 6 shows the scatter plot of actual against forecast residual values for the linguistic decision tree fusion of the 10 k-plane hyperplanes.

5 Conclusions

A model fusion method has been proposed based on linguistic decision trees and the LID3 learning algorithm. This approach has been successfully applied to the forecasting of the residuals for tidal surge data from Whitby to Sheerness in the UK. For this problem decision trees where successfully applied in order to aggregate linear models resulting in an improvement on an earlier fuzzy Bayesian forecasting model [6].

Acknowledgements. This research has been funded by the EPSRC FRMRC EP/FO20511/1. The authors would also like to thank the BODC for supplying the tidal surge data and Dr Kevin Horsburgh at POL for his help and advice.

References

1. Abrahart, R.J., See, L.: Multi-model Data Fusion for River Flow Forecasting: An Evaluation of Six Alternative Methods Based on two Contrasting Catchments. Hydrol. Earth Syst. Sci. 6(4), 655–670 (2002)
2. Bradley, P.S., Mangasarian, O.L.: k-Plane Clustering. J. Global Optimization 16, 23–32 (2000)
3. Bell, C.: POLTIPS.3-Tidal Prediction Software. Application Group, Proudman Oceanographic Laboratory (2005), http://www.pol.ac.uk/appl/poltipsw.html
4. Lawry, J.: A Framework for Linguistic Modelling. Artificial Intelligence 155, 1–39 (2004)
5. Lawry, J.: Modelling and Reasoning with Vague Concepts. Springer, New York (2006)
6. Randon, N.J., Lawry, J., Horsburgh, K., Cluckie, I.D.: Fuzzy Bayesian Modelling of Sea-Level Along the East Coast of Britain. IEEE Trans. Fuzzy Systems 16(3), 725–738 (2008)
7. Qin, Z., Lawry, J.: Decision Tree Learning with Fuzzy Labels. Inform. Sci. 172, 91–129 (2005)
8. Quinlan, J.R.: Induction of Decision Trees. Machine Learning 1, 81–106 (1986)
9. Winkler, R.L., Makridakis, S.: The Combination of Forecasts. J. Roy. Stat. Soc. Ser. A 146(2), 150–177 (1983)

Set-Valued Square Integrable Martingales and Stochastic Integral

Shoumei Li

Abstract. In this paper, we firstly introduce the concept of set-valued square integrable martingales. Secondly, we give the definition of stochastic integral of a stochastic process with respect to a set-valued square integrable martingale, and then prove the representation theorem of this kind of integral processes. Finally, we show that the stochastic integral process is a set-valued sub-martingale.

1 Introduction

In classical stochastic analysis, stochastic integral is one of the most important concepts. Stochastic integrals mean the integrals of a stochastic process with respect to firstly a Brownian motion, then a square integrable martingale, and more general a semimartingale. They are many important applications such as optimal control (e.g. [31]), mathematical finance (e.g. [13]) and so on. Recently, the theory of set-valued stochastic processes has been developed quickly due to the measurements of various uncertainties arising from not only the randomness but also from the impreciseness in some situations.

Concerning set-valued stochastic analysis, it may relate two kind integrals: stochastic integral of a set-valued stochastic process with respect to some kind of single-valued processes (e.g. a Brownian motion), and stochastic integral of a single-valued process with respect to some kind of set-valued processes (e.g. a set-valued martingale). For the first type, Kisielewicz introduced the definition of the stochastic integral of a set-valued stochastic process with respect to a Brownian motion in [15]. More related works have [1], [2], [8], [14]—[19], [27], [38], [41] and so on. There are also some work of Lebesgue

Shoumei Li
Department of Applied Mathematics, Beijing University of Technology,
Beijing 100124, P.R. China
e-mail: lisma@bjut.edu.cn

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 411–417.
springerlink.com

integral of a set-valued stochastic process with respect to time t, readers may refer to [20], [21], [39] and their related references.

Concerning set-valued martingales, it was first introduced by Van Cutsem in the case of convex compact values in [7]. Hiai and Umegaki gave more general definition of conditional expectation of a set-valued random variable in [10] so that the theory of set-valued martingales could be developed deeply and extensively. There are many works in this area, for instances, [4], [9]—[12], [23]—[29], [32]—[34], [37], [40], [41].

But there are no so many works on stochastic integral of a stochastic process with respect to a set-valued martingale. So far, we only find the reference [35]. In their paper, Qi and Wang gave a definition of integral of a stochastic process with respect to a set-valued square integrable martingale by essential convex closure. But they even did not discuss whether the result of integral is measurable or not. So it is difficult to have further applications. On the other hand, it is necessary to develop this kind of theory. For example, it is well-known that the method of equivalent martingale measure is play an very important role in option pricing. But if the financial market is not complete, the equivalent martingale measures are not unique in the dynamic pricing model. In this paper, we shall give a new definition of the stochastic integral of a classical stochastic process with respect to a set-valued martingale so that the result of the stochastic integral process is still measurable. Then we shall discuss some properties and a representation theorem of the stochastic integral processes.

We organize our paper as follows: in Section 2, we shall introduce some necessary notations, definitions and results about set-valued stochastic processes. In Section 3, we shall give a new definition of stochastic integral of a predictable stochastic process with respect to a set-valued square integrable martingale, prove the representation theorem and discuss some properties of set-valued stochastic integral, especially set-valued submartingale property. Since the page limitation, we have to omit the proofs of results. If you are interested in the proofs, please refer to our paper [22].

2 Set-Valued Martingales and Square Integrable Matingales

Throughout this paper, assume that $\mathbb{R}$ is the set of all real numbers, $I = [0,T]$, $\mathbb{N}$ is the set of all natural numbers, $\mathbb{R}^d$ is the d-dimensional Euclidean space with usual norm $\|\cdot\|$, $\mathscr{B}(E)$ is the Borel σ-field of the metric space E, $(\Omega, \mathscr{A}, (\mathscr{A}_t)_{t\in I}, \mu)$ is a complete filtration probability space, the σ-field filtration $\{\mathscr{A}_t : t \in I\}$ satisfies the usual conditions (i.e. complete, non-decreasing and right continuous). Let $L^p[\Omega, \mathscr{A}_t, \mu; \mathbb{R}^d]$ be the set of $\mathbb{R}^d$-valued $\mathscr{A}_t$-measurable random variables ξ with $E[\|\xi\|^p] < \infty$ $(1 \le p < \infty)$, and write $\|\xi\|_p = [E[\|\xi\|^p]]^{\frac{1}{p}}$. When $\mathscr{A}_t$ is replaced by $\mathscr{A}$, $L^p[\Omega, \mathscr{A}, \mu; \mathbb{R}^d]$ can be written as $L^p[\Omega; \mathbb{R}^d]$ for short.

Now we review notations and concepts of set-valued stochastic processes.

Assume that $\mathbf{K}(\mathbb{R}^d)$ is the family of all nonempty closed subsets of $\mathbb{R}^d$, and $\mathbf{K}_c(\mathbb{R}^d)$ is the family of all nonempty closed convex subsets of $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, A is a nonempty subset of $\mathbb{R}^d$, define the distance between x and A

$$d(x,A) = \inf_{y\in A} \|x-y\|.$$

The *Hausdorff metric* for two bounded sets A and B is defined as

$$d_H(A,B) = \max\{\sup_{a\in A} d(a,B), \sup_{b\in B} d(b,A)\},$$

and define $\|A\|_{\mathbf{K}} = d_H(\{0\},A) = \sup_{a\in A}\|a\|$.

If $F : (\Omega,\mathscr{A}) \to \mathbf{K}(\mathbb{R}^d)$ satisfies that for any open set $O \subseteq \mathbb{R}^d$, $F^{-1}(O) = \{\omega \in \Omega : F(\omega)\cap O \neq \emptyset\} \in \mathscr{A}$, then F is called ($\mathscr{A}$-) measurable (or a set-valued random variable, random set, multivalued function (e.g. [5], [10], [26]). If $\mathscr{F}$ is a sub-σ-field of $\mathscr{A}$, Let

$$S_F^p(\mathscr{F}) = \{f \in L^p[\Omega,\mathscr{F},\mu;\mathbb{R}^d] : f(\omega) \in F(\omega) \text{ a.e. } \omega \in \Omega\}.$$

When $\mathscr{F} = \mathscr{A}$, it is written S_F^p for short.

A set-valued random variable $F : \Omega \to \mathbf{K}(\mathbb{R}^d)$ is called *integrable* if S_F^1 is non-empty. F is *L^p-bounded* if and only if the real-valued random variable $\|F\|_{\mathbf{K}} \in L^p[\Omega;\mathbb{R}]$. If F is L^1-bounded, then F is also called *integrably bounded.* Let $L^p[\Omega,\ \mathscr{F},\mu;\mathbf{K}(\mathbb{R}^d)]$ be the family of all $\mathbf{K}(\mathbb{R}^d)$-valued L^p-bounded $\mathscr{F}$-measurable random variables. Similarly, we have notations $L^p[\Omega,\mathscr{F},\mu;\ \mathbf{K}_c(\mathbb{R}^d)]$.

Definition 1. *A non-empty set $\Gamma \subseteq L^p[\Omega,\mathscr{F},\mu;\mathbb{R}^d]$ is called* decomposable with respect to the sub-σ-field $\mathscr{F}$*, if for any $f,g \in \Gamma$, any $U \in \mathscr{F}$, we have $I_U f + I_{U^c} g \in \Gamma$.*

Firstly, we know that for any set-valued random variable $F \in L^p[\Omega,\mathscr{F},\mu;\ \mathbf{K}(\mathbb{R}^d)]$, $S_F^p(\mathscr{F})$ is decomposable with respect to $\mathscr{F}$. We also have the following opposite result.

Theorem 1 (cf. [10] or [26]). *Let Γ be a nonempty closed subset of $L^p[\Omega,\mathscr{F},\mu;\ \mathbb{R}^d]$. Then there exists an $\mathscr{F}$-measurable set-valued random variable F such that $\Gamma = S_F^p(\mathscr{F})$ if and only if Γ is decomposable with respect to $\mathscr{F}$. Furthermore, Γ is bounded if and only if F is integrably bounded, and Γ is convex if and only if F is convex.*

$F = \{F(t) : t \in I\}$ is called *a set-valued stochastic process* if $F : I \times \Omega \to \mathbf{K}(\mathbb{R}^d)$ is a set-valued function such that for any fixed $t \in I$, $F(t,\cdot)$ is a set-valued random variable. A set-valued process $F = \{F(t) : t \in I\}$ is called *adapted with respect to the filtration* $\{\mathscr{A}_t : t \in I\}$, if $F(t)$ is measurable with respect to $\mathscr{A}_t$ for each $t \in I$, and denoted by $\{F(t),\mathscr{A}_t : t \in I\}$.

Now we start to recall set-valued martingale and set-valued square integrable martingale.

Definition 2. *A set-valued stochastic process $F = \{F(t), \mathscr{A}_t : t \in I\}$ is called a set-valued martingale if*
(i) $F = \{F(t), \mathscr{A}_t : t \in I\}$ is adapted and for any $t \in I$, $F(t)$ is L^1-bounded;
(ii) for any $t \geq s$, $t, s \in I$, $E[F(t)|\mathscr{A}_s] = F(s)$, a.e.$(\mu)$.

Definition 3. *An adapted $\mathbb{R}^d$-valued stochastic process $f = \{f(t), \mathscr{A}_t : t \in I\}$ is called an L^p-*martingale selection *of $F = \{F(t), \mathscr{A}_t : t \in I\}$ if*
(1) For any $t \in I$, $f(t) \in L^p[\Omega, \mathscr{A}_t, \mu; \mathbb{R}^d]$, $f(t, \omega) \in F(t, \omega)$, a.e.;
(2) $\{f(t), \mathscr{A}_t : t \in I\}$ is a martingale.

Definition 4. *A set-valued martingale $F = \{F(t), \mathscr{A}_t : t \in I\}$ is called square integrable, if $\sup\limits_{t\in I} E[\|F(t)\|^2_{\mathbf{K}}] < \infty$.*

Note that a set-valued square integrable martingale is L^2-bounded. Furthermore, if $\mathscr{A}$ is μ-separable, we can prove the following Theorem.

Theorem 2. *Assume that $F = \{F(t), \mathscr{A}_t : t \in I\}$ is a set-valued square integrable martingale taking values in $\mathbf{K}_c(\mathbb{R}^d)$, and it is lower semicontinuous, then there exists a continuous square integrable martingale selection of F.*

Concerning more definitions and more results of the conditional expectations of set-valued random variables and set-valued martingales, readers may refer to the excellent paper [10] or the books [26] and [41].

Next we shall discuss stochastic integral of a stochastic process with respect to a set-valued square integrable martingale. We assume that $\mathscr{A}$ is μ-separable, $F = \{F(t), \mathscr{A}_t : t \in I\}$ is a set-valued square integrable martingale taking values in $\mathbf{K}_c(\mathbb{R}^d)$ and $\mathbf{CMS}(F) \neq \emptyset$, which $\mathbf{CMS}(F)$ is the set of $\mathbb{R}^d$-valued continuous square integrable martingale selections of F, in the following section.

3 Stochastic Integral a Stochastic Process with Respect to a Set-Valued Square Integrable Martingale

Definition 5. *Assume that $F = \{F(t), \mathscr{A}_t : t \in I\}$ is a set-valued square integrable martingale and $F(0) = 0$ a.e., g is a predictable bounded stochastic process. For any $\omega \in \Omega$, $t \in I$, $(A)\int_0^t g(s, \omega) dF(s, \omega)$ is defined as the set*

$$\Big\{ \int_0^t g(s, \omega) df(s, \omega) : f = \{f(t) : t \in I\} \in \mathbf{CMS}(F) \Big\},$$

$(A)\int_0^t g(s, \omega) dF(s, \omega)$ is said to be the Aumann type stochastic integral of g with respect to the set-valued square integrable martingale F.

Theorem 3. *Assume that $F=\{F(t),\mathscr{A}_t:t\in I\}$ is a convex set-valued square integrable martingale, g is a predictable bounded stochastic process, and for any $t\in I$ and $\omega\in\Omega$, $\Gamma(t,\omega)=:(A)\int_0^t g(s,\omega)dF(s,\omega)$ defined above. Then for any $t\in I$, $\Gamma(t)=:\int_0^t g(s)dF(s)$ is a non-empty convex subset of $L^2[\Omega,\mathscr{A}_t,\mu;\mathbb{R}^d]$.*

Remark 1. In [35], authors introduced a definition by taking convex closure of $\Gamma(t)$. However, they did not discuss whether the integral is a set-valued random variable or not. It is natural to hope that the result of integral is a set-valued stochastic process taking values in $\mathbf{K}_c(\mathbb{R}^d)$ rather than in $L^2[\Omega,\mathscr{A}_t,\mu;\mathbb{R}^d]$. According to Theorem 1, $\Gamma(t)$ should be decomposable with respect to $\mathscr{A}_t$ and closed if we want it to decide an $\mathscr{A}_t$-measurable set-valued random variable. Unfortunately, it is not true in general. Hence we will take the decomposable closure of $\Gamma(t)$ to modify the definition 5.

Theorem 4. *Assume that $F=\{F(t),\mathscr{A}_t:t\in I\}$ is a set-valued square integrable martingale, g is a predictable bounded stochastic process, and $\Gamma(t,\omega)=\int_0^t g(s,\omega)dF(s,\omega)$, then for any $t\in I$, there exists $M_t(g)\in\mathscr{M}[\Omega,\mathscr{A}_t,\mu;\mathbf{K}_c(\mathbb{R}^d)]$ such that*

$$S^2_{M_t(g)}(\mathscr{A}_t)=\overline{de}_{\mathscr{A}_t}\Gamma(t),$$

where the decomposable closure $\overline{de}_{\mathscr{A}_t}$ is taken in $L^2[\Omega,\mathscr{A}_t,\mu;\mathbb{R}^d]$ (cf. [38]).

Now we may give our modified definition of stochastic integral with respect to a set-valued square integral martingale.

Definition 6. *The set-valued stochastic process $M(g)=\{M_t(g):t\in I\}$ defined in Theorem 4 is called stochastic integral of g with respect to a set-valued square integral martingale F, and denoted as $M_t(g)=(M)\int_0^t gdF$.*

Now we start the study of representation theorem of stochastic integral of g with respect to F. We need to prove the following Lemma.

Lemma 1. *Assume that $F=\{F(t),\mathscr{A}_t:t\in I\}$ is a set-valued square integrable martingale, then there exists a sequence $\{f^n:n\in\mathbb{N}\}\subseteq\mathbf{CMS}(F)$, such that for every $t\in I$,*

$$S^2_{M_t(g)}(\mathscr{A}_t)=\overline{de}_{\mathscr{A}_t}\Big\{\int_0^t g(s)df^n(s):n\in\mathbb{N}\Big\},\tag{4.2}$$

where the closure is taken in L^2.

Theorem 5. *(Castaing representation theorem) Assume that $F=\{F(t),\mathscr{A}_t:t\in I\}$ is a set-valued square integrable martingale, and g is a predictable bounded stochastic process, then there exists a sequence of $\mathbb{R}^d$-valued martingales $\{f^i=\{f^i(t):t\in I\}:i\geq 1\}\subseteq\mathbf{CMS}(F)$ such that for any $t\in I$,*

$$M_t(g)(\omega)=\mathrm{cl}\Big\{\int_0^t g(s,\omega)df^i(s,\omega):i\geq 1\Big\},\quad a.e.\ \omega\in\Omega.$$

Now we give the following property of the stochastic integral $M_t(g)$.

Theorem 6. *Assume that $F = \{F(t), \mathscr{A}_t : t \in I\}$ is a set-valued square integrable martingale and g is a predictable bounded stochastic process, then the stochastic integral $\{M_t(g), \mathscr{A}_t : t \in I\}$ is a set-valued submartingale.*

Acknowledgements. This research is partially supported by PHR(IHLB), Beijing Educational Committee, NSFC (No.10771010), P.R. China.

References

1. Ahmed, N.U.: Nonlinear Stochastic differential inclusions on Banach space. Stoch. Anal. Appl. 12, 1–10 (1994)
2. Aubin, J.P., Prato, G.D.: The viability theorem for stochastic differenrial inclusions. Stoch. Anal. Appl. 16, 1–15 (1998)
3. Aumann, R.: Integrals of set valued functions. J. Math Anal. Appl. 12, 1–12 (1965)
4. Bagchi, S.: On a.s. convergence of classes of multivalued asymptotic martingales. Ann. Inst. H. Poincaré Probab. Statist. 21, 313–321 (1985)
5. Castaing, C., Valadier, M.: Convex analysis and measurable multifunctions. Lecture Notes in Math., vol. 580. Springer, Berlin (1977)
6. Da Prato, G., Frankowska, H.: A stochastic Filippov theorem. Stoch. Anal. Appl. 12, 409–426 (1994)
7. Van. Cutsem, B.: Martingales de multiapplications à valeurs convexes compactes. C. R. Math. Acad. Sci. Paris 269, 429–432 (1969)
8. Jung, E.J., Kim, J.H.: On set-valued stochastic integrals. Stoch. Anal. Appl. 21, 401–418 (2003)
9. Hess, C.: On multivalued martingales whose values be unbounded: martingale selectors and Mosco convergence. J. Multivariate Anal. 39, 175–201 (1991)
10. Hiai, F., Umegaki, H.: Integrals, conditional expectations and martingales of multivalued functions. J. Multivariate Anal. 7, 149–182 (1977)
11. Hu, S., Papageorgiou, N.S.: Handbook of Multivalued Analysis. Kluwer Academic Publishers, Dordrecht (1997)
12. de Korvin, A., Kleyle, R.: A convergence theorem for convex set valued supermartingales. Stoch. Anal. Appl. 3, 433–445 (1985)
13. Karatzas, I.: Lectures on the mathematics of finance. American Mathematical Society, Providence (1997)
14. Kim, B.K., Kim, J.H.: Stochastic integrals of set-valued processes and fuzzy processes. J. Math. Anal. Appl. 236, 480–502 (1999)
15. Kisielewicz, M.: Set valued stochastic integrals and stochastic inclusions. Discuss. Math. 13, 119–126 (1993)
16. Kisielewicz, M.: Set-valued stochastic integrals and stochastic inclusions. Stoch. Anal. Appl. 15, 783–800 (1997)
17. Kisielewicz, M.: Weak compactness of solution sets to stochastic differential inclusions with non-convex right-hand sides. Stoch. Anal. Appl. 23, 871–901 (2005)
18. Kisielewicz, M., Michta, M., Motyl, J.: Set valued approach to stochastic control, part I: existence and regularity properties. Dynam. Systems Appl. 12, 405–432 (2003)

19. Kisielewicz, M., Michta, M., Motyl, J.: Set valued approach to stochastic control, part II: viability and semimartingale issues. Dynam. Systems Appl. 12, 433–466 (2003)
20. Li, J., Li, S.: Set-valued stochastic Lebesgue integral and representation theorems. Int. J. Comput. Intell. Syst. 1, 177–187 (2008)
21. Li, J., Li, S., Ogura, Y.: Strong solution of Itô type set-valued stochastic differential equation. Acta Math. Sinica (to appear, 2010)
22. Li, S., Li, J., Li, X.: Stochastic integral with respect to set-valued square integrable martingales. J. Math. Anal. Appl. (2010), doi:10.1016/j.jmaa.2010.04.040
23. Li, S., Ogura, Y.: Convergence of set valued sub- and super-martingales in the Kuratowski-Mosco sense. Ann. Probab. 26, 1384–1402 (1998)
24. Li, S., Ogura, Y.: Convergence of set valued and fuzzy valued martingales. Fuzzy Sets Syst. 101, 453–461 (1999)
25. Li, S., Ogura, Y.: A convergence theorem of fuzzy-valued martingales in the extended Hausdorff metric H_∞. Fuzzy Sets Syst. 135, 391–399 (2003)
26. Li, S., Ogura, Y., Kreinovich, V.: Limit theorems and applications of set-valuded and fuzzy sets-valued random variables. Kluwer Academic Publishers, Dordrecht (2002)
27. Li, S., Ren, A.: Representation theorems, set-valued and fuzzy set-valued Itô intergal. Fuzzy Sets Syst. 158, 949–962 (2007)
28. Luu, D.Q.: Representations and regularity of multivalued martingales. Acta Math. Vietnam. 6, 29–40 (1981)
29. Luu, D.Q.: Applications of set-valued Radon-Nikodym theorms to convergence of multivalued L^1-amarts. Math. Scand. 54, 101–114 (1984)
30. Molchanov, I.: Theory of Random Sets. Springer, London (2005)
31. Oksendal, B.: Stochastic Differential Equations. Springer, London (1995)
32. Papageorgiou, N.S.: On the theory of Banach space valued multifunctions. 1. integration and conditional expectation. J. Multivariate Anal. 17, 185–206 (1985)
33. Papageorgiou, N.S.: A convergence theorem for set valued multifunctions. 2. set valued martingales and set valued measures. J. Multivariate Anal. 17, 207–227 (1985)
34. Papageorgiou, N.S.: On the conditional expectation and convergence properties of random sets. Trans. Amer. Math. Soc. 347, 2495–2515 (1995)
35. Qi, Y., Wang, R.: Set-valued stochastic integral of bounded predictable processes w.r.t. square integrable martingale. Comm. Appl. Math. Comput. 2, 77–81 (1998)
36. Shreve, S.E.: Stochastic Calculus for Finance. Springer, London (2004)
37. Wang, Z.P., Xue, X.: On convergence and closedness of multivalued martingales. Trans. Amer. Math. Soc. 341, 807–827 (1994)
38. Zhang, J., Li, S., Mitoma, I., Okazaki, Y.: On set-valued stochastic integrals in an M-type 2 Banach space. J. Math. Anal. Appl. 350, 216–233 (2009)
39. Zhang, J., Li, S., Mitoma, I., Okazaki, Y.: On the solution of set-valued stochastic differential equation in M-type 2 Banach space. Tohoku Math. J. 61, 417–440 (2009)
40. Zhang, W., Gao, Y.: A convergence theorem and Riesz decomposition for set valued supermartingales. Acta Math. Sinica 35, 112–120 (1992)
41. Zhang, W., Li, S., Wang, Z., Gao, Y.: An introduction of set-valued stochastic processes. Science Press, Beijing (2007)

Smooth Transition from Mixed Models to Fixed Models

María José Lombardía and Stefan Sperlich

Abstract. In multi-level regression, such as small area studies, and in panel data studies, using a fixed effect for each region leads to models that are flexible but that have poor estimation accuracy; they are over-parameterized. We bridge the gap between Fixed Effects Models, Mixed Effects Models and Partial Linear Models by a flexible modeling of area effects. The transition from Mixed Effects Models to Semiparametric Mixed Effects Models and Fixed Effects Models is achieved by progressively relaxing the smoothness assumption on the semiparametric area specific impact. The methodology is illustrated with a complete simulation study and applied for a small area analysis of tourist expenditures in Galicia.

Keywords: Semi-mixed effects models, Semiparametric regression, Multilevel models, Small area statistics, Panel data analysis.

1 Introduction

For a response $Y_{dj} \in \mathbb{R}$ and covariates $\boldsymbol{X}_{dj} \in \mathbb{R}^p$, including the intercept, consider a generalized linear Mixed Effects Model (MEM) with known link g

$$E\left[Y_{dj}|\boldsymbol{u}_d, \boldsymbol{X}_{dj}\right] = g\left\{\boldsymbol{X}^t_{dj}\boldsymbol{\beta} + \boldsymbol{Z}^t_{dj}\boldsymbol{u}_d\right\}, \quad d = 1, \ldots, D;\ j = 1, \ldots, n_d, \tag{1}$$

with $\boldsymbol{Z}_{dj} \subseteq \boldsymbol{X}_{dj}$ of dimension ρ, $\boldsymbol{\beta} \in \mathbb{R}^p$ the fixed effect, and $\boldsymbol{u}_d \in \mathbb{R}^\rho$ the i.i.d. unobservable random effect with mean zero and unknown variances-covariance matrix $\boldsymbol{\Sigma}_u$. The latter has to be estimated. Suppose to have sample

María José Lombardía
Universidad de Coruña, 15071 - A Coruña, Spain
e-mail: maria.jose.lombardia@udc.es

Stefan Sperlich
Georg-August Universität Göttingen, 37073 Göttingen, Germany
e-mail: stefan.sperlich@wiwi.uni-goettingen.de

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 419–425.
springerlink.com

size $n=\sum_{d=1}^{D} n_d$, where D is the number of areas (domains or groups) with the typical assumption that $D \to \infty$ at rate $O(n)$. In panel data analysis i may be time and d the individual. A crucial assumption for the existing methodology is that $\boldsymbol{X}_{dj}$ and $\boldsymbol{u}_d$ are independent and that $g(\cdot)$ is known. Note that, if g is the identity, model (1) includes the nested-error model ($\boldsymbol{Z}_{dj}=1$ and $u_d \in \mathbb{R}$), the random regression coefficient model ($\boldsymbol{Z}_{dj}=\boldsymbol{X}_{dj}$), and the Fay-Herriot model (only area specific information, [3]); see [11] for a summary.

Today, mixed effects models are popular in many areas of statistics, especially in small area statistics, see [7] or [12] for reviews; for panel data analysis [1], and [4] for a typical example. They are widely applied in biomedical, forestry, agricultural, economic and social science studies. Although the different research areas favor different terminology, like small area statistics, multi-level or simply mixed effects models, the statistical problems of modeling, estimation and testing are basically the same; the differences arise mainly in the subsequent inferences. For example, in biometrics they serve to analyze data with repeated measurements; in panel data analysis they account for possible heterogeneity over the cross sectional samples; in small area statistics they serve to improve the prediction of area-level parameters, while in econometrics they improve the calculation of macro indices from micro-data. Apart from the increasing interest in multi-level modeling (see [5]), they have also become popular in economics for data matching, i.e. to impute a certain factor for the individuals in the sample of interest with the aid of a different sample (see [2], for a recent example in poverty mapping). At the end, they all have in common that they try to account for a certain clustering, may it be due to space, time or individuals over time in panels, climate, administrative area or districts, villages or even large families, genetic groups or species.

More recently, mixed effects models have entered the nonparametric world; see [10], [6], and [13]. However the asymptotic theory for estimation in semiparametric mixed models was developed only recently. [8] and [9] introduced an estimation procedure for generalized partial linear mixed effects models, specification tests with bootstrap procedures, and provided asymptotic theory for these methods.

Thus, for a more flexible modeling we may also allow some covariates to enter the model nonparametrically. To ease the notation let us call these variables $\boldsymbol{T} \in \mathbb{R}^q$ and be different from the variables $\boldsymbol{X}$ which enter the model linearly. Then we have a generalized Partial Linear mixed effects Model (PLM) of the form

$$E\left[Y_{dj}|\boldsymbol{X}_{dj},\boldsymbol{T}_{dj}\right] = g\left\{\boldsymbol{X}_{dj}^t\boldsymbol{\beta}+\gamma(\boldsymbol{T}_{dj}+\boldsymbol{Z}_{dj}^t\boldsymbol{u}_d)\right\}, \tag{2}$$

with $d=1,\ldots,D,\ j=1,\ldots,n_d$ and a nonparametric function $\gamma: \mathbb{R}^q \to \mathbb{R}$.

The MEM is often motivated by the fact that it allows for efficient estimation of the fixed part, but makes also use of the random effects for prediction. This seems to outperform other parametric models in predicting and efficient estimation. When predicting, the additional variance that results from assuming this effect to be random, is only slightly larger than the variance of a fixed effect estimate based on small samples, and this deficiency is easily compensated by the efficient estimation of $\boldsymbol{\beta}$. However, this improved prediction in the mean is illusory if the somewhat unrealistic assumption of independence between area effects and the covariates, as well as the unobserved individual effects, is not met. Thus, even when a MEM leads to a better sample fit, it does so at the cost of producing biased estimates, and consequently bad out-of-sample prediction. Furthermore, methods to do valid inference have not yet been developed. All the currently available methods for testing or prediction intervals are clearly inconsistent if the assumption of independence is violated. This deficiency is not shared by the Fixed Effects Model (FEM)

$$E\left[Y_{dj}|\boldsymbol{X}_{dj}\right] = g\left\{\boldsymbol{X}_{dj}^{t}\boldsymbol{\beta} + c_d\right\}, \quad d = 1,\ldots,D;\ j = 1,\ldots,n_d, \tag{3}$$

with c_d being an area (domain or group) specific fixed effect without the assumption of independence from the individual effects $\boldsymbol{X}_{dj}$.

We studied many applications where the random effects represented the effect of either a region, a climate type, a socio-economic group or the proband group in biostatistics. In almost all cases the independence assumption was hardly credible. This causes endogeneity giving inconsistent estimates for $\boldsymbol{\beta}$ and potentially woeful out-of-sample prediction performance. The affirmation, for the purpose of estimation the FEM, and for prediction the MEM, would be the right model is unfortunately wrong. For example, the FEM does not allow to include covariates which do not or hardly vary with i (*time* in panel data) for given d. A prediction with MEM when the unrealistic independence assumption is violated performs only well for in-sample prediction, and parameter estimates can not be interpreted.

We therefore propose to use a flexible modeling of area effects that allows to change continuously from a MEM (Eq. 1) without area specific covariates to a Semiparametric Mixed Effects Model (SMEM, Eq. 4) with a smooth area specific mean and a random effect, up to the other extreme, an FEM (Eq. 3). Where SMEM is

$$E\left[Y_{dj}|\boldsymbol{X}_{dj},\boldsymbol{W}_{d},u_d\right] = g\left\{\boldsymbol{X}_{dj}^{t}\boldsymbol{\beta} + \eta_v(\boldsymbol{W}_d) + u_d\right\}\ , \tag{4}$$

with $\eta_v : \mathbb{R}^q \to \mathbb{R}$ an unknown nonparametric function with a given "slider" v.

We bridge the gap between FEM, MEM and PLM by a flexible modeling of area effects. The transition from MEM to SMEM and FEM is achieved by progressively relaxing the smoothness assumption on the semiparametric area specific impact: we start with the highest degree of smoothness (a constant) yielding to a random effects model, and end up with the lowest degree (interpolation of the area effects) yielding a fixed effects model. This way one can resolve all problems at once: model, and thus explain, the area or group effect, and dispose of the “independence assumption” problem. One obtains consistent estimates and valid inference. This is achieved without loosing the advantages of MEM, and without running into the problems we would face in a FEM. It should be emphasized that it nests MEM, FEM, and PLM. Consequently, it outperforms them all in estimation and prediction.

We apply of our model class in the context of small area statistics predicting average tourist expenditures in the 53 counties of Galicia, a region in the Northwest of Spain. As with the rest of the country, tourism is one of the most important sources of revenue. Therefore, official statistics and politics have a strong interest in acquiring information about the expenditure behaviour of tourists. Presently, the Galician Statistical Institute (IGE) is focusing its efforts on extending their statistics to county level, and to the level of the so-called *comarcas* of which 53 exist in Galicia.

Table 1 Descriptive statistics: mean, standard deviation, and median.

The dependent variable (Y)				
lexp	ln of total expenditure per day & cap.	4.064	.6464	4.086
Variables of the individuals (X)				
sex	= 1 if male	.4774	.4995	.0000
age1	= 1 if strictly younger than 29	.2340	.4233	.0000
age2	= 1 if $29 \leq$ age ≤ 65	.7057	.4557	1.000
single	= 1 if single	.4094	.4917	.0000
child	= 1 if children ≤ 16 years old	.2792	.4486	.0000
ngal	= 1 if not from Galicia	.7453	.4357	1.000
educ	= 1 if academic	.4981	.5000	.0000
stud	= 1 if student	.1226	.3280	.0000
self	= 1 if self-employed	.1000	.3000	.0000
pilgr	= 1 if pilgrim	.1189	.3236	.0000
family	= 1 visit family, friends, etc.	.3868	.4870	.0000
stay	measured in days	16.74	17.71	10.00
Variables of the comarca (W)				
lpopd	ln of population density	3.276	.8068	3.156
ftrail	= 1 if French pilgrim trail	.0440	.0913	.0000
coast	= 1 if coast	.0839	.1122	.0000

Table 2 Coefficients estimates with their bootstrap standard errors.

	FEM		*SMEM*		*MEM*	
	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.	$\hat{\beta}$	S.E.
sex	-.0284	.0447	-.0242	.0448	-.0327	.0409
age1	.2428	.1199	.2249	.1188	.2271	.1017
age2	.2665	.1080	.1978	.0966	.2145	.0879
single	-.0402	.0613	-.0745	.0568	-.0543	.0526
child	.0003	.0565	-.0166	.0501	-.0164	.0486
ngal	.2288	.0560	.2377	.0565	.2474	.0471
educ	.0648	.0487	.0481	.0454	.0517	.0410
stud	-.2219	.1011	-.2212	.0963	-.2312	.0829
self	.0809	.0786	.1288	.0740	.1131	.0721
pilgr	-.7004	.0910	-.6926	.0752	-.6918	.0683
family	-.1798	.0544	-.1478	.0487	-.1689	.0443
stay	-.0047	.0014	-.0045	.0014	-.0044	.0011
$\hat{\sigma}_u^2$			.0297	.0171	.0650	.0123

The presented study uses the set of variables described in Table 1. We included all three area variables in η_v to account for interactions. Here we show some preliminary results. We examine the coefficient estimates, together with the bootstrap estimates of the standard errors, see Table 2. In the bootstrap we used a pilot bandwidth for the pre-estimation, and 400 bootstrap replications.

2 Conclusions

We have introduced a new class of semi-mixed effects models that combines fixed effects, mixed effects and partial linear models. Nesting these models it can benefit from the advantages each model offers, and at the same time mitigate or even avoid its shortcomings. Our SMEM allows for a smooth transition from FEM to MEM, i.e. our class contains the continuum between them including also the PLM. Under the wrong assumption of independence the model is estimated with a serious bias in the MEM, and the variance of the estimates is larger than that in our semiparametric alternative. Moreover, we do not only offer consistent estimators, but also outperform the nested models FEM, PLM and MEM by construction. That this holds also true for finite samples, is exactly the strength of the proposed class as it demonstrates that we have successfully combined the advantages of these models to find a compromise that avoids the pitfalls of each extreme.

Further, although the construction of the MEM would favor it's performance in calculating macro or area parameters (rather than in estimating individual effects), the simulations show that SMEM is superior in terms of both out-of-sample and in-sample prediction. It is clear that SMEM is always better for consistent estimation and modeling with respect to interpretability. The example of analyzing tourist expenditures underpins this finding.

Finally, a consistent bootstrap arms us with a valid and feasible procedure to do statistical inference. FEM and PLM based bootstrap will suffer from a large variance in practice, whereas the SMEM is consistent and has small variance. In contrast, applying bootstrap in MEM when the independence assumption is violated is inconsistent as it is based on a wrong model and therefore leads to wrong conclusions.

Acknowledgements. We gratefully thank Walter Zucchini, Jean Opsomer, Thomas Kneib, and Bernd Fitzenberger for helpful discussion and comments. This work has been supported by grants MTM2008-03010 from the Spanish "Ministerio de Ciencia e Innovación" and Deutsche Forschungsgemeinschaft FOR916, and the DE2009-0030 and DAAD-acciones integradas "Semiparametric Mixed Effects Models".

References

1. Diggle, P.J., Heagerty, P., Liang, K.-L., Zeger, S.: Analysis of Longitudinal Data, 2nd edn. Oxford Statistical Science Series, vol. 25. Oxford University Press, Oxford (2002)
2. Elbers, C., Lanjouw, J.O., Lanjouw, P.: Micro-level Estimation of Poverty and Inequality. Econometrica 71, 355–364 (2003)
3. Fay, R.E., Herriot, R.A.: Estimates of Income for Small Places. An Application of James-Stein Procedures to Census Data. J. Amer. Statist. Assoc. 74, 269–277 (1979)
4. Ghosh, M., Nangia, N., Kim, D.H.: Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach. J. Amer. Statist. Assoc. 91, 1423–1431 (1996)
5. Goldstein, H.: Multilevel Statistical Models, 3rd edn. Edward Arnold, London (2003)
6. Hamilton, J.D.: A Parametric Approach to Flexible Nonlinear Inference. Econometrica 69, 537–573 (2001)
7. Jiang, J., Lahiri, P.: Mixed Model Prediction and Small Area Estimation. Test 15, 1–96 (2006)
8. Lin, X., Carroll, R.: Semiparametric Regression for Clustered Data Using Generalized Estimating Equations. J. Amer. Statist. Assoc. 96, 1045–1056 (2001)
9. Lombardía, M.J., Sperlich, S.: Semiparametric Inference in Generalized Mixed Effects Models. J. Roy. Statist. Soc. Ser. B 70, 913–930 (2008)
10. Opsomer, J., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J.: Nonparametric Small Area Estimation Using Penalized Spline Regression. J. Roy. Statist. Soc. Ser. B 70, 265–286 (2008)

11. Prasad, N.G.N., Rao, J.N.K.: The Estimation of the Mean Squared Error of Small-Area Estimators. J. Amer. Statist. Assoc. 85, 163–171 (1990)
12. Rao, J.N.K.: Small Area Estimation. John Wiley and Sons, Inc., New-York (2003)
13. Tutz, G.: Generalized Semiparametrically Structured Mixed Models. Comput. Statist. Data Anal. 46, 777–800 (2001)

Mixture Models with a Black-Hole Component

Nicholas T. Longford and Pierpaolo D'Urso

Abstract. We define a class of mixture models in which a component comprises an assortment of units that are not associated with any proper distribution. The models, motivated by the EM algorithm and fitted by its simple adaptation, are illustrated on several examples with large samples, one of them about transactions of residential properties in Wellington, New Zealand, in 2006.

Keywords: Black-hole component, Black-hole density, Clustering, EM algorithm, Outliers.

1 Introduction

Mixture models ([4]) have a wide range of applications. They are used for partitioning a population into subpopulations (groups) and, more generally, for estimating and approximating distributions that do not have a simple form. For example, a mixture of three normal distributions with distinct means and variances can have up to three modes and a variety of asymmetries in the tails and shoulders of the distribution. The density of such a mixture around the mode(s) can differ substantially from the normal. Mixture models are relatively easy to fit with the EM algorithm ([2]). Let f_k, $k = 1,\ldots,K$, be the densities of the $K > 1$ components of the model. Each density may involve some parameters, such as the mean μ_k and variance σ_k^2 of the normal distribution. The K densities may have different forms and may have some,

Nicholas T. Longford
SNTL and Universitat Pompeu Fabra, Barcelona, Spain
e-mail: NTL@sntl.co.uk

Pierpaolo D'Urso
Sapienza Università di Roma, Rome, Italy
e-mail: Pierpaolo.Durso@uniroma1.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 427–435.
springerlink.com

but not all, parameters in common. The EM algorithm comprises iterations, each with two steps. Prior to the first iteration, initial values are set for all the model parameters; they include the parameters involved in the densities f_k and the marginal probabilities p_k of belonging to components, which add up to unity: $p_1 + \cdots + p_K = 1$. The indicator of the component, $I_i(k)$, equal to unity if unit i belongs to component k and to zero otherwise, is regarded as the missing data in the EM algorithm. In the E-step of the algorithm, the conditional probability of belonging to component k, $\hat{r}_{ik} = \mathrm{E}\{I_i(k) \,|\, \hat{\theta}, \mathbf{x}\}$, given the data $\mathbf{x}$ and the current values of the parameter estimates, is evaluated for every unit $i = 1, \ldots, n$:

$$\hat{r}_{ik} = \frac{\hat{p}_k \hat{f}_k(x_i)}{\hat{p}_1 \hat{f}_1(x_i) + \cdots + \hat{p}_K \hat{f}_K(x_i)} \,. \qquad (1)$$

The circumflexes ˆ indicate estimation. To avoid a clutter of indices, we omit the iteration number; the right-hand side is associated with iteration $t-1$ and the left-hand side with iteration t. In the M-step, the parameters of the densities f_k and the marginal probabilities p_k are estimated by the algorithm that would be applied if the assignment indicators $I_i(k)$ were known. As these indicators are are merely estimated, their estimated conditional expectations $\hat{r}_{ik}$ are substituted for them. For example, when f_k is normal with mean and variance involved in no constraints,

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{\hat{r}_{\cdot k}} \sum_{i=1}^{n} \hat{r}_{ik} x_i \\ \hat{\sigma}_k^2 &= \frac{1}{\hat{r}_{\cdot k}} \sum_{i=1}^{n} \hat{r}_{ik} \left(x_i - \hat{\mu}_k^{(t-1)} \right)^2 , \end{aligned}$$

where $\hat{r}_{\cdot k} = \hat{r}_{1k} + \cdots + \hat{r}_{nk}$. The marginal probabilities p_k are estimated as the averages $\hat{r}_{\cdot k}/n$. In the expression for $\hat{\sigma}_k^2$ we indicate that the estimate $\hat{\mu}_k$ from the previous iteration should be applied.

In a mixture of symmetric unimodal distributions with finite expectations and variances, it is natural to regard the estimated expectation $\hat{\mu}_k$ of each component as a focus or centroid and the estimated standard deviation $\hat{\sigma}_k$ as the reciprocal of the force of attraction to component k. When unit i is in a relatively large distance from $\hat{\mu}_k$ its assignment to component k is questionable even when the unit is in a large distance also from the other estimated foci $\hat{\mu}_h$, $h \neq k$. This motivates our proposal for an additional (*black-hole*) component that would have no focus, but would contain all the units that do not fit well within any of the other (*proper*) components.

2 Black-Hole Component

The expression (1) can be interpreted as a competition of the mixture components for unit i. The total prize of unity is split depending on the apparent relevance of the components for the unit, moderated by the estimated marginal probabilities $\hat{p}_k$. Component k is said to be a poor winner (in the contest for unit i) when $\hat{r}_{ik}$ is large, but the value of the estimated density $f_k(x_i)$ is relatively small. To avoid having (many) poor winners, we introduce another component (contestant), which is not subject to the rules of a proper density ($\int_{-\infty}^{+\infty} f = 1$). It has no focus and is associated with no distribution. We specify it by a non-negative function f_0 that is non-increasing in an interval $(-\infty;T)$ and non-decreasing in $(T;+\infty)$ for some real T. This function is used in (1) as if it were the density of an additional component:

$$\hat{r}_{ik} = \frac{\hat{p}_k \hat{f}_k(x_i)}{\hat{p}_0 f_0(x_i) + \hat{p}_1 \hat{f}_1(x_i) + \cdots + \hat{p}_K \hat{f}_K(x_i)},$$

$k = 0,1,\ldots,K$. An example of such an improper density is

$$f_0(x) = H_1 \left[1 - \exp\left\{-H_2(x-T)^2\right\}\right], \tag{2}$$

where $H_1 > 0$ and $H_2 > 0$; H_1 is the limit of $f_0(x)$ at ±∞. We refer to T, where f_0 attains its minimum, as the *anti-focus*, to the density in (2) as *anti-normal* and, in general, to component 0 and its density as the *black-hole* component and density, respectively. The other (original) components and densities are referred to as *proper*. The black-hole component provides a generalisation similar to the fuzzy clustering proposed by Dave ([1]), in which the degrees of membership r_{ik} of component k add up to values smaller than unity for some or all units i. The shortfalls $1 - r_{i1} - \ldots - r_{iK}$ correspond to the probabilities r_{i0} of the black-hole component.

The black-hole density f_0 need not be symmetric and does not have to have the same (or any) limit as x diverges to ±∞. For a fixed dataset, increasing f_0 results in greater probability p_0 of component 0. The function f_0 may be equal to zero at T or in an interval that includes T. In this interval, the component 0 does not compete with the other components; $r_{ik} = 0$ when x_i is in this interval. With increasing distance of x_i from T, component 0 becomes a more potent competitor, and for very large $|x_i|$, when $f_1,\ldots,f_K$ are all very small, it is a clear winner. The behaviour (shape) of f_0 for such values of x is immaterial.

A black-hole component can be interpreted as having its focus at infinity. If we regard the value of a symmetric unimodal density (e.g., of the normal distribution) as a measure of proximity to the focus, then the black-hole density can also be interpreted as such, with its focus at ±∞; units in greater 'distance' from ±∞ have smaller values of f_0. The density f_0 does not have to be continuous and can be set to a constant throughout the real axis.

3 Examples

This section illustrates the properties of mixture models with a black-hole component. The first example is based on simulated data, and in the following

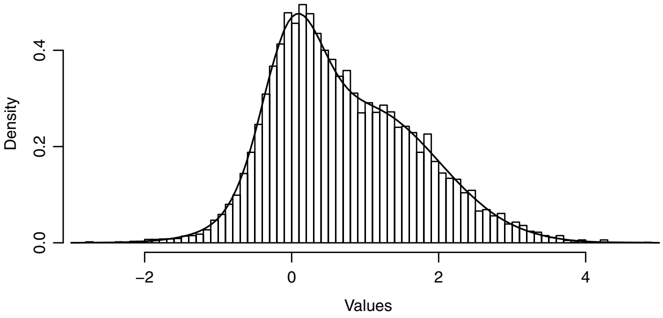


Fig. 1 The data and its generating distribution; Example 1.

two mixtures with a black-hole component are fitted to real datasets. We denote by $\mathcal{N}(\mu,\sigma^2)$ the normal distribution with mean μ and variance σ^2.

Example 1. We generated a sample of 10 000 units from the mixture of the normal distributions $\mathcal{N}(0,0.16)$ and $\mathcal{N}(1,1)$, with respective probabilities 0.3 and 0.7. For orientation, the empirical and exact densities of the mixture are plotted in Figure 1.

We fit the mixture model with two proper and a black-hole component with constant density $f_0 \equiv D$ for D in the range $(0.1, 0.5)$. The results are displayed in Table 1. They confirm that with increasing D the black-hole component becomes more attractive. The proper components are competitive only at close proximity to their respective foci $\hat{\mu}_k$, $k = 1,2$, and so the fitted standard deviations $\hat{\sigma}_k$ decrease with D. For $D = 0.1$, the black-hole component is unattractive ($\hat{p}_0 \doteq 0$), and for $D = 0.5$ the proper components are unattractive ($\hat{p}_0 \doteq 1$). Convergence of the adapted EM algorithm appears not to present any problems. We applied no methods for speeding up the convergence and, except for $D = 0.34$, required between 66 and 405 iterations to satisfy the criterion that the norm of the difference between two consecutive solutions (vectors of the estimated means and variances) is smaller than 10^{-4}. For $D = 0.34$, 817 iterations were required.

The marginal probability of the second component, associated with a greater variance, decreases faster than for the first component. This can be interpreted as the black-hole component attracting or 'recruiting' more units from the component with greater variance. We cannot assess the appropriateness of these solutions by any diagnostic methods, because a mixture with a black-hole component cannot be simulated. But the example shows that such a mixture can be fitted even when it is not appropriate; the data were generated by a model with $D = 0$.

Next we fit mixture models with two proper normal components and one anti-normal black-hole density given by (2), with $T = 0$, $H_2 = 0.2$ and $H_1 = 0.1, 0.2, \ldots, 2.0$. The fitted black-hole probability $\hat{p}_0$ increases rapidly

Table 1 Model fits with constant black-hole density D.

	Mean (μ)		*St. deviation* (σ)		*Probability* (p)		
D	1	2	1	2	1	2	0
0.10	0.004	1.019	0.415	1.003	0.311	0.689	0.000
0.14	0.002	1.035	0.422	0.989	0.320	0.676	0.004
0.18	0.034	1.292	0.464	0.777	0.449	0.469	0.082
0.22	0.045	1.200	0.401	0.639	0.383	0.354	0.263
0.26	0.060	1.009	0.330	0.506	0.270	0.210	0.520
0.30	0.066	0.719	0.263	0.383	0.159	0.097	0.744
0.34	−0.067	0.286	0.146	0.198	0.044	0.075	0.881
0.38	0.066	0.338	0.157	0.094	0.045	0.016	0.939
0.42	0.141	0.175	0.137	0.131	0.019	0.003	0.978
0.46	0.152	0.156	0.044	0.046	0.001	0.000	0.998
0.50	0.122	0.122	0.003	0.003	0.000	0.000	1.000

for small H_1, and then more slowly; $\hat{p}_0 = 0.000$ for $H_1 = 0.1$ and $\hat{p}_0 = 0.445$ for $H_1 = 1.0$, but $\hat{p}_0 = 0.573$ for $H_1 = 2.0$ and $\hat{p}_0 = 0.860$ for $H_1 = 25$. For $H_1 \in (0.1, 0.3)$, the attractiveness of the first component increases. The estimated standard deviation $\hat{\sigma}_1$ increases for small H_1, because the focus of the second component, $\hat{\mu}_2$, increases and leaves the first component attractive in a wider range. However, for values of $H_1 > 2.0$, the black-hole component takes over and the proper components become less attractive. Full listing of the results can be obtained from the authors.

The dependence of p_0 on H_2 is also as expected; smaller H_2 is associated with smaller estimate $\hat{p}_0$; for example, with H_1 fixed at 25.0, p_0 is equal to 0.612, 0.860 and 0.912 for the respective values $H_2 = 0.02, 0.2$ and 0.5. The estimated variances are decreasing functions of H_1 and H_2; uniformly greater values of the improper density function f_0 make the black-hole component more attractive, and only units very close to the foci of the proper components are attracted to them.

Example 2. The data for this example are records of six blood tests on 345 patients with liver disorder. They are extracted from the University of California Irvine (UCI) Machine Learning Repository, where they are stored as dataset 'BUPA liver disorders' in `archive.ics.uni.edu/ml/machine-learning-databases/liver-disorders` We explore the variables mean corpuscular volume (*mcv*) and alkaline phosphatase (*alkPh*). Their histograms are drawn in Figure 2. The sample correlation of the two variables is 0.044, so we lose little by analysing the two variables separately.

For *mcv*, the estimates of $\hat{\mu}$ depend on D very weakly, and a single observation, No. 224, is assigned to the black hole with near certainty for a wide range of settings of D. For example, for $D = 0.001$, when $\hat{\mu} = 90.16$, $\hat{\sigma}^2 = 16.74$ and $\hat{p} = 0.970$, $\hat{r}_{224,1} < 10^{-6}$, and next in the order of size is $\hat{r}_{69,1} = 0.425$. The largest possible value of $\hat{r}_{i1}$ is 0.9898. Values very close to it are attained by

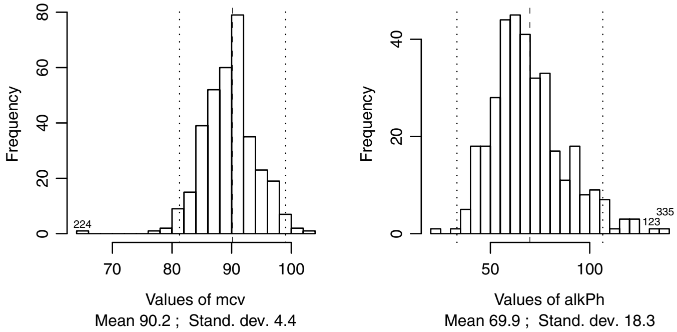


Fig. 2 Histograms of the variables *mcv* and *alkPh* in the BUPA liver disorders data. The vertical dashes indicate the sample mean $\hat{\mu}$ and the vertical dots are drawn at $\hat{\mu} \pm 2\hat{\sigma}$. The order No.s of the subjects mentioned in the text are marked near their respective values of *mcv* and *alkPh*.

the majority of the observations; 210 of them (61%) have values greater than 0.9850. For $D = 0.003$, $\hat{\mu} = 90.19$, $\hat{\sigma}^2 = 15.25$ and $\hat{p} = 0.926$; the black hole is more attractive. The values of all $\hat{r}_{i1}$ are smaller than with $D = 0.001$ and seven observations have $\hat{r}_{i1} < 0.5$: $\hat{r}_{224,1} < 10^{-7}$, followed by observations with $\hat{r}_{i1}$ equal to 0.14, 0.21, 0.26, and so on.

For *alkPh*, much smaller values of D should be used, because the values of *alkPh* are dispersed much more than for *mcv*. For $D = 0.0002$, we have $\hat{\mu} = 69.01$, $\hat{\sigma}^2 = 287.38$ and $\hat{p} = 0.969$. There are two clear outliers, subjects No. 335 and 123, with $\hat{r}_{i1}$ equal to 0.029 and 0.070; they have the largest values of *alkPh*, 138, 134, respectively. The values of $\hat{r}_{i1}$ are smaller than 0.5, but only by a narrow margin, for three other observations, all of them in the right-hand tail.

Example 3. This example analyses the sale prices of single-household residential properties in the City of Wellington, the capital of New Zealand, in 2006. In New Zealand, every residential property has a valuation, an official estimate of the value of the property; see Longford ([3]) for details. After discarding duplicate entries, properties with a lot of land, and other problematic transactions, we have 5201 records of capital values (CV), the official valuation, and sale prices (SP), both in New Zealand dollars (NZ$). To simplify the example and to focus on mixture modelling, we ignore other variables, such as floor and land area and the year of construction, which would be relevant in another context. As is common for variables for income, prices and other monetary values, we analyse CV and SP on the log-scale.

The data is summarised in Figure 3. Two mixture components can be discerned, but also several outliers. We fit the mixture model with two bivariate normal components and a constant black-hole component with the constant D set to 0, 0.02, ..., 0.2. The model fits are displayed in Table 2.

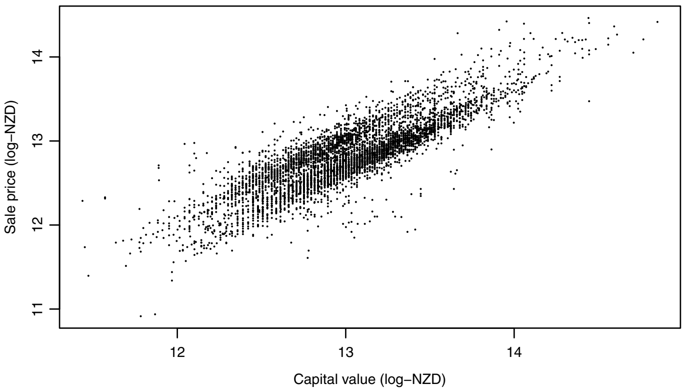


Fig. 3 The capital values and sale prices of residential properties sold in 2006 in Wellington City, New Zealand. Both variables are on the log scale.

Table 2 Model fits to the transactions of residential properties in Wellington City, New Zealand, in 2006. Two bivariate log-normal and a uniform black-hole component.

D	*Proper components* $\mathcal{N}_2(\mu,\Sigma)$							
	Means		*Variances*					
	log-CV	log-SP	log-CV	log-SP	*Cor.*	*Prob.*	$\hat{p}_0$	*Iter.*
0.00	13.035	12.786	0.112	0.086	0.871	0.496	0.000	24
	12.968	12.833	0.251	0.247	0.852	0.504		
0.02	13.130	12.783	0.152	0.153	0.971	0.504	0.023	26
	12.858	12.842	0.153	0.153	0.915	0.473		
0.04	13.129	12.782	0.149	0.148	0.971	0.503	0.030	25
	12.856	12.840	0.149	0.148	0.916	0.467		
0.06	13.127	12.781	0.146	0.147	0.971	0.501	0.036	24
	12.855	12.839	0.145	0.144	0.916	0.463		
0.08	13.127	12.781	0.145	0.145	0.972	0.499	0.041	25
	12.854	12.838	0.142	0.141	0.917	0.460		
0.10	13.126	12.781	0.143	0.140	0.972	0.498	0.046	25
	12.853	12.838	0.140	0.139	0.917	0.456		
⋮								⋮
0.20	13.125	12.781	0.137	0.132	0.972	0.490	0.065	23
	12.852	12.835	0.136	0.130	0.917	0.445		

Note: log-CV — the logarithm of capital value; log-SP — the logarithm of the sale price.

Without a black-hole component (with $D = 0$), we obtain a component with a relatively high log-mean of CV and low log-mean of SP, 13.035 and 12.786, respectively, which convert to NZ\$458 200 and NZ\$357 200. The fitted means for the other component correspond to NZ\$428 500 (CV) and NZ\$374 400 (SP). The second component is associated with much greater variances (0.251 and 0.247 *vs.* 0.112 and 0.086, for log-CV and log-SP, respectively). The results for a black-hole component with $D = 0.02$ differ substantially. The two foci for log-CV are in a greater distance ($13.130 - 12.858 = 0.272$ *vs.* 0.067 for $D = 0$), and the variances are almost identical. The within-component correlations are increased (by $0.971 - 0.871 = 0.100$ and 0.063), suggesting that the black-hole component recruits transactions for which CV and SP are not closely related. As we increase the black-hole density (D), the log-means are changed only slightly, the variances decrease very gradually, and the correlations are changed imperceptibly. As expected, the probability of the black-hole component increases with D. The results for $0.10 < D < 0.18$ can be obtained by linear interpolation with precision. The two proper components have nearly identical marginal probabilities for $D = 0$. For greater D, they decrease, but the probability of the first component (which has higher log-mean CV) decreases much more slowly. The black-hole component is more attractive for some of the transactions that had originally (with $D = 0$) higher probabilities $\hat{r}_{2k}$ for the second component. The column on the right-hand side gives the number of iterations required to achieve precision to four decimal places. No problems with convergence arise.

We obtain very similar results for a wide range of values of D. The black-hole component is useful to dismiss the hypothesis that the variances of the two (proper) components differ substantially. With a black-hole component, these variances are very similar; the first component has a greater correlation than the other.

4 Discussion

We defined a class of mixture models with an improper (black-hole) component. The presented examples are for normal component distributions, but extensions to other distributions, both continuous and discrete (and multivariate), present no conceptual difficulties.

In the examples with real data, the black hole can be interpreted as the component that contains the outliers, so its probability p_0 should always be small. The black-hole density may be set so as to match the anticipated value of p_0.

A mixture model specified with two or more black-holes is unlikely to be useful in practice because a mixture of black holes can be regarded as a single black hole.

Acknowledgements. The idea for this manuscript originates from the first author's visit at the Sapienza University of Rome, Italy, in 2009. Partial support by the Grant SEJ–2006–13537 from the Spanish Ministry of Science and Technology is acknowledged. The dataset for Example 3 was extracted from a database obtained from Headway Systems Ltd., Christchurch, New Zealand.

References

1. Dave, R.N.: Characterization and detection of noise in clustering. Pattern Recogn. Lett. 12, 657–664 (1991)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–38 (1977)
3. Longford, N.T.: A house price index defined in the potential outcomes framework. Working Paper 1175, Universitat Pompeu Fabra, Barcelona, Spain (2009)
4. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)

On the Preservations of Contractions: An Application to Stochastic Orders

M.C. López-Díaz and M. López-Díaz

Abstract. We obtain conditions for a function to preserve contractions, which do not involve convex sums of orthogonal matrices. The verification of such conditions implies, in general, a considerable saving of time compared with checking conditions already known. An application to stochastic orders of the above results is developed.

Keywords: Contraction, Loewner order, Orthogonal matrix, Dispersion order.

1 Introduction

Stochastic orders can be defined as partial order relations on a set of probabilities associated with a measurable space which have been applied successfully in many fields like reliability theory, economy, biology, medicine, genetics, statistical physics, decision theory, queueing systems, scheduling problems, etc.

Roughly speaking, a stochastic order aims to rank probabilities in accordance with an appropriate criterion. Different criteria have been considered for such rankings, like variability, dispersion, location, dependence, majorization, etc.

The reader is referred for instance to the monographs of [4] and [5] for an introduction to stochastic orderings.

M.C. López-Díaz
Departamento de Matemáticas, Universidad de Oviedo, E-33007 Oviedo, Spain
e-mail: cld@uniovi.es

M. López-Díaz
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,
E-33007 Oviedo, Spain
e-mail: mld@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 437–442.
springerlink.com

Most of the stochastic orders are developed for sets of probabilities on the space $(\mathbb{R}^n, \mathscr{B})$, where $\mathscr{B}$ stands for the usual Borel σ-field on $\mathbb{R}^n$.

A criterion for ranking probabilities on such a space is dispersion, which attempts to determine which vector induces a more dispersive probability distribution on $(\mathbb{R}^n, \mathscr{B})$, according to an appropriate viewpoint.

One of such orderings is the *strong dispersion order*, which is defined as follows: given $\mathbf{X}$ and $\mathbf{Y}$, $\mathbb{R}^n$-valued random vectors, it is said that $\mathbf{X}$ is not more dispersive that $\mathbf{Y}$ in the strong dispersion ordering if there exists a mapping $k : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{X} \sim_{st} k(\mathbf{Y})$, where $\sim_{st}$ stands for the stochastic equality, and k is a contraction (see [3]). This relation will be denoted by $\mathbf{X} \preceq_{SD} \mathbf{Y}$. Note that this means that the distribution of $\mathbf{X}$ can be obtained by means of a contraction of the random vector $\mathbf{Y}$.

The definition of the strong dispersion order involves contractions. We analyze the problem of the preservation of the strong dispersion ordering through transformations, studying the preservation of contractions.

Some conditions for preserving the strong dispersion ordering through transformations were developed in [2]. Namely, if $\mathbf{X}$ and $\mathbf{Y}$ are random vectors with $\mathbf{X} \preceq_{SD} \mathbf{Y}$, that is, $\mathbf{X} \sim_{st} k(\mathbf{Y})$ for certain contraction $k : \mathbb{R}^n \to \mathbb{R}^n$, and $h : \mathbb{R}^n \to \mathbb{R}^n$ is a measurable mapping, conditions on h and k such that $h(\mathbf{X}) \preceq_{SD} h(\mathbf{Y})$ are proposed.

Such conditions are the following:

1) k must be continuously differentiable,
2) there exists h^{-1},
3) h must be differentiable,
4) h^{-1} must be differentiable,
5) $h \circ k \circ h^{-1}$ must be continuously differentiable,
6) the Jacobian matrix of h, J_h, must be weakly orthogonal in each point of its domain. In [2] the authors call weakly orthogonal a matrix A if for any orthogonal matrix Γ, the eigenvalues of $A^t A \Gamma^t (A^t A)^{-1} \Gamma$ are all equal to 1, and assert that it is equivalent to the existence of a constant α such that αA is orthogonal,
7) h must be strongly attracted by k. This means that

$$J_h(k(x)) J_h^{-1}(x) = \sum_{i=1}^{M(x)} \alpha_i(x) \Gamma_i(x)$$

with $\alpha_i(x) > 0$ for all x in the domain of h, $\Gamma_i(x)$ is an orthogonal matrix for each i and for each x in such a domain, $1 \le i \le M(x)$, and $\sum_{i=1}^{M(x)} \alpha_i(x) = 1$.

Perhaps the most difficult condition to analyze in practice is 7) since it involves the existence of a convex sum of orthogonal matrices which depends on each point of the domain of h.

We obtain conditions for the preservation of the strong dispersion order, which in some frequent cases are very easy to check, avoiding the condition on the strong attraction required in [2].

2 Preliminaries

Some concepts and results that will be used to obtain conditions for the preservation of the strong dispersion order through transformations are introduced.

Let us denote by $\mathscr{M}_{n\times p}$ the vector space of all $n\times p$ matrices with elements in $\mathbb{R}$. The identity matrix in $\mathscr{M}_{n\times n}$ will be denoted by I_n. An invertible matrix A in $\mathscr{M}_{n\times n}$ will be called *weakly orthogonal* if there exists a positive real number α such that αA is an orthogonal matrix. Recall that an invertible matrix A in $\mathscr{M}_{n\times n}$ is called orthogonal if $A^t = A^{-1}$.

In [2] the authors defined weakly orthogonal matrices by means of eigenvalues. A matrix A in $\mathscr{M}_{n\times n}$ is called weakly orthogonal if the eigenvalues of $A^tA\Gamma^t(A^tA)^{-1}\Gamma$ are all equal to 1 for all Γ in $\mathscr{M}_{n\times n}$ orthogonal matrix.

We have obtained the equivalence of A being weakly orthogonal and $A^tA\Gamma^t(A^tA)^{-1}\Gamma$ having all eigenvalues equal to 1 for every Γ orthogonal matrix, that is:

Let A be an invertible matrix in $\mathscr{M}_{n\times n}$, then A is a weakly orthogonal matrix, that is, there exists a positive real number α such that αA is an orthogonal matrix, if and only if the eigenvalues of $A^tA\Gamma^t(A^tA)^{-1}\Gamma$ are all equal to 1 for all Γ in $\mathscr{M}_{n\times n}$ orthogonal matrix.

Conditions for the preservation of the strong dispersion order through transformations will be obtained using the Loewner ordering of symmetric matrices.

The *Loewner ordering of symmetric matrices* is given by $A \leq_L B$ if and only if $B-A$ is a nonnegative definite matrix.

In [3] the following results are proved:

- *A continuously differentiable function $k:\mathbb{R}^n \to \mathbb{R}^n$ is a contraction if and only if*
$$J_k(x)^t J_k(x) \leq_L I_n$$
for all x in $\mathbb{R}^n$ where $J_k(x)$ is the Jacobian matrix of k.
- *For two continuously differentiable functions $f,g:\mathbb{R}^n \to \mathbb{R}^n$ with f invertible, the condition*
$$J_g(x)^t J_g(x) \leq_L J_f(x)^t J_f(x)$$
for all x in $\mathbb{R}^n$ is necessary and sufficient for the function $g\circ f^{-1}$ to be a contraction.

In [1] the following result is proved:

- *The closed convex hull of O_n in the vector space $\mathscr{M}_{n\times n}$, O_n being the group of $n\times n$ orthogonal matrices, is $\{A\in\mathscr{M}_{n\times n} \,|\, A^tA \leq_L I_n\}$.*

Clearly, as a consequence of the results above, we have the following consecuence which is used in [2] for obtaining conditions on the preservation of the strong dispersion order through transformations:

A continuously differentiable function $k : \mathbb{R}^n \to \mathbb{R}^n$ is a contraction if and only if

$$J_k(x) = \sum_{i=1}^{m(x)} a_i(x) P_i(x)$$

where $a_i(x) > 0$, $\sum_{i=1}^{m(x)} a_i(x) = 1$ and $P_i(x)$ is an orthogonal matrix, for all $i \in \{1,\ldots,m(x)\}$ and for all x in $\mathbb{R}^n$.

Now, let us consider the next definition which is introduced in [2]: A differentiable function $h : \mathbb{R}^n \to \mathbb{R}^n$ is called weakly orthogonal if $J_h(x)$ is a weakly orthogonal matrix for all x in $\mathbb{R}^n$.

Note that if a differentiable function h is weakly orthogonal, there exists a function $\alpha : \mathbb{R}^n \to (0,+\infty)$ such that $\alpha(x)J_h(x)$ is an orthogonal matrix for all x in $\mathbb{R}^n$.

Now let us introduce a new definition which will be relevant in our work: A differentiable function $h : \mathbb{R}^n \to \mathbb{R}^n$ is called *dominated weakly orthogonal by a function* $k : \mathbb{R}^n \to \mathbb{R}^n$, if there exists a function $\alpha : \mathbb{R}^n \to (0,+\infty)$ such that $\alpha(x)J_h(x)$ is an orthogonal matrix with $\alpha(x) \leq \alpha(k(x))$ for all x in $\mathbb{R}^n$.

3 Main Results

In this section we obtain conditions for the preservation of the strong dispersion ordering through transformations without involving convex sums of orthogonal matrices. Our main result is:

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors in $\mathbb{R}^n$ with $\boldsymbol{X} \leq_{SD} \boldsymbol{Y}$. Suppose that there exists $k : \mathbb{R}^n \to \mathbb{R}^n$ a continuously differentiable contraction such that $\boldsymbol{X} \sim_{st} k(\boldsymbol{Y})$. Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be a continuously differentiable and invertible function which is dominated weakly orthogonal by k. Then $h(\boldsymbol{X}) \leq_{SD} h(\boldsymbol{Y})$.

Hence conditions for the preservation of the strong dispersion ordering through transformations which have been obtained are the following:

1) there exists the map h^{-1},
2) h must be continuously differentiable,
3) k must be continuously differentiable,
4) h must be dominated weakly orthogonal by k.

These conditions are easy to check in general since they do not involve convex sums of orthogonal matrices.

Let us compare the above condition 4) with the condition where h is strongly attracted by k which is required in [2]. As we show, 4) is equivalent to the conditions where h is weakly orthogonal and where h is strongly attracted by k.

Note that to check if h is strongly attracted by k implies dealing with convex sums of orthogonal matrices. In particular $J_h(k(x))J_h^{-1}(x)$ must be a convex sum of orthogonal matrices for every x. It is equivalent to verifying that all eigenvalues of $(J_h(k(x))J_h^{-1}(x))^t J_h(k(x))J_h^{-1}(x)$ are lower than or equal to 1 for every x. Verifying this condition is much more difficult in general than checking if h is dominated weakly orthogonal by k.

We have seen the following equivalence:

Let $g:\mathbb{R}^n \to \mathbb{R}^n$ be a function and let $h:\mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function which is weakly orthogonal. Then h is dominated weakly orthogonal by g if and only if h is strongly attracted by g.

If we analyze the conditions for checking the preservation of the strong dispersion ordering through transformations which have been obtained in this manuscript and in [2], we have the following table:

Reference [2]	This manuscript	Condition
Yes	Yes	k contraction continuously differentiable
Yes	Yes	h differentiable
Yes	Yes	h weakly orthogonal
Yes	Yes	there exists h^{-1}
Yes	No	$h \circ k \circ h^{-1}$ continuously differentiable
Yes	No	h^{-1} differentiable
Yes	An easier condition to check not involving convex sums of orthogonal matrices	h strongly attracted by k
No	Yes	h continuously differentiable

Acknowledgements. This work has been supported by Grants MTM2007-67884, MTM2008-01519 and IB08-147.

References

1. Eaton, M.L.: On group induced orderings, monotone functions, and convolution theorems. In: Tong, Y.L. (ed.) Proceedings of the Symposium on Inequalities in Statistics and Probability (Lincoln, Nebraska, 1982), Inst. Math. Statist., Hayward, CA. IMS Lecture Notes - Monographs Series, vol. 5 (1984)

2. Fernández-Ponce, J.M., Rodríguez-Griñolo, R.: Preserving multivariate dispersion: an application to the Wishart distribution. J. Multivariate Anal. 97, 1208–1220 (2006)
3. Giovagnoli, A., Wynn, H.P.: Multivariate dispersion orderings. Stat. Probab. Lett. 22, 325–332 (1995)
4. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. John Wiley & Sons, Chichester (2000)
5. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer, New York (2007)

A New Multivariate Stochastic Order, Main Properties

Miguel López-Díaz

Abstract. Different multivariate extensions of the bidirectional order can be considered. In this paper we propose a multivariate stochastic order which under mild conditions is an extension of the bidirectional order. The new relation is a proper stochastic order in the sense that it satisfies reflexivity, transitivity and antisymmetric properties. Moreover, the new order is integral. A maximal generator of it is obtained, which is used to obtain important properties of the order as a Strassen type theorem. We also obtain a characterization of the order by construction of random vectors on the same probability space. Different properties of the order are studied as well as connections with other stochastic orders and conditions, which in conjunction with the new order lead to the stochastic equality.

Keywords: Bidirectional order, Stochastic order, Strassen's Theorem.

1 Introduction

The bidirectional stochastic order for univariate distributions has been stated and studied in detail in a recent paper (see [3]).

This stochastic order is defined as follows: let X and Y be random variables, then X is not superior to Y in the bidirectional order, denoted by $X \preceq_{bd} Y$, if $X_+ \preceq_{st} Y_+$ and $X_- \preceq_{st} Y_-$, where $\preceq_{st}$ stands for the usual stochastic order, $a_+ = \max\{a, 0\}$ and $a_- = \max\{-a, 0\}$ with $a \in \mathbb{R}$. This definition is inspired in the univariate version of symmetric stochastic order proposed in [1].

Different characterizations of the bidirectional order have been developed in [3], as for instance:

Miguel López-Díaz
Dpto. de Estadística e I.O. y D.M., Universidad de Oviedo, E-33007 Oviedo, Spain
e-mail: mld@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 443–448.
springerlink.com

- $X \preceq_{bd} Y$ if and only if $F_X - F_Y$ pivots on 0, that is, $F_X(x) \leq F_Y(x)$ for all $x < 0$, and $F_X(x) \geq F_Y(x)$ for all $x \geq 0$, where F_W stands for the distribution function of the random variable W,
- $X \preceq_{bd} Y$ holds if and only if in any interval of the set $\{(-\infty,-t),(t,\infty) : t \geq 0\}$, the random variable Y deposits as much probability as the random variable X,
- $X \preceq_{bd} Y$ if and only if there exist random variables $\widetilde{X}, \widetilde{Y}$, defined on the same probability space, with $X \sim_{st} \widetilde{X}$ and $Y \sim_{st} \widetilde{Y}$, satisfying that $\widetilde{X}_+ \leq \widetilde{Y}_+$ and $\widetilde{X}_- \leq \widetilde{Y}_-$.

Different multivariate extensions of this univariate stochastic order can be proposed. In this paper we analyze one of them.

2 The New Order: Main Properties

We describe the new order, for such a purpose we need the following sets and notations. Given $z \in \mathbb{R}$, we define the set

$$P_z = \begin{cases} [z,\infty) & \text{if } z \geq 0, \\ (-\infty,z] & \text{if } z < 0. \end{cases}$$

If $\mathbf{z} = (z_1,z_2,\ldots,z_d) \in \mathbb{R}^d$, then let $P_{\mathbf{z}} = P_{z_1} \times P_{z_2} \times \ldots \times P_{z_d}$.

Let $\mathbf{X}$ and $\mathbf{Y}$ be $\mathbb{R}^d$-valued random vectors. It will be said that $\mathbf{X}$ is not superior to $\mathbf{Y}$ in the quadrant ordering if

$$P(\mathbf{X} \in P_{\mathbf{z}}) \leq P(\mathbf{Y} \in P_{\mathbf{z}})$$

for any $\mathbf{z} \in \mathbb{R}^d$. Such a condition will be denoted by $\mathbf{X} \preceq_{quad} \mathbf{Y}$.

It is possible to show that $\preceq_{quad}$ is a stochastic order on the set of probabilities on $(\mathbb{R}^d, \mathscr{B}_d)$, where $\mathscr{B}_d$ stands for the usual Borel σ-algebra on $\mathbb{R}^d$, that is, it satisfies the reflexive, transitive and antisymmetric properties.

We should indicate that the quadrant order is integral (see [4] or [5] for integral stochastic orders). This is a direct consequence of the definition of the order. Let $\mathscr{D} = \{I_{P_{\mathbf{z}}} : \mathbf{z} \in \mathbb{R}^d\}$, trivially $\mathbf{X} \preceq_{quad} \mathbf{Y}$ if and only if

$$\int_{\mathbb{R}^d} f \, dP_{\mathbf{X}} \leq \int_{\mathbb{R}^d} f \, dP_{\mathbf{Y}}$$

for all $f \in \mathscr{D}$, that is, the class $\mathscr{D}$ is a generator of the order.

A maximal generator of the quadrant order can be obtained. Important properties of integral stochastic orders can be derived by means of basic properties of the functions of maximal generators.

Note that the concept of maximal generator of an integral stochastic ordering is associated with the so-called weight function, which determines the space of functions which is considered. Such a function is a measurable mapping $b : \mathbb{R}^d \to [1,\infty)$, which induces the so-called b-norm (see [4] or [5]). We will consider the usual function $b = 1$. In this way, our space of functions will be the class of bounded functions.

We introduce the following notations and definitions.

Given $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, the notation $\mathbf{x} \perp \mathbf{z}$ will stand for $x_i \perp z_i$ for all $1 \le i \le d$, where $x_i \perp z_i$ means that $\begin{cases} x_i \ge z_i & \text{if } z_i \ge 0, \\ x_i \le z_i & \text{if } z_i < 0. \end{cases}$ We should observe that $\mathbf{x} \in P_{\mathbf{z}}$ if and only if $\mathbf{x} \perp \mathbf{z}$.

We will say that a map $f : \mathbb{R}^d \to [0,\infty)$ is a quadrant function if $f(\mathbf{y}) \ge f(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\mathbf{y} \perp \mathbf{x}$.

From now on we will denote by $\mathscr{G}$ the class of bounded quadrant functions. Note that $\mathscr{D} \subset \mathscr{G}$.

It is possible to prove that the maximal generator of the quadrant order, let us denote it by $\mathscr{R}_{quad}$, is the class of bounded quadrant functions, that is, $\mathscr{R}_{quad} = \mathscr{G}$.

Some consequences can be obtained by means of the above result. The first one is a Strassen type theorem:

Let P_1, P_2 be probabilities on the space $(\mathbb{R}^d, \mathscr{B}_d)$. The following conditions are equivalent:

i) $P_1 \preceq_{quad} P_2$,

ii) there exists a transaction kernel Q from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that

$$P_2(A) = \int_{\mathbb{R}^d} Q(\mathbf{x}, A)\, dP_1,$$

and for all $\mathbf{x} \in \mathbb{R}^d$ the probability $Q(\mathbf{x}, \cdot) : \mathscr{B}_d \to \mathbb{R}$ verifies that

$$\int_{\mathbb{R}^d} f(\mathbf{s})\, dQ(\mathbf{x}, \cdot) \ge f(\mathbf{x})$$

for all $f \in \mathscr{R}_{quad}$.

By means of the above result we can construct a characterization of the quadrant order by construction of random vectors on the same probability space. Thus:

Let $\mathbf{X}, \mathbf{Y}$ be random vectors. Then $\mathbf{X} \preceq_{quad} \mathbf{Y}$ if and only if there exist random vectors, $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}$, defined on the same probability space such that $\widetilde{\mathbf{Y}} \perp \widetilde{\mathbf{X}}$ a.s., with $\mathbf{X} \sim_{st} \widetilde{\mathbf{X}}$ and $\mathbf{Y} \sim_{st} \widetilde{\mathbf{Y}}$.

As a consequence of the above result we obtain the following statement:

Let $\mathbf{X}, \mathbf{Y}$ be random vectors with $\mathbf{X} \preceq_{quad} \mathbf{Y}$. Then there exist random vectors, $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}$, defined on the same probability space such that

$$(\widetilde{X}_{1\bullet_1}, \widetilde{X}_{2\bullet_2}, \ldots, \widetilde{X}_{d\bullet_d}) \le (\widetilde{Y}_{1\bullet_1}, \widetilde{Y}_{2\bullet_2}, \ldots, \widetilde{Y}_{d\bullet_d}) \ a.s.$$

for any $\bullet_1, \bullet_2, \ldots, \bullet_d$ belonging to the set $\{+,-\}$, where $\leq$ represents the usual componentwise order, with $\mathbf{X} \sim_{st} \widetilde{\mathbf{X}}$ and $\mathbf{Y} \sim_{st} \widetilde{\mathbf{Y}}$.

We can obtain the converse of the above result under mild conditions. Thus:

Let $\mathbf{X}, \mathbf{Y}$ be random vectors whose components have no atom at the origin. Then $\mathbf{X} \preceq_{quad} \mathbf{Y}$ if and only if there exist random vectors, $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}$, defined on the same probability space such that

$$(\widetilde{X}_{1\bullet_1}, \widetilde{X}_{2\bullet_2}, \ldots, \widetilde{X}_{d\bullet_d}) \leq (\widetilde{Y}_{1\bullet_1}, \widetilde{Y}_{2\bullet_2}, \ldots, \widetilde{Y}_{d\bullet_d}) \; a.s.$$

for any $\bullet_1, \bullet_2, \ldots, \bullet_d$ belonging to the set $\{+,-\}$, where $\mathbf{X} \sim_{st} \widetilde{\mathbf{X}}$ and $\mathbf{Y} \sim_{st} \widetilde{\mathbf{Y}}$.

Other properties of the quadrant order are the following:

1. The quadrant order is closed under mixtures.
2. The quadrant order is not closed under convolutions.
3. Let $\mathbf{X}, \mathbf{Y}$ be random vectors, let $I = \{i_1, i_2, \ldots, i_l\} \subset \{1, 2, \ldots, d\}$. If $\mathbf{X} \preceq_{quad} \mathbf{Y}$, then $\mathbf{X}_I \preceq_{quad} \mathbf{Y}_I$, that is, the quadrant order is preserved by marginalization.
4. Let $\mathbf{X}_j, \mathbf{Y}_j$, $1 \leq j \leq m$, be independent random vectors, where for each j, $\mathbf{X}_j$ and $\mathbf{Y}_j$ have the same dimension. If $\mathbf{X}_j \preceq_{quad} \mathbf{Y}_j$, $1 \leq j \leq m$, then $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m) \preceq_{quad} (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_m)$, that is, the quadrant order is preserved under the conjunction of independent vectors.
5. The stochastic order $\preceq_{quad}$ is closed with respect to identical concatenation, that is, if $\mathbf{X}, \mathbf{Y}$ are random vectors with $\mathbf{X} \preceq_{quad} \mathbf{Y}$, then $(\mathbf{X}_K, \mathbf{X}_L) \preceq_{quad} (\mathbf{Y}_K, \mathbf{Y}_L)$ for all $K, L \subset \{1, 2, \ldots, d\}$.
6. The quadrant order is closed with respect to the weak convergence for limits whose components have no atom at the origin.

3 Relations with Other Stochastic Orders

In this section we briefly describe some relations of the new order with other stochastic orders and state conditions, which in conjunction with the quadrant order guarantee the stochastic equallity.

Firstly we show that under mild conditions, the univariate version of the quadrant order is the bidirectional order.

Let X, Y be random variables with no atom at the origin. Then $X \preceq_{quad} Y$ if and only if $X \preceq_{bd} Y$.

Different extensions of the univariate usual stochastic order to the multivariate case have been stated, among them the multivariate stochastic order ($\mathbf{X} \preceq_{st} \mathbf{Y}$ if $E(f(\mathbf{X})) \leq E(f(\mathbf{Y}))$ for all bounded increasing mappings $f : \mathbb{R}^d \to \mathbb{R}$), the *upper orthant order* ($\mathbf{X} \preceq_{uo} \mathbf{Y}$ if $\overline{F}_{\mathbf{X}}(\mathbf{z}) \leq \overline{F}_{\mathbf{Y}}(\mathbf{z})$ for all $z \in \mathbb{R}^d$, where $\overline{F}_{\mathbf{W}}$ stands for the survival function of the random vector $\mathbf{W}$), and the *lower orthant order* ($\mathbf{X} \preceq_{lo} \mathbf{Y}$ if $F_{\mathbf{X}}(\mathbf{z}) \geq F_{\mathbf{Y}}(\mathbf{z})$ for all $\mathbf{z} \in \mathbb{R}^d$, where $F_{\mathbf{W}}$ stands for the distribution function of the random vector $\mathbf{W}$) (see for instance [6]).

There are not general implications between the quadrant order and any of the above orders. However we can state the following results under additional conditions.

Let $\mathbf{X},\mathbf{Y}$ be random vectors such that $P(\mathbf{X} \in P_{\mathbf{0}}) = 1$, where $\mathbf{0} = (0,\ldots,0)$. Then $\mathbf{X} \preceq_{st} \mathbf{Y}$ implies that $\mathbf{X} \preceq_{quad} \mathbf{Y}$.

Let $\mathbf{X},\mathbf{Y}$ be random vectors such that $\overline{F}_{\mathbf{X}}(\mathbf{0}) = 1$. Then it holds that $\mathbf{X} \preceq_{uo} \mathbf{Y}$ if and only if $\mathbf{X} \preceq_{quad} \mathbf{Y}$.

Let $\mathbf{X},\mathbf{Y}$ be random vectors such that $P(X_1 < 0,\ldots,X_d < 0) = 1$. Then it holds that $\mathbf{Y} \preceq_{lo} \mathbf{X}$ if and only if $\mathbf{X} \preceq_{quad} \mathbf{Y}$.

For random vectors with the same copula, whose components have no atom at the origin, the quadrant order reduces to compare marginals with this order.

Let $\mathbf{X},\mathbf{Y}$ be absolutely continuous random vectors with a common copula, whose components have no atom at the origin. Then it holds that $\mathbf{X} \preceq_{quad} \mathbf{Y}$ if and only if $X_i \preceq_{quad} Y_i$ for all $1 \leq i \leq d$.

Other connections we can state are the following:

1. Let $\mathbf{X},\mathbf{Y}$ be random vectors with $\mathbf{X} \preceq_{quad} \mathbf{Y}$. It holds that
$$(|X_1|,|X_2|,\ldots,|X_d|) \preceq_{uo} (|Y_1|,|Y_2|,\ldots,|Y_d|).$$
2. Let $\mathbf{X},\mathbf{Y}$ be random vectors such that
$$(|X_1|,|X_2|,\ldots,|X_d|) \text{ and } (|Y_1|,|Y_2|,\ldots,|Y_d|)$$
have the same copula. If $\mathbf{X} \preceq_{quad} \mathbf{Y}$ then it holds that
$$(|X_1|,|X_2|,\ldots,|X_d|) \preceq_{st} (|Y_1|,|Y_2|,\ldots,|Y_d|).$$
3. Let $\mathbf{X},\mathbf{Y}$ be random vectors with independent components such that X_i and Y_i have symmetric distribution with respect to 0 and no atom at such a point. Then $\mathbf{X} \preceq_{quad} \mathbf{Y}$ if and only if $(|X_1|,|X_2|,\ldots,|X_d|) \preceq_{st} (|Y_1|,|Y_2|,\ldots,|Y_d|)$.

   Let $(\preceq_{icx})$ $\preceq_{cx}$ stand for the (increasing) convex order, that is, $(\mathbf{X} \preceq_{icx} \mathbf{Y})$ $\mathbf{X} \preceq_{cx} \mathbf{Y}$ if $Ef(\mathbf{X}) \leq Ef(\mathbf{Y})$ for all (increasing) convex functions $f : \mathbb{R}^d \to \mathbb{R}$ such that the expectations exist (see for instance [6]). The following results show connections of the new order with these orders.

4. Let $\mathbf{X},\mathbf{Y}$ be random vectors with independent components and $\mathbf{X} \preceq_{quad} \mathbf{Y}$. If $E\mathbf{X} \leq E\mathbf{Y}$, then $\mathbf{X} \preceq_{icx} \mathbf{Y}$. Moreover, if $E\mathbf{X} = E\mathbf{Y}$ then $\mathbf{X} \preceq_{cx} \mathbf{Y}$.
5. Let $\mathbf{X},\mathbf{Y}$ be random vectors with independent components without atom at the origin. Then $\mathbf{X} - \mathbf{X}' \preceq_{quad} \mathbf{Y} - \mathbf{Y}'$ if and only if $X_i \preceq_w Y_i$, $1 \leq i \leq d$, where $\preceq_w$ stands for the weak dispersive order (see [2] for the weak dispersive order).

In relation to conditions which, in conjunction with the quadrant order, lead to stochastic equality, we can enounce:

1. Let $\mathbf{X},\mathbf{Y}$ be random vectors having a same copula with $\mathbf{X} \preceq_{quad} \mathbf{Y}$ and $E|X_i| = E|Y_i|$ for all $1 \leq i \leq d$. Then $\mathbf{X} \sim_{st} \mathbf{Y}$.
2. Let $\mathbf{X},\mathbf{Y}$ be random vectors having a same copula with $\mathbf{X} \preceq_{quad} \mathbf{Y}$. If $E(X_i^j) = E(Y_i^j)$ for all $1 \leq i \leq d$ and $1 \leq j \leq 2$. Then $\mathbf{X} \sim_{st} \mathbf{Y}$.

We obtain the following example for random vectors with normal distributions. Let $\mathbf{X} \sim_{st} N(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}}), \mathbf{Y} \sim_{st} N(\mu_{\mathbf{Y}}, \Sigma_{\mathbf{Y}})$ be random vectors with the same copula. Then $\mathbf{X} - \mu_{\mathbf{X}} \preceq_{quad} \mathbf{Y} - \mu_{\mathbf{Y}}$ if and only if $\sigma_{X_i} \leq \sigma_{Y_i}$.

Acknowledgements. The author is indebted to the Spanish Ministry of Science and Innovation since this research is financed by Grant MTM2008-01519.

References

1. Cascos, I., Molchanov, I.: A stochastic order for random vectors and random sets based on the Aumann expectation. Statist. Probab. Lett. 63, 295–305 (2003)
2. Giovagnoli, A., Wynn, H.P.: Multivariate dispersion orderings. Statist. Probab. Lett. 22, 325–332 (1995)
3. López-Díaz, M.: A stochastic order for random variables with applications. Aust. N. Z. J. Stat. 52, 1–16 (2010)
4. Müller, A.: Stochastic orders generated by integrals: a unified study. Adv. Appl. Prob. 29, 414–428 (1997)
5. Müller, A., Stoyan, D.: Comparison methods for stochastic models and risks. John Wiley and Sons, Chichester (2002)
6. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer, New York (2007)

ANOVA for Fuzzy Random Variables Using the R-package SAFD

M. Asunción Lubiano and Wolfgang Trutschnig

Abstract. Due to the important role as central summary measure of a fuzzy random variable (FRV), statistical inference procedures about the mean of FRVs have been developed during the last years. 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 doing a bootstrap test for the equality of means of two or more FRVs. The corresponding algorithm will be described and applied to both real-life and simulated data.

Keywords: ANOVA, Fuzzy random variable, R package.

1 Introduction

The concept of a Fuzzy Random Variable (FRV) in Puri & Ralescu's sense [13] was introduced as a notion combining both randomness (stochastic uncertainty) and fuzziness (imprecision), where imprecision means non-statistical uncertainty due to the inaccuracy of human knowledge or the inexactness of measurements. FRVs can also be seen as natural generalisation of random sets [10].

The mean of a FRV [13] has been defined as natural extension of the Aumann-expectation of random sets - it is a fuzzy-valued quantity measuring

M. Asunción Lubiano
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,
33007 Oviedo, Spain
e-mail: `lubiano@uniovi.es`

Wolfgang Trutschnig
European Centre for Soft Computing, Edificio Científico-Technológico,
33600 Mieres, Spain
e-mail: `wolfgang.trutschnig@softcomputing.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 449–456.
springerlink.com

(summarizing) the ‘central tendency’ of the FRV. Various ways of testing about the mean of FRV can be found in the literature, most of them are based on statistics analogous to the classical ones using different metrics on the class $\mathscr{F}_c(\mathbb{R})$ of fuzzy numbers ([2, 3, 4, 7, 11, 12]).

In the above-mentioned papers the considered bootstrap statistic is just the numerator of the extension of the classical statistic. Nevertheless simulations studies have shown that considering the studentized version dividing by the estimation of the variance of the involved random elements also works nicely. Therefore we will consider the quotient when developing bootstrap testing procedures for the multi-sample hypothesis test.

2 Preliminaries

Throughout the paper we will work with the class $\mathscr{F}_c(\mathbb{R})$ of *fuzzy numbers*, i.e. mappings $\widetilde{U} : \mathbb{R} \to [0,1]$ such that for each $\alpha \in [0,1]$ the *α-level set* $U_\alpha := \{x \in \mathbb{R} : U(x) \geq \alpha\}$ is a nonempty compact interval $[\underline{U}_\alpha, \overline{U}_\alpha]$ in $\mathbb{R}$. For $\widetilde{U} \in \mathscr{F}_c(\mathbb{R})$ we define the functions *mid* and *spread* for every $\alpha \in [0,1]$ by

$$\mathsf{mid}\,(\widetilde{U})(\alpha) := \frac{1}{2}(\underline{U}_\alpha + \overline{U}_\alpha) \quad \text{and} \quad \mathsf{spr}\,(\widetilde{U})(\alpha) := \frac{1}{2}(\overline{U}_\alpha - \underline{U}_\alpha).$$

Given two fuzzy numbers $\widetilde{U}, \widetilde{V} \in \mathscr{F}_c(\mathbb{R})$, the *sum* $\widetilde{U} \oplus \widetilde{V}$ of $\widetilde{U}$ and $\widetilde{V}$ is defined as the fuzzy number $\widetilde{U} \oplus \widetilde{V} \in \mathscr{F}_c(\mathbb{R})$ such that for each $\alpha \in [0,1]$

$$(\widetilde{U} \oplus \widetilde{V})_\alpha = \left\{ y + z : y \in \widetilde{U}_\alpha, z \in \widetilde{V}_\alpha \right\}$$

holds. Analogously, given a fuzzy number $\widetilde{U} \in \mathscr{F}_c(\mathbb{R})$ and a real number γ, the *product* $\gamma \odot \widetilde{U}$ of $\widetilde{U}$ and γ is defined as the fuzzy number $\gamma \cdot \widetilde{U} \in \mathscr{F}_c(\mathbb{R})$ such that for each $\alpha \in [0,1]$:

$$(\gamma \odot \widetilde{U})_\alpha = \left\{ \gamma \cdot y : y \in \widetilde{U}_\alpha \right\}.$$

Remark 1. Since the above mentioned operators are in fact levelwise Minkowski operations $(\mathscr{F}_c(\mathbb{R}), \oplus, \odot)$ only has a semilinear structure.

One of the most important aspects of the (statistical) analysis of fuzzy data is the usage of a suitable distance on the family $\mathscr{F}_c(\mathbb{R})$, a distance that is both not too hard to calculate and which reflects the intuitive meaning of fuzzy sets. One good choice is the metric D_θ^φ introduced by Bertoluzza et al. [1], which can also be generalized to the multivariate setting without losing any nice property (see Trutschnig et al. [14]). D_θ^φ can be expressed in terms of the squared Euclidean distances between the mids and the squared Euclidean distances between the spreads of the interval level sets of the fuzzy numbers involved. Let φ be a probability density on $[0,1]$ with $\varphi(\alpha) > 0$ for almost every $\alpha \in [0,1]$. Define $D_\theta^\varphi : \mathscr{F}_c(\mathbb{R}) \times \mathscr{F}_c(\mathbb{R}) \to [0, +\infty)$ by

$$D_\theta^\varphi(\widetilde{U},\widetilde{V})^2 := \int_{[0,1]} \left[\mathrm{mid}\,(\widetilde{U}) - \mathrm{mid}\,(\widetilde{V})\right]^2 + \theta \cdot \left[\mathrm{spr}\,(\widetilde{U}) - \mathrm{spr}\,(\widetilde{V})\right]^2 \varphi(\alpha) d\alpha.$$

The parameter θ plays the role of a weight of the (distance between the) spreads against the (distance between the) mids. The absolutely continuous measure $d\mu = \varphi d\alpha$ serves as a weight measure of the different α-levels, i.e. we can assign weights according to our interpretation of the importance of the α-levels. In the R-package SAFD (see Section 3) μ has been chosen to be the Lebesgue measure on $[0,1]$, i.e. $\varphi \equiv 1$

Given a probability space $(\Omega, \mathscr{A}, P)$ a FRV $\widetilde{\mathscr{X}}$ is a mapping $\widetilde{\mathscr{X}} : \Omega \to \mathscr{F}_c(\mathbb{R})$ that is Borel-measurable w.r.t. the Borel σ-field generated by the metric D_θ^φ on $\mathscr{F}_c(\mathbb{R})$. A FRV $\widetilde{\mathscr{X}}$ is said to be *integrably bounded* if $\underline{X}_0, \overline{X}_0 \in L^1(\Omega, \mathscr{A}, P)$. If $\widetilde{\mathscr{X}}$ is integrably bounded then the *expectation (or mean)* of $\widetilde{\mathscr{X}}$ is the unique element $\mathbb{E}(\widetilde{\mathscr{X}}) \in \mathscr{F}_c(\mathbb{R})$ such that

$$\left(\mathbb{E}(\widetilde{\mathscr{X}})\right)_\alpha = [\mathbb{E}(\min \mathscr{X}_\alpha), \mathbb{E}(\max \mathscr{X}_\alpha)]$$

holds for every $\alpha \in [0,1]$.
The D_θ^φ*-variance* of a FRV $\widetilde{\mathscr{X}}$ is defined as

$$\mathrm{Var}(\widetilde{\mathscr{X}}) = \mathbb{E}\left(\left[D_\theta^\varphi\left(\widetilde{\mathscr{X}}, \mathbb{E}(\widetilde{\mathscr{X}})\right)\right]^2\right)$$

whenever this quantity is finite. It is easy to see that variance and expectation fulfil the usual property of Fréchet expectation (also see [8]), i.e.

$$\mathbb{E}\left(\left[D_\theta^\varphi\left(\widetilde{\mathscr{X}}, \mathbb{E}(\widetilde{\mathscr{X}})\right)\right]^2\right) = \inf_{\widetilde{A} \in \mathscr{F}_c(\mathbb{R})} \mathbb{E}\left(\left[D_\theta^\varphi\left(\widetilde{\mathscr{X}}, \widetilde{A}\right)\right]^2\right).$$

3 The R-package SAFD: Basic Features and How to Test Equality of Means with It

The aim of the R-package SAFD (Statistical Analysis of Fuzzy Data) is to provide some basic functions for statistics with one-dimensional fuzzy data. The package allows to work with polygonal fuzzy numbers, being represented as data frames with columns x and alpha (equidistant alpha levels in $[0,1]$). SAFD contains functions for the basic operations on the class of fuzzy numbers (sum, scalar product, mean, Hukuhara difference) as well as for calculating (Bertoluzza-) distance, sample variance, sample covariance, sample correlation, and the Dempster-Shafer (levelwise) histogram. Moreover a function to simulate fuzzy random variables and a function to do linear regression given trapezoidal fuzzy data is included. For more information see the reference manual at `http://cran.r-project.org/web/packages/SAFD/index.html`.

Apart from the above mentioned functions SAFD also provides a bootstrap test for the equality of means - the implemented procedure works as follows: Let $\widetilde{\mathscr{X}_1}, \ldots, \widetilde{\mathscr{X}_k}$ be independent FRVs ($k \geq 2$). For each $i \in \{1, \ldots, k\}$ consider a sample $\widetilde{X}_{i1}, \ldots, \widetilde{X}_{in_i}$ of $\widetilde{\mathscr{X}_i}$. Denote the sample mean of group i by $\overline{\overline{\widetilde{X}}}_{i\cdot}$ and the overall sample mean by $\overline{\overline{\widetilde{X}}}_{\cdot\cdot}$, i.e.

$$\overline{\overline{\widetilde{X}}}_{i\cdot} = \frac{1}{n_i} \cdot \left(\widetilde{X}_{i1} + \ldots + \widetilde{X}_{in_i} \right) \quad \text{and}$$

$$\overline{\overline{\widetilde{X}}}_{\cdot\cdot} = \frac{1}{n} \cdot \left(\widetilde{X}_{11} + \ldots + \widetilde{X}_{kn_k} \right) = \frac{n_1}{n} \cdot \overline{\overline{\widetilde{X}}}_{1\cdot} + \ldots + \frac{n_k}{n} \cdot \overline{\overline{\widetilde{X}}}_{k\cdot}$$

whereby $n = n_1 + \ldots + n_k$ is the overall sample size.

The implemented bootstrap test for $H_0 : \mathbb{E}(\widetilde{\mathscr{X}_1}) = \ldots = \mathbb{E}(\widetilde{\mathscr{X}_k})$ against $H_1 : \mathbb{E}(\mathscr{X}_{i_1}) \neq \mathbb{E}(\mathscr{X}_{i_2})$ for some $i_1 \neq i_2$ works as follows (also see [5]):

Algorithm. Implemented multi-sample bootstrap test.

Step 1. Compute the value of the statistic

$$T = \frac{\displaystyle\sum_{i=1}^{k} n_i \left[D_\theta^\varphi \left(\overline{\overline{\widetilde{X}}}_{i\cdot}, \overline{\overline{\widetilde{X}}}_{\cdot\cdot} \right) \right]^2}{\displaystyle\sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[D_\theta^\varphi \left(\widetilde{X}_{ij}, \overline{\overline{\widetilde{X}}}_{i\cdot} \right) \right]^2}$$

Step 2. Compute the bootstrap populations by adding to each sample the sum of the means of the other ones; for this purpose we set

$$\widetilde{y}_{ij} = \widetilde{x}_{ij} + \left(\overline{\overline{\widetilde{X}}}_{1\cdot} + \ldots + \overline{\overline{\widetilde{X}}}_{(i-1)\cdot} + \overline{\overline{\widetilde{X}}}_{(i+1)\cdot} + \ldots + \overline{\overline{\widetilde{X}}}_{k\cdot} \right)$$

and for each i define a FRV $\widetilde{\mathscr{Y}_i}$ that assume the above values $\widetilde{y}_{ij}$, $j = 1 \cdots n_i$, with the corresponding relative frequencies.

Step 3. Draw bootstrap samples $(\widetilde{Y}_{i1}^*, \ldots, \widetilde{Y}_{in_i}^*)$ from $\widetilde{\mathscr{Y}_i}$ for each $i = 1 \ldots k$.

Step 4. Compute the value of the bootstrap statistic

$$T^* = \frac{\displaystyle\sum_{i=1}^{k} n_i \left[D_\theta^\varphi \left(\overline{\widetilde{Y}_{i\cdot}^*}, \overline{\widetilde{Y}_{\cdot\cdot}^*} \right) \right]^2}{\displaystyle\sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[D_\theta^\varphi \left(\widetilde{Y}_{ij}^*, \overline{\widetilde{Y}_{i\cdot}^*} \right) \right]^2}$$

Step 5. Repeat steps 3 and 4 a large number B of times.

Step 6. Compute the bootstrap p-value as the portion of values in $\{T_1^*, \ldots, T_B^*\}$ being not smaller than T and return the p-value.

4 Illustrative Example 1

Within a study about the progress of reforestation in a given area of Asturias (Spain) the INDUROT (Institute of Natural Resources and Zoning of the University of Oviedo) wanted to quantify the "mean quality" of the three main species of trees used in the reforestation: birch (*Betula celtiberica*), sessile oak (*Quercus petraea*) and rowan (*Sorbus aucuparia*). The available information given was a (randomly collected) sample of $n_1 = 133$ birches, $n_2 = 109$ sessile oaks and $n_3 = 37$ rowans (see Table 1).

Table 1 Contingency table of the quality of the three species of trees

	Quality of tree								
Species	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_7	x_9
birch	4	6	13	15	44	9	23	11	8
sessile oak	17	10	19	13	32	9	7	1	1
rowan	1	0	7	8	9	4	5	0	3

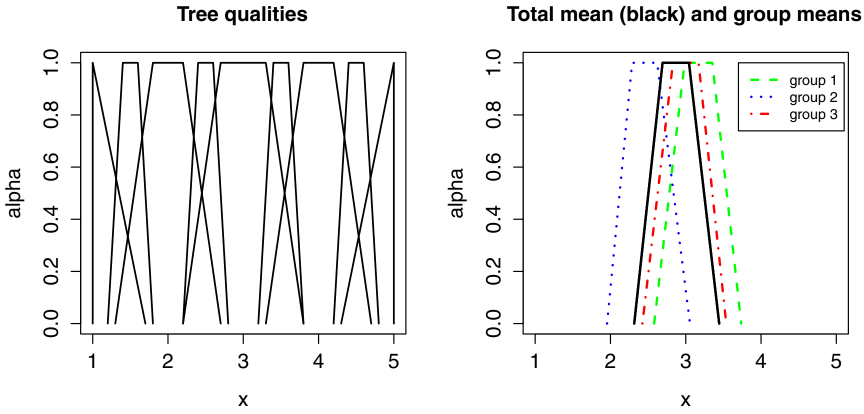


Fig. 1 Nine different tree qualities and group means

Each tree was assigned a trapezoidal fuzzy number that models the experts subjective judgements/perceptions of the tree quality on a scale from 0 to 5 (0 meaning very bad quality to 5 meaning very good quality). Thereby the 1-cut is the interval in which the expert thinks the quality is contained and the support (0-cut) is the interval in which the expert is highly willing to accept that the quality is contained. The left part of Figure 1 depicts the nine different values of the quality which were considered by the experts.

In order to check whether or not the mean quality of the three species of trees is equal we can directly use the function `btestk.mean` and the dataset `Trees.RData` in the SAFD package:

```
> data(Trees); sel<-c(1,2,3)
> btestk.mean(Trees,sel,1000)
```

The group means as well as the overall mean of the sample are depicted in the right part of Figure 1.
For $\theta = 1/3$ we get $T = 11.611$ and an associated p-value of 0. Consequently (for all standard significance levels) we can conclude that the 'mean quality' is not the same for all three species.
Since the sample mean of groups one and three seem to be similar we can rerun the test for equality of means only for these two groups:

```
> btestk.mean(Trees,c(1,3),1000)
[1] 0.246
```

Consequently we will not reject the hypothesis of equality of means for groups one and three. Figure 2 depicts the ecdf of the bootstrap statistic $T^{\star}$ and the corresponding value of T in the second test.

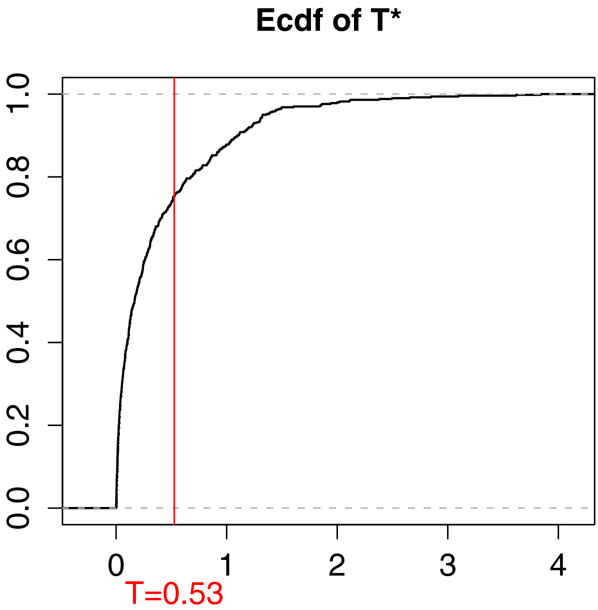


Fig. 2 Ecdf of the bootstrap statistic $T^{\star}$

5 Illustrative Example 2

The R package SAFD also offers the possibility to generate samples of fuzzy random variables - essentially the procedure described in [6] has been implemented. As second example we generated three samples of size 20 of three different FRV having the same mean. The true mean of the FRVs as well as the group means are depicted in Figure 3, the returned *p*-value is 0.536.

```
> data(XX); V<-translator(XX[[3]],20)
> YY<-ZZ<-WW<-list(length=10)
> for(i in 1:10){
+   YY[[i]]<-generator(V,,,) +
ZZ[[i]]<-generator(V,pertV=list(dist="unif",par=c(-3,3)),, +
pertR=list(dist="chisq",par=c(1))) +
WW[[i]]<-generator(V,pertV=list(dist="unif",par=c(-3,3)), +
pertL=list(dist="chisq",par=c(1)),) }
> XXX<-list(YY,ZZ,WW)
> A<-btestk.mean(XXX,sel=c(1,2,3),1000)
[1] 0.536
```

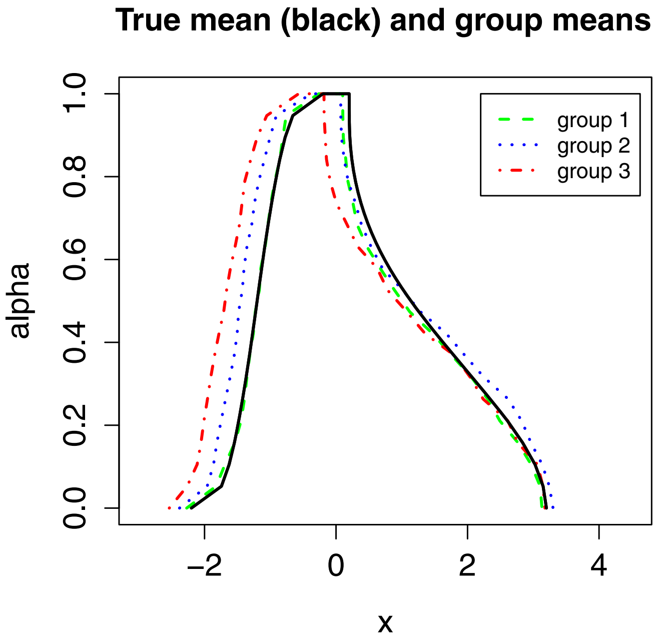


Fig. 3 Group means and true mean in the second example

6 Conclusions

The bootstrap test for the equality of means contained in the R-package SAFD provides useful results for small or medium sample sizes. The R package will be extended and improved in the future, in particular the speed of the bootstrap tests has to be increased.

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.: On a new class of distances between fuzzy numbers. Mathware Soft. Comput. 2, 71–84 (1995)
2. Colubi, A.: Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data. Fuzzy Sets Syst. 160(3), 344–356 (2009)
3. Gil, M.A., Montenegro, M., González-Rodríguez, G., Colubi, A., Casals, R.: Bootstrap approach to the multi-sample test of means with imprecise data. Comput. Statist. Data Anal. 51(1), 148–162 (2006)
4. González-Rodríguez, G., Montenegro, M., Colubi, A., Gil, M.A.: Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst. 157, 2608–2613 (2006)
5. González-Rodríguez, G., Colubi, A., D'Urso, P., Montenegro, M.: Multi-sample test-based clustering for fuzzy random variables. Internat. J. Approx. Reason. 50(5), 721–731 (2009)
6. González-Rodríguez, G., Colubi, A., Trutschnig, W.: Simulation of fuzzy random variables. Inform. Sci. 179(5), 642–653 (2009)
7. Körner, R.: An asymptotic α-test for the expectation of random fuzzy variables. J. Statist. Plann. Inference 83, 331–346 (2000)
8. Körner, R., Näther, W.: On the variance of random fuzzy variables. In: Bertoluzza, C., Gil, M.A., Ralescu, D.A. (eds.) Statistical Modeling, Analysis and Management of Fuzzy Data, pp. 22–39. Physica-Verlag, Heidelberg (2002)
9. Lubiano, M.A., Gil, M.A., López-Díaz, M., López-García, M.T.: The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable. Fuzzy Sets Syst 11, 307–317 (2000)
10. Molchanov, I.: Theory of Random Sets. Springer, London (2009)
11. Montenegro, M., Colubi, A., Casals, M.R., Gil, M.A.: Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59, 31–49 (2004)
12. Montenegro, M., Casals, M.R., Lubiano, M.A., Gil, M.A.: Two-sample hypothesis tests of means of a fuzzy random variable. Inform. Sci. 133(1-2), 89–100 (2001)
13. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
14. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inform. Sci. 179(23), 3964–3972 (2009)

Probability Tree Factorisation with Median Free Term

Irene Martínez, Carmelo Rodríguez, and Antonio Salmerón

Abstract. We study the decomposition of probability trees as a product of factors with reduced domains. We introduce exact and approximate factorisation techniques with respect to a median of proportional sub-trees as an alternative to the factorisation with average free term.

1 Introduction

A Bayesian network can be seen as a representation of a multivariate probability distribution, such that the conditional independence relations induced by the distribution are encoded by the structure of the network, according to the *d*-separation criterion. This feature allows the computation of posterior distributions (also called *probability propagation*) without actually handling the joint distribution over all the variables in the problem.

Recent advances in propagation have come along with methods that incorporate the ability of dealing with factorised representations of the potentials associated with the join tree, as Lazy [5] and Lazy-penniless propagation [4].

A particular feature of the Lazy-penniless algorithm is that is uses probability trees [1, 2] to represent probabilistic potentials. Probability trees are usually more compact than probability tables and, what is more important, provide a flexible way to reduce the space required to store a probabilistic potential, approximating it by pruning some of the branches of the trees.

Irene Martínez
Dpt. Languages and Computation, University of Almería, 04120 Almería, Spain
e-mail: irene@ual.es

Carmelo Rodríguez and Antonio Salmerón
Dpt. Statistics and Applied Mathematics, University of Almería,
04120 Almería, Spain
e-mail: crt,antonio.salmeron@ual.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 457–465.
springerlink.com

A *potential* ϕ over a random vector $\mathbf{X}$ is a mapping $\phi : \Omega_{\mathbf{X}} \to \mathbb{R}_0^+$, where $\Omega_{\mathbf{X}}$ is the support of $\mathbf{X}$. We will consider only discrete variables with a finite number of cases, and the *size* of a potential ϕ is defined on $\Omega_{\mathbf{X}}$, will be $|\Omega_{\mathbf{X}}|$.

A *probability tree* [1, 2, 8] is a directed labeled tree, where each internal node represents a variable and each leaf node represents a probability value. Each internal node has one outgoing arc for each state of the variable associated with that node. Each leaf contains a non-negative real number. The *size* of a tree $\mathscr{T}$, denoted as $\mathsf{size}(\mathscr{T})$, is defined as its number of leaves.

A probability tree $\mathscr{T}$ on variables $\mathbf{X}_I = \{X_i | i \in I\}$ represents a potential $\phi : \Omega_{\mathbf{X}_I} \to \mathbb{R}_0^+$ if for each $\mathbf{x}_I \in \Omega_{\mathbf{X}_I}$ the value $\phi(\mathbf{x}_I)$ is the number stored in the leaf node that is reached by starting from the root node and selecting the child corresponding to coordinate x_i for each internal node labeled with X_i.

Probability propagation relies on the combination and marginalisation operations, but the complexity is determined by the combination. For instance, consider the situation in which we are about to delete a variable X_i in order to send a message between two nodes of the join tree. The first step is to combine the potentials (probability trees in this case) containing X_i. The result will be, in the worst case, a potential of size equal to the product of the sizes of the trees that took part in the combination. A gain in efficiency could be achieved if we managed to decompose each tree containing X_i as a product of two trees (factors) of lower size, one of them containing X_i and the other not containing it. Then, the product would be actually carried out over potentials (trees) with reduced domains and thus, the complexity of probability propagation could decrease.

2 Exact Factorisation of Probability Trees

Now assume that the next variable to marginalise out is X, and we find it in the tree shown in Fig. 1. Within context $W = 0$, all the children of X are proportional. Thus, it is possible to factorise the tree as a product of two trees, where the size of each factor is lower than the size of the original tree (see Fig. 2), in such a way that one of the factor keeps the information regarding X and the other contains the information irrelevant to X [6, 7].

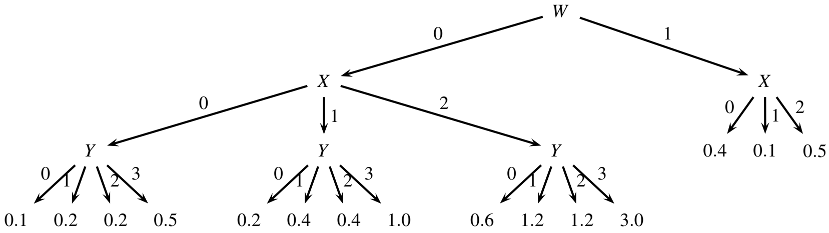
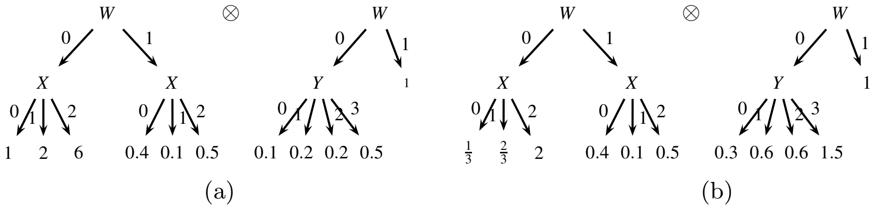


Fig. 1 A probability tree proportional below X for context $(W = 0)$.

Let $\mathscr{T}$ be a probability tree. Let $(\mathbf{X}_C = \mathbf{x}_C)$ be a configuration of variables leading from the root node in $\mathscr{T}$ to a variable X. We say that $\mathscr{T}$ is *proportional below X within context* $(\mathbf{X}_C = \mathbf{x}_C)$ if there is a $x_i \in \Omega_X$ such that for every $x_i, x_j \in \Omega_X$, $\exists \pi_{ij} > 0$ such that $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)} = \pi_{ji} \cdot \mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_i)}$, where $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x)}$ denotes the sub-tree of $\mathscr{T}$ reached following the path determined by configuration $(\mathbf{X}_C = \mathbf{x}_C, X = x)$. The values $\boldsymbol{\pi} = \{\pi_{ij}\}$ are called *proportionality factors*.

Let $\mathscr{T}$ be a probability tree proportional below X within context $(\mathbf{X}_C = \mathbf{x}_C)$, with proportionality factors $\boldsymbol{\pi}$. The *core term* of $\mathscr{T}$, denoted by $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x_i, \boldsymbol{\alpha})$ is the tree obtained by replacing sub-tree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_i)}$ in $\mathscr{T}$ by constant 1 and any other sub-tree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)}$ by constant π_{ji}.

The *free term* of $\mathscr{T}$, denoted by $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x_i)$ is the tree obtained from $\mathscr{T}$ by replacing sub-tree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ by $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_i)}$ and any other sub-tree $\mathscr{T}^{R(\mathbf{X}_D=\mathbf{x}_D)}$ by 1 for any context inconsistent with $(\mathbf{X}_C = \mathbf{x}_C)$. It holds that $\mathscr{T} = \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi}) \times \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x)$.



(a)

(b)

Fig. 2 Decomposition of the tree in figure 1 with respect to variable X.

Instead of factorising with respect to an arbitrary subtree, in [7], it is proposed to factorise with respect to an average of the proportional subtrees, with the aim of avoiding the high values that appear in the leaves of the core term of the factorisation (see Fig. 2(b)). Let $\mathscr{T}$ be a probability tree proportional below X within context $(\mathbf{X}_C = \mathbf{x}_C)$, with proportionality factors $\boldsymbol{\pi}$. The *exact factorisation of $\mathscr{T}$ with average free term* is $\mathscr{T} = \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x) \times \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi})$, where the free term, $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x)$, is computed as the tree obtained from $\mathscr{T}$ by replacing $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ by $\bar{\mathscr{T}}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ and replacing $\mathscr{T}^{R(\mathbf{X}_D=\mathbf{x}_D)}$ by a 1 for every context $(\mathbf{X}_D = \mathbf{x}_D)$ incompatible with $(\mathbf{X}_C = \mathbf{x}_C)$, and where $\bar{\mathscr{T}}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ is

$$\bar{\mathscr{T}}^{R(\mathbf{X}_C=\mathbf{x}_C)} = \frac{1}{|\Omega_X|} \sum_{x_j \in \Omega_X} \mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)}. \tag{1}$$

The core term, $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi})$, is the tree obtained from $\mathscr{T}$ by replacing each tree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)}$ by a constant π_j given by:

$$\pi_j = \frac{1}{\bar{\pi}_{\cdot j}} = \frac{|\Omega_X|}{\sum_{k: x_k \in \Omega_X} \pi_{kj}} \quad . \tag{2}$$

2.1 Factorisation with Median Free Term

By a motivation similar to those given in [7], a factorisation with median free term could be considered as well. The median is more robust than average, for example for outliers and moreover, a better accuracy could be obtained for some divergence measures.

Definition 1. *Let $\mathscr{T}$ be a probability tree proportional below X for context $(\mathbf{X}_C = \mathbf{x}_C)$ with proportionality factors $\boldsymbol{\pi}$. We define the* exact factorisation of $\mathscr{T}$ with median free term *as $\mathscr{T} = \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x) \times \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi})$, where the free term, $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x)$, is computed as the tree obtained from $\mathscr{T}$ by replacing $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ by $\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ and replacing $\mathscr{T}^{R(\mathbf{X}_D=\mathbf{x}_D)}$ by a 1 for every context $(\mathbf{X}_D = \mathbf{x}_D)$ incompatible with $(\mathbf{X}_C = \mathbf{x}_C)$, and where $\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ is the tree given by*

$$\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C)} = \begin{cases} \mathscr{T}^{(\frac{n+1}{2})}, & \text{if } n \text{ is odd;} \\ \frac{1}{2}(\mathscr{T}^{(\frac{n}{2})} + \mathscr{T}^{(\frac{n}{2}+1)}), & \text{if } n \text{ is even,} \end{cases} \tag{3}$$

with $n = |\Omega_X|$, and for each $x_j \in \Omega_X$, $\mathscr{T}^j$ represents the subtree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)}$ and $\{\mathscr{T}^{(j)} : x_j \in \Omega_X\}$ is the ordered sequence of the trees $\{\mathscr{T}^j : x_j \in \Omega_X\}$[1].

The core term, $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi})$, is the tree obtained from $\mathscr{T}$ by replacing each tree $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)}$ by a constant π_j given by:

$$\pi_j = \begin{cases} \dfrac{1}{\pi_{kj}} = \pi_{jk}, & \text{if } n \text{ odd; } \mathscr{T}^k = \mathscr{T}^{(\frac{n+1}{2})} \\ \dfrac{2}{\pi_{kj} + \pi_{(k+1)j}} = \dfrac{2\pi_{jk}\pi_{j(k+1)}}{\pi_{jk} + \pi_{j(k+1)}}, & \text{if } n \text{ even, } \mathscr{T}^k = \mathscr{T}^{(\frac{n}{2})}. \end{cases} \tag{4}$$

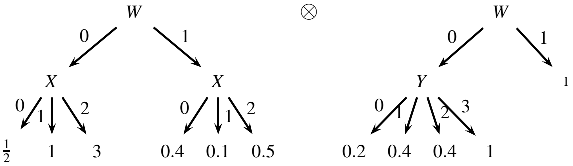


Fig. 3 Median factorisation of the tree in Fig. 1 .

Remark 1. In general, not every set of trees can be ordered. In this case, we define the median tree, $\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C)}$, of a set of trees, $\{\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j)} : x_j \in \Omega_X\}$, with the same structure as a tree with the same structure and each leaf, $\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C, l_i)}$, is the median of the respective leaves of the subtrees, that is,

$$\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C, l_i)} = Median\left\{\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C, X=x_j, l_i)}, x_j \in \Omega_X\right\} \tag{5}$$

[1] A set of proportional trees can be fully ordered, and therefore the median tree given by 3 is well defined.

It is immediately noticed that, in the case of proportional trees, both definitions given by (3) and (5) coincide.

Theorem 1. *Let $\mathscr{T}$ be a probability tree proportional below X for context $(\mathbf{X}_C = \mathbf{x}_C)$ such that $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C,X=x_i)} = \pi_{ij} \cdot \mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C,X=x_j)} \quad x_i,x_j \in \Omega_X$. Then, the tree $\mathscr{T}_{Me}^{R(\mathbf{X}_C=\mathbf{x}_C)}$ defined as in (3) is proportional to all the factors $\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C,X=x_i)}$ and it holds that*

$$\mathscr{T} = \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi}) \times \mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x) \tag{6}$$

where $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x, \boldsymbol{\pi})$ and $\mathscr{T}(\mathbf{X}_C = \mathbf{x}_C, X = x)$ are as in Definition 1.

3 Approximate Factorisation

The problem of approximate factorisation can be stated as follows. Let $\mathscr{T}_1$ and $\mathscr{T}_2$ two sub-trees which are siblings for a given context (i.e. both sub-trees are children of the same node), such that both have the same size and their leaves contain only positive numbers. The goal of the *approximate factorisation* is to find a tree $\mathscr{T}_2^*$ with the same structure than $\mathscr{T}_2$, such that $\mathscr{T}_2^*$ and $\mathscr{T}_1$ become proportional, under the restriction that the potential represented by $\mathscr{T}_2^*$ must be as close as possible to the one represented by $\mathscr{T}_2$. Then, $\mathscr{T}_2$ can be replaced by $\mathscr{T}_2^*$ and the resulting tree that contain $\mathscr{T}_1$ and $\mathscr{T}_2$ can be decomposed, as it would become proportional or partially proportional for the given context.

Two main issues are the determination of the proportionality factor, π, and measuring the accuracy of the approximation. Both are connected, as in general, different divergence measures would result in different values for π.

The problem of approximate factorisation was introduced in [6], in which the formulae for computing the proportionality factors according to several divergence measures were given. A probability tree $\mathscr{T}$ is δ*-factorisable* within context $(\mathbf{X}_C = \mathbf{x}_C)$, with proportionality factors $\boldsymbol{\pi}$ with respect to a divergence measure $\mathscr{D}$ if for each $x_j, x_i \in \Omega_X$ $\exists \pi_{ji} > 0$ such that $\mathscr{D}(\mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C,X=x_j)}, \pi_{ji} \cdot \mathscr{T}^{R(\mathbf{X}_C=\mathbf{x}_C,X=x_i)}) \leq \delta$. Parameter $\delta > 0$ is called *tolerance of the approximation*. Observe that proportional and partially proportional trees for context $(\mathbf{X}_C = \mathbf{x}_C)$ are δ-factorisable, with $\delta = 0$.

For carrying out the approximate factorisation, in this paper we follow the methodology established in [7] and factorise with respect to the median free term. We have found two ways of factorising in an approximate:

Strategy 1: Given a tree $\mathscr{T}$ δ-factorisable below X for context $(\mathbf{X}_C = \mathbf{x}_C)$, with respect to a divergence measure $\mathscr{D}$, the approximately proportional sub-trees are computed according to the chosen method, and the exact factorisation with median free term is applied. We shall refer to this strategy as **approximate factorisation with median free term (MAF)**.

Strategy 2: Given a tree $\mathscr{T}$, δ-factorisable below X for context $(\mathbf{X}_C = \mathbf{x}_C)$, an median tree is computed from the original factors, and the constants of

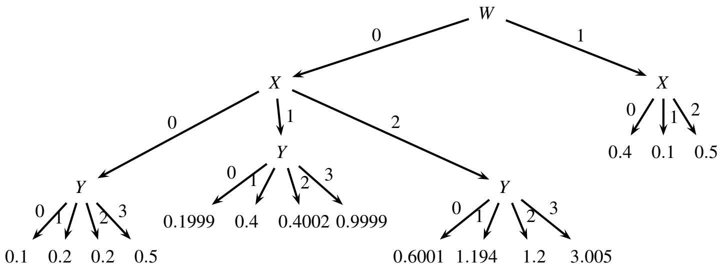


Fig. 4 A probability tree δ-factorisable below X for context $W = 0$.

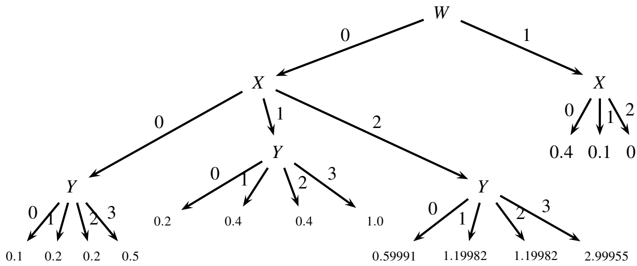


Fig. 5 Approximation of the sub-trees below X for context $W = 0$ by the invariant potential method, for which the proportionality factors are 1, 2 and 5.9991.

the core term are obtained by any approximation method, with respect to the average tree. We shall refer to this strategy as **approximate direct factorisation with median free term (MDAF)**.

Fig. 4 shows a probability tree δ-factorisable below X for context $W = 0$. Fig. 5 and 6 show the sub-trees of X for context $W = 0$ approximately proportional according to the invariant potential method, and the approximate factorisation resulting by applying Strategy 1 to the tree in Fig. 4.

Fig. 7 shows the median tree of the sub-trees δ-proportional for context $W = 0$ from the tree in Fig. 4. Fig. 8 shows the approximate direct

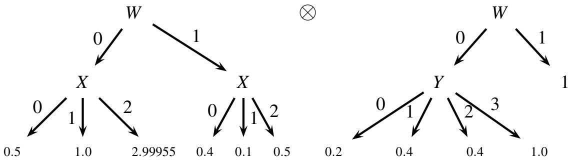


Fig. 6 Approximate factorisation with median free term of the tree in Fig. 4 according to the invariant potential method.

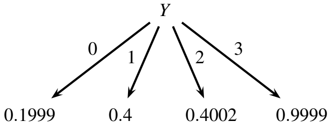


Fig. 7 Median tree of the sub-trees δ-proportional below X for context $W = 0$ in the tree in Fig. 4.

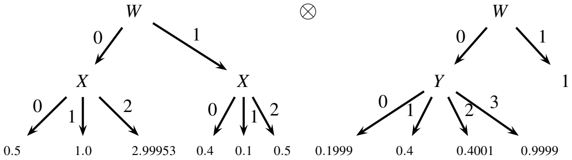


Fig. 8 Approximate direct factorisation with median free term of the tree in Fig. 4 using the method of minimum χ^2 divergence.

Table 1 Proportion (%) of experiments where MDAF is more accurate. Scenario (i).

Method	MAXD	MAD	D_χ	ND_χ	MSE	WMSE	Hellinger
Inv Pot	71.5	68.1	71	71	69.2	68.7	70.6
Min D_χ	71.3	68.1	70.9	70.9	69.1	68.5	70.1
Min MSE	71.3	68.1	70.6	70.6	68.9	68.3	70.4
Min WMSE	71.4	68.2	70.7	70.8	69.3	68.7	70.7
Null DKL	71.8	67.7	70.6	70.7	69	68.6	70.3
Wa	71.4	68	70.6	70.7	68.9	68.4	70.2
Min Hell	71.2	67.7	70.6	70.8	69	68.6	70.5

factorisation with median free term of the tree in Fig. 4 using the method of minimum χ^2 divergence.

We have carried out a simulation over randomly generated trees with various features, and in each case, we have annotated which factorisation strategy (approximate factorisation or approximate direct factorisation) provides the best results. More precisely, we have conducted 10000 runs, each one with a set of m sub-trees (m generated at random between 2 and 102) with n leaves (n generated at random between 2 and 52), considering three scenarios: (i) Each leaf with a random real number between 0 and 10. Table 1 shows the proportion of runs in which the error of the direct approximation is lower, for the different divergence measures considered. (ii) The leaves in each tree, instead of containing random numbers, contain real numbers in increasing order (Table 2). (iii) The leaves are generated in such a way that the resulting trees are δ-factorisables with $\delta = 0.01$. (Tab. 3).

Table 2 Proportion (%) of experiments where MDAF is more accurate. Scenario (ii).

Method	MAXD	MAD	D_χ	ND_χ	MSE	WMSE	Hellinger
Inv Pot	99.7	99.9	98.5	98.5	99.9	99.9	99.9
Min D_χ	99.7	99.9	99.9	99.9	99.9	100	99.9
Min MSE	99.3	99.9	97.9	97.9	99.9	99.9	99.9
Min WMSE	99.4	99.9	95.2	95.2	99.9	99.9	99.9
Null DKL	99.7	99.9	99.5	99.5	99.9	99.9	99.9
Wa	99.8	99.9	99.9	99.9	99.9	99.9	99.9
Min Hell	99.6	99.9	98.6	98.6	99.9	99.9	99.9

Table 3 Proportion (%) of experiments where MDAF is more accurate. Scenario (iii).

Method	MAXD	MAD	D_χ	ND_χ	MSE	WMSE	Hellinger
Inv Pot	38	68.6	75.8	75.8	70.3	68.8	75.8
Min D_χ	37.9	68.6	75.8	75.8	70.3	68.7	75.8
Min MSE	34.7	61.6	76.8	76.8	64.4	43.8	76.8
Min WMSE	34.2	62	77.9	77.9	65	41.3	77.9
Null DKL	38	68.6	75.8	75.8	70.3	68.9	75.8
Wa	38	68.6	75.8	75.8	70.3	68.9	75.8
Min Hell	37.9	68.6	75.8	75.8	70.3	68.8	75.8

4 Conclusions

We have introduced a methodology for decomposing probability trees with respect to a median free term. Using the median can produce more robust decompositions that using the mean in some cases. Also, it can increase the accuracy of the decomposition if the mean absolute error is used as divergence measure.

Acknowledgements. This work has been supported by the Spanish Ministry of Science and Innovation, under project TIN2007-67418-C03-02 and by EFRD (FEDER) funds.

References

1. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Horvitz, E., Jensen, F. (eds.) Proceedings of the 12th Conf. on Uncertainty in Artificial Intelligence, UAI 1996, Portland, Oregon, USA, pp. 115–123. Morgan Kaufman Publishers, San Francisco (1996)
2. Cano, A., Moral, S.: Propagación exacta y aproximada con árboles de probabilidad. In: Actas De La VII Conferencia De La Asociación Española Para La Inteligencia Artificial, CAEPIA 1997, Málaga, Spain, pp. 635–644 (1997)
3. Cano, A., Moral, S., Salmerón, A.: Penniless propagation in join trees. Internat. J. Intell. Syst. 15, 1027–1059 (2000)
4. Cano, A., Moral, S., Salmerón, A.: Lazy evaluation in Penniless propagation over join trees. Networks 39, 175–185 (2002)

5. Madsen, A., Jensen, F.: Lazy propagation: a junction tree inference algorithm based on lazy evaluation. Artificial Intelligence 113, 203–245 (1999)
6. Martínez, I., Moral, S., Rodríguez, C., Salmerón, A.: Approximate factorisation of probability trees. In: Godo, L. (ed.) ECSQARU 2005. LNCS (LNAI), vol. 3571, pp. 51–62. Springer, Heidelberg (2005)
7. Martínez, I., Moral, S., Rodríguez, C., Salmerón, A.: Approximate decomposition of probabilistic potentials in Bayesian networks. Technical Report (2010)
8. Salmerón, A., Cano, A., Moral, S.: Importance sampling in Bayesian networks using probability trees. Comput. Statist. Data Anal. 34, 387–413 (2000)

Comparison of Random Variables Coupled by Archimedean Copulas

I. Montes, D. Martinetti, S. Díaz, and S. Montes

Abstract. Random variables are used to model many processes in real life and in many cases we have to choose among those processes. The usual way to compare random variables is the classical stochastic dominance. One drawback of stochastic orders is that it is not always possible (or easy) to use them, arising then the need for alternative models in many situations. A new model has been recently been developed and it was called statistical preference. This method allows to compare every pair of alternatives. It also takes into consideration the possible dependence between the alternatives. In this contribution we analyze the possible relationship between statistical preference and stochastic dominance of continuous random variables. We focus on random variables whose joint cumulative distribution function is obtained by Archimedean copulas.

1 Introduction

Stochastic dominance is used in investment decision-making under uncertainty. In economics, investments usually involve some risk. Then, random variables are employed to model these assets and therefore a criterion to choose among those random variables is necessary. In decision-making the comparison is carried out by pairs. Given two random variables, stochastic dominance is the most usual way to compare them. It has been widely applied and also studied in depth (see among others [4, 9]). However, it is not exhaustive. It does not always allow to fix a preferred random variable. There are pairs of random variables for which no one of them stochastically dominates the other one. In addition to this, stochastic dominance is not easy to

I. Montes, D. Martinetti, S. Díaz, and S. Montes
Dept. Statistics and O. R., University of Oviedo, 33007 Oviedo, Spain
e-mail: imontes@spi.uniovi.es,
martinettidavide,diazsusana,montes@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 467–474.
springerlink.com

be computed in some cases. And it does not take into account the possible dependence between the random variables confronted.

Statistical preference is a new way of comparing random variables. It is obtained departing from a probabilistic relation defined over the pair of random variables. Let us recall that a probabilistic relation defined over a set of alternatives is a binary relation that takes values in the unit interval and such that the sum of the values of the relation acting over any pair of elements and over the transposed pair is 1. De Schuymer et al. (see for example [2, 3]) defined a probabilistic relation over a set of random variables. And from such a relation, statistical preference was defined.

The first natural step when a new definition appears is to confront it with the classical definitions. In this work we confront stochastic dominance and statistical preference. We do not restrict our study to independence, but we admit possible dependence between the random variables. In fact, we will consider the general framework of random variables coupled by means of Archimedean copulas (see [10]).

The work is organized as follows: after this introduction, in Section 2 we give some previous notions. First of all we recall the different types of stochastic dominance. We also recall the notion of copula. Section 3 is devoted to statistical preference. In Section 4 we compare statistical preference and stochastic dominance. Finally, in Section 5 we provide some conclusions.

2 Previous Notions

A classical way to compare random variables is the stochastic dominance. In this section we recall the definition of stochastic dominance of any degree, but we pay special attention to the two most usual definitions: first and second degree stochastic dominance.
After that we recall the notion of copula, that plays an important role in this contribution.

2.1 Stochastic Dominance

The best known way to compare random variables is the first degree stochastic dominance.

Definition 1. *A random variable with cumulative distribution function F_X* stochastically dominates in first degree *a random variable Y with cumulative distribution function F_Y, denoted as $X \geq_{FSD} Y$, if for all real t it holds that $F_X(t) \leq F_Y(t)$.*

This definition is quite restrictive: for many pairs of random variables no one of them stochastically dominates the other one in first degree.

This lead to less restrictive definitions of stochastic comparison.

Definition 2. *A random variable with cumulative distribution function F_X* stochastically dominates in second degree *a random variable Y with cumulative distribution function F_Y, denoted as $X \geq_{SSD} Y$ if it holds that*

$$\int_{-\infty}^{x} F_X(t)dt \leq \int_{-\infty}^{x} F_Y(t)dt$$

for all $x \in \mathbb{R}$.

Both types of dominance can be characterized by the mean of the variables confronted (see for example [4]). It holds that $X \geq_{FSD} Y$ if and only if $E[u(X)] \geq E[u(Y)]$ for every non-decreasing function u. It holds that $X \geq_{SSD} Y$ if and only if $E[u(X)] \geq E[u(Y)]$ for every non-decreasing concave function u. It is easy to realize from this result that first degree stochastic dominance is a stronger condition than second degree stochastic dominance.

Despite first and second degree stochastic dominances are the most employed ones, stochastic dominance can be defined for any degree n.

Definition 3. *Let X and Y be two random variables with cumulative distribution functions F_X and F_Y, respectively. Let $F_X^1 = F_X$, $F_Y^1 = F_Y$ and recursively define*

$$F_X^{n+1}(x) = \int_{-\infty}^{x} F_X^n(t)dt, \quad F_Y^{n+1}(y) = \int_{-\infty}^{y} F_Y^n(t)dt, \qquad \text{for } n \in \{1,2\ldots\}.$$

Random variable X stochastically dominates in n^{th} degree *Y, denoted as $X \geq_{nSD} Y$, if it holds that*

$$F_X^n(x) \leq F_Y^n(x), \text{ for all } x \in \mathbb{R}.$$

In general it holds that for any pair of random variables X and Y,

$$X \geq_{nSD} Y \quad \Rightarrow \quad X \geq_{mSD} Y, \qquad \forall n \leq m.$$

Example 1. Thus, consider the random variables X and Y with uniform distribution $U(3,5)$ and $U(1,4)$, respectively. It is obvious that $X \geq_{FSD} Y$, and therefore $X \geq_{nSD} Y$ for all $n \in \{1,2,\ldots\}$.

2.2 Copulas

Stochastic dominance does not take into account the possible dependence between the random variables compared. It is well known that when the random variables are independent, the joint cumulative distribution function $F_{X,Y}$ is obtained as the product of the marginal cumulative distribution functions

$$F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y).$$

It is also known that in general, the joint cumulative distribution function of two random variables X and Y, can be expressed by a copula whose arguments are the marginal distribution functions of the original variables (Sklar's theorem).

Definition 4. *A* copula *is an operator* $C : [0,1]^2 \to [0,1]$ *satisfying*

- $C(x,0) = C(0,x) = 0$ *for all* $x \in [0,1]$,
- $C(x,1) = C(1,x) = x$ *for all* $x \in [0,1]$,
- *the property of moderate growth:*

  $$C(x_1,y_1) + C(x_2,y_2) \geq C(x_1,y_2) + C(x_2,y_1)$$

  for every $(x_1,x_2,y_1,y_2) \in [0,1]^4$ *such that* $x_1 \leq x_2$ *and* $y_1 \leq y_2$.

By the Fréchet-Hoeffding bounds inequality, for every copula C and every $(x,y) \in [0,1]^2$,

$$W(x,y) \leq C(x,y) \leq M(x,y),$$

where the first one, M, is the minimum operator ($M(x,y) = \min(x,y)$) and the second one, W, represents the Łukasiewicz operator ($W(x,y) = \max(x+y-1,0)$). M is also referred to as the Fréchet-Hoeffding upper bound and W as the Fréchet-Hoeffding lower bound. A third important copula is the product copula, $\Pi(x,y) = x \cdot y$. As commented before, the product is the copula employed when the random variables are independent. In the other two cases, the random variables are called comonotonic and countermonotonic, respectively (see, for instance, [1, 11]).

These three copulas are also particular cases of Archimedean copulas.

Definition 5. *An* Archimedean copula *is a function* C *from* $[0,1]^2$ *to* $[0,1]$ *given by*

$$C(x,y) = \varphi^{[-1]}(\varphi(x) + \varphi(y))$$

where φ *(the generator of* C*) is a continuous, strictly decreasing function from* $[0,1]$ *to* $[0,\infty]$ *such that* $\varphi(1) = 0$ *and where* $\varphi^{[-1]}$ *denotes the pseudo-inverse of* φ*:*

$$\varphi^{[-1]}(t) = \begin{cases} \varphi^{-1}(t), & 0 \leq t \leq \varphi(0), \\ 0, & \varphi(0) \leq t \leq \infty. \end{cases}$$

3 Statistical Preference

As commented in the introduction, stochastic dominance leads to a partial order in a set of univariate random variables. De Schuymer et al. [2, 3] introduced a different way of stochastic comparison, that allows to order any pair of random variables. They defined a probabilistic relation D over the pairs of random variables (X,Y) as follows:

$$D(X,Y) = \Pr(X > Y) + \frac{1}{2}\Pr(X = Y)$$

It clearly holds that D is a probabilistic relation, that is, it takes values in $[0,1]$ and $D(X,Y)+D(Y,X)=1$ for every pair of random variables X and Y.

D can be seen as a graded variant of stochastic orders because it allow us to give degrees of preference. However, sometimes it is necessary to dissolve the fuzziness inside this relation and obtain a crisp comparison for random variables from it. It is very simple by means of the 1/2-cut of this relation.

Definition 6. [2] *Given two random variables X and Y, the variable X is* statistically preferred *to Y if $D(X,Y) \geq \frac{1}{2}$. We will denote it $X \geq_{SP} Y$.*

As a consequence of this definition, we say that two random variables X and Y are statistically indifferent if $D(X,Y) = \frac{1}{2}$.

Example 2. Let us consider X and Y uniformly distributed on the intervals $[0,20]$ and $[1,4]$, respectively. It is easy to check that there are not first degree stochastic dominance neither second degree stochastic dominance. Thus, they are not comparable with that method. However, X intuitively seems (the most of the times) greater than Y. We have that $D(X,Y) = \frac{7}{8}$ and therefore $X \geq_{SP} Y$. Thus, we have found finally a way to compare X and Y.

Thus, statistical preference is a relation on a set of random variables which avoids the pointwise comparison of performance functions. It also avoids specific reference points and it allows us to induce a total order relation on the set of random variables, which is called statistical preference. Another problem of stochastic dominance is that it does not take into account the possible relationship between the random variables being compared. Statistical preference does consider the connection between the random variables. By the Sklar's theorem, the statistical preference of one random variable X over another one Y depends on the copula that connects them.

4 Statistical Preference versus Stochastic Dominance

We have already studied in detail the connection between statistical preference and first and second degree stochastic dominance for independent random variables [6]. The case of comonotonicity and countermonotonicity was applied for comparing two fitness values of two knowledge bases [7].

Now we consider a more general case: we assume that the random variables can be coupled by any Archimedean copula. In that case, for any pair of continuous random variables, we have proven that first degree stochastic dominance is a stronger condition than statistical preference. Due to space limitation, we will not detail the proof of the following results.

Theorem 1. *Let X and Y be two continuous random variables. Let C be an archimedean copula and $F_{X,Y}(x,y) = C(F_X(x), F_Y(y))$. Then*

$$X \geq_{FSD} Y \quad \Rightarrow \quad X^Y.$$

Thus, this theorem is a more general than the given by De Schuymer et al. [2], where it was presented for the particular case of independent random variables, although only proven for the continuous case. The proof for the discrete case is a little more complicated but the result still holds.

Remark 1. The converse of previous theorem does not hold. Even more, we can find a pair of random variables X and Y and a copula C such that:

- X and Y are coupled by C.
- X is statistically preferred to Y.
- There does not exist an integer number n, with $n \geq 1$, such that X stochastically dominates in n^{th} degree Y.

For example, it is sufficient to consider $X \equiv U(0,10)$ and $Y \equiv U(1,2)$. Then it holds that $Q(X,Y) > \frac{1}{2}$ when X and Y are independent, comonotonic or countermonotonic. Therefore, in those cases X^Y. However, it is not possible that $X \geq_{nSD} Y$ for any degree n.

Then, statistical preference is not as restrictive as first degree stochastic dominance in general. However, for some particular probability distributions they are equivalent [6].

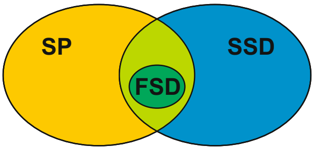
From Theorem 1, we could think that any stochastic order implies statistical preference. However this implication is not fulfilled for any other stochastic order different from the first degree order. This is a consequence that second degree stochastic dominance does not guarantee the statistical preference when dealing with independent random variables (see [6]) and this is a stronger condition than stochastic dominance of order n for $n \geq 3$. Thus, for any $n \geq 2$, there exist two continuous random variables X and Y and an Archimedean copula C such that:

- X and Y are coupled by C.
- X stochastically dominates in n^{th} degree Y.
- X is not statistically preferred to Y.

Apart from Theorem 1, we have proven that no other implication among n-degree stochastic dominance and statistical preference holds in general. Thus, for instance, we know that second degree stochastic dominance and statistical preference are not connected in general. We also know that each of them by itself is not enough to obtain first degree stochastic dominance. We could prove that both conditions together neither guarantee first degree stochastic dominance. For these reason, we can find two continuous random variables X and Y and an Archimedean copula C such that:

- X and Y are coupled by C.
- X is statistically preferred to Y.
- X stochastically dominates in second degree Y
- X does not stochastically dominate in first degree Y.

The previous results can be graphically summarized:



As a completion of the study presented in [5], we have conducted an experiment in order to investigate if these relationships are natural for the human reasoning and we have found very supporting results.

In the case of independent continuous random variables, it was also proven that $X \geq_{SP} Y$ is equivalent to $E_X(F_X) \leq E_X(F_Y)$, where $E_X(\cdot)$ denotes the expectation with respect to the variable X. From here, Theorem 1 is trivial in the particular case of independence. We have generalized that result for archimedean copulas:

Theorem 2. *Let X and Y be two continuous random variables and let C be an archimedean copula with generator φ. Then X^Y if and only if*

$$E_X\left(\left[\left(\varphi^{[-1]}\right)'(\varphi(F_X(x))+\varphi(F_Y(x)))-\left(\varphi^{[-1]}\right)'(2\varphi(F_X(x)))\right]\varphi'(F_X(x))\right)\geq 0.$$

Remark 2. If X and Y are independent, $\varphi(x) = -\ln(x)$, and simplifying in the previous expression we obtain that that condition is equivalent to $E_X(F_X) \leq E_X(F_Y)$.

Thus, we have obtained a new characterization of the statistical preference for continuous random variables in the particular case they are coupled by means of Archimedean copulas. In general, we had obtained (see [8]) the following equivalence: $X^Y \Longleftrightarrow Me_{X-Y} \geq 0$.

5 Conclusion

Statistical preference is a new way of comparing random variables. It seems to be a clear alternative to classical stochastic dominance. It takes into account the possible dependence between the random variables compared, and it allows to order any pair of random variables. In this contribution we study the connection between statistical preference and stochastic dominance. If the random variables to be compared are independent, it is known that statistical preference is a weaker condition than first degree stochastic dominance. In this contribution we have considered pairs of random variables whose joint cumulative distribution function is obtained by an archimedean copula acting over the marginal distribution functions of the random variables. We have compared first degree stochastic dominance and statistical preference and we have proven that in any case, first degree stochastic dominance is a stricter way of comparing continuous random variables than statistical preference. We expect an analogous result for discrete random variables and we are already working on it.

Acknowledgements. The research reported on in this paper has been partially supported by Project FEDER-MEC-MTM2007-61193.

References

1. De Meyer, H., De Baets, B., De Schuymer, B.: On the transitivity of the comonotonic and countermonotonic comparison of random variables. J. Multivariate Anal. 98, 177–193 (2007)
2. De Schuymer, B., De Meyer, H., De Baets, B.: Cycle-transitive comparison of independent random variables. J. Multivariate Anal. 96, 352–373 (2005)
3. De Schuymer, B., De Meyer, H., De Baets, B., Jenei, S.: On the cycle-transitivity of the dice model. Theory Decis. 54, 261–285 (2003)
4. Levy, H.: Stochastic dominance: Investment Decision Making Under Uncertainty. Kluwer Academic Publishers, Boston (1998)
5. Levy, M., Levy, H.: Testing for risk aversion: a stochastic dominance approach. Econom. Lett. 71, 233–240 (2001)
6. Montes, I., Martinetti, D., Díaz, S.: On the Statistical Preference as a Pairwise Comparison for Random Variables. In: Proceedings of the EUROFUSE Workshop on Preference Modelling and Decision Analysis, Pamplona, Spain (2009)
7. Montes, I., Martinetti, D., Díaz, S., Montes, S.: Statistical preference as a comparison method of two imprecise fitness values. In: Proceedings of the XV Congreso Español sobre Tecnologías y Lógica Fuzzy, ESTYLF 2010, Huelva, Spain (2010)
8. Montes, S., Martinetti, D., Montes, I., Díaz, S.: Min-transitivity of graded comparisons for random variables. In: Proceedings of the 2010 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2010, Barcelona, Spain (2010)
9. Müller, A., Stoyan, D.: Comparison Methods for Stochastic Models and Risks. J. Wiley and Sons, Chichester (2002)
10. Nelsen, R.: An Introduction to Copulas. Lecture Notes in Statistics, vol. 139. Springer, New York (2006)
11. Yang, J., Cheng, S., Zhang, L.: Bivariate copula decomposition in terms of comonotonicity, countermonotonicity and independence. Insurance Math. Econom. 39, 267–284 (2006)

Two-Way Analysis of Variance for Interval-Valued Data

Takehiko Nakama, Ana Colubi, and M. Asunción Lubiano

Abstract. We establish two-way analysis of variance (ANOVA) for interval-valued data. Each observation is assumed to be a compact convex interval, and the two-way ANOVA determines whether to reject null hypotheses about the effects of two factors on the observed intervals. The Minkowski support function is used to obtain a metric for intervals and to transform them to Hilbert-space-valued functions. We derive test statistics that are appropriate for testing the null hypotheses, and we develop a bootstrap scheme for approximating the p-values of the observed test statistics.

1 Introduction and Summary

A wide variety of statistical procedures have been established for analyzing data in which each observation is nominal, ordinal, or numerical. However, there are many practical problems that require dealing with observations that represent inherently imprecise, uncertain, or linguistic characteristics. In such cases, intervals and fuzzy sets are more effective in encoding observations. For instance, if one reports a perceived length of an object in a perceptual study, imprecise responses such as “between 6 and 10” or “about 8” may reflect the perceived length better than real-valued responses.

Various statistical tests have been developed to quantitatively analyze fuzzy data. For instance, Körner [10], Montenegro et al. [14], and González-Rodríguez et al. [9] developed one-sample methods for hypothesis testing

Takehiko Nakama
European Center for Soft Computing, Edificio Científico-Technológico,
33600 Mieres, Spain
e-mail: nakama@jhu.edu

Ana Colubi and M. Asunción Lubiano
Departamento de Estadística, I.O. y D.M., Universidad de Oviedo,
33007 Oviedo, Spain
e-mail: colubi,lubiano@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 475–482.
springerlink.com

about the fuzzy population mean. Montenegro et al. [13] and González-Rodríguez et al. [8] established a two-independent-sample test of equality of fuzzy means, and González-Rodríguez et al. [8] developed a paired-sample test of the same type. These are considered extensions of classical t tests to fuzzy data. Gil et al. [5] and Corral et al. [4] developed a multiple-sample test of equality of fuzzy means; this is one-way analysis of variance (ANOVA) for fuzzy data.

In this paper, we establish two-way ANOVA for interval-valued data. A two-way layout is designed to examine the effects of two factors, which each involve at least two levels. This experimental design offers two main advantages compared to combining two one-way layouts in testing the effects of two factors. Conducting one two-way-layout experiment is usually more cost-effective than conducting two one-way-layout experiments since the two-way layout uses the same observations to compare the levels of one factor as are used to compare the levels of the other factor. The other advantage is that the two-way layout allows us to examine the effect of interaction between the two factors. There are cases where the interaction between the two factors is significant even though the main effect of each factor is not significant. Thus the two-way layout is an important experimental design, and two-way ANOVA is an essential technique for analyzing the resulting observations.

The rest of this paper is organized as follows. In Section 2.1, we describe fundamentals of interval arithmetic. As in most of the previous studies that developed statistical methods for fuzzy data, we consider intervals in $\mathbb{R}^d$ that are compact and convex. We use the Minkowski support function to establish a metric for intervals and to transform them to Hilbert-space-valued functions; see Section 2.2. As described in Section 2.3, this approach allows us to derive convergence results for random intervals from the strong law of large numbers and central limit theorems for Hilbert-space-valued random variables. Two-way ANOVA for interval-valued data is formulated in Section 3. We describe test statistics that can be used to test null hypotheses about the main effect of each factor and about the interaction between the two factors. We develop a bootstrap scheme for approximating the distributions of the test statistics and their p-values.

Space limitations on this paper force us to omit proofs of our theorems presented in this paper. We will provide them in our full-length paper.

2 Preliminaries

2.1 Interval Arithmetic

As mentioned in Section 1, we consider intervals in $\mathbb{R}^d$ that are compact and convex. Let $\mathscr{K}_c(\mathbb{R}^d)$ denote the set of such intervals. Addition and scalar multiplication are the basic arithmetic operations on intervals. For $A, B \in \mathscr{K}_c(\mathbb{R}^d)$, the (Minkowski) addition of A and B is defined by

$$A+B := \{a+b \,|\, a \in A,\ b \in B\}.$$

For $\lambda \in \mathbb{R}$, we define $\lambda A := \{\lambda a \,|\, a \in A\}$. Notice that in general, $A - A \neq 0$, where 0 denotes the additive identity (i.e., the interval that contains only 0); except for 0, each element in $\mathscr{K}_c(\mathbb{R}^d)$ has no additive inverse.

2.2 Metric and Transformation of Intervals to Functions

We will follow the approach introduced by Puri and Ralescu [15, 16, 17] and map the space of $\mathscr{K}_c(\mathbb{R}^d)$ described in Section 2.1 to a closed convex cone of a Hilbert space by means of the (Minkowski) support function. Let $\mathbb{S}^{d-1}$ denote the unit sphere in $\mathbb{R}^d$. For each $A \in \mathscr{K}_c(\mathbb{R}^d)$, the support function s_A of A is defined by

$$s_A(u) := \sup_{a \in A} \langle a, u \rangle \qquad \forall\ u \in \mathbb{S}^{d-1}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$. The interval

$$[-s_A(-u),\ s_A(u)]$$

is the projection of A onto the line spanned by u.

With this support function, we can define a metric for intervals in $\mathscr{K}_c(\mathbb{R}^d)$. Several metrics can be used in our study. For concreteness, we use a family of metrics recently proposed by Trutschnig et al. [18]. Let $\mathscr{H}_s$ denote the space of all support functions that can be derived from $\mathscr{K}_c(\mathbb{R}^d)$. For each $f \in \mathscr{H}_s$, define $\operatorname{mid} f(u) := \frac{f(u)-f(-u)}{2}$ and $\operatorname{spr} f(u) := \frac{f(u)+f(-u)}{2}$; thus $\operatorname{mid} f(u)$ and $\operatorname{spr} f(u)$ denote the middle point and half of the length of the interval $[-f(-u),\ f(u)]$, respectively. Let $\lambda_{\mathbb{S}^{d-1}}$ denote the Lebesgue measure on $\mathbb{S}^{d-1}$. For each $f, g \in \mathscr{H}_s$, define

$$\langle f, g \rangle_\theta := [\![\operatorname{mid} f, \operatorname{mid} g]\!] + \theta [\![\operatorname{spr} f, \operatorname{spr} g]\!], \tag{2}$$

where $\theta \in (0,1]$ and

$$[\![f, g]\!] := \int_{u \in \mathbb{S}^{d-1}} f(u) g(u) d\lambda_{\mathbb{S}^{d-1}}(u).$$

The parameter θ specifies the relative importance of $[\![\operatorname{spr} f, -\operatorname{spr} g]\!]$ compared to $[\![\operatorname{mid} f, -\operatorname{mid} g]\!]$. We define the metric D_θ for intervals in $\mathscr{K}_c(\mathbb{R}^d)$ by

$$D_\theta(A, B) := ||s_A - s_B||_\theta, \tag{3}$$

where

$$||s_A - s_B||_\theta := \sqrt{\langle s_A - s_B, s_A - s_B \rangle_\theta}. \tag{4}$$

This is an extension of the metric introduced by Bertoluzza, Corral, and Salas [1] for intervals in $\mathscr{K}_c(\mathbb{R})$

Let $\mathscr{H} := \{f : \mathbb{S}^{d-1} \to \mathbb{R} \,|\, ||f||_\theta < \infty\}$. Then (2) is an inner product for $\mathscr{H}$, and the resulting inner product space is a Hilbert space. The support function (1) isometrically embeds the space $\mathscr{K}_c(\mathbb{R}^d)$ in a closed convex cone of the Hilbert space. We will obtain important results for our two-way ANOVA based on this isometric embedding.

2.3 Random Intervals and Hilbert-Space-Valued Random Variables

As in previous studies (see, for instance, Gil et al. [6]), random intervals will be considered compact convex random sets as follows. Given a probability space $(\Omega, \mathscr{F}, P)$, we consider a measurable space $(\mathscr{K}_c(\mathbb{R}^d), \mathscr{F}_{\mathscr{K}})$ where $\mathscr{F}_{\mathscr{K}}$ denotes the σ-field in $\mathscr{K}_c(\mathbb{R}^d)$ generated by the topology induced by any metric that belongs to the family defined at (3). We treat each random interval as a mapping $\mathscr{X}$ that is measurable $\mathscr{F}/\mathscr{F}_{\mathscr{K}}$. Fundamental probabilistic notions for real-valued random variables, such as distributions and independence of random variables, remain the same for these random intervals.

Regarding the Hilbert space $\mathscr{H}$ described in Section 2.2, consider a σ-field $\mathscr{F}_{\mathscr{H}}$ in $\mathscr{H}$ generated by the topology induced by the metric

$$D_\theta^{\mathscr{H}}(f,g) := ||f-g||_\theta,$$

where $||\cdot||_\theta$ is defined at (4). Let $\mathscr{X}'$ denote a mapping that is measurable $\mathscr{F}_{\mathscr{K}}/\mathscr{F}_{\mathscr{H}}$. Then the composition $\mathscr{X}' \circ \mathscr{X}$ is measurable $\mathscr{F}/\mathscr{F}_{\mathscr{H}}$, and this is a Hilbert-space-valued random variable. The strong law of large numbers and central limit theorems exist for Hilbert-space-valued random variables (see, for example, Laha and Rohatgi [11], Ledoux and Talagrand [12], Colubi et al. [3]). Using the isometric embedding described in Section 2.2, we can thus derive various convergence results for random intervals from those for the corresponding Hilbert-space-valued random variables. In fact, the results described in Section 3 are obtained in this manner.

We define the expectation $\widetilde{E}\mathscr{X}$ of a random interval $\mathscr{X}$ to be its Aumann integral (see, for instance, Puri and Ralescu [17]):

$$\widetilde{E}\mathscr{X} := \left\{ \int X(\omega)dP(\omega) \,\middle|\, X : \Omega \to \mathbb{R}^d,\ X \in L^1(\Omega, \mathscr{F}, P),\ X \in \mathscr{X} \text{ a.e.} \right\}.$$

Using the metric (3), we define the variance of $\mathscr{X}$ by

$$\mathrm{Var}(\mathscr{X}) := E[D_\theta(\mathscr{X}, \widetilde{E}\mathscr{X})]^2.$$

3 Two-Way ANOVA for Random Intervals

Suppose that factors 1 and 2 have I and J levels, respectively. We let $\mathscr{X}_{ijk}$ denote the kth interval-valued observation under the ith level of factor 1 and the jth level of factor 2. We let n_{ij} denote the number of observations under this condition and let n denote the total number of observations: $n := \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$. In two-way ANOVA for these random intervals, we consider the following model:

$$\mathscr{X}_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}. \tag{5}$$

Here the parameters of the model, α_i, β_j, and δ_{ij} $(1 \le i \le I,\ 1 \le j \le J)$, are all compact and convex intervals (they all belong to $\mathscr{K}_c(\mathbb{R}^d)$), and ε_{ijk} denote random intervals, which are the random components of the model. We assume that $\varepsilon_{ijk} \in \mathscr{K}_c(\mathbb{R}^d)$ almost everywhere and that ε_{ijk} are independent and identically distributed with finite variance. In the model, α_i and β_j represent the main effects of factors 1 and 2, respectively, and δ_{ij} represent the interaction between the two factors. For all i, j, and k, let $\varepsilon := E\varepsilon_{ijk}$. Then we have

$$E\mathscr{X}_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon.$$

In order to determine the significance of the main effects and interaction, we test the following null hypotheses:

$$\begin{aligned} H_0^{(1)} &: \ \alpha_1 = \alpha_2 = \cdots = \alpha_I. \\ H_0^{(2)} &: \ \beta_1 = \beta_2 = \cdots = \beta_J. \\ H_0^{(1,2)} &: \delta_{1,1} = \delta_{1,2} = \cdots = \delta_{IJ}. \end{aligned}$$

First we describe how to test $H_0^{(1)}$. Define

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

$$\overline{\mathscr{X}}_{ij\cdot} := \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} \mathscr{X}_{ijk}, \quad T_n^{(1)} := \frac{\sum_{i=1}^{I}(\sum_{j=1}^{J} n_{ij})(D_\theta(\overline{\mathscr{X}}_{i\cdot\cdot}, \overline{\mathscr{X}}_{\cdots}))^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(D_\theta(\mathscr{X}_{ijk}, \overline{\mathscr{X}}_{ij\cdot}))^2}.$$

For each i and j, suppose that $n_{ij}/n \to p_{ij} > 0$ as $n \to \infty$. Then we can prove the following theorem:

Theorem 1. *If $H_0^{(1)}$ does not hold, then $P(T_n^{(1)} \le t) \to 0$ as $n \to \infty$ for any $t \in \mathbb{R}$.*

This theorem can be proved based on the corresponding results for the $\mathscr{H}$-valued random variables described in Section 2.3, which can be derived from the strong law of large numbers and central limit theorems for Hilbert-space-valued random variables. Therefore, if we let z_α denote the $100(1-\alpha)$ percentile of the distribution of $T_n^{(1)}$ under $H_0^{(1)}$, then we can reject $H_0^{(1)}$ with significance level α if the observed value of $T_n^{(1)}$ is greater than z_α.

The distribution of $T_n^{(1)}$ is unknown. However, the Giné-Zinn bootstrap central limit theorem (Giné and Zinn [7]) is applicable in this case due to the isometric embedding of $\mathscr{K}_c(\mathbb{R}^d)$ into a convex cone of $\mathscr{H}$ described in Section 2.2. Thus we establish a bootstrap scheme for approximating the distribution of $T_n^{(1)}$ and the value of z_α. For each i and j, define

$$\varGamma_{(-i,-j)} := \sum_{1\le i'\le I: i'\ne i} \; \sum_{1\le j'\le J: j'\ne j} \overline{\mathscr{X}}_{i'j'\cdot}\ ,$$

and consider the set

$$S_{ij} := \{\mathscr{X}_{ijk} + \varGamma_{(-i,-j)} \,|\, 1 \le k \le n_{ij}\}.$$

We can use $\bigcup_{i=1}^{I}\bigcup_{j=1}^{J} S_{ij}$ as the bootstrap population. For each i and for each j, we generate n_{ij} bootstrap observations $\mathscr{Y}_1^{(i,j)}, \mathscr{Y}_2^{(i,j)}, \ldots, \mathscr{Y}_{n_{ij}}^{(i,j)}$ by simple random sampling from S_{ij} and compute

$$\overline{\mathscr{Y}} := \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \mathscr{Y}_k^{(i,j)}, \quad \overline{\mathscr{Y}}i := \frac{1}{\sum_{j=1}^{J} n_{ij}} \sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \mathscr{Y}_k^{(i,j)},$$

$$\overline{\mathscr{Y}}ij := \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} \mathscr{Y}_k^{(i,j)}.$$

With these quantities, define

$$T_n^{*(1)} := \frac{\sum_{i=1}^{I}(\sum_{j=1}^{J} n_{ij})(D_\theta(\overline{\mathscr{Y}}i, \overline{\mathscr{Y}}))^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(D_\theta(\mathscr{Y}_k^{(i,j)}, \overline{\mathscr{Y}}ij))^2}.$$

Then the distribution of $T_n^{*(1)}$ is the bootstrap approximation of that of $T_n^{(1)}$ under $H_0^{(1)}$. Therefore, we generate a large number of realizations of $T_n^{*(1)}$ and reject $H_0^{(1)}$ with significance level α if $T_n^{(1)} > z_\alpha^*$, where z_α^* denotes the $100(1-\alpha)$ fractile of the realizations of $T_n^{*(1)}$.

We can test $H_0^{(2)}$ and $H_0^{(1,2)}$ analogously. We continue to assume that $n_{ij}/n \to p_{ij} > 0$ as $n \to \infty$ for each i and j. Define

$$\overline{\mathscr{X}}_{\cdot j \cdot} := \frac{1}{\sum_{i=1}^{I} n_{ij}} \sum_{i=1}^{I} \sum_{k=1}^{n_{ij}} \mathscr{X}_{ijk},$$

$$T_n^{(2)} := \frac{\sum_{j=1}^{J} (\sum_{i=1}^{I} n_{ij}) (D_\theta(\overline{\mathscr{X}}_{\cdot j \cdot}, \overline{\mathscr{X}}_{\cdots}))^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (D_\theta(\mathscr{X}_{ijk}, \overline{\mathscr{X}}_{ij\cdot}))^2}.$$

$$T_n^{(1,2)} := \frac{\sum_{j=1}^{J} \sum_{i=1}^{I} n_{ij} (D_\theta(\overline{\mathscr{X}}_{ij\cdot}, \overline{\mathscr{X}}_{\cdots}))^2}{\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (D_\theta(\mathscr{X}_{ijk}, \overline{\mathscr{X}}_{ij\cdot}))^2}.$$

Then we have the following theorems, which are analogous to Theorem 1:

Theorem 2. *If $H_0^{(2)}$ does not hold, then $P(T_n^{(2)} \leq t) \to 0$ as $n \to \infty$ for any $t \in \mathbb{R}$.*

Theorem 3. *If $H_0^{(1,2)}$ does not hold, then $P(T_n^{(1,2)} \leq t) \to 0$ as $n \to \infty$ for any $t \in \mathbb{R}$.*

Therefore, we can use $T_n^{(2)}$ and $T_n^{(1,2)}$ as test statistics to determine whether to reject $H_0^{(2)}$ and $H_0^{(1,2)}$, respectively. Again, the bootstrap technique can be used to approximate the distributions of the statistics under the null hypotheses. Since the procedures are entirely analogous to those described for testing $H_0^{(1)}$, we omit details.

4 Discussion

To our knowledge, our study is the first to extend classical two-way ANOVA to interval-valued data. We established a bootstrap approach to approximating the distribution and the *p*-value of each test statistic. The effectiveness of bootstrap techniques in hypotheses testing for fuzzy data has been empirically demonstrated (e.g., Montenegro et al. [14], González-Rodríguez et al. [9], Gil et al. [5]; also see Colubi [2]), and we intend to conduct empirical studies to examine our bootstrap scheme for the two-way ANOVA.

Currently we are extending the two-way ANOVA to factorial ANOVA for interval-valued data and for fuzzy data. We hope that our study will help to further facilitate rigorous statistical analyses of imprecise data.

Acknowledgements. This research has been partially supported by the following grants: (i) the Spanish Ministry of Science and Innovation Grants, MTM2009-09440-C02-01 and MTM2009-09440-C02-02; (ii) the Principality of Asturias Grants, IB09-042C1 and IB09-042C2; and (iii) the COST Action, IC0702.

References

1. Bertoluzza, C., Corral, N., Salas, A.: On a new class of distances between fuzzy numbers. Math. Soft Comput. 2, 71–84 (1995)
2. Colubi, A.: Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy- and real-valued data. Fuzzy Sets Syst. 160, 344–356 (2009)

3. Colubi, A., López-Díaz, M., Domínguez-Menchero, J.S., Gil, M.A.: A generalized strong law of large numbers. Probab. Theory Related Fields 114, 401–417 (1999)
4. Corral, N., González-Rodríguez, G., López-García, M.T., Ramos, A.B.: Multiple-sample test for fuzzy random variables. In: Abstracts of the 56th Session of the International Statistical Institute, ISI 2007, Lisbon, Portugal, p. 165 (2007)
5. Gil, M.A., Montenegro, M., González-Rodríguez, G., Colubi, A., Casals, M.R.: Bootstrap approach to the multi-sample test of means with imprecise data. Comput. Statist. Data Anal. 51, 148–162 (2006)
6. Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval data. Comput. Statist. Data Anal. 51, 3002–3015 (2007)
7. Giné, E., Zinn, J.: Bootstrapping general empirical measures. Ann. Probab. 18, 851–869 (1990)
8. González-Rodríguez, G., Colubi, A., Gil, M.A., D'Urso, P.: An asymptotic two dependent samples test of equality of means of fuzzy random fuzzy random variables. In: Proceedings of the 17th Conference of IASC-ERS, COMPSTAT 2006, Roma, pp. 689–695 (2006)
9. González-Rodríguez, G., Montenegro, M., Colubi, A., Gil, M.A.: Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data. Fuzzy Sets Syst. 157, 2608–2613 (2006)
10. Körner, R.: An asymptotic α-test for the expectation of random fuzzy variables. J. Statist. Plann. Inference 83, 331–346 (2000)
11. Laha, R.G., Rohatgi, V.K.: Probability Theory. Wiley, New York (1979)
12. Ledoux, M., Talagrand, M.: Probability in Banach Spaces: Isometry and Processes. Springer, Berlin (1991)
13. Montenegro, M., Casal, M.R., Lubiano, M.A., Gil, M.A.: Two-sample hypothesis tests of means of a fuzzy random variable. Inform. Sci. 133(1-2), 89–100 (2001)
14. Montenegro, M., Colubi, A., Casal, M.R., Gil, M.A.: Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika 59, 31–49 (2004)
15. Puri, M.L., Ralescu, D.A.: Differentials of fuzzy functions. J. Math. Anal. Appl. 91, 552–558 (1983)
16. Puri, M.L., Ralescu, D.A.: The concept of normality for fuzzy random variables. Ann. Probab. 11, 1373–1379 (1985)
17. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
18. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread Inform. Sci 179, 3964–3972 (2009)

Estimating the Variance of a Kernel Density Estimation

Bilal Nehme, Olivier Strauss, and Kevin Loquin

Abstract. This article proposes an interval-valued extension of kernel density estimation. We show that the imprecision of this interval-valued estimation is highly correlated with the variance of the density estimation induced by the statistical variations of the set of observations.

Keywords: Kernel density estimation, Maxitive kernel, Imprecise expectation.

1 Introduction

The Parzen-Rosenblatt density estimation is a well known nonparametric way of estimating the probability density function (pdf) underlying a finite set of observations. Since the convergence of this estimation towards the true density is only guaranteed for a infinite number of observations, it can be of prime interest to have a measure of the statistical error of this estimation (e.g. its variance). Such a measure cannot be directly computed when the pdf has to be estimated with a single set of observations. One can use resampling techniques, like Jackknife or Bootstrap [4], to perform this estimation. However, those methods can lead to computationally very expensive solutions.

In this paper, we propose a very novel approach for computing this estimation error. This approach is based on an extension of the Parzen-Rosenblatt

Bilal Nehme and Olivier Strauss
Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), 34095 Montpellier Cedex 5 , France
e-mail: Bilal.Nehme@lirmm.fr,Olivier.Strauss@lirmm.fr

Kevin Loquin
Télécom ParisTech (Ecole Nationale Supérieure des Télécommunications), F-75634 Paris Cedex 13, France
e-mail: kevin.loquin@gmail.com

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 483–490.
springerlink.com

method that leads to an interval-valued estimation of the pdf. Such an extension have been used in the past [8] for quantifying the effect of the input random noise on the output of a filtering process. It is based on replacing the summative kernel, on which is based the estimation, by a maxitive kernel [7], i.e. a possibility distribution. In this case, however, the Parzen-Rosenblatt estimator has to be reformulated to comply with the maxitive-based estimation extension.

2 Preliminarys Concepts

This section aims at presenting some preliminaries that are necessary to build the interval-valued pdf estimation we propose. Let Ω be a subset of $\mathbb{R}$, $\mathscr{P}(\Omega)$ the collection of all Lebesgue measurable subsets of Ω and $s:\Omega\to\mathbb{R}$ a bounded L_1 function associated to a distribution in the meaning of Schwartz [12].

We call **summative kernel** [7] a function $\kappa:\Omega\longrightarrow\mathbb{R}^+$ such that $\int_\Omega \kappa(x)dx=1$. It defines a probability measure on Ω denoted P_κ : $\forall A\in\mathscr{P}(\Omega)$, $P_\kappa(A)=\int_A \kappa(x)dx$. Let $\mathscr{K}(\Omega)$ be the set of summative kernels on Ω.

We call **maxitive kernel** [7] a function $\pi:\Omega\longrightarrow[0,1]$ such that $\sup_{x\in\Omega}\pi(x)=1$. It defines two dual confidence measures on Ω: a possibility measure Π_π and a necessity measure N_π by: $\forall A\in\mathscr{P}(\Omega)$, $\Pi_\pi(A)=\sup_{x\in A}\pi(x)$ and $N_\pi(A)=1-\sup_{x\notin A}\pi(x)$. Based on [2], a maxitive kernel π defines a convex set $\mathscr{M}(\pi)$ of summative kernels [6]:

$$\mathscr{M}(\pi)=\{\kappa\in\mathscr{K}(\Omega)/\forall A\in\mathscr{P}(\Omega),N_\pi(A)\le P_\kappa(A)\le\Pi_\pi(A)\}. \tag{1}$$

Let Δ be a positive real value and $x\in\Omega$, a summative kernel κ_Δ^x can be derived from another summative kernel κ by: $\forall u\in\Omega$, $\kappa_\Delta^x(u)=\frac{1}{\Delta}\kappa(\frac{u-x}{\Delta})$. In the same way, a maxitive kernel π_Δ^x can be defined from a maxitive kernel π by: $\forall u\in\Omega$, $\pi_\Delta^x(u)=\pi(\frac{u-x}{\Delta})$. Δ is called the bandwidth of the kernel.

2.1 Derivative of a Summative Kernel

Kernel used in density estimation are usually unimodal, symmetric with a bounded support and having a first derivative. Let us denote $\mathscr{K}'(\Omega)$ the subset of those kernels on Ω.

Property 1. Let $\kappa\in\mathscr{K}'(\Omega)$ and $\Delta\in\mathbb{R}^+$, the first derivative $d\kappa_\Delta$ of the kernel κ_Δ can be written as a linear combination of two summative kernels η_Δ^+ and η_Δ^- [9]:

$$\forall u\in\Omega,-d\kappa_\Delta(u)=a_\Delta\left(\eta_\Delta^+(u)-\eta_\Delta^-(u)\right), \tag{2}$$

where a_Δ is a constant value defined by $a_\Delta=\int_\Omega \max(0,-d\kappa_\Delta(u))$ and $\eta_\Delta^+(u)=\frac{d\kappa_\Delta^+(u)}{a_\Delta}$, $\eta_\Delta^-(u)=\frac{d\kappa_\Delta^-(u)}{a_\Delta}$ with $d\kappa_\Delta^+=max(0,d\kappa_\Delta)$, $d\kappa_\Delta^-=max(0,-d\kappa_\Delta)$. Note that, by construction, $a_\Delta=\frac{a}{\Delta}$, with $a=\int_\Omega \max(0,-d\kappa(u))du$.

2.2 Derivative in the Sense of Distributions

The convolution of a L_1 function s by a summative kernel κ, denoted $\widehat{s_\kappa} = s \star \kappa$ is given by [5]:

$$\widehat{s_\kappa}(x) = \big(s \star \kappa\big)(x) = \int_\Omega s(u)\kappa(x-u)du = \int_\Omega s(u)\kappa^x(u)du = \langle s, \kappa^x \rangle\,, \qquad (3)$$

κ^x being the function κ translated in x, and $\langle .,. \rangle$ being the dot product defined for L_1 functions. The value $\widehat{s_\kappa}(x)$ can also be viewed as $\mathbb{E}_{\kappa^x}$, the expectation of s according to the neighborhood of x defined by the kernel κ .

If the summative kernel κ is differentiable, it can be seen as a test function [12]. It is thus possible to link ds, the derivative of s in the sense of distributions, to $d\kappa$, the derivative of κ in the sense of functions by [5]:

$$\langle ds, \kappa^x \rangle = \int_\Omega ds(u)\kappa^x(u)du = -\int_\Omega s(u)d\kappa^x(u)du = \langle s, -d\kappa^x \rangle\,. \qquad (4)$$

2.3 Reformulation of the Parzen-Rosenblatt Density Estimator

Let $(x_1,...,x_n)$ be a sample coming from the same random variable X with density function f. The Parzen-Rosenblatt kernel estimate [10, 11] of the density f in every point $x \in \Omega$ is given by:

$$\widehat{f^n_{\kappa_\Delta}}(x) = \frac{1}{n\Delta}\sum_{i=1}^{n} \kappa(\frac{x-x_i}{\Delta}) = \frac{1}{n}\sum_{i=1}^{n} \kappa^x_\Delta(x_i). \qquad (5)$$

Property 2. The estimation $\widehat{f^n_{\kappa_\Delta}}$ in every point $x \in \Omega$ can be interpreted as the expectation of the empirical distribution e_n according to a neighborhood of x defined by the summative kernel κ_Δ:

$$\widehat{f^n_{\kappa_\Delta}}(x) = \mathbb{E}_{\kappa^x_\Delta}(e_n) = \langle e_n, \kappa^x_\Delta \rangle\,. \qquad (6)$$

with $e_n = \frac{1}{n}\sum_{i=1}^{n} \delta^{x_i}$ and δ^{x_i} is the impulse Dirac translated in x_i.

In the same manner, an estimate of the cumulative distribution function F_{η_Δ}, associated with the random variable X, can be obtained by computing the expectation of the empirical distribution function E_n according to a neighborhood of x defined by the summative kernel η_Δ:

$$\widehat{F^n_{\eta_\Delta}}(x) = \mathbb{E}_{\eta^x_\Delta}(E_n) = \langle E_n, \eta^x_\Delta \rangle\,, \qquad (7)$$

with $E_n(x) = \frac{1}{n}\sum_{i=1}^{n} H(x-x_i)$ and H being the Heaviside function defined by $H(x) = 1$ if $x \geq 0$ and 0 elsewhere. Since e_n is the derivative of E_n in the sense of distributions [12], the Parzen-Rosenblatt estimator can be rewritten, for all $x \in \Omega$, as:

$$\widehat{f}^n_{\kappa_\Delta}(x) = \langle e_n, \kappa^x_\Delta \rangle = \langle dE_n, \kappa^x_\Delta \rangle = \langle E_n, -d\kappa^x_\Delta \rangle . \tag{8}$$

Theorem 1. *Let $\kappa_\Delta \in \mathscr{K}'(\Omega)$, whose first derivative $d\kappa_\Delta$ can be decomposed in: $\forall u \in \Omega, -d\kappa_\Delta(u) = a_\Delta \left(\eta^+_\Delta(u) - \eta^-_\Delta(u)\right)$, with $a_\Delta \in \mathbb{R}^+$ and $\left(\eta^+_\Delta, \eta^-_\Delta\right) \in \mathscr{K}(\Omega)$, then, for all $x \in \Omega$, $\widehat{f}^n_{\kappa_\Delta}(x) = a_\Delta \left(\widehat{F}^n_{\eta^+_\Delta}(x) - \widehat{F}^n_{\eta^-_\Delta}(x)\right)$.*

Proof. According to (2) and (7), we have:

$$\widehat{f}^n_{\kappa_\Delta}(x) = a_\Delta \left(\langle E_n, \eta^{x+}_\Delta \rangle - \langle E_n, \eta^{x-}_\Delta \rangle\right) = a_\Delta \left(\widehat{F}^n_{\eta^+_\Delta}(x) - \widehat{F}^n_{\eta^-_\Delta}(x)\right) . \qquad \square$$

3 Interval-Valued Estimation

A maxitive kernel based imprecise estimate of the cumulative distribution function has been proposed in [6]. It is defined for all $x \in \Omega$ by:

$$\overline{\underline{F}}^n_{\pi_\Delta}(x) = \left[\underline{F}^n_{\pi_\Delta}(x), \overline{F}^n_{\pi_\Delta}(x)\right] = \overline{\underline{\mathbb{E}}}_{\pi^x_\Delta}(E_n) = \left[\underline{\mathbb{E}}_{\pi^x_\Delta}(E_n), \overline{\mathbb{E}}_{\pi^x_\Delta}(E_n)\right], \tag{9}$$

where π is a maxitive kernel, $\Delta \in \mathbb{R}^+$ a bandwidth and $\overline{\underline{\mathbb{E}}}_\pi(.)$ is the imprecise expectation based on the maxitive kernel π [6].

The computation of the lower and the upper bounds of the imprecise cumulative distribution estimator, defined by (9), is given in [6] by:

$$\overline{\mathbb{E}}_{\pi^x_\Delta}(E_n) = \mathbb{C}_{\Pi_{\pi^x_\Delta}}(E_n) = \frac{1}{n}\sum_{i=1}^{n} \left(\pi^x_\Delta(x_i) H(x_i - x) + H(x - x_i)\right), \tag{10}$$

$$\underline{\mathbb{E}}_{\pi^x_\Delta}(E_n) = \mathbb{C}_{N_{\pi^x_\Delta}}(E_n) = \frac{1}{n}\sum_{i=1}^{n} \left((1 - \pi^x_\Delta(x_i)) H(x - x_i)\right), \tag{11}$$

$\mathbb{C}_{\Pi_{\pi^x_\Delta}}(E_n)$ (resp. $\mathbb{C}_{N_{\pi^x_\Delta}}(E_n)$) being the Choquet integral of E_n with respect to the possibility measure $\Pi_{\pi^x_\Delta}$ (resp. the necessity measure $N_{\pi^x_\Delta}$). As shown in [6] when $\kappa \in \mathscr{M}(\pi)$, then $\forall \Delta \in \mathbb{R}^+$, $\forall x \in \Omega$, $\widehat{F}^n_{\kappa_\Delta}(x) \in \overline{\underline{F}}^n_{\pi_\Delta}(x)$.

3.1 Interval-Valued Estimation of the Probability Density Function

The idea underlying the maxitive based imprecise estimation of the density is the following: instead of dominating the summative kernel on which is based the density estimation like in (9), we will dominate the summative kernels involved in the decomposition (2) of its derivative.

Let $(x_1, \ldots, x_n)$ be a set of n observations, f the pdf underlying the observation process and E_n the empirical distribution function associated with this set of observations. Let $\kappa_\Delta \in \mathscr{K}'(\Omega)$ be a summative kernel, whose derivative $-d\kappa_\Delta$ can be decomposed in: $a_\Delta \left(\eta^+_\Delta - \eta^-_\Delta\right)$, $a_\Delta \in \mathbb{R}^+$ and $\left(\eta^+_\Delta, \eta^-_\Delta\right) \in \mathscr{K}(\Omega)$. Let π^+ (rsp. π^-) be the most specific maxitive kernel dominating η^+ (rsp. η^-) [7].

Definition 1. *A Parzen-Rosenblatt-like imprecise estimator of the pdf underlying a set of observations, whose empirical cumulative is E_n, is defined by:*

$$\forall x \in \Omega, \overline{\underline{f}}^n_{(\kappa_\Delta)}(x) = a_\Delta \left(\overline{\underline{\mathbb{E}}}_{\pi_\Delta^{+x}}(E_n) \ominus \overline{\underline{\mathbb{E}}}_{\pi_\Delta^{-x}}(E_n) \right), \tag{12}$$

where $\ominus$ is the Minkowski difference [1].

The question concerns now the properties of the obtained imprecise estimation. We will first denote $\mathscr{D}(a, \Delta, (\pi^+, \pi^-))$ a subset of $\mathscr{K}'(\Omega)$ defined by:

$$\mathscr{D}\left(a, \Delta, (\pi^+, \pi^-)\right) = \left\{ \begin{array}{c} \upsilon \in \mathscr{K}'(\Omega), \exists\ \xi^+ \in \mathscr{M}(\pi_\Delta^+) \text{ and } \xi^- \in \mathscr{M}(\pi_\Delta^-), \\ \text{such that } -d\upsilon = a_\Delta\left(\xi^+ - \xi^-\right) \end{array} \right\}$$

where a_Δ, Δ, a, π_Δ^+ and π_Δ^- have been previously defined.

The interval-valued estimation, defined by (12),verifies the following property:

Property 3. Let $\overline{\underline{f}}^n_{(\kappa_\Delta)}$ be the interval-valued estimation of the pdf defined by Equation (12), then:

$$\forall x \in \Omega, \forall \varphi \in \mathscr{D}\left(a, \Delta, (\pi^+, \pi^-)\right), \widehat{f}^n_\varphi(x) \in \overline{\underline{f}}^n_{(\kappa_\Delta)}(x). \tag{13}$$

Remark 1. The reverse property of expression (13) is not true, i.e.:

$$\exists y \in \overline{\underline{f}}^n_{(\kappa_\Delta)}(x), \forall \varphi \in \mathscr{D}\left(a, \Delta, (\pi^+, \pi^-)\right), y \neq \widehat{f}^n_\varphi(x).$$

3.2 Integrated Imprecision of the Interval-Valued Estimation

It would have been nice if the imprecision of the interval-valued density estimate we propose had decreased with Δ and $\frac{1}{n}$. Unfortunately, as we prove here, the integral of the imprecision of $\overline{\underline{f}}^n_{(\kappa_\Delta)}$ depends neither on n nor on Δ. To prove this property, we need the following theorem:

Theorem 2. *Let π_Δ be a maxitive kernel. Let $\varepsilon^n_{\pi_\Delta}(x) = \overline{F}^n_{\pi_\Delta}(x) - \underline{F}^n_{\pi_\Delta}(x)$ be the imprecision at x of the interval-valued estimation $\overline{\underline{F}}^n_{\pi_\Delta}(x)$, defined by (9), then: $\int_\Omega \varepsilon^n_{\pi_\Delta}(x)dx = \rho(\pi_\Delta) = \Delta\ \rho(\pi)$, with $\rho(\pi) = \int_\Omega \pi(x)dx$ being the granulosity of the maxitive kernel π [7], i.e. its degree of imprecision.*

Proof. According to (10) and (11), we have $\varepsilon^n_{\pi_\Delta}(x) = \frac{1}{n}\sum_{i=1}^n(\pi_\Delta^x(x_i) - \mathbb{1}_{x=x_i})$. Since $\int_\Omega \mathbb{1}_{x=x_i}dx = 0$, $\forall i \in \{1, \ldots, n\}$, we obtain: $\int_\Omega \varepsilon^n_{\pi_\Delta}(x)dx = \frac{1}{n}\sum_{i=1}^n \rho(\pi_\Delta) = \Delta\ \rho(\pi)$. □

Theorem 3. *Let $\kappa_\Delta \in \mathscr{K}'(\Omega)$ be a summative kernel. Let $\zeta^n_{(\kappa_\Delta)}(x) = \overline{f}^n_{(\kappa_\Delta)}(x) - \underline{f}^n_{(\kappa_\Delta)}(x)$ be the imprecision at x of the interval-valued estimation $\overline{\underline{f}}^n_{(\kappa_\Delta)}$ defined by (12), then $\int_\Omega \zeta^n_{(\kappa_\Delta)}(x)dx$ is a constant value that we call α.*

Proof. According to (12) and by theorem 2 we have:

$$\int_\Omega \zeta^n_{(\kappa_\Delta)}(x)dx = a_\Delta \left(\int_\Omega (\overline{F}^n_{\pi^+_\Delta}(x) - \underline{F}^n_{\pi^+_\Delta}(x))dx + \int_\Omega (\overline{F}^n_{\pi^-_\Delta}(x) - \underline{F}^n_{\pi^-_\Delta}(x)) \right) dx,$$
$$= a\left(\rho(\pi^+) + \rho(\pi^-)\right) = \alpha.$$

□

The main consequence of theorem 3 is that the defined imprecise estimator cannot converge to the true density, i.e. when $\Delta \to 0$ and $n\Delta \to \infty$, $\left(\overline{f}^n_{(\kappa_\Delta)} - \underline{f}^n_{(\kappa_\Delta)}\right) \not\to 0$.

4 Link between Imprecision and Variance

This section is dedicated to an experiment showing that the imprecision $\zeta^n_{(\kappa_\Delta)}(x)$ of the interval-valued estimate $\overline{\underline{f}}^n_{(\kappa_\Delta)}(x)$ can be used to estimate $var(\widehat{f}^n_{\kappa_\Delta}(x))$, the variance of the Parzen-Rosenblatt estimate of f via the kernel κ_Δ. First, as shown by numerous other works, theoretically $var(\widehat{f}^n_{\kappa_\Delta}(x))$ decreases when n and Δ increases. In fact, as stated in [13]:

$$\forall x \in \Omega, \quad var(\widehat{f}^n_{\kappa_\Delta}(x)) \approx (n\Delta)^{-1} f(x) R(\kappa_\Delta), \tag{14}$$

with $R(\kappa_\Delta) = \int_\Omega \kappa_\Delta(x)^2 dx$. Since the integral of $\zeta^n_{(\kappa_\Delta)}$ depends neither on n nor on Δ, the direct value of $\zeta^n_{(\kappa_\Delta)}(x)$ cannot be used directly to estimate $var(\widehat{f}^n_{\kappa_\Delta}(x))$ but should be multiplied by a factor $\gamma(n,\Delta)$ that depends on both n and Δ. Let us suppose this relation to be linear, i.e.:

$$var(\widehat{f}^n_{\kappa_\Delta}(x)) = \mathbb{E}\big(\gamma(n,\Delta)\ \zeta^n_{(\kappa_\Delta)}(x)\big). \tag{15}$$

Thus, by integrating expression (15), we directly obtain $\gamma(n,\Delta) = \frac{R(\kappa_\Delta)}{\alpha n \Delta}$, with $\alpha = \int_\Omega \zeta^n_{(\kappa_\Delta)}(x)dx$. The experiment we report here aims at testing whether $\big(\gamma(n,\Delta)\ \zeta^n_{(\kappa_\Delta)}(x)\big)$ is correlelated or not with $var(\widehat{f}^n_{\kappa_\Delta}(x))$. It is based on simulating a random process whose underlying pdf is a mixture of two Gaussian distributions of mean 3 (resp. 8) and variance 1 (resp. 4). We use the symmetric summative kernel defined by $\kappa_\Delta(x) = \frac{1}{2\Delta}(1 + cos(\frac{|x|\pi}{\Delta}))\mathbb{1}_{[-\Delta,\Delta]}(x)$. The computation of the different values associated with this kernel are: $\eta^+_\Delta(x) = \eta^-_\Delta(x) = \frac{\pi}{2\Delta}(cos(\frac{|x|\pi}{\Delta}))\mathbb{1}_{[\frac{-\Delta}{2},\frac{\Delta}{2}]}$, $a = 1$, $\alpha \approx 0.7268$ and $R(\kappa_\Delta) = \frac{3}{4\Delta}$. The value of Δ is fixed to $\Delta = 1$, while the number of observations varies from $n = 1000$ to $n = 10000$. For each values of n, we compute 400 different sets of observation. We then estimate both $var(\widehat{f}^n_{\kappa_\Delta}(x))$ and $\mathbb{E}\big(\gamma(n,\Delta)\ \zeta^n_{(\kappa_\Delta)}(x)\big)$ on 500 equally spaced samples of the reference subset $\Omega = [-5, 20]$.

Fig. 1 shows the result of this experiment by plotting $var(\widehat{f}^n_{\kappa_\Delta})$ versus $\mathbb{E}\big(\gamma(n,\Delta)\ \zeta^n_{(\kappa_\Delta)}\big)$. As can be seen on Fig. 1, the correlation between $var(\widehat{f}^n_{\kappa_\Delta})$

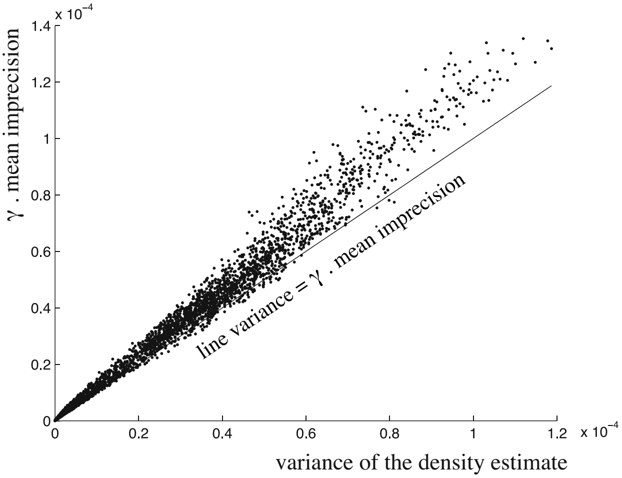


Fig. 1 The cloud of values $\mathbb{E}\big(\gamma(n,\Delta)\ \zeta^{n}_{(\kappa_\Delta)}\big)$ versus $var(\widehat{f}^{n}_{\kappa_\Delta})$.

and $\mathbb{E}\big(\gamma(n,\Delta)\ \zeta^{n}_{(\kappa_\Delta)}\big)$ is high (correlation coefficient $r \approx 0.995$). However, the cloud of the computed values is close but rather above the theoretical line materializing equation (15) on Fig. 1. This bias can be explained first by the fact that relation (14) is an approximation and second by the fact that the dependence is possibly not exactly linear. However, the numerous experiments we carried out show that $\big(\gamma(n,\Delta)\ \zeta^{n}_{(\kappa_\Delta)}\big)$ provide a good estimation of $var(\widehat{f}^{n}_{\kappa_\Delta})$.

5 Conclusion

The interval-valued nonparametric extension of the kernel density estimation improves on the traditional approach by providing an estimation of the error induced by the statistical variation of the set of observations with a significant increase of the computational complexity.

Future work should focus on the relation between the median of this interval-valued density and the true density (convergence if any ?) and propose a modification of expression (12) that leads to an interval-valued density whose imprecision decreases with the bandwidth of the kernel or when the number of observation increases. We are now working on comparing this approach with the classical approach based on confidence intervals [3].

Acknowledgements. This work is in part supported by VOODDO (2008-2011), a project of ANR and the region of Languedoc Roussillon, France.

References

1. Danilov, V., Koshevoy, G.: Cores of cooperative games, superdifferentials of function, and the minkowski difference of sets. Math. Anal. Appl. 247, 1–14 (2000)
2. Dubois, D., Prade, H.: When upper probabilities are possibility measures. Fuzzy Sets Syst 49(1), 65–74 (1992)
3. Giné, E., Nickl, R.: Confidence bands in density estimation. Ann. Statist. 38, 1122–1170 (2010)
4. Huber, P.: Robust Statistics. Wiley, New York (1981)
5. Loquin, K.: De l'utilisation des noyaux maxitifs en traitement de l'information. PhD thesis, LIRMM, Université Montpellier II (2008)
6. Loquin, K., Strauss, O.: Imprecise functional estimation: the cumulative distribution case. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Handling Variability and Imprecision. Advances in Soft Computing, vol. 48, pp. 175–182. Springer, Heidelberg (2008)
7. Loquin, K., Strauss, O.: On the granularity of summative kernels. Fuzzy Sets Syst. 159, 1952–1972 (2008)
8. Loquin, K., Strauss, O.: Noise quantization via possibilistic filtering. In: Proceedings of the International Symposium on Imprecise Probability: Theory and Applications, ISIPTA 2009, Durham, UK, pp. 297–306 (2009)
9. Nehme, B., Strauss, O.: Towards an interval-valued estimation of the density. In: Proceedings of the 2010 IEEE International Conference on Computational Intelligence (accept, 2010)
10. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statist. 33, 1065–1076 (1962)
11. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832–837 (1956)
12. Schwartz, L.: Théorie des distributions. Hermann, Paris (1950)
13. Silvermann, B.: Density Estimation for Statistics and Data Analysis. Chapman and Hall, London (1986)

Uncertainty Invariance Transformation in Continuous Case

María José Pardo and David de la Fuente

Abstract. In this work, a general procedure for transforming a possibility distribution into a probability density function, in the continuous case, is proposed, in a way that the resulting distribution contains the same uncertainty as the original distribution. A significant aspect of this approach is that it makes use of Uncertainty Invariance Principle which is itself a general procedure for going from an initial representation of uncertainty to a new representation.

Keywords: Possibility/Probability Transformation, Uncertainty Invariance Principle, Uncertainty Measures, Consistency Principle.

1 Introduction

It has been carried out a wide review of the existent possibility → probability transformations in the literature, in the continuous case, and we have only found transformations in [1], [2], [6], [7], [9], [14], [18] and [23]. The transformations in [2], [6] and [18] are based on the Principle of Insufficient Reason, but obtained from different methods, so the calculated probability density function is similar, and it is the transformation more used by other authors who need to incorporate uncertainty in their works ([15], [16], [21]). In the

María José Pardo
Department of Applied Economic IV, University of the Basque Country,
48015 Bilbao, Spain
e-mail: mjose.pardo@ehu.es

David de la Fuente
Department of Business Administration, University of Oviedo,
33204 Gijón, Spain
e-mail: david@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 491–498.
springerlink.com

whole literature that was analyzed, we have not found any work that develops a possibility → probability transformation with uncertainty invariance in the continuous case, except in [23]. His transformation meets the Principle of Uncertainty Invariance through a parameter β, like the new proposed transformation. The advantage of the developed method in this work is that the computational cost is lower.

The transformations with uncertainty invariance in the discrete case are extensively studied in [8] and [12]. These transformations are basically guided by two principles: i) the conservation of uncertainty; namely, the uncertainty contained in the possibility distribution is equal to the probability uncertainty expressed in terms of Shannon entropy [22]; ii) an appropriate scale that defines some mechanical transformation of data. In [12] are proposed different scale transformations: *ratio scale*, *interval scale* and *log-scale*.

2 Uncertainty Invariance Principle

The Principle of Uncertainty Invariance, introduced in [12], ensures that the total amount of information supplied by each model is kept unchanged.

In order to transform the representation of a problem in a theory T1, into an equivalent representation in other theory T2 using the Principle of Uncertainty Invariance, the following aspects must be met:

- The amount of uncertainty associated with the situation must be maintained when we move from T1 to T2. In order to achieve this, the uncertainty measure in both theories will be studied and the transformation developed, so if $U(\mu)$ is the uncertainty measure of the possibility distribution μ and $H(f)$ is the uncertainty measure of the probability density function f, then both must be the same, $U(\mu) = H(f)$.
- The degree of belief in T1 will be converted into its compensation in T2 by means of an appropriate scale. This condition guarantees that certain features of μ, which are considered essential in the given context (e.g. the order), are maintained in the transformation. In measurement theory, these transformations are known as scales. Therefore, the proposed transformation must be $f(x) = f_{scale}\mu(x)$, where f_{scale} is some function standing for the scale preservation. The different scale transformations, in the discrete case, proposed in [12] have been studied, and the only transformation that maintains the Principle of Uncertainty Invariance is the *log-interval scale* transformation, which is formulated as: $p_i = \alpha \cdot \pi_i^{\beta}$.

3 Uncertainty Measures in the Continuous case

In this section, a study of different measures of uncertainty in the continuous case has been carried out.

Uncertainty Measures of Probability Density Functions

In [3] the concept of *differential entropy* is defined, which is the entropy of a continuous random variable, and it is similar in some way to the Shannon entropy [22]. The *differential entropy* $H(f)$ of a continuous random variable X with probability density function $f(x)$ is defined as: $H(f) = -\int_{\mathbb{R}} f(x) \ln f(x)\, dx$.

The uncertainty measures should reach the maximum value when the distribution is uniform [11]. So, when X is a uniformly distributed random variable in $[a,b]$ then the *differential entropy* is $H(f) = \ln(b-a)$ and the range is $(-\infty, \ln(b-a)]$.

Uncertainty Measures of Possibility Distributions

The Principle of Uncertainty Invariance will be respected if it is chosen the uncertainty measure that reflects the kind of uncertainty equivalent to the *differential entropy*. Therefore the range of the uncertainty measure of the possibility distribution must be $(-\infty, \ln(b-a)]$. When the possibility distribution is uniform then the uncertainty measures should reach the maximum, $U(\mu) = \ln(b-a)$. In the analyzed literature there are uncertainty measures of possibility distributions in continuous case in [4], [10], [13], [17], [19], [23], [24] (this is only a list of the most representative works). For all of them, the range and the value when the possibility distribution is uniform have been calculated and only [10] and [23] verify the conditions, so they reflect the kind of uncertainty equivalent to the *differential entropy*. Furthermore, it is verified that when space is one-dimensional then the expression of the two measures is the same. This uncertainty measure for a possibility distribution $\mu(x)$ using the notion of α-cut is: $U(\mu) = \int_0^1 \ln |A_\alpha|\, d\alpha$, where $A_\alpha = \left[A_\alpha^L, A_\alpha^U\right]$ is the α-cut of $\mu(x)$ and $|A_\alpha| = A_\alpha^U - A_\alpha^L$ is the cardinality of A_α.

4 Uncertainty Invariance Transformation in the Continuous Case

At this point, the only thing that has to be defined is the scale transformation, for example adapting the transformation in [12] to the continuous case. In this way, we define the transformation of the possibility distribution $\mu(x)$ into the probability density function $f(x)$ as: $f(x) = \alpha \cdot \mu(x)^\beta$.

Now, considering the probabilistic normalization condition, ($\int f(x)\, dx = 1$) the parameter α is calculated:

$$f(x) = \alpha \cdot \mu(x)^\beta = \frac{\mu(x)^\beta}{\int \mu(x)^\beta\, dx} \tag{1}$$

And the parameter β is the solution to the equation: $U(\mu) = H(f)$.

If $\beta = 1$, we find the transformation that Lee and Li have proposed in [14]. They believe that when the only known information is the membership function $\mu(x)$ for a fuzzy event, then a very reasonable probability density function is the proportional probability density function, which means that $f(x) = \alpha \cdot \mu(x)$.

Properties of the Uncertainty Invariance Transformation

In the continuous case, besides the uncertainty preservation and preservation of scale, the Uncertainty Invariance Transformation meets the following properties [20]: i) Bijectiveness, if T is the possibility → probability transformation, then the inverse mapping T^{-1} sets for the inverse probability → possibility transformation; ii) Strong preference preservation, if the element x is preferred over another element y according to the possibility distribution, then this preference is maintained in the probabilistic setting; iii) Ignorance preservation, if the complete ignorance state occurs in one formalism, it should be translated into its counterpart in another formalism (if all elements of the universe of discourse have possibility one then the probability density function is formulated as a uniform probability density function over the set of alternatives X); and iv) Symmetry preservation, the general shape of the probabilistic data should be the same as its possibilistic counterpart. Usually, the strong preference preservation entails symmetry preservation.

Comparison of the Uncertainty Invariance Transformation with the Transformation by Dubois et al. [6]

The transformation of Dubois et al. [6] has been chosen because it is the most used method by the researchers who need to transform. We have also compared the method developed in this work with the transformations proposed in [9] and [14], but the results are not included here in order to shorten this study. To carry out the comparison, we have set as a starting point that the possibility distribution is determined by a triangular fuzzy number (TFN), because the TFNs are simple and easy to manipulate in mathematical calculus; they are also very used in practical situations and in the design, modelling and simulation of systems. The comparison has also been carried out when a trapezoidal fuzzy number is transformed, but due to the limitation of space the analysis of the results is not detailed in the work. Moreover, the comparison can be performed with other possibility distributions. In order to compare the two methods the possibility distribution must be known. Therefore, in this work more general results can not be obtained.

First, the TFN is going to be transformed following the method proposed in this work, and then following the method developed in [6]. In both cases, the *differential entropy* and the *consistency degree* between the possibility distribution and the probability density function have been calculated.

The Possibility/Probability Consistency Principle starts from the existence of two kind of information: the possibilistic and the probabilistic ones, both over a variable X. Since it is possible to understand the simultaneous existence of both kinds of information, it is then when the question about the relation between them arises and the transformations between possibilistic data to probabilistic data and vice versa are possible. This principle is a result of the fact that "what is possible" influences and is influenced by "what is probable". For this situation, in [25] is established, in the discrete case, the Principle of Possibility/Probability Consistency and in [2] is calculated, in the continuous case, the *consistency degree* or measure between $\mu(x)$ and $f(x)$: $\gamma = \int_{\mathbb{R}} \mu(x) f(x)\, dx$. Thus, "*the greater* γ *becomes the better the consistency between the two concepts* [20]".

Let be $\widetilde{A}$ a TFN, $\widetilde{A} = [a-d_1, a, a+d_2]$ with membership function:

$$\mu_{\widetilde{A}}(x) = \begin{cases} \dfrac{x-(a-d_1)}{d_1} & if \;\; a-d_1 \le x \le a \\ \dfrac{(a+d_2)-x}{d_2} & if \;\; a \le x \le a+d_2 \end{cases} \tag{2}$$

with α-cut $A_\alpha = [a-d_1(1-\alpha), a+d_2(1-\alpha)]$, and if $D = d_1+d_2$ then $|A_\alpha| = (d_1+d_2)(1-\alpha) = D(1-\alpha)$ is the cardinality and $U\left(\mu_{\widetilde{A}}(x)\right) = \ln D - 1$ is the uncertainty measure of $\widetilde{A}$.

Transformation of a TFN with Uncertainty Preservation

The following steps are carried out:

- The probability density function is calculated:

$$f(x) = \frac{\mu_{\widetilde{A}}(x)^\beta}{\int_{-\infty}^{\infty} \mu_{\widetilde{A}}(x)\, dx} \tag{3}$$

  The term of denominator is: $\int_{-\infty}^{\infty} \mu_{\widetilde{A}}(x)\, dx = \dfrac{D}{\beta+1}$, then:

$$f(x) = \frac{\beta+1}{D} \mu_{\widetilde{A}}(x)^\beta \tag{4}$$

- The *differential entropy* of the probability density function is calculated:

$$H(f(x)) = -\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx = \frac{\beta}{\beta+1} - \ln \frac{\beta+1}{D} \tag{5}$$

- Finally, the parameter β is computed from $U\left(\mu_{\widetilde{A}}(x)\right) = H(f(x))$, obtaining $\beta = 5.3054$. And substituting this result in (4) is:

$$f(x) = \frac{6.3054}{D} \mu_{\widetilde{A}}(x)^{5.3054} \tag{6}$$

- *Differential entropy* and *consistency degree*:

  $H(f(x)) = \ln D - 1$ (equal to $U\left(\mu_{\widetilde{A}}(x)\right)$) and $\gamma = \beta + 1/\beta + 2 = 0.863$.

Transformation of a TFN Following the Method of Dubois et al. [6]

If the probability density function calculated from the transformation of a TFN following [6] is denoted by $f_D(x)$, then:

$$f_D(x) = \int_0^{\mu_{\tilde{A}}(x)} \frac{d\alpha}{|A_\alpha|} = \frac{-1}{D} \begin{cases} \ln \dfrac{a-x}{d_1} & if \ \ a-d_1 \leq x \leq a \\ \ln \dfrac{x-a}{d_2} & if \ \ a \leq x \leq a+d_2 \end{cases} \tag{7}$$

- *Differential entropy* and *consistency degree*:

$H(f_D(x)) = \ln D - 1+\text{gamma} = \ln D - 1 + 0.577$ and $\gamma_D = 0.75$.

Analysis of Results when a TFN is Transformed

The method developed in [6] bases its transformation on the Principle of Insufficient Reason and since it does not verify the Principle of Uncertainty Invariance, the probability density function $f_D(x)$ contains more uncertainty than the original possibility distribution $\mu_{\tilde{A}}(x)$. Our method meets such principle and no information is added or removed in the transformation. Therefore, $H(f_D(x)) > H(f(x))$, that means that the *differential entropy* $H(f_D(x))$ of the probability density function $f_D(x)$ is greater than the measure of uncertainty $U\left(\mu_{\tilde{A}}(x)\right)$ of the original possibility distribution $\mu_{\tilde{A}}(x)$, which at the same time is equal to the *differential entropy* $H(f(x))$ of the probability density function $f(x)$.

Regarding the *consistency degree* between both distributions, the probability density function $f(x)$ has a *consistency degree* with the possibility distribution $\mu_{\tilde{A}}(x)$ equal to $\gamma = 0.863$, which is greater than the *consistency degree* of the probability density function $f_D(x)$, equal to $\gamma_D = 0.75$ and therefore, $f(x)$ is more consistent and coherent with $\mu_{\tilde{A}}(x)$ than $f_D(x)$. Both results are also obtained when a trapezoidal fuzzy number is transformed.

Consistency Degree by Dubois and Prade [5]

In order to complete the study of the method proposed in this work, we are going to prove that the basic premise of consistency between the possibility distribution and the probability density function or consistency degree by Dubois and Prade [5], from the form $\mu(A) \geq f(A), \forall A$, is verified. In [23] it is proved that the transformation by Dubois et al. [6] meets the consistency degree by Dubois and Prade.

Let be $\mu(A) = \max_{x \in A} \mu(x)$. The set A can be formed at the most by $A = [a-d_1, x_1] \cup [x_2, a+d_2]$, with $x_1 < x_2$ and $\mu(A) = \mu(x_1) = \mu(x_2)$. Then $f(A)$ is:

$$f(A) = \int_{a-d_1}^{x_1} f(x)\,dx + \int_{x_2}^{a+d_2} f(x)\,dx = \tag{8}$$

$$= \frac{(x_1-(a-d_1))^{\beta+1}}{Dd_1^{\beta}} + \frac{((a+d_2)-x_2)^{\beta+1}}{Dd_2^{\beta}}$$

As $\mu(x_1)=\mu(x_2)$ then $\frac{x_1-(a-d_1)}{d_1}=\frac{(a+d_2)-x_2}{d_2}$, and with $\beta=5.3054$ is $\frac{x_1-(a-d_1)}{d_1} > \left(\frac{x_1-(a-d_1)}{d_1}\right)^{\beta}$, and $\frac{x_1-x_2+D}{D}<1$ and we come immediately to $\mu(A)\geq f(A), \forall A$. Following a similar reasoning, it is proved that if a trapezoidal fuzzy number is transformed with the proposed method, then the consistency degree by Dubois and Prade is also verified.

5 Conclusions

After carrying out a broad study of the different possibility → probability transformations that are in the literature in order to transform data in the continuous case, and also the uncertainty measures in both theories, a new possibility → probability transformation for the continuous case have been developed, which meets significant requirements so that this kind of transformation is accepted by the researchers. The main requirement that verifies is the Principle of Uncertainty Invariance, which guarantees that no information is unconsciously added or removed when changing the mathematical framework from which a concrete phenomenon is formalized. And it verifies the consistency degree of Dubois and Prade [5] and the properties of bijectiveness, strong preference preservation, ignorance preservation and symmetry preservation. The proposed transformation have been compared with the transformation from Dubois et al. [6], both of a TFN and of a trapezoidal fuzzy number, and in this cases the proposed transformation provides a probability density function that is more consistent and coherent with the initial possibility distribution, and maintains the uncertainty without adding or removing information. In a future work, this transformation will be used to develop the fuzzy negative exponential density function.

Acknowledgements. This research is supported by MICINN, reference number TIN2009-14374-C03-01.

References

1. Baudrit, C., Dubois, D.: Practical representations of incomplete probabilistic knowledge. Comput. Stat. Data An. 51, 86–108 (2006)
2. Chanas, S., Nowakowski, M.: Single value simulation of fuzzy variable. Fuzzy Set Syst. 25, 43–57 (1988)
3. Cover, T., Thomas, J.A.: Elements of Information Theory. John Wiley and Sons, New York (1991)
4. De Luca, A., Termini, S.: A definition on a nonprobabilistic entropy in the setting of fuzzy sets theory. Inform. Contr. 20, 301–312 (1972)

5. Dubois, D., Prade, H.: On several representations of an uncertain body of evidence. In: Gupta, M.M., Sanchez, E. (eds.) Fuzzy Information and Decision Processes. North-Holland, New York (1982)
6. Dubois, D., Prade, H., Sandri, S.: On possibility/probability transformations. In: Lowen, R., Roubens, M. (eds.) Fuzzy Logic. Kluwer Academic Publishers, Dordrecht (1993)
7. Florea, M.C., Jousselme, A.L., Grenier, D., Bossé, E.: Approximation techniques for the transformation of fuzzy sets into random sets. Fuzzy Set Syst. 159, 270–288 (2008)
8. Geer, J.F., Klir, G.J.: A mathematical analysis of information preserving transformations between probabilistic and possibilistic formulations of uncertainty. Int. J. Gen. Syst. 20, 143–176 (1992)
9. Gupta, C.P.: A note on the transformation of possibilistic information into probabilistic information for investment decision. Fuzzy Set Syst. 56, 175–182 (1993)
10. Higashi, M., Klir, G.J.: Measures of uncertainty and information based on possibility distributions. Int. J. Gen. Syst. 9, 43–58 (1982)
11. Kapur, J.N.: Measures of information and their applications. John Wiley and Sons, New Delhi (1994)
12. Klir, G.J.: A principle of uncertainty and information invariance. Int. J. Gen. Syst. 17, 249–275 (1990)
13. Klir, G., Wierman, M.: Uncertainty-Based Information: Elements of Generalized Information Theory. Physica-Verlag, New York (1999)
14. Lee, E.S., Li, R.J.: Comparison of fuzzy numbers based on the probability measure of fuzzy events. Comput. Math. Appl. 15, 887–896 (1988)
15. Lee, K.D., Llinas, J.: Hybrid model for intent estimation. In: Proceedings of the Sixth International Conference of Information Fusion, FUSION 2003, Queensland Australia, vol. 1(2), pp. 1215–1222 (2003)
16. Lee, K.D., Llinas, J., Adleson, R.: Hybrid framework for fusion 2+ in multi-multi air engagement. In: Dasarathy, B.V. (ed.) Proceedings of SPIE Multisensor, Multisource Information Fusion: Architectures, Algorithms and Applications 2005, vol. 5813, pp. 86–95 (2005)
17. Li, X., Liu, B.: Maximum entropy principle for fuzzy variables. Int. J. Uncertain. Fuzz. 15, 43–52 (2007)
18. Negi, D.S., Lee, E.S.: Analysis and simulation of fuzzy queues. Fuzzy Set Syst. 46, 321–330 (1992)
19. Nuñez, J., Kutalik, Z., Cho, K.H., Wolkenhauer, O.: Level sets and minimum volume sets of probability density functions. Int. J. Approx. Reason. 34, 25–47 (2003)
20. Oussalah, M.: On the probability/possibility transformations: a comparative analysis. Int. J. Gen. Syst. 29, 671–718 (2000)
21. Rebiasz, B.: Fuzziness and randomness in investment project risk appraisal. Comput. Oper. Res. 34, 199–210 (2007)
22. Shannon, C.E.: The mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423 (1948)
23. Wonneberger, S.: Generalization of an invertible mapping between probability and possibility. Fuzzy Set Syst. 64, 229–240 (1994)
24. Yager, R.R.: Entropy and especificity in a mathematical theory of evidence. Int. J. Gen. Syst. 9, 249–260 (1983)
25. Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Set Syst. 1, 3–28 (1978)

Detection of Outliers Using Robust Principal Component Analysis: A Simulation Study

C. Pascoal, M.R. Oliveira, A. Pacheco, and R. Valadas

Abstract. Outlier detection is an important problem in statistics that has been addressed in a variety of research areas and applications domains. In this paper, we tackle this problem using robust principal component analysis. We consider different robust estimators along with the classical estimator of principal components and develop a simulation study to compare the envisage outlier detection methods in two different scenarios: semi-supervised, where we have a training set composed only by regular observations, and an unsupervised scenario, where nothing is known about the class (regular or outlier) of each training observation.

Keywords: Outlier detection, Robustness, Principal component analysis, Simulation.

1 Introduction

The development of new technology, opposite to the benefits it brings to the daily life of human beings, originates new types of threats, e.g. criminal activities on electronic commerce. In general, these threats are atypical observations from the main bulk of data, leading to the need of applying outlier detection methods to identify it.

C. Pascoal, M.R. Oliveira, and A. Pacheco
CEMAT and Department of Mathematics, Instituto Superior Técnico,
Technical University of Lisbon, 1049-001 Lisboa, Portugal
e-mail: claudiapascoal@ist.utl.pt,rsilva,apacheco@math.ist.utl.pt

R. Valadas
Instituto de Telecomunicações, Instituto Superior Técnico,
Technical University of Lisbon, 1049-001 Lisbon, Portugal
e-mail: rui.valadas@ist.utl.pt

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 499–507.
springerlink.com

A recent review of the large body of literature on outlier detection is done in [4], which provides a structured and comprehensive overview of the research on anomaly detection covering a variety of application domains. Despite the existence of other recent methodologies [3, 4, 12], our emphasis will be only on the use of robust Principal Component Analysis (PCA) for outlier detection, extending the work in [11, 13].

Reference [11] uses a principal component classifier in an intrusion detection problem where the training data is only composed of regular observations (semi-supervised case). These authors construct a predictive intrusion model from major and minor principal components of regular instances. They argue that this new method outperforms the k-nearest neighbor method, the density-based local outliers (LOF) approach, and the outlier detection algorithms based on Canberra metric. Using both major and minor PCs, the method would be able to detect, respectively: extreme observations with large values and observations that do not conform to the common correlation structure.

ROBPCA was used in [13] to distinguish regular observations from outliers in unsupervised learning. The authors applied it to traffic flow of cars in a city. They argue that the method is an useful tool for distinguishing the regular from the abnormal traffic flow patterns caused by accidents and loop detector faults, reducing the human effort in finding potential anomalies in the traffic flow.

The contents of the paper are the following. In Section 2 we briefly review PCA and in Section 3 we explain how it can be used in outlier detection problems. In Section 4 we develop a simulation study to analyze and compare the performance of different robust principal component approaches on the classification of observations, in semi-supervised learning, and also to the case where the investigator does not have any prior information about the class (regular or outlier) each observation belongs to (unsupervised learning). Finally, in Section 5 we present some conclusions.

2 Principal Components Analysis

PCA seeks to maximize the variance of uncorrelated linear combinations of the original variables [8], called principal components (PCs). If a small number of PCs explain a large proportion of the total variance of the p original variables, then PCA can be successfully used as a dimension reduction technique. Given the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_p)^t$, with expected value $\boldsymbol{\mu}$, the j-th principal component is defined as the linear combination, $Z_j = \boldsymbol{\alpha}^t_j(\mathbf{X} - \boldsymbol{\mu})$, such that $\boldsymbol{\alpha}^t_j \boldsymbol{\alpha}_j = 1$, Z_j has maximum variance and is uncorrelated with the previous PCs (for $j \geq 2$). It can easily be proved that the loadings, $\boldsymbol{\alpha}_j$, and the variances of the PCs, $\lambda_j = \mathrm{Var}(Z_j)$, are, respectively, the eigenvectors and eigenvalues of the covariance matrix of $\mathbf{X}$, where the eigenvalues are arranged by decreasing order of magnitude.

In real applications, $\boldsymbol{\alpha}_j$ and λ_j have to be estimated. Given $\mathbf{x}_i$, the values associated with the random vector $\mathbf{X}$ on subject i, the score of subject i on the j^{th} PC is given by $z_{ij} = \hat{\boldsymbol{\alpha}}_j^t(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\alpha}}_j$ denote the estimates of the expected value and the loadings, and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^t$. Several estimation methods have been proposed in the literature. The classical method (called CPCA, in the following sections) uses the eigenvalues and eigenvectors of the sample covariance matrix to estimate λ_j and $\boldsymbol{\alpha}_j$. Even though this is an easy and fast estimation method, it is known that the presence of atypical observations can strongly influence the estimation process, leading to biased estimates. Several robust estimation methods have been proposed to overcome this limitation. In this paper, we are going to consider some of these methods.

Locantore et al. [10] proposed a simple methodology that performs classical PCA to the projected centered data onto the $p-$sphere with unitary radius (this will be referred as SPHE). The method requires low computation time and has good performance, reasons that justify it still being used.

The Projection Pursuit (PP) approach [5, 6] seeks for projections of multivariate data that reveal interesting structures of the data. In the case of PCA, we look for the univariate projection of the data that maximizes a robust scale and proceed looking for the second direction, orthogonal to the first, with maximum robust scale, and so on [1, 5, 6]. Several variants of these ideas have been explored leading to different estimation methods. In this paper we will consider two of these alternatives: one proposed in [5], and referred as PCAPROJ, and another one proposed in [6], called PCAGRID.

One of the most popular approaches to construct robust principal components is the method proposed in [7], ROBPCA. This method combines PP ideas with robust scatter matrix estimation. It is known that the method produces accurate estimates for non-contaminated datasets and robust ones for contaminated data. Besides this, it is computationally fast and can be applied to datasets with more variables than observations.

Recently, [1] described a new robust principal component method (WCPCA) intended to handle high-dimensional data eventually containing atypical observations. Its good performance associated with its computational efficiency are considerable advantages to take into account.

3 Detection of Outliers Using Principal Components

PCA can be used to detect atypical observations and, as such, it can be regarded as a classification procedure. Given an observation, it belongs to one of the following classes: (i) regular, non-outlier or non-atypical; or (ii) anomaly, outlier or atypical[1].

Usually, in a classification problem the data is divided in two sets: the training and the test set. The training set is used to train the classifier and the

[1] In some domains these observations are also known as “discordant”, “exceptions”, “aberrations”, “surprises”, “peculiarities”, or “contaminants”; vide [4].

test set to estimate error rates. In Telecommunications, several authors have detected anomalies using CPCA, admitting they have a training set that contains only regular observations [9], a scenario which is called semi-supervised learning. However, this is an unrealistic assumption since the available methods to determine if an observation is an anomaly are not 100% trustful. Thus, the investigator can not guarantee that the training set is composed only of regular observations, and this is the principal reason for the unsatisfactory performance of the method [2]. A more realistic situation is the one where the class (regular or outlier) to which the training observation belongs to is unknown. This is called unsupervised learning.

Traditionally, statisticians pay more attention to the estimation process, using all the available data to estimate the classifier. Thus, the observations used to train are also used to classify leading to optimistic estimates of the error rates. Henceforth, results and discussions are mainly oriented towards describing and comparing the performances of the estimation methods and less to the performance of the classifiers and the estimation of error rates (cf. e.g. [1]).

In the semi-supervised scenario the method is described by the following steps:

1. Estimate the first k PCs and their variances, λ_j, $j = 1,\ldots,k$, based on the data set composed only of regular observations;
2. Given a new observation, $\mathbf{x}_0$, project it into the subspace obtained in 1.; the projection is henceforth denoted by $\mathbf{z}_0$;
3. Calculate the Mahalanobis or Score Distance, $\mathrm{SD}(\mathbf{x}_0) = \left(\sum_{j=1}^{k} z_{0j}^2/\lambda_j\right)^{1/2}$, of the projected observation;
4. If the scores are normally distributed, classify $\mathbf{x}_0$ as an outlier if $\mathrm{SD}(\mathbf{x}_0) > \left(\chi^2_{k,1-\alpha}\right)^{1/2}$ (vide [7]). Otherwise, classify $\mathbf{x}_0$ as an outlier if $\mathrm{SD}(\mathbf{x}_0) > Q_{1-\alpha}$ where $Q_{1-\alpha}$ is the $1-\alpha$ quantile of the score distances calculated for the training set [11] and α, known as false alarm rate, is the probability of an observation being classified as outlier when in fact is a regular observation.

The previous procedure only uses the major (or highest) k principal components. By contrast, [11] proposes a modified procedure that also uses minor (or lowest) PCs to detect outliers. Let α_M (α_m) denote the probability that the major (minor) k (r) PCs detect an outlier, when in fact it is a regular observation. The procedure proposed in [11] starts with steps $1.-4.$ with $\alpha = \alpha_M$. Note that in 1. we need to estimate all the p PCs. The procedure is then completed with the following two steps:

5. Calculate the score distance of the projection of the new observation into the space spanned by the lowest r PCs, $\mathrm{SD}^{\mathrm{m}}(\mathbf{x}_0) = \left(\sum_{j=p-r+1}^{p} z_{0j}^2/\lambda_j\right)^{1/2}$;
6. Classify $\mathbf{x}_0$ as an outlier if $\mathrm{SD}^{\mathrm{M}}(\mathbf{x}_0) > Q_{1-\alpha_{\mathrm{M}}}$ or $\mathrm{SD}^{\mathrm{m}}(\mathbf{x}_0) > q_{1-\alpha_{\mathrm{m}}}$ where $Q_{1-\alpha_{\mathrm{M}}}$ ($q_{1-\alpha_{\mathrm{m}}}$) is the $1-\alpha_{\mathrm{M}}$ ($1-\alpha_{\mathrm{m}}$) quantile of the score distances calculated for the k major (r minor) PCs obtained from the training set [11].

Considering that the major and minor PCs are independent, then the global false alarm rate is $\alpha = \alpha_M + \alpha_m - \alpha_M \alpha_m$, and α_M and α_m should be chosen to reflect the relative importance of the types of outliers we would like to detect. However, as we are not interested in distinguishing the importance of different types of outliers, we choose $\alpha_M = \alpha_m$.

This methodology can be seen as unrealistic since it demands a training set with only regular observations, which in pratice can be difficult to obtain. The unsupervised learning approach constitutes a more realistic scenario. In this case, no information about the label of the data in the training set is available. In our simulation study, we admit that the labels for the test set are always known, which leads to more trustful estimates of the error rates and fairer comparisons between the scenarios under study.

For the unsupervised case the procedure is:

1. Estimate the first k PCs of the training set and their variances, λ_j, $j = 1,\ldots,k$;
2. Calculate the score distance and the orthogonal distance of the training set observations projected into the space spanned by the first k PCs determined in 1. The Orthogonal Distance is defined by $\mathrm{OD}(x_i) = \left\|(\mathbf{x}_i - \boldsymbol{\mu}) - \mathbf{P}_{p,k}\mathbf{z}_i\right\|$, $i = 1,\ldots,n$, where $\mathbf{P}_{p,k}$ is the $p \times k$ matrix having by columns the loadings of the first k PCs.
3. Classify a new observation $\mathbf{x}_0$ as an outlier if $\mathrm{SD}(\mathbf{x}_0) > \left(\chi^2_{k,1-\alpha}\right)^{1/2}$ ([7]) or $\mathrm{OD}(\mathbf{x}_0) > (\hat{\mu} + \hat{\sigma}\Phi^{-1}(1-\alpha))^{3/2}$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimates of the mean, μ, and standard deviation, σ, of the orthogonal distances obtained from the training set and $\Phi(\cdot)$ denotes the standard normal distribution function. Thresholds based on 3 different pairs of estimates are considered:
   a. The location and scale univariate MCD estimates, $\hat{\mu}_{\text{MCD}}$ and $\hat{\sigma}_{\text{MCD}}$ [7];
   b. $\hat{\mu}_{\text{MEDIAN}}$ and $\hat{\sigma}_{\text{MAD}}$;
   c. $\hat{\mu}_{\text{MEDIAN}}$ and $\hat{\sigma}_{\text{Q}}$ [1].

Even though the different thresholds may seem similar and their differences irrelevant, they play an important role in the performance of anomaly detection methods.

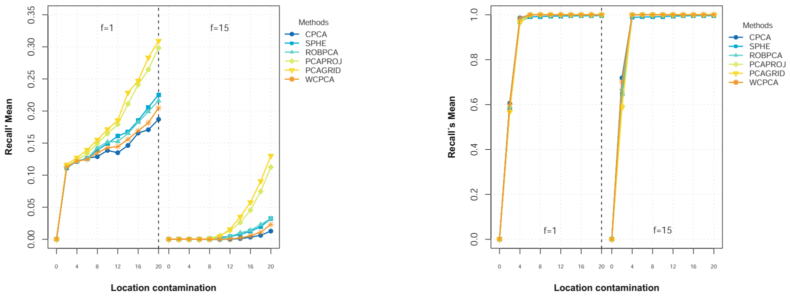
In the next section, we present our simulation study and discuss the results in terms of measures of the classification procedure. For that, we use: (i) Recall, the probability that an observation is classified as outlier when in fact is an outlier; and (ii) False positive rate, the probability that and observation is classified as outlier, when in fact is a regular observation (also known as false alarm rate).

4 Comparing the Methods by a Simulation Study

A simulation study was develop to compare the performance of the outlier detection methods based on PCA. In future work, other (robust and non-robust) alternatives should be considered as, e.g., the ones presented in [3] and [12].

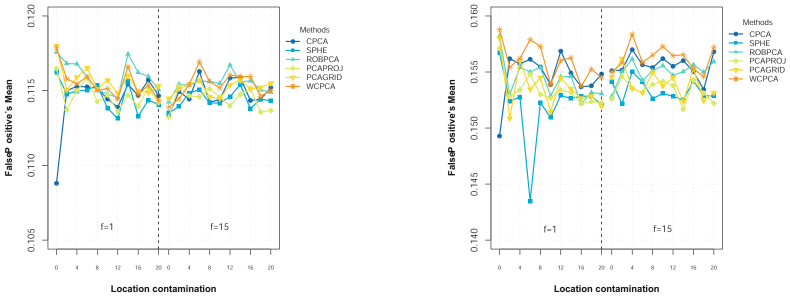
For the simulation study we follow a setup similar to Case 1 in [1]. We consider 6 estimation methods previously mentioned (CPCA, SPHE, ROBPCA, PCAPROJ, PCAGRID, WCPCA) and the semi-supervised scenario was subdivided in two cases: using the two major PCs and using the two major and the minor PCs.

To construct different contaminated data sets, we randomly generate 500 samples of size $n = 100$ from: $(1-\varepsilon)N_p(\mathbf{0},\boldsymbol{\Sigma}) + \varepsilon N_p(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}/f)$, $i = 0,1,\ldots,10$, $f = 1,15$, $\varepsilon = 0.1$, $\boldsymbol{\mu}_i = (0,0,0,2i)^t$, $i = 0,1,\ldots,10$, and $\boldsymbol{\Sigma} = \mathrm{diag}(8,4,2,1)$. The



(a) (b)

Fig. 1 Average of the estimated recall obtained by the semi-supervised approach with (a) 2 major PCs and (b) 2 major and 1 minor PCs, for 20 contaminated scenarios using the CPCA, SPHE, ROBPCA, PCAPROJ, PCAGRID, WCPCA, ($\alpha = 0.1$).



(a) (b)

Fig. 2 Average of the estimated false positive rate obtained by the semi-supervised approach with (a) 2 major PCs and (b) 2 major and 1 minor PCs, for 20 contaminated scenarios using CPCA, SPHE, ROBPCA, PCAPROJ, PCAGRID, WCPCA, ($\alpha = 0.1$).

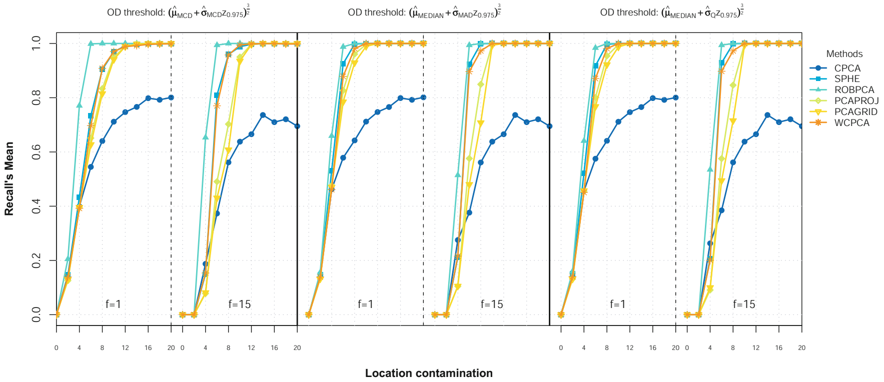


Fig. 3 Average of the estimated recall obtained by the unsupervised approach considering 3 different thresholds for the orthogonal distance, for 20 contaminated scenarios using CPCA, SPHE, ROBPCA, PCAPROJ, PCAGRID, WCPCA, $(\alpha = 0.1)$.

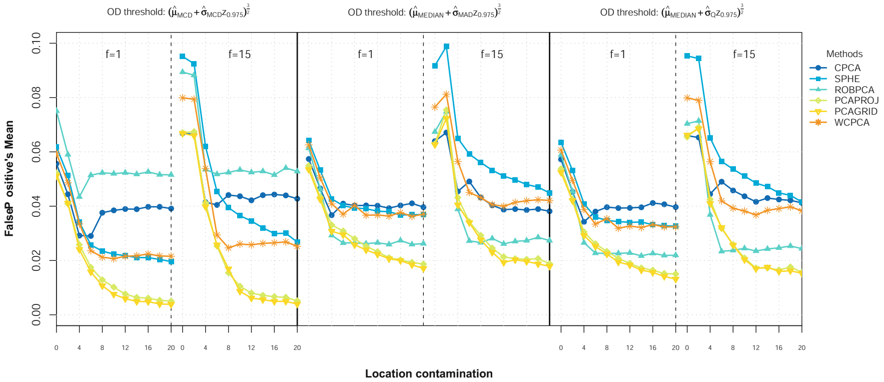


Fig. 4 Average of the estimated false positive rate obtained by the unsupervised approach considering 3 different thresholds for the orthogonal distance, for 20 contaminated scenarios using CPCA, SPHE, ROBPCA, PCAPROJ, PCAGRID, and WCPCA, $(\alpha = 0.1)$.

considered value for the global false alarm rate is $\alpha = 0.1$. Note that $i = 0$ and $f = 1$ implies that the data is not contaminated.

Taking into account the results obtained for the simulation study using the semi-supervised approach presented in Fig. 1 and 2, it is easy to notice that the minor PC seems to have an important role on the detection of outliers, with its use leading to higher values of the estimated recall, reaching one. The inclusion of one minor PC implies that almost all outliers are detected even in cases with soft contamination, e.g. $f = 15$ and $\mu = (0,0,0,4)^t$. The drawback of this option is a slightly higher estimated false positive rate.

For the unsupervised approach, ROBPCA has the highest and CPCA the lowest estimated recall, vide Fig. 3. From the thresholds point of view, the choice μ_{MEDIAN} and σ_{MAD} lead to greater or equal estimated recall for all methods except for ROBPCA, where the parameters estimated based on the univariate MCD constitute a better choice. If we want to choose a method to detect outliers with the smallest estimated false positive rate, PCAGRID and PCAPROJ are the best methods, vide Fig. 4. In this case, the best threshold is the one using MCD estimators, except for ROBPCA where the choice should be the one using μ_{MEDIAN} and σ_{Q}.

5 Conclusions

In this paper we develop a simulation study to compare the performance of several robust principal component methods to detect outliers. In order to do so, several contamination schemes are considered, and recall and false positive rate estimates were the two measures used to summarize the results and identify the best outlier detection method.

Taking into account the results, if the investigator has a training set with only regular observations we should choose a classifier based on major and minor PCs. However, a more complete and exhaustive study should be done in order to justify this decision and the choice of the number of major and minor PCs to use needs to be addressed. As expected, if the training set does not contain outliers, the best estimates for recall are obtained with CPCA. Nevertheless, all other robust methods produced very high values for estimated recall. Moreover, for soft contamination schemes, WCPCA slightly outperforms the other robust methods. If no information is available about the class of each training observation, the results are slightly worse. That is, the classifiers have more difficulties in detecting outliers in the presence of soft contamination. However, under more severe contamination all robust methods considered reached an estimated recall of 1. Moreover, ROBPCA lead to the highest and CPCA to the lowest estimated recall. This means that training with outliers using non-robust methods has a major negative impact in the performance of the classifier.

References

1. Branco, J., Pires, A.M.: Robust Principal Component Analysis for High-Dimensional Data (submitted for publication, 2010)
2. Casas, P., Fillatre, L., Vaton, S., Nikiforov, I.: Volume Anomaly Detection in Data Networks: an Optimal Detection Algorithm vs. the PCA Approach. In: Valadas, R. (ed.) FITraMEn 2008. LNCS, vol. 5464, pp. 96–113. Springer, Heidelberg (2008)

3. Cerioli, A.: Multivariate Outlier Detection with High Breakdown Estimators. J. Amer. Statist. Assoc. 105(4), 147–156 (2010)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. ACM Computing Surveys 41(3), 1–58 (2009)
5. Croux, C., Ruiz-Gazen, A.: High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited. J. Multivariate Anal. 95, 206–226 (2005)
6. Croux, C., Filzmoser, P., Oliveira, M.R.: Algorithms for Projection-Pursuit Robust Principal Component Analysis. Chemometrics Intell. Laboratory Syst. 87(2), 218–225 (2007)
7. Hubert, M., Rousseeuw, P.J., Branden, K.V.: ROBPCA: a New Approach to Robust Principal Component Analysis. Technometrics 47, 64–79 (2005)
8. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
9. Lakhina, A., Crovella, M., Diot, C.: Diagnosing Network-Wide Traffic Anomalies. In: Proceedings of 3rd ACM SIGCOMM workshop on Network and system support for games, Portland, Oregon, USA, pp. 219–230. ACM, New York (2004)
10. Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L.: Robust Principal Component Analysis for Functional Data. Test 8(1), 1–73 (1999)
11. Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., Chang, L.W.: A Novel Anomaly Detection Scheme Based on Principal Component Classifier. In: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, ICDM 2003, Melbourne, FL, USA, pp. 172–179. IEEE Computer Society, Los Alamitos (2003)
12. Willems, G., Joe, H., Zamar, R.: Diagnosing Multivariate Outliers Detected by Robust Estimators. J. Comput. Graph. Statist. 18(1), 73–91 (2009)
13. Xuexiang, J., Yi, Z., Li, L., Jianming, H.: Robust PCA-based Abnormal Traffic Flow Pattern Isolation and Loop Detector Fault Detection. Tsinghua Sci. Technol. 13(6), 829–835 (2008)

Exploiting Sparse Dependence Structure in Model Based Classification

Tatjana Pavlenko and Anders Björkström

Abstract. Sparsity patterns discovered in the data dependence structure were used to reduce the dimensionality and improve performance accuracy of the model based classifier in a high dimensional framework.

Keywords: Classification, High dimensionality, Sparsity, Lasso, Variable selection.

1 Introduction

We focus on the classification problem which is concerned with the allocation of a given object, $\mathbf{X}$, represented by a set of features $(X_1,\dots,X_p)$, to one of known classes, Π_i, $i = 1,\dots,c$. Let $\mathscr{Y} : \mathbb{R}^p \to \{1,\dots,c\}$ be a non-randomized decision rule with Borel measurable decision regions $\Omega_i \subset \mathbb{R}^p$, $\Omega_i = \mathscr{Y}^{-1}(i)$ corresponding to class Π_i. In the model-based setting, we assume that classes, Π_is, are represented by the densities, $f(\mathbf{x};\theta^i)$, and a *priori* probability of Ω_i is π_i. Then the optimal decision rule minimizing the posterior misclassification probability $\mathscr{E}(\mathscr{Y}, f(\mathbf{x};\theta)) = \sum_{j=1}^{c} \pi_j P(\mathscr{Y} \neq j|\mathbf{x} \in \Omega_j)$ is based on the Bayes theorem

$$P(\mathscr{Y} = i|\mathbf{x},\theta^i) = \frac{\pi_i f(\mathbf{x};\theta^i)}{\sum_{j=1}^{c} \pi_j f(\mathbf{x};\theta^j)} \tag{1}$$

and assigns the observed $\mathbf{x}$ to the class Π_i for which a class posterior probability $P(\mathscr{Y} = i|\mathbf{x};\theta^i)$ is maximum.

Tatjana Pavlenko
Dept. of Statistics, Stockholm University, Sweden
e-mail: Tatjana.Pavlenko@stat.su.se

Anders Björkström
Dept. of Mathematics, Stockholm University, Sweden
e-mail: bjorks@math.se.se

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 509–517.
springerlink.com

The performance accuracy of the sample based classifier (1) is known to be poor in a high-dimensional framework, i.e. if the sample size n and dimensionality p are both large, and the decline in accuracy is especially pronounced if $p > n$. To overcome this problem, we suggest a two-stage procedure which first groups the feature variables into blocks using the sparsity patterns of the underlying dependence structure, and then exploits the Lasso technique [3] for selection the variables with highest classification potential at the block level. This approach allows for substantial global dimensionality reduction while maintaining the misclassification probability at a certain desired level. To capture the high-dimensionality of the model (1) in an asymptotic sense, we allow the number of variables $p = p_n$ and the parameters $\theta^i = \theta^i_n$ depend on the sample size n. To indicate that parameters can change with n we turn to the sequence of classification problems $\{p, n_i, f(\mathbf{x};\theta), \mathscr{E}, \phi_i\}_n, \quad n = 1, 2, \ldots,$ instead of the isolated one, and allow p grow to infinity together with $n \to \infty$ in such a way that $p/n_i \to \phi_i < \infty$, $i = 1, \ldots, c$. It is in particular possible for the following analysis that the number of variables is much larger than the number of observations, $p \gg n_i$.

The principle difference of this approach from the standard asymptotics can be demonstrated by the following example: Let $X_1, \ldots, X_n$ be a sample of independent observations from $\mathscr{N}_p(0, \sigma^2 I)$. The square length of $\bar{X}$, i.e. square of its ℓ_2-norm is $||\bar{x}||_2^2 = \sum_{i=1}^p (\bar{x}^{(i)})^2 = \sum_{i=1}^p \left(\sum_{j=1}^n x_j^{(i)}/n\right)^2 \to \phi\sigma^2 > 0 \quad \text{as} \quad p, n \to \infty$ unlike the standard asymptotic where $||\bar{x}||_2^2$ would converge to 0.

2 Covariance Structure and Classification Accuracy

To exploit the covariance structure for reducing the dimensionality, we first turn to the binary classifier and model each class conditional distribution as $X \sim \mathscr{N}_p(\mu_i, \Sigma)$ given the class variable $\mathscr{Y}$, $i = 0, 1$. Then the Bayesian classifier (1) is equivalent to

$$\mathscr{L}(\mathbf{x}; \mu_0, \mu_1, \Sigma) = \Big(\mathbf{x} - \frac{1}{2}(\mu_1 + \mu_0)\Big)' \Sigma^{-1}(\mu_1 - \mu_0) \lesseqgtr \ln\frac{\pi_0}{\pi_1} \to \mathbf{x} \in \begin{cases} \Omega_0 \\ \Omega_1 \end{cases} \qquad (2)$$

which is known as the Anderson-Fisher discriminant that preserves the ordering of class posterior probabilities and can therefore be used instead of them for classification. The posterior misclassification probability of this rule can be computed as

$$\mathscr{E}_{opt} = P(\mathscr{L}(\mathbf{X}; \mu_0, \mu_1, \Sigma) \leq 0 | \mathbf{X} \in \Omega_1) = \boldsymbol{\Phi}\Big(-\frac{1}{2}\sqrt{\delta(\mu_0, \mu_1, \Sigma)}\Big), \qquad (3)$$

where $\boldsymbol{\Phi}$ is the Gaussian cumulative distribution function and $\delta(\mu_0, \mu_1, \Sigma) = (\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)$ is the Mahalanobis distance between the classes. Observe that (3) directly relates the classification accuracy to the structure of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

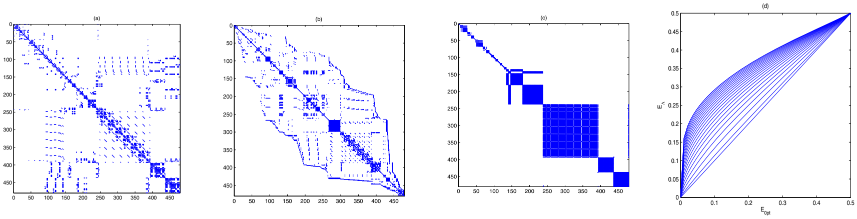


Fig. 1 Sparse covariance matrix using spy-plot technique, non-zero patterns are colored in blue in (a)-(c) (a) A model of covariance with sparsity patterns, number of non-zero elements is about 3% (b) The same covariance structure after the reverse Cuthill-McKee reordering algorithm. (c) A model of the block-diagonal approximation. (d) Classification accuracy for $b_i = 1$ and τ going from 1 to 20.

An example of sparsity patterns discovered in the covariance structure is illustrated by the panel (a) of Figure 1. We apply Cuthill-McKee reordering algorithm, a permutation invariant transform of $\boldsymbol{\Sigma}$ (see e.g. [1] and references therein) that moves all non-zero elements closer to the diagonal thereby reducing the bandwidth of the original matrix. Observe that the permutation invariance of the covariance structure with respect to indexing the class predictor variables is a key feature associated with the suggested variable selection technique. The panel (b) of Figure 1 shows an example of such a transform of $\boldsymbol{\Sigma}$. and in the panel (c) we infer a block-diagonal approximation of $\boldsymbol{\Sigma}$ with a given number of sub-matrices $\boldsymbol{\Sigma}_{[b_i]}$ having various degree of remaining within-block sparsity where $i = 1, \ldots, b$ The main advantage of the block-dependence structure in a high-dimensional framework will be shown in Sections 4 and 5.

It turns out that the block-diagonal approximation of the covariance matrix while reducing the dimensionality does not lead to a serious decline of the performance accuracy of (2) assuming that the true $\boldsymbol{\Sigma}$ is sparse and well-conditioned. An impression of the increase of the misclassification probability induced is given by an *extreme* case of the approximation where $\boldsymbol{\Sigma}$ in the classifier $\mathscr{L}(\mathbf{x}; \mu_i, \boldsymbol{\Sigma})$ is replaced $\boldsymbol{\Lambda} = \mathrm{Diag}[\Sigma]$, i.e. all off-diagonal elements of $\boldsymbol{\Sigma}$ are replaced by zero. This is a special case of the block structure with $b_{[i]} = 1$, $i = 1, \ldots, p$ For a Gaussian class conditional model the misclassification probability $P(\mathscr{L}(\mathbf{X}; \mu_i, \boldsymbol{\Sigma}) \leq 0 | \mathbf{X} \in \boldsymbol{\Omega}_1)$ is then given by

$$\mathscr{E}_\Lambda = \boldsymbol{\Phi}\Big(-\frac{1}{2} \frac{(\mu_1 - \mu_0)' \Lambda^{-1} (\mu_1 - \mu_0)}{\sqrt{(\mu_1 - \mu_0)' \Lambda^{-1} \Sigma \Lambda^{-1} (\mu_1 - \mu_0)}} \Big), \tag{4}$$

and $\mathscr{E}_\Lambda > \mathscr{E}_{\mathrm{opt}}$. Now we notice that $\mathscr{E}_{\mathrm{opt}}$ is a monotone decreasing function of the Mahalanobis distance, i.e. $\delta(\mu_0, \mu_1, \boldsymbol{\Sigma}) = -2\boldsymbol{\Phi}^{-1}(\mathscr{E}_{\mathrm{opt}})$, and hence $\delta(\mu_0, \mu_1, \boldsymbol{\Sigma})$ and $\mathscr{E}_{\mathrm{opt}}$ provide equivalent information about the classification performance. Then we evaluate the relative accuracy of $\mathscr{E}_\Lambda$ to $\mathscr{E}_{\mathrm{opt}}$ by computing the quot of the arguments of $\mathscr{E}_\Lambda$ and $\mathscr{E}_{\mathrm{opt}}$. By denoting $\mu = \boldsymbol{\Lambda}^{-1/2}(\mu_1 - \mu_0)$ and $\boldsymbol{\Psi} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Sigma} \boldsymbol{\Lambda}^{-1/2}$, and after some rearrangements we get

$$Q^2 = \frac{\arg(\mathscr{E}_{\text{opt}})}{\arg(\mathscr{E}_{\Lambda})} = \frac{\mu'\Psi^{-1}\mu \cdot \mu'\Psi\mu}{(\mu'\mu)^2}.$$

Now using the matrix version of Cauchy inequality for a symmetric positive defined $p \times p$ matrix $\mathscr{A}$ and any $p \times 1$ vector u (see [5])

$$(u'\mathscr{A}^{-1}u)(u'\mathscr{A}u) \leq \frac{[\alpha_{\min}(\mathscr{A}) + \alpha_{\max}(\mathscr{A})]^2}{4 \cdot \alpha_{\min}(\mathscr{A})\alpha_{\max}(\mathscr{A})} \cdot (u'u)^2, \tag{5}$$

where $\alpha_{\min}(\mathscr{A})$ and $\alpha_{\min}(\mathscr{A})$ are maximal and minimal eigenvalues of $\mathscr{A}$, respectively, we obtain $Q^2 \leq (1 + \tau(\Psi))^2/4\tau(\Psi)$, where $\tau(\Psi) = \alpha_{\max}(\Psi)/\alpha_{\min}(\Psi)$ is the condition number of the matrix Ψ. Further, by monotonicity of $\Phi(\cdot)$ the upper bound of the misclassification probability $\mathscr{E}_{\Lambda}$ is found as

$$\mathscr{E}_{\Lambda} \leq \Phi\Big(-\frac{1}{2}\frac{\sqrt{4\tau\delta(\mu_0,\mu_1,\Sigma)}}{1+\tau}\Big) \tag{6}$$

Now, recall that for a reasonable decision rule we must stipulate that $0 < const \leq \delta(\mu_0,\mu_1,\Sigma) \leq \infty$ in (3). Thus, if the eigenvalues of Σ are close to 0 or ∞ as $p \to \infty$ so that $\tau \to \infty$ then $\arg\mathscr{E}_{\Lambda}$ in goes to 0 resulting in the upper bound of misclassification probability of 1/2. However, numerical analysis of (6) shows that for a moderate values of τ an increase of $\mathscr{E}_{\Lambda}$ is not that pronounced event for the $b_{[i]} = 1$ for all i, as illustrated in Figure 1(d). Hence when approximating Σ with the structure illustrated in Figure 1(c), the decline in the performance accuracy will even be less observed due to capturing more true entire dependence.

3 Sparsity Conditions and Approximation Accuracy

To demonstrate the importance of the sparsity assumption, we consider the classifier (2) with $\Sigma = \mathscr{M}_p(i,j) = \sigma^2\mathbf{1}_p\mathbf{1}'_p + (1-\sigma^2)I_p$, where $\mathbf{1}_p$ is $p \times 1$ vector of ones and I_p is the $p \times p$ identity matrix. This means that all the features are assumed to be equally dependent exhibiting thereby no sparsity patterns in the covariance structure ($\mathscr{M}_p$ is zero-sparse). Then the p eigenvalues of $\mathscr{M}_p$ are $\lambda_1 = 1 + (p-1)\sigma^2$, and $\lambda_i = 1 - \sigma^2$ for the remaining $i = 2,\ldots,p$ so that $\tau(\mathscr{M}_p) \to \infty$ as $p \to \infty$, and thus by applying the bounding condition(5) we see that for large enough values of p the resulting asymptotic misclassification $\mathscr{E}_{\Lambda}$ tends to 1/2 so that the classification does no better than a pure guess work. Such a decline in the classification accuracy would be anticipated: since the true covariance matrix is zero-sparse the block-diagonal approximation totally ignores the underlying dependence in the classification procedure. On the other hand, inducing a certain degree of sparsity on Σ in (3) makes it possible to bound $\mathscr{E}_{\Lambda}$ away from 1/2 when using the approximation. To show this we let $\Sigma = T_p(i,j) = p^{-1/2}1_{\{|i-j|\leq 2\}}$ in the classifier (2) and find the eigenvalues λ_k of T_p to equal $1 + 2p^{-1/2}\cos(k\pi(p+1)^{-1})$ for $k = 1,\ldots,p$, so that $\tau(T_p) \to const$ as $p \to \infty$ which ensures that the corresponding classifier does not degenerate. Observe that the matrix T_p in this example is a special case of the symmetric Toeplitz structure of the order 2 and the degree of sparsity induced on Σ is of order $\mathscr{O}(p)$.

The observed phenomena can intuitively be explained as a *trade-off* that occurs with the block-diagonal covariance approximation in a way that a small increase in the misclassification probability can be traded for an essential dimensionality reduction that led to better classification results.

4 Block-Wise Variable Selection with the Lasso

The discovered sparsity and block-diagonal covariance structure can be related to the grouping of feature variables and, for a Gaussian model, zero patterns in the covariance structure can be equated with the variables independence. This means that observed vector $\mathbf{x}$ is decomposable into a family of b independent groups (blocks) so that for each given b_i $\mathbf{x}_i \in R^{b_i}$. Furthermore, since Σ is a block-diagonal, its inverse is also block-diagonal with each diagonal block equal to $\Sigma_{[b_i]}^{-1}$ so that the within-block dependence structure is preserved. For a non-Gaussian $f(\mathbf{x};\theta^j)$,+ the sparsity can be related the *asymptotic* independence, that is, for a large p the joint density of any two blocks $\mathbf{x}_i = (x_{i1},\dots,x_{im})$ and $\mathbf{x}_j = (x_{j1},\dots,x_{jm})$ such that $\mathbf{x}_i \cap \mathbf{x}_j = \varnothing$ can be approximated by the product of densities. The disjointness assumption is essential to our classification approach, as we see below.

The approximation of Σ represented in Figure 1(c) has another advantage of allowing a flexible amount of sparsity within the discovered blocks. We focus on those groups of variables for which $\Sigma_{[b_i]}$ remains sparse and show that the within-block Lasso can be very useful as a second stage of our procedure suggested for discovering the variables with high predictive potential.

Firstly, to embed the Lasso technique in the classification framework we turn to the logistic regression model where the class variable $\mathscr{Y}$ is assumed to be a binary response whose conditional distribution given a specific block, $\mathbf{x}_i$, is modelled as a Bernoulli random variable with parameter $\exp(\beta'\mathbf{x})/(1+\exp(\beta'\mathbf{x}))$, where β is the parameter vector corresponding to the ith block. Then, by the technique discussed in [3], the ℓ_1-penalized maximum loglikelihood estimate for β can be found as $\hat{\beta}_\lambda = \arg\min_\beta \Big(l(\beta;y) + \lambda||\beta||_{\ell_1}\Big)$ where $l(\beta;y) = \sum_{j=1}^n \Big(y_i\beta'\mathbf{x}_i - \log(1+\exp(\beta'\mathbf{x}_i))\Big)$ is the log-likelihood function constructed by n observations $(y_j;[\mathbf{x}_i]_j)_{j=1}^n$, and $\lambda \geq 0$ is the tuning parameter that controls the amount of penalization.

In our numerical examples we set class priors $\pi_0 = \pi_1 = 1/2$, fix $b_i = 160$, $n_1 = n_2 = n = 100$ and let the mean vectors μ_0 and μ_1, covariance matrix $\Sigma_{[b_i]}$ and the penalty parameter λ vary in the experiments. To encourage the block-wise sparsity we let most of the coordinates of the block shift vector $\Delta = \mu_1 - \mu_0$ be zero which means that the most of predictor variables in the set $\mathbf{x_i}$ are irrelevant for accurate classification. In some of our simulations we have used $\Delta = 0$, that is, $\mu_0 = \mu_1$, in order to see how the Lasso performs when *no* feature is informative for predicting $\mathscr{Y}$. Due to sparse data structure this is a situation we expect to meet for many of the blocks. In addition to the *number* of nonzero components in Δ, one can also vary the *absolute value* of Δ. As for the second factor, covariance matrix, we generate two types of $\Sigma_{[b_i]}$ structures, one with conditional number $\tau = 10$ and the other one with $\tau = 10^3$. We let the eigenvalues λ_is of $\Sigma_{[b_i]}$ progress in equal steps from 0.2 to 2 for the former case, and from 0.02 to 20 in the latter case. For this setting, the Lasso based technique for the ith block is described in the following algorithm:

```
Step 1: Generate $n$ independent outcomes of $\mathscr{Y} \in Be(\pi_1)$.
Step 2: For each observed $y_j$ generate a $(n \times b_i)$-matrix $X$ where the $j$:th row
    $X_{j,\cdot} \sim \mathscr{N}(\mu_y, \Sigma_{[b_i]})$, $y = \mathscr{Y}_j$, $j = 1, \ldots, n$
Step 3: Use glmnet with $X$ and $y$ as input, to compute a sequence $\hat{\beta}_\lambda$ for a range
    of $\lambda$.
Step 4: Find $\hat{\lambda}_5 = arg\max_\lambda \hat{\beta}_\lambda$ for which the number of nonzero in $\beta_\lambda$ is at least 5.
Step 5: Check whether the relevant feature is among those 5.
Step 6: Repeat steps 1-5 many times and check how often the method succeeds
    in finding the correct feature.
```

Observe that this algorithm can also serve as a block selector: if the resulting sequence $\hat{\beta}_\lambda$ is similar to the left panel in Figure 2, this indicates that the block can be discarded from the model as non-informative. In the case of ill-conditioned $\Sigma_{[b_i]}$, i.e.for $\tau = 10^3$ one may apply PLS technique to construct the block linear predictor $\sum_{k=1}^{b_i} c_k x_k$ and check its predictive potential for $\mathscr{Y}$ as described in e.g. [2].

Typically, the estimated Lasso traces look like the three graphs in Figure 2, depending on whether the true Δ has zero, one or two nonzero components. There are zero, one or two coefficients of the vector $\hat{\beta}_\lambda$ that are significantly larger than the others, and they remain (stably) non-zero for higher values of the tuning parameter λ. This hints that the Lasso trace is a useful way to identify those feature variables that are highly relevant for predicting $\mathscr{Y}$. However, simulations do not always yield plots that belong distinctly to the appropriate type. Even when no feature is relevant, we get a pattern presented

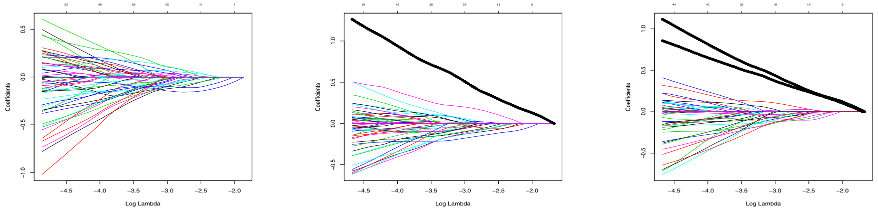


Fig. 2 Lasso coefficients $\hat{\beta}_\lambda$ vs $\log(\lambda)$ for $j = 1, \ldots, 160$, in a case when non (left panel), one (middle panel), or two (right panel) of the features are highly predictive for $\mathscr{Y}$.

Table 1 Percentages of times when 0, 1, or 2 of the relevant features are among the 5 captured, for various combinations of τ and (Δ_1, Δ_2).

$\tau =$	10			10^3		
$(\Delta_1, \Delta_2) =$	(0.3,0.3)	(0.3, 1.0)	(1.0, 1.0)	(0.3,0.3)	(0.3, 1.0)	(1.0, 1.0)
0	75%	12%	2%	82%	22%	4%
1	24%	80%	23%	18%	72%	32%
2	1%	8%	75%	0%	6%	64%
Sum	100%	100%	100%	100%	100%	100%

in the middle panel of Figure 2 with roughly 20 % probability. When one feature is relevant, the chance for the corresponding trace to stand out in the plot increases with $||\Delta||_2$, but is far from 100 % even when $||\Delta||_2 \approx 1$. After visual inspection of a number of trace plots, we have concluded that only when $1.5 \leq ||\Delta||_2 \leq 2$ we can be fairly certain that the relevant feature will stand out in the plot. The necessary size of $||\Delta||_2$ appears to be smaller if the condition number of $\Sigma_{[b_i]}$ is $\tau = 10^3$ than when $\tau = 10$. Since irrelevant features will sometimes also yield traces that stand out from the majority, we do not want to increase the penalty parameter λ further than to have at least 5 x-variables remaining in the model.

By simulations, we have explored the probability that the relevant feature(s) will be included among the 5 captured, for various different types of Δ and $\Sigma_{[b_i]}$. When only one component of Δ is nonzero, this feature is found in 14% of all cases when $\Delta = 0.3$, and in 90% of all cases when $\Delta = 1.0$, in the well-conditioned case. In the ill-conditioned case, the relevant feature is found in 10% of the cases when $\Delta = 0.3$, and 75% when $\Delta = 1.0$. Not surprisingly, the chance to identify the correct component increases with the distance between μ_1 and μ_2. The case $\tau = 10$ performs somewhat better than $\tau = 10^3$.

When two components of Δ are nonzero (say, Δ_1 and Δ_2), none, one or both of these may be included in the model among the 5 captured. Table 1 shows in how many out of 100 cases these respective outcomes were observed, for various condition number τ.

5 Generalised Additive Classifier

The block-wise Lasso variable selection is especially efficient for sparse high-dimensional data because the ℓ_1 penalty allows for controlling the number of nonzero $\hat{\beta}_\lambda$ coefficients when forming a dense block of class predictor variables. This in turn makes it possible to constrain the size $\tilde{b}_i \ll b_i$ of preprocessed blocks, so that for some $0 \leq m < 1$, $\max_i \tilde{b}_i = \mathscr{O}(n^m)$, $i = 1, \ldots, \tilde{b}$, and even bound the number of selected blocks, $\tilde{b} < b$ by $\lim_{n \to \infty} \tilde{b}/n = \kappa \leq 1$. Jointly, these two asymptotic assumptions control global rate of growth of the number highly predictive variables $\tilde{p} = \mathscr{O}(n)$ in the sequence of classification problems. Henceforth, when analysing the performance property of the classifier based on $\tilde{b}$ selected blocks we assume that these constraints are fulfilled.

Observe that the Lasso variable selection in combination with the asymptotic independence of the resulting blocks allows for factorization of the class conditional density $f(\mathbf{x},\theta) = \Pi_{i=1}^{\tilde{b}} f(\mathbf{x}_i,\theta_i)$, so that using the set of n observations from each class Π_j the corresponding binary classifier can be estimated as $\mathscr{L}(\mathbf{x};\theta) = \sum_{i=1}^{\tilde{b}} \mathscr{L}_i(\mathbf{x}_i;\hat{\theta}_i^0,\hat{\theta}_i^1)$ where $\hat{\theta} = (\hat{\theta}_i^0,\hat{\theta}_i^1)$ satisfies the standard set of *good* asymptotic properties. To investigate the performance properties of this classifier we modify the initial sequence of classification problems discussed in Introduction rewriting it as $\{\tilde{p}, b_i, n_j, \kappa, \mathscr{L}(\mathbf{x};\theta), \mathscr{D}, \mathscr{E}_j\}_k$, where $j = 0,1$, $k = 1,2,\ldots$ and $\mathscr{D} = \sum_{i=1}^{\tilde{b}} \mathscr{D}_i = \sum_{i=1}^{\tilde{b}} \int \ln \frac{f(\mathbf{x}_i;\theta_i^1)}{f(\mathbf{x}_i;\theta_i^0)}[f(\mathbf{x}_i;\theta_i^1) - f(\mathbf{x}_i;\theta_i^0)]d\mathbf{x}_i$ is the Jeffrey's distance between two distributions. Observe that $\mathscr{D}$ is the analog of $\delta(\mu_0,\mu_1,\Sigma)$ for a non-Gaussian case and $0 < \mathscr{D} < \infty$. Given the asymptotic block independence, the classifier $\mathscr{L}(\mathbf{x};\theta)$ is a special case of the *Generalized Additive Model*, [4] we can state convergence of the sum towards a normal distribution and then get a closed form expression for the misclassification probabilities $\mathscr{E}$ in terms of the first and second moments of $\mathscr{L}(\mathbf{x},\hat{\theta})$. The details of this estimation technique are presented [6]. Furthermore, in the high-dimensional setting, i.e. when $\tilde{p} \sim n$, each sample-based term $\mathscr{L}(\mathbf{x}_i;\hat{\theta})$ is essentially overestimated with the local bias is of order $\mathscr{O}(\tilde{b}_i/n)$. Accumulating this bias over $\tilde{b}$ blocks naturally leads to a loss in the classification accuracy. Given that $\tilde{b}/n \rightarrow \kappa$ and $\kappa < 1$ this bias can be captured and essentially reduced by *down-weighting* the corresponding block input. We consider a weighted classifier $\mathscr{L}_\omega(\mathbf{x};\hat{\theta}) = \sum_{i=1}^{\tilde{b}} \omega_i \mathscr{L}_i(\mathbf{x};\theta)$ where weights $\omega_i := \omega_i\left(\frac{n\hat{\mathscr{D}}_i}{2}\right)$ represent the estimate of the i block distance input towards the Jeffrey's distance between the classes. Given the asymptotic normality of the weighted classifier $\mathscr{L}_\omega(\mathbf{x};\hat{\theta})$ and assuming that that $\mathscr{E}_1 = \mathscr{E}_2$ the optimal in a sense of minimum misclassification probability choice of ω is given by

$$\arg\min_{\omega(u)} \mathscr{E}_\omega = \omega_0(u) = \frac{\int \gamma_i^2 \chi(u;\tilde{b}_i+2,\gamma^2)dH(\gamma_i^2)}{u\int \gamma^2 \chi(u;\tilde{b}_i,\gamma_i^2)dH(\gamma_i^2)},$$

where $\chi(u;\tilde{b}_i,\gamma_i^2)$ is the density of the non-central χ^2 distribution with $\tilde{b}_i$s degrees of freedom and non-centrality parameter γ^2 representing the true ith block impact towards the Jeffrey's distance, and $H(u) = \frac{1}{\tilde{b}}\sum_{i=1}^{\tilde{b}} 1_{\{\frac{n\mathscr{D}_i}{2},\infty\}}(u)$; see detailed proof of this result in [6]. Since $\omega_0(u)$ is a decreasing function of u the suggested procedure provides desirable down-weighting.

References

1. Bai, Y., Wart, R.C.: Parallel block tridiagonalization of real symmetric matrices. J. Parallel Distrib. Comput. 68(5), 703–715 (2008)
2. Baumgartner, R., Somorjai, R.L.: Data complexity assessment in undersampled classification of high-dimensional biomedical data. Pattern Recogn. Lett. 27(12), 1383–1389 (2006)

3. Friedman, J., Hastie, T., Tibshirani, R.: Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software 33, 1–22 (2010)
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)
5. Marshall, A.W., Olkin, I.: Matrix versions of the Cauchy and Kantorovich inequalities. Aequationes Math. 40(1), 89–93 (1990)
6. Pavlenko, T., von Rosen, D.: On the optimal weighting of high-dimensional Bayesian networks. Adv. Appl. Stat. 4(3), 357–377 (2004)

Independence Tests for Uncertain Data with a Frequentist Method

Simon Petitrenaud

Abstract. In this paper, the analysis of contingency tables and the problem of independence of two variables are tackled, when information concerning one or both variables is missing or uncertain. In a first approach, the relations between the variables are described with the belief function theory. A second approach, a "frequentist" one, takes into account all the possible probabilistic distributions of contingency tables. With the use of classical statistical tests as the *chi*-square test, several dependence criteria adapted to the data are proposed and analyzed. A simulation and a real data example are presented to illustrate the methods.

1 Introduction

The analysis of contingency tables and independence tests of two variables have always received considerable attention in statistics [1, 5, 10]. In some cases, for some observations, information concerning one or the other variable are *missing* or *uncertain*. Statistical analysis with incomplete data have been largely developed in literature [4, 5, 9, 10]. In most cases, tests of independence are based on an EM algorithm with missing data but generally the studies are restricted to binary data.

To the best of our knowledge, no works have been made on the test of independence for data with *uncertain* observations. In [6], we proposed to use the belief functions [7, 8] since it allows an easier data description. But this method was restrictive, because we did not manage to keep all information in the decision process. In this article, we come back to the probabilistic formalism, and we propose a frequentist method that takes into account the

Simon Petitrenaud
Laboratoire d'Informatique, Université du Maine, 72085 Le Mans Cedex, France
e-mail: simon.petit-renaud@lium.univ-lemans.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 519–526.
springerlink.com

set of all probabilistic distributions of contingency tables in order to capture the more information as possible.

Firstly (in Section 2), we will briefly recall the classical independent tests used. Section 3 defines generalized contingency tables and presents existing methods in the particular case of missing data. In Section 4 we recall the belief function method. In Section 5, we describe the frequentist approach and we define several dependence criteria adapted to uncertain data and a confidence measure concerning this dependence. Several experiences showing the relevance of these criteria and measures are introduced in Section 6 and Section 7 concludes the article.

2 Classical Independence Tests

Let X and Y be two variables respectively described by a set of I categories $\mathscr{X} = \{x_1,\dots,x_I\}$ and J categories $\mathscr{Y} = \{y_1,\dots,y_J\}$. Let us denote by p_{ij} the joint probability $p_{ij} = P_{XY}(x_i,y_j)$ and $p_{i.} = P_X(x_i)$, $p_{.j} = P_Y(y_j)$ the corresponding marginal probabilities. We suppose we have a sample of n couples of variables $(X_k,Y_k)_{k=1}^n$ coming from (X,Y). The variable N_{ij} represents the simultaneous occurrences of the categories (x_i,y_j). Let us denote by $f_{ij} = \frac{n_{ij}}{n}$, the joint frequency of the couple (x_i,y_j), that estimates p_{ij} and by $f_{i.} = \sum_{j=1}^J f_{ij}$, $f_{.j} = \sum_{i=1}^I f_{ij}$ the marginal frequencies that respectively estimate the $p_{i.}$ and the $p_{.j}$. The independence condition in probability between X and Y, denoted H_0, is defined by:

$$H_0 : p_{ij} = p_{i.}p_{.j} \; (i = 1,\dots,I; j = 1,\dots,J). \tag{1}$$

Let us consider the independence test:“H_0: X and Y are independent, versus H_1: X and Y are not independent”. We can measure the independence degree between the two variables with the famous Pearson's χ^2 test. The χ^2 distance, computed with the distribution of the joint frequencies $f = \{f_{ij}\}_{i=1,\dots I; j=1,\dots,J}$ is defined by:

$$D(f) = n\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}}. \tag{2}$$

As $n \to +\infty$, $D(f)$ converges in law to a $\chi^2((I-1)(J-1))$ distribution if H_0 is true. Let $\alpha \in]0,1[$ be the risk of rejecting hypothesis H_0 by mistake. Let $K_{1-\alpha}$ be the quantile of order $1-\alpha$ of the χ^2 distribution with $(I-1)(J-1)$ degrees of freedom. If $D(f) \leq K_{1-\alpha}$, we could accept H_0 at the significance level α, otherwise, we should reject H_0 at this level. Then, in some cases, for some observations, information concerning one or the other variable are *missing* or *uncertain* and it is necessary to adapt independence tests to these situations.

3 Independence Tests for Generalized Contingency Tables

3.1 Generalized Contingency Tables

In this section, we propose to generalize the notion of contingency table to the case when the variables X and Y are not always perfectly observed. Let us assume that again we have a sample of size n of (X,Y), but now, the observations are not necessarily precise but take the form of a subset $C \in Rect(X,Y) = \{(A,B) | A \subset \mathscr{X}, B \subset \mathscr{Y}\}$, whose number of occurrences is denoted by n_C. In the rest of the paper, for simplicity sake, $n_{\{x_i\}\times\{y_j\}}$, $n_{\mathscr{X}\times\{y_j\}}$ and $n_{\{x_i\}\times\mathscr{Y}}$ are respectively denoted by n_{ij}, $n_{\mathscr{X}j}$ and $n_{i\mathscr{Y}}$. Table 1 gives an example of that sort of generalized contingency table. For obvious reasons linked to the simultaneous knowledge of the variables, the set of "useful" subsets C is restricted to $Rect(X,Y)$. The table size is then: $(2^I - 1, 2^J - 1)$. If we assume that available information is given by this kind of table, what can we say about the independence of X and Y?

Table 1 Example of a generalized contingency table ($I = 2$, $J = 3$)

	y_1	y_2	y_3	$\{y_1,y_2\}$	$\{y_1,y_3\}$	$\{y_2,y_3\}$	$\mathscr{Y} = \{y_1,y_2,y_3\}$
x_1	33	5	12	2	8	2	1
x_2	11	9	4	3	2	5	0
$\mathscr{X} = \{x_1,x_2\}$	1	1	1	0	0	0	1

3.2 EM Approach for Missing Data

In this section, we briefly recall a popular solution in the particular case of *missing* data, where the sets A and B are *singletons* or *the whole set* $\mathscr{X}$ and $\mathscr{Y}$ respectively. In this case, it is possible to iteratively determine the estimates $\hat{p}_{ij}$ of p_{ij} by a formula combining the two steps of the Expectation-Maximization (EM) algorithm. At each iteration t, the frequencies $p_{ij}^{(t+1)}$ are computed as follows [4]:

$$p_{ij}^{(t+1)} = f_{ij} + \frac{n_{i\mathscr{Y}}}{n} \frac{p_{ij}^{(t)}}{\sum_{j=1}^{J} p_{ij}^{(t)}} + \frac{n_{\mathscr{X}j}}{n} \frac{p_{ij}^{(t)}}{\sum_{i=1}^{I} p_{ij}^{(t)}} \; (i = 1,\ldots,I; j = 1,\ldots,J). \tag{3}$$

The frequencies $p^{(t)} = \left(p_{ij}^{(t)}\right)$ converge to the estimates $\hat{p} = (\hat{p}_{ij})$. Then we can easily define independence tests using the statistic $D(\hat{p})$ (cf. Equation 2) or likelihood ratio tests. Many variants of the EM algorithm have been proposed in literature [5, 9].

4 Belief Function Approach

4.1 Belief Functions

The following section recalls and discuss the proposed solution developed in [6], when information is described as a belief function. First, we very briefly recall some notions of the belief function theory [7, 8]. Let Ω be a finite set, the uncertainty representation is made by the means of the concept of belief function, defined as a function m from 2^{Ω} to $[0,1]$ such as: $\sum_{A\subseteq\Omega} m(A) = 1$. The quantity $m(A)$ represents the belief exactly allowed to proposition A. The plausibility function Pl, which quantifies the maximal belief which might be allowed to a proposition, is defined by: $Pl(A) = \sum_{B\cap A\neq\emptyset} m(B)$, $\forall A \subset \Omega$. The function m describes the state of a belief. Once a structure m is defined, it is possible to transform it to a probability distribution, particularly for decisional aspects. One of these distributions, called *pignistic* probability and denoted by P_*, consists in sharing equitably the mass of a subset of Ω between its elements. So it is defined for all $\omega \in \Omega$ by [8]:

$$P_*(\{\omega\}) = \sum_{A\subset\Omega} \frac{m(A)}{|A|(1-m(\emptyset))}\delta_A(\omega), \tag{4}$$

where $|A|$ is the cardinal of A, $\delta_A(\omega) = 1$ if $\omega \in A$ and $\delta_A(\omega) = 0$ if $\omega \notin A$.

4.2 Uncertain Observations and Pignistic Criterion

A generalized contingency table with uncertain observations defined in Section 3.1 can also be associated to a normalized belief function m in $\Omega = \mathscr{X} \times \mathscr{Y}$ which generalizes the joint frequency of (X,Y) to uncertain observations: $m(C) = \dfrac{n_C}{n}, \quad C \in Rect(X,Y)$. In the example of Table 1, $m(\{(x_1,y_1),(x_1,y_3)\}) = 0.08$. Instead of trying to capture all the uncertainty of the data, we proposed a simple dependence criterion using belief functions, based on a representative distribution, the pignistic distribution:

$$P_*(\{(x_i,y_j)\}) = \sum_{\{(A,B)|x_i\in A, y_j\in B\}} \frac{m(A\times B)}{|A\times B|}\ (i=1,\ldots,I; j=1,\ldots,J). \tag{5}$$

Then we can obtain an independence indicator when computing the χ^2 statistic $D(P_*)$ for this particular distribution (cf. Equation 2). This method generalizes the classical cases when the observations are accurate and complete but has the disadvantage of not checking the principle of neutrality towards independence of the belief function measuring the total ignorance ($m(\Omega) = 1$). This function expresses the total lack of information concerning the relationship between the variables X and Y whereas the associated pignistic distribution leads to the exact independence. Several definition of independence for belief functions are defined in [2, 3]. Consequently, other criteria, in particular based on the plausibility function of m and its projections on $\mathscr{X}$ and $\mathscr{Y}$, are discussed in [6], but they have not been found enough relevant.

5 Frequentist Approach

5.1 Analysis of All Probabilistic Distributions

Now, if we want to keep and use the information of the complete contingency table $\mathscr{T}$ of $(2^I-1, 2^J-1)$ size as best as possible, we can notice that it corresponds to $\mathscr{N}$ classical tables $(T_q)_{q=1}^{\mathscr{N}}$ of size (I,J) which have not been specified. For example, in Table 1, there are 4 ways of distributing the 3 elements of the cell $\{x_1,x_2\}\times\{y_2\}$ in the two classical cells:$\{x_1,y_2\}$ and $\{x_2,y_2\}$. The total number of possible tables (counted with the probability that they come true) is:

$$\mathscr{N} = \prod_{A\in\Omega} |A|^{n_A}, \tag{6}$$

where n_A is the number of occurrences of A. In the example of Table 1, $\mathscr{N} = 2^{25} * 6 \simeq 2 * 10^8$. We suppose that the distribution of the n_A occurrences of a cell A among the corresponding classical cells has a multinomial distribution with parameters n_A and the *uniform* proportions $\frac{1}{|A|}$. Then, the distribution of the T_q is the result of a mixing of multinomial distributions. The uncertainty concerning the independence of the two variables X and Y appears at two levels: the dependence measure of these variables, for example defined by the χ^2 statistic of a given random distribution and the uncertainty concerning this measure itself, which is due to the multiplicity of possible distributions. The idea is to take into account all the $\mathscr{N}$ possible distributions of the n individuals in the IJ cells of the "real" table of singletons (x_i,y_j). Each table $(T_q)_{q=1}^{\mathscr{N}}$ corresponds to a distribution defined by the joint frequency f_q, and leads to a χ^2 distance $D_q = D(f_q)$. The complete set of all possible distributions $\mathscr{D} = (D_1,\dots,D_{\mathscr{N}})$ is then built and studied. It is possible to find a representative value such as the mean $\overline{D}$ or the median D_{Me} of the family $\mathscr{D}$. More generally, if D_- and D_+ denote respectively the minimum and the maximum of the family $\mathscr{D}$, three cases occur:

- if $D_- > K_{1-\alpha}$, the independence hypothesis is rejected at level $1-\alpha$;
- if $D_+ < K_{1-\alpha}$, the independence hypothesis is accepted at level $1-\alpha$;
- if $K_{1-\alpha} \in [D_-, D_+]$, we cannot conclude *a priori*, because some distributions accept H_0 but other ones reject it.

In general cases, when $I > 2$ or $J > 2$, D_- and D_+ are not analytically defined but are iteratively computed, since the use of an optimization program with linear constraints generally lead to a local minimum or maximum. But when $I = J = 2$, the analytic computation of D_+ is straightforward:

$$D_+ = \max\left\{D(f), D(g)\right\}, \tag{7}$$

where $f_{11} = \frac{n_{11}+n_{\{x_1\}\times\mathscr{Y}}+n_{\mathscr{X}1}}{n}$, $f_{12} = \frac{n_{12}}{n}$, $f_{21} = \frac{n_{21}}{n}$, and $f_{22} = \frac{n_{22}+n_{2\mathscr{Y}}+n_{\mathscr{X}2}}{n}$ and $g_{12} = \frac{n_{12}+n_{1\mathscr{Y}}+n_{\mathscr{X}2}}{n}$, $g_{11} = \frac{n_{11}}{n}$, $g_{22} = \frac{n_{22}}{n}$, and $g_{21} = \frac{n_{21}+n_{2\mathscr{Y}}+n_{\mathscr{X}1}}{n}$.

5.2 Dependence Degree of Variables

Since the interval $[D_-, D_+]$ may be large, the rule derived in Section 5.1 leads, either to binary decisions, or in most cases to a lack of decision as soon as the proportion of uncertain observations becomes important. We propose to refine the decision process by introducing the notion of dependence degree. We define a dependence degree ($\in [0,1]$) between the variables X and Y, as the proportion of possible distributions f_q for whom H_0 would be rejected:

$$C(\mathscr{D}) = \sum_{q=1}^{\mathscr{N}} 1_{\{D_q > K_{1-\alpha}\}}(f_q). \tag{8}$$

Thanks to the empirical distribution function $\widehat{F_D}$ of $\mathscr{D}$, it is possible to build robust fluctuation intervals of D: $[d_{\frac{\beta}{2}}, d_{1-\frac{\beta}{2}}]$ containing a proportion of $1-\beta$ of the values, with d_γ, the quantile of order γ of $\widehat{F_D}$. As we have seen earlier, $\mathscr{N}$ may be very large. In order to make computations possible in a reasonable time, we can estimate this distribution by randomly building a representative sample in a Monte-Carlo way. Finally, in order to measure the uncertainty degree concerning the independence, we define a confidence index which is expressed in the form of an interval $\in [0,1]$:

$$CI_{1-\beta}(\mathscr{D}) = [F_{\chi^2}(d_{\frac{\beta}{2}}), F_{\chi^2}(d_{1-\frac{\beta}{2}})], \tag{9}$$

in which F_{χ^2} is the χ^2 cumulative distribution function with $(I-1)(J-1)$ degrees of freedom. The idea is the following: the more uncertain the distribution, the largest the interval. If we take the two extreme cases, the total ignorance leads to a very large confidence index, and for a single distribution f, $CI_{1-\beta}(f)$ is reduced to a singleton. In the case of total ignorance, $\mathscr{D}$ has a multinomial distribution with uniform proportions $\frac{1}{IJ}$, thus approximately a χ^2 distribution with $(I-1)(J-1)$ degrees of freedom. Therefore, $CI_{1-\beta}(\mathscr{D}) = [\frac{\beta}{2}; 1-\frac{\beta}{2}]$, which makes the decision obviously unreliable.

6 Experiments

In this section, we illustrate the proposed measures within two situations. We take an error risk $\alpha = 0.05$ and the parameter $\beta = 0.05$. In each case, the distribution of $\mathscr{D}$ is simulated with a sample of size 10000.

Example 1 (Analysis of crime data.). The crime data, given in Table 2, have been taken as an example of real incomplete categorical data sets by several authors [9]. The sample size n of the data is 641 and the data originally come from the National Crime Survey by the U.S. Bureau of the Census. In this example, the number of missing data is quite important (80), so the size $\mathscr{N}$ of possible distributions $\mathscr{D}$ is extremely large ($\approx 2*10^{24}$). All situations may *a priori* occur, since $D_- = 1.9$ and $D_+ \approx 69.5 \gg K_{0.95}(1) = 3.84$. But these extreme values are very rare, and we can see in Figure 1 that almost all values are greater that 3.84. The reference pignistic value $D(P_*)$ is equal to 21.77. Center values of $\mathscr{D}$ are very close: $\overline{D} = 21.86$ and $D_{me} = 21.77$. So, the independence null hypothesis H_0 is unquestionably rejected. Similar results can be found in [9, 4]. The EM algorithm in Equation (3) leads to $D(\hat{p}) = 33.40$.

Table 2 Example 1: Victimization status of housing units occupants.

	y_1	y_2	$\mathscr{Y} = \{y_1, y_2\}$
x_1	392	55	33
x_2	76	38	9
$\mathscr{X} = \{x_1, x_2\}$	31	7	0

Example 2. Let us take the Example 1 (cf. Figure 1). If we make the decision with the pignistic measure, H_0 is rejected: $D(P_*) = 9.44 > K_{0.95}(2) \approx 5.99$. Here again, center values of $\mathscr{D}$ are close: $\overline{D} = 9.66$ and $D_{me} = 9.45$. If we take all the distributions into account, the decision is not as clear as in the first case, since $1 - \alpha \in CI_{1-\beta}(\mathscr{D}) = [0.902; 1]$, but we can say that X and Y have 91.5% of chances to be dependent ($C(\mathscr{D}) = 0.915$).

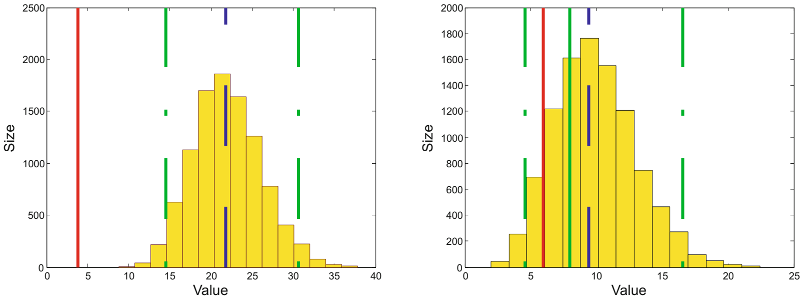


Fig. 1 **Left**: Example 1 (crime data) and **Right**: Example 2. Distribution of the distances D, $D(P_*)$ $(--)$, $[d_{0.025}, d_{0.975}]$, (-.-), decision threshold (-)

7 Conclusion

In this article, we analyzed the independence of two variables X and Y in the case when observations may be uncertain or incomplete. We proposed criteria generalizing independence tests for two variables as extension of the classical χ^2 test. In a first approach, the joint frequency of the variables is expressed in the form of a belief function m and the decision is made with the pignistic transformation of m. A second approach, a "frequentist" one, taking all the possible probabilistic distributions of contingency tables into account, has been presented.

Experiments have been made in little sized tables (2×2 or 3×2) but the benefit of our methods should increase with their dimension, as a next study should show. In a future work, we will also consider the extension of the independence test when the data are not only uncertain, but also fuzzy. Finally, it should be interesting to study the impact of this sort of uncertain data to the similar problem of estimating and testing the multinomial parameters $p_{ij} = P_{XY}(x_i, y_j)$ of the joint distribution.

References

1. Agresti, A.: Categorical Data Analysis. Wiley, New York (1990)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief Function Independance: I. The Marginal Case. Internat. J. Approx. Reason. 29, 47–70 (2002)
3. Couso, I.: Independence Concepts in Evidence Theory. In: Proceedings of the 5th International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2007, Prague, Czhec Republic, pp. 125–133 (2007)
4. Chen, T., Fienberg, S.E.: Two-dimensional contingency tables with both completely and partially cross-classified data. Biometrics 30, 629–642 (1974)
5. Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley, New York (2002)
6. Petitrenaud, S.: Tests d'indépendance pour observations incomplétes ou incertaines. In: Rencontres francophones sur la Logique Floue et ses Applications, Lens (2008)
7. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
8. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligence 66, 191–234 (1994)
9. Takai, K., Kano, Y.: Test of Independence in a 2 × 2 Contingency Table with Nonignorable Nonresponse via Constrained EM Algorithm. Comput. Statist. Data Anal. 52, 5229–5241 (2008)
10. Taneichi, N., Sekiya, Y.: Improved Transformed Statistics for the Test of Independance in $r \times s$ Contingency Tables. J. Multivariate Anal. 98, 1630–1657 (2007)

Why Imprecise Regression: A Discussion

Henri Prade and Mathieu Serrurier

Abstract. Machine learning, and more specifically regression, usually focuses on the search for a precise model, when precise data are available. It is well-known that the model thus found may not exactly describe the target concept, due to the existence of learning bias. In order to overcome the problem of learning models having an illusory precision, a so-called imprecise regression method has been recently proposed for non-fuzzy data. The goal of imprecise regression is to find a model that offers a good trade-off between faithfulness w.r.t. data and (meaningful) precision. In this paper, we propose an improved version of the initial approach. The interest of such an approach with respect to classical regression is discussed in the perspective of coping with learning bias. This approach is also contrasted with other fuzzy regression approaches.

1 Introduction

Fuzzy regression methods have been proposed for now more than twenty years. Apparently, the motivations for such extended forms of regression have been either to generalize regression to fuzzy data, or to describe envelopes for the data by associating each input with an interval covering output data. This second type of regression (often termed possibilistic regression) yields interval representations even when input and output data are non-fuzzy. This suggests that possibilistic regression does not serve exactly the same purpose as classical regression. Still the purpose of possibilistic regression has never been fully laid bare (beyond the informal idea of coverage of the data). Classical least square regression has both a geometrical and a statistical justification that coincide. Indeed the regression line is supposed to pass through the "middle" of the "cloud" of data points. Moreover, in the statistical view, the

Henri Prade and Mathieu Serrurier
IRIT, Universite Paul Sabatier, Toulouse, France
e-mail: prade@irit.fr,serrurie@irit.fr

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 527–535.
springerlink.com

regression curve is interpreted as the mean of the normal probability distribution for the output given an input vector. This interpretation requires that the data variations obey a Gaussian law with fixed standard deviation. Imprecise regression, whose a preliminary form has been proposed in [19], may be considered as being midway between possibilistic regression (due to its coverage concern) and least square regression (due to an uncertainty interpretation). Indeed, learning biases involve different kinds of uncertainty due to noisy data on the one hand and to the description language on the other hand. Noise may be handled by a probabilistic representation (if the distribution type is known), while description language is pervaded with epistemic uncertainty. In imprecise regression, we propose to describe the data by means of possibility distributions, which can be seen as providing a representation of a family of probabilities [6]. Possibility distributions are indeed well known as a tool for representing epistemic uncertainty. The paper is structured as follows. First we present an improved version of the imprecise regression approach with an illustrative example. Next, we discuss the differences between least square regression and imprecise regression. The last sections are devoted to comparisons with the related literature, and to concluding remarks.

2 Imprecise Regression

The new setting. The goal of imprecise regression [19] is to overcome the learning biases by considering them as factors that have impact on the precision of the models rather than as a boundary to the effectiveness of the learning process. Knowing that the representation of the examples and the hypothesis correspond necessarily to an incomplete view of the world, we will search for imprecise hypotheses that take into account this incompleteness. Thus, given a set of crisp data, we will search for a model that is as precise as possible and which provides a faithful description of the data. When the imprecision tends to 0, we obtain a crisp hypothesis that describes the concept exactly. In a formal way, imprecise regression allows us to represent the imprecision associated with the model by taking into account the incompleteness of the information provided by the data and the chosen representation space for the hypotheses.

A regression database is a set of m pairs $(\overrightarrow{x}_i, y_i)$, $1 \leq i \leq m$, where $\overrightarrow{x}_i \in \mathbb{R}^n$ is a vector of n input variables and $y_i \in \mathbb{R}$ is the real output variable. An imprecise fuzzy function F is a function from $\mathbb{R}^n$ to $(\mathbb{R} \rightarrow [0,1])$ that associates a distribution on the possible values of the output to the input vector $\overrightarrow{x}$. The goal of imprecise regression is to find the fuzzy function $F(\overrightarrow{x})$ that maximizes the evaluation function:

$$Info(F) = -\sum_{i=1}^{m} \pi_i(y_i) * log(\frac{Area(\pi_i) + Area(A_{(\pi_i(y_i))})}{2 * Area(\pi_{max})}) \tag{1}$$

where $\pi_i = F(\overrightarrow{x}_i)$ and $A_{(\pi_i(y_i))}$ is the $\pi_i(y_i)$-level cut of the fuzzy set A having π_i as membership function. π_{max} is the maximal binary possibility distribution, fixed a priori by the user, that covers all the data. Note that the evaluation function has been improved with respect to the one initially used in [19], which had more parameters to be tuned in order to obtain good results. This is an information measure which may be viewed as a counterpart of the Kullback-Liebler one in the possibilistic setting. It measures how much the distribution π_i increases information about the output w. r. t. π_{max}. By maximizing together the accuracy of the imprecise function, here estimated through the terms $\pi_i(y_i)$'s, and its precision, estimated by $Area(\pi_i) + Area(A_{(\pi_i(y))})$, we ensure a trade-off between accuracy and precision of the model. Note that we use $Area(\pi_i) + Area(A_{(\pi_i(y))})$ rather than $2 * Area(\pi_i)$ in order to avoid that the maximum be systematically reached for a rectangular possibility distribution. Indeed the second term $Area(A_{(\pi_i(y))})$ creates a synergy between gradualness and precision (area). Using this second term only would not properly handle the data that are out of the core of the distribution. Maximum is reached when the function describes exactly the data without imprecision. Since the learning bias may prevent reaching this maximum, the function will describe both the general tendency of the data and the variations around it.

Algorithm. In the following, we consider imprecise regression functions of the form $F_{a,b,c,d}(\overrightarrow{x}) = T_{f_a(\overrightarrow{x}),f_b(\overrightarrow{x}),f_c(\overrightarrow{x}),f_d(\overrightarrow{x})}$ which associate a trapezoidal fuzzy set to a vector of input variables, although the framework would be applicable to any kind of membership functions. Functions f_a, f_b, f_cet f_d can be linear functions of the form $f(\overrightarrow{x} =< x_1, \ldots, x_n >) = a_0 + a_1 * x_1 + \ldots + a_n * x_n$, or kernel functions (e.g. Gaussian functions in our application) $f(\overrightarrow{x}) = a_0 + a_1 * K(s_1, x) + \ldots + a_n * K(s_k, x)$, where s_1, s_k are support vectors which are computed previously by using a clustering algorithm. Finding optimal f_a, f_b, f_c, f_d constitutes a hard problem which is not solvable by classical optimization methods. We propose to solve the problem by using a simulated annealing algorithm [16]. The goal of simulated annealing is to determine the function F which maximizes the measure defined by the Equation 1. In order to use simulated annealing, we first need to define the neighborhood V of a function F :

$$V(F_{a,b,c,d}) =< V(f_a), V(f_b), V(f_c), V(f_d) > \quad (2)$$

The neighborhood of a linear or kernel function is obtained by randomly adding or removing a fixed small value to all the coefficients. The use of a fixed small value for the variations is due to the fact that simulated annealing is designed for discrete exploration of the space.

Illustration. We illustrate the approach on the learning of Fitt's law [10]. Fitt's law is used for predicting the movement time for an interaction given its difficulty. The interaction considered here is the pointing of a circular target with mouse device from a fixed starting point. The input variable is the interaction difficulty which depends on the movement amplitude and the target size. The output variable is the time for interaction with success. The database used here is obtained by means of an experiment where 20 different users that performed 100 interaction tasks, each dispatched in a dozen of different difficulty degrees. Results are presented in Figure 1.

Such a data are usually processed in the following way. An average point is computed for each pair (difficulty degree, user). Then a linear regression is performed and the line obtained is usually found to be acceptable in statistical terms. If we consider all the data as in Figure 1, the direct use of classical regression is not appropriate for at least two reasons. First, the residual values are too high for being statistically accepted. Second, the data are not distributed around the general tendency in a normal way (the distribution are not even symmetrical for physical reasons, and the distribution is generally multi modal due to the existence of different kinds of users), which makes problem at least for least square regression. Using imprecise regression enables us to overcome these limitations (in Figure 1, the possibility distribution is described in the third dimension). Observe that what is obtained is not strictly linear, not symmetrical and has a varying spread. We can also notice that imprecise regression leaves naturally away the outliers which are numerous in the considered dataset.

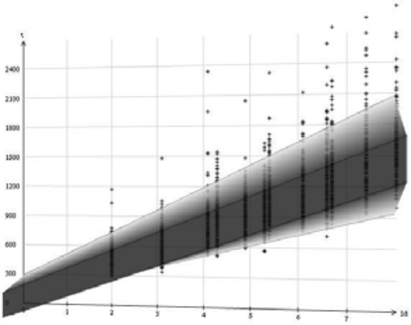


Fig. 1 Imprecise non-linear regression for Fitt's law using Gaussian kernel.

3 Discussion

Least square regression can be considered from at least two different viewpoints. In the first view, the problem considered amounts to determine a line, and more generally a curve, that fits a set of data points in the sense that the curve is as close as possible to each point. This leads to minimize the sum of some evaluations based on the distances of the points to the curve. Gauss and Legendre have shown that minimizing the squares of the distance leads to a solvable linear system of equations in the case of linear regression. This point of view turns to coincide with a particular statistical view of the problem. Namely, the points are viewed as a result of a measurement whose error behaves as a Gaussian noise. In this second view, regression amounts to consider the conditional probability of the output variable given the input variable(s). Assuming that the error follows a normal distribution, the regression curve is nothing but the locus of the means of the distributions associated with input vectors. Finding the optimal curve that maximizes the likelihood in the Bayesian sense corresponds to find the curve that minimizes the sum of the squares of the distances. Robust regression methods no longer require Gaussian distributions and allow for variance depending on the input variables. However these methods still assume that the type of the distribution is known.

When learning from crisp data, different important kinds of biases have to be considered. The first one, called description language bias, comes from the data description itself. More precisely, the language used constitutes a bound for the descriptive power of the data and leads to an incomplete view of the world. The second one, called sample bias, refers to the fact that the available data are limited and may be irregularly distributed, and thus provides only an incomplete view of the world. The last one, called noise bias, reflects that the data may be pervaded with noise (measurement noise, variability, ...). One advantage of the least square regression, in the case where the normal distribution has the same variance everywhere, is that we can easily describe the variations around the general tendency. Then, statistical criteria can be used in order to check that the description is acceptable. If not, the hypothesis learned has to be rejected. Non parametric approaches can then be used for describing the dispersion around the curve. However, it also requires the choice of a particular kind of probability distribution (whose parameters may vary) and a large amount of data. Since it is a local estimation problem with respect to the data in the neighborhood, this approach is not suitable for prediction or for handling sparse data. The Bayesian view of regression corresponds to a treatment of the noise bias. Moreover, it assumes that the distribution of the noise is known, which is not always the case.

Imprecise regression which rather associates a conditional possibility distribution to the input does not suppose a particular kind of probability distribution, but rather implicitly handles a family of conditional probabilities since a possibility distribution is an exact or approximate way for describing such a family [6]. Thus, this takes care of the description language bias. The description language bias means that there exist hidden variables that influence the conditional probabilities. This corresponds to epistemic uncertainty whose handling requires some kind of imprecise probability models. Note also that generalizing the geometrical view of linear regression by only requiring that the data points be close to *one* among two, three or four lines (associating interval, triangular or trapezoidal fuzzy numbers with input data), would not really cover *all* the data, since data points that are not outliers would remain outside the fuzzy numbers scope, which will not be satisfactory.

The sample bias which also refers to epistemic uncertainty is not really handled by any type of regression. Indeed, when there is a lack of input data in between two areas where data are available, the imprecise regression may yield a more narrow trapezoid in the area where data are lacking, while one may have expected larger trapezoids there for reflecting the larger amount of ignorance. This is due to the fact that the absence of data induces less constraints for the optimization process and may lead to some over-fitting. Another major bias is the complexity of the hypothesis space. Indeed, due to the limitation of the hypothesis language and the complexity of the algorithms, it is rarely possible to find the hypothesis that describes exactly the concept we want to learn. One of the major machine learning theorems [22] shows that these biases lead to a bound on the effectiveness of the learning method used. It has been shown that too complex hypotheses language (e.g., polynomials of too high degree in regression, or the use of too many input variables w.r.t. the number of examples in linear regression) may lead to learn models whose precision is illusory. This remains true with imprecise regression.

As for classical regression, imprecise regression is well suited for interpolation. However, when considering extrapolation, as in time series, extra information such as the general shape of the function has to be taken into account in the model.

4 Related Works

In this section, we emphasize the differences between imprecise regression and the different proposals in fuzzy regression. A first type of fuzzy regression approach assumes that we start with fuzzy data, which means that the output values are fuzzy and maybe also the input values. Then, a fuzzy representation is searched for describing such data [3, 4]. Diamond's method is based on the extension of least square error minimization using a metrics on fuzzy sets. Such a fuzzy extension of classical regression has been extensively studied [18] in the linear case [8, 9, 17, 11]. Non-linear approaches have been also proposed by using neural networks [7, 13], SVM's [12] and genetic algorithms [2]. The major advantage of the least square method is that it appears to be a natural mathematical extension of crisp regression. In this context, when data inputs and output are not fuzzy, fuzzy least square regression reduces to standard least square regression, thus leading to a non fuzzy result. This constitutes a major difference with our approach. In fact, imprecise regression aims at being faithful to the distribution of the data, and associates a fuzzy representation with crisp input and output data. In contrast, fuzzy least square regression propagates the imprecision/fuzziness of the data. Moreover, the use of set distances extended to fuzzy sets do really not agree with the view of a fuzzy number as a possibility distribution restricting the possible values of a real-valued variable (since the distance between two identical sets is always 0 even if these sets are not singletons).

A second type of approach, named possibilistic regression, has been initially proposed by Tanaka [20], and is reminiscent of quantile regression. The goal of this approach is to associate the data with a pair of upper and lower regression functions, while minimizing the total spread of the output coverage. In the original method, the lower bound and the upper bound of the regression function are computed separately. Clearly, this method can be used with crisp data and/or with fuzzy data. A linear model [20] has been initially proposed and then extended [21] to non-linear possibilistic regression in [14] by using neural networks. The main disadvantage of this method is that it is very sensitive to outliers (even if it may be somewhat controlled [15, 23] by using SVM's together with outliers tolerance). Indeed, the optimal upper (resp. lower) bound function is basically the function that is immediately above (resp. below) the whole set of output data. Thus, outliers may affect to a large extent the function that is learnt. Recently, Bisserier et al. [1] have proposed a slightly different approach also aiming at providing a fuzzy coverage of the dataset. These authors look for a fuzzy linear regression model, represented by means of fuzzy parameters. In fact, it reduces to an interval regression problem by considering the support of these fuzzy parameters (supposed to be symmetrical triangle fuzzy numbers).

At first glance, imprecise regression may seem to be very close to possibilistic regression. First, the two approaches deal with crisp data. Second, they use separate functions in order to represent fuzzy sets or intervals. However, the two approaches differ both at the theoretical level and at the algorithmic level. Possibilistic regression aims at finding the most precise function that is totally accurate with respect to all the examples up to some fixed outliers tolerance. On the contrary, the goal of imprecise regression is to find the function that has the better trade-off between data faithfulness and precision in order to take into account epistemic uncertainty associated with the learning problem. This is why imprecise regression is less sensitive to outliers than possibilistic regression. Moreover, the lower bound and the higher bound of the function in possibilistic regression are learned separately. Thus, when dealing with crisp data, possibilistic regression can only produce intervals rather that genuine fuzzy sets. Imprecise regression quality measure is global, and all the functions that describe the fuzzy sets are learnt together. It allows us to learn models that can represent the imprecision by any kind of fuzzy sets (here trapezoidal) in a coherent way. Moreover, imprecise regression is liable to receive an interpretation in terms of imprecise probabilities (family of probability measures).

5 Conclusion

In this paper we have contrasted different extensions of regression methods. The choice of a particular method depends on the problem at hand. If the shape of the law to be learnt is known and the data are just noisy due to the measurement device (think for instance of the Hooke's law which states that the extension of a spring is proportional to the force applied to it), then classical regression is enough for the task (e.g. identifying the value of the force constant in Hooke's law for a given spring). On the contrary, as illustrated in our Fitt's law example, if the law to be learnt is not completely determined by the input variables available, imprecise regression allows us to capture the epistemic uncertainty associated with the model. A further line of research is obviously the study of the interpretation of imprecise regression in terms of imprecise probabilities. Possibilistic regression seems to be more suitable when the possible values of the output variable are represented by intervals. The application of imprecise regression to fuzzy data is also worth considering. This would basically amount to replace the degree of membership $\pi_i(y_i)$ in Eq. 1 by the possibilistic lower or the upper expectation of the compatibility of the fuzzy output value with respect to π_i [5].

References

1. Bisserier, A., Boukezzoula, R., Galichet, S.: An interval approach for fuzzy linear regression with imprecise data. In: Proceedings of the IFSA/EUSFLAT 2009 Conference, IFSA/EUSFLAT 2009, Lisbon, Portugal, pp. 1305–1310 (2009)
2. Buckley, J., Feuring, T.: Linear and non-linear fuzzy regression: Evolutionary algorithm solutions. Fuzzy Sets Syst. 112, 381–394 (2000)
3. Celmins, A.: Least squares model fitting to fuzzy vector data. Fuzzy Sets Syst. 22(3), 245–269 (1987)
4. Diamond, P.: Fuzzy least squares. Inform. Sci. 46(3), 141–157 (1988)
5. Dubois, D., Prade, H.: Possibility theory, pp. 125–126. Plenum Press, New York (1988)
6. Dubois, D., Prade, H.: When upper probabilities are possibility measures. Fuzzy Sets Syst. 49, 65–74 (1992)
7. Dunyak, J.P., Wünsche, D.: Fuzzy regression by fuzzy number neural networks. Fuzzy Sets Syst. 112(3), 371–380 (2000), doi:10.1016/S0165-0114(97)00393-X
8. D'Urso, P.: An "orderwise" polynomial regression procedure for fuzzy data. Fuzzy Sets Syst. 130, 1–19 (2002)
9. D'Urso, P.: Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput. Stat. Data Anal. 42(1-2), 47–72 (2003), doi:10.1016/S0167-9473(02)00117-2
10. Fitts, P.: The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47, 381–391 (1954)
11. González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M.A.: Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets Syst. 160, 357–370 (2009)
12. Hong, D.H., Hwang, C.: Support vector fuzzy regression machines. Fuzzy Sets and Systems 138(2), 271–281 (2003), doi:10.1016/S0165-0114(02)00514-6
13. Ishibuchi, H., Nii, M.: Fuzzy regression using asymmetric fuzzy coefficients and fuzzified neural networks. Fuzzy Sets Syst. 119(2), 273–290 (2001), doi:10.1016/S0165-0114(98)00370-4
14. Ishibuchi, H., Tanaka, H.: Fuzzy regression analysis using neural networks. Fuzzy Sets Syst. 50, 57–65 (1992)
15. Jenga, J.T., Chuang, C.C., Su, S.F.: Support vector interval regression networks for interval regression analysis. Fuzzy Sets Syst. 138(2), 283–300 (2003)
16. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220, 671–680 (1983)
17. Modarres, M., Nasrabadi, E., Nasrabadi, M.M.: Fuzzy linear regression models with least square errors. Appl. Math. Comput. 15, 873–881 (2003)

18. Näther, W.: Regression with fuzzy random data. Comput. Stat. Data Anal. 51, 235–252 (2006)
19. Serrurier, M., Prade, H.: A general framework for imprecise regression. In: Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2007, London, UK, pp. 1597–1602 (2007)
20. Tanaka, H.: Fuzzy data analysis by possibilistic linear models. Fuzzy Sets Syst. 24(3), 363–376 (1987)
21. Tanaka, H., Guo, P.: Possibilistic data analysis for operations research. Physica-Verlag, Heidelberg (1999)
22. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
23. Xu, S.Q., Luo, Q.Y., Xu, G.H., Zhang, L.: Asymmetrical interval regression using extended epsilon-SVM with robust algorithm. Fuzzy Sets Syst. 160(7), 988–1002 (2009)

Power Analysis of the Homoscedasticity Test for Random Fuzzy Sets

Ana Belén Ramos-Guajardo, Gil González-Rodríguez,
Manuel Montenegro, and María Teresa López

Abstract. Some tools for testing hypotheses about the variance of random fuzzy sets are already available. Asymptotically correct procedures for the k-sample homoscedasticity tests have been recently developed. However, the power of such procedures has not been analyzed yet. In this paper, some studies about the power function of the asymptotic procedure for the homoscedasticity test are presented. The theoretical analysis is carried out by considering the capability of the test under local alternatives. Finally, the behavior of the power function is illustrated by means of simulation studies.

Keywords: Homoscedasticity test, Random fuzzy sets, Power function, Local alternatives.

1 Introduction

Random fuzzy sets (RFS's for short) in Puri & Ralescu's sense [8] were introduced to model random mechanisms generating imprecisely-valued data which can be described by means of fuzzy sets. The stochastic variability of the fuzzy values of an RFS can be measured by means of the *Fréchet variance* defined in terms of a generalized metric. This measure quantifies the squared error of approximating RFSŠ values by means of its fuzzy mean.

The problem of testing about the variance of an RFS has been previously analyzed. The one-sample test has been developed in [7] in a particular class of fuzzy sets by using the large sample theory. These studies have been extended

Ana Belén Ramos-Guajardo and Gil González-Rodríguez
European Centre for Soft Computing, 33600 Mieres, Spain
e-mail: `anabelen.ramos@softcomputing.es,gil.gonzalez@softcomputing.es`

Manuel Montenegro and María Teresa López
Dpto. de Estadística, I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain
e-mail: `mmontenegro@uniovi.es,mtlopez@uniovi.es`

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 537–544.
springerlink.com

to a more general setting in [9] by considering other asymptotic and bootstrap techniques.

Additionally, a test for analyzing the equality of variances of k RFS's has been developed in [10] inspired by the classical Levene's test [6]. The theoretical results developed in this context have been mainly focussed on the type I error (more precisely on the test size). However, the analysis of the power function has not been addressed yet. In this work, the capability of the test is analyzed by using a sequence of alternatives converging to the null hypothesis as the sample size increases (also called *local alternatives*).

In Section 2, some preliminaries about RFS's are gathered. The homoscedasticity test for RFS's is introduced in Section 3. The analysis of the power of the test under local alternatives is developed in Section 4. In Section 5, some simulations of the power function are carried out. Finally, some concluding remarks and open problems are presented in Section 6.

2 Preliminaries

Let $\mathscr{K}_c(\mathbb{R}^p)$ be the family of all non-empty compact convex subsets of $\mathbb{R}^p$ and let $\mathscr{F}_c(\mathbb{R}^p)$ denote the class of the fuzzy sets $U : \mathbb{R}^p \to [0,1]$ such that the α-levels of U, $U_\alpha \in \mathscr{K}_c(\mathbb{R}^p)$ for all $\alpha \in [0,1]$, where $U_\alpha = \{x \in \mathbb{R}^p | U(x) \geq \alpha\}$ if $\alpha \in (0,1]$ and $U_0 = \mathrm{cl}\{x \in \mathbb{R}^p | U(x) > 0\}$.

The usual arithmetic between elements of $\mathscr{F}_c(\mathbb{R}^p)$ is based on Zadeh's extension principle [12], and it agrees levelwise with the Minkowski addition and the product by a real number for compact convex sets.

Let $\|\cdot\|_2$ denote the usual functional L_2-norm. In [11] a distance between two fuzzy sets U and V has been introduced defined as:

$$D_\theta^\varphi(U,V) = \sqrt{\int_{[0,1]} \left(\|\mathrm{mid}_{U_\alpha} - \mathrm{mid}_{V_\alpha}\|_2^2 + \theta \|\mathrm{spr}_{U_\alpha} - \mathrm{spr}_{V_\alpha}\|_2^2\right) d\varphi(\alpha)},$$

where $\mathrm{mid}_{V_\alpha}(u) = \dfrac{s_{V_\alpha}(u) - s_{V_\alpha}(-u)}{2}$, and $\mathrm{spr}_{V_\alpha}(u) = \frac{s_{V_\alpha}(u)+s_{V_\alpha}(-u)}{2}$, with s_{V_α} being the support function of V_α, which is the mapping $s_{V_\alpha} : \mathbb{S}^{p-1} \to \mathbb{R}$ such that $s_{V_\alpha}(u) = \sup\limits_{w \in V_\alpha} \langle u, w \rangle$ for all $u \in \mathbb{S}^{p-1}$.

The value $\theta > 0$ (often assumed to belong to $(0,1]$) determines the relative weight of the distance of the generalized spread against the distance of the generalized mid. The mapping φ is an absolutely continuous measure with positive mass function on the unit interval weighting the importance of each level [11].

Given the probability space $(\Omega, \mathscr{A}, P)$, an RFS is a Borel measurable mapping [2, 11]. This definition is equivalent to the one by Puri & Ralescu [4, 8].

Let $\|\cdot\|$ be the usual norm in $\mathbb{R}^p$. If $\sup_{x \in \mathscr{X}_0} \|x\| \in L^1(\Omega, \mathscr{A}, P)$, then the *expected value of an RFS* is defined in terms on the Aumann integral [1] as the unique fuzzy set $E(\mathscr{X}) \in \mathscr{F}_c(\mathbb{R}^p)$, such that for all $\alpha \in [0,1]$,

$$E(\mathscr{X}_\alpha) = \left\{ \int_\Omega f(w)dP(w) \,|\, f : \Omega \to \mathbb{R}, f \in L^1(\Omega, \mathscr{A}, P), f \in \mathscr{X}_\alpha \; a.s. - [P] \right\}.$$

If $\sup_{x \in \mathscr{X}_0} \|x\| \in L^2(\Omega, \mathscr{A}, P)$, the *Fréchet-type variance* [3] is defined as

$$\sigma^2_{\mathscr{X}} = E\big(D^\varphi_\theta(\mathscr{X}, E(\mathscr{X}))\big)^2.$$

3 Homoscedasticity Test for RFS's

Consider k populations and k independent associated RFS's, $\mathscr{X}_1, \ldots, \mathscr{X}_k$. From each $\mathscr{X}_i$, a simple random sample $\{\mathscr{X}_{i1}, \ldots, \mathscr{X}_{in_i}\}_{j=1}^{n_i}$ is drawn, where the total sample size equals N.

The aim is to test the hypotheses

$$\begin{cases} H_0 : \sigma^2_{\mathscr{X}_1} = \ldots = \sigma^2_{\mathscr{X}_k} \quad \text{vs.} \\ H_1 : \exists\, i, j \in \{1, \ldots, k\} \text{ s.t. } \sigma^2_{\mathscr{X}_i} \neq \sigma^2_{\mathscr{X}_j} \end{cases} \tag{3.1}$$

The sample mean associated with the i-th variable and the total sample mean are defined as usual on the basis of the fuzzy arithmetic. In addition:

- The *variance in the i-th sample* is defined as $\widehat{\sigma}^2_{\mathscr{X}_i} = \dfrac{\sum_{j=1}^{n_i} D^\varphi_\theta(\mathscr{X}_{ij}, \overline{\mathscr{X}_{i\cdot}})}{n_i}$.
- The *quasi-variance in the i-th sample* (unbiased and consistent estimate of the population variance as shown in [5]) is $\widehat{S}^2_{\mathscr{X}_i} = n_i \widehat{\sigma}^2_{\mathscr{X}_i}/(n_i - 1)$.

Inspired by the classical Levene's theory [6], the following statistic has been considered in [10] to test the proposed null hypothesis:

$$T_{(n_1,\ldots,n_k)} = \frac{\displaystyle\sum_{i=1}^k n_i \left(\widehat{\sigma}^2_{\mathscr{X}_i} - \frac{1}{N} \sum_{l=1}^k n_l \widehat{\sigma}^2_{\mathscr{X}_l} \right)^2}{\displaystyle\sum_{i=1}^k \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\left(D^\varphi_\theta\left(\mathscr{X}_{ij}, \overline{\mathscr{X}_{i\cdot}} \right) \right)^2 - \widehat{\sigma}^2_{\mathscr{X}_i} \right]^2}.$$

Under the null hypothesis H_0, if $\sup_{x \in \mathscr{X}_0} \|x\| \in L^4(\Omega, \mathscr{A}, P)$ and $n_i/N \to p_i \in (0,1)$ as $n_i \to \infty$ for all $i \in \{1, \ldots, k\}$, then

$$T_{(n_1,\ldots,n_k)} \xrightarrow{\mathscr{L}} \sum_{i=1}^k \left(y_i - \sum_{l=1}^k \sqrt{p_i p_l}\, y_l \right)^2 \Big/ \sum_{i=1}^k \sigma^2_{(D^\varphi_\theta(\mathscr{X}_i, E(\mathscr{X}_i)))^2},$$

where $(y_1, \ldots, y_k)^T \equiv \mathscr{N}_k\Big(\mathbf{0}, \Sigma\Big)$ with covariance matrix

$$\sum = diag\Big(\sigma^2_{(D^\varphi_\theta(\mathscr{X}_1, E(\mathscr{X}_1)))^2}, \ldots, \sigma^2_{(D^\varphi_\theta(\mathscr{X}_k, E(\mathscr{X}_k)))^2}\Big).$$

Thus, the following asymptotic procedure is obtained: for all $\beta \in [0,1]$, the size of the test that rejects H_0 whenever $T_{(n_1,\ldots,n_k)} > t_{1-\beta}$ converges to β, where

$t_{1-\beta}$ is the $[1-\beta]$-quantile of the asymptotic distribution of $T_{(n_1,\ldots,n_k)}$ under H_0, converges to β.

4 Power Analysis Under Local Alternatives

Local alternatives are a sequence of alternative hypotheses which converge to the null one as the sample size increases and determine the sensitivity of the test under small deviations from the null hypothesis.

Let $\mathscr{X}_1 \ldots \mathscr{X}_k$ be k independent RFS's verifying $H_0 : \sigma^2_{\mathscr{X}_1} = \ldots = \sigma^2_{\mathscr{X}_k}$. For $n_i \in \mathbb{N}$ and $i \in \{1,\ldots,k\}$, let $\{\mathscr{X}_{ij}\}_{j=1}^{n_i}$ be a random sample obtained from $\mathscr{X}_i$, and consider a 'correction' $\{\mathscr{X}_{ij}^{[n_i]}\}_{j=1}^{n_i}$ of $\{\mathscr{X}_{ij}\}_{j=1}^{n_i}$ defined as follows:

$$\mathscr{X}_{1j}^{[n_1]} = \sqrt{1+\frac{a_{n_1}}{\sqrt{n_1}}}\mathscr{X}_{1j} \quad \text{for } j \in \{1,\ldots,n_1\}$$

$$\mathscr{X}_{ij}^{[n_1]} = \mathscr{X}_{ij} \quad \text{for } j \in \{1,\ldots,n_i\}, i \in \{2,\ldots,k\}$$

in order to obtain RFS's which variances are given by

$$\sigma^2_{\mathscr{X}_1^{[n_1]}} = \left(1+\frac{a_{n_1}}{\sqrt{n_1}}\right)\sigma^2_{\mathscr{X}_1}, \text{ and } \sigma^2_{\mathscr{X}_i^{[n_i]}} = \sigma^2_{\mathscr{X}_i} \quad \text{for } i \in \{2,\ldots,k\}$$

where a_{n_1} is a sequence belonging to the interval $(0,\infty)$. Thus, if $a_{n_1} \to \infty$ and $a_{n_1}/\sqrt{n_1} \to 0$ as $n_1 \to \infty$, then the sequence of variances $\sigma^2_{X_1^{[n_1]}}$ converges pointwise to $\sigma^2_{\mathscr{X}_1}$ as the sample size n_1 tends to ∞. Consider the 'corrected' statistic:

$$T^* = \frac{\sum_{i=1}^k n_i \left[\widehat{\sigma}^2_{\mathscr{X}_i^{[n_i]}} - \frac{1}{N}\sum_{l=1}^k n_l \widehat{\sigma}^2_{\mathscr{X}_l^{[n_l]}}\right]^2}{\sum_{i=1}^k \frac{1}{n_i}\sum_{j=1}^{n_i}\left[\left(D_\theta^\varphi\left(\mathscr{X}_{ij}^{[n_i]}, \overline{\mathscr{X}_{i\cdot}^{[n_i]}}\right)\right)^2 - \widehat{\sigma}^2_{\mathscr{X}_i^{[n_i]}}\right]^2}.$$

In Theorem 1 it is shown that the power of the test under the proposed local alternatives converges to 1.

Theorem 1. *Let $\mathscr{X}_1,\ldots,\mathscr{X}_k$ be k RFS's verifying H_0. For $n_i \in \mathbb{N}$ and $i \in \{1,\ldots,k\}$, let $\{\mathscr{X}_{ij}\}_{j=1}^{n_i}$ be a random sample obtained from $\mathscr{X}_i$, and consider a 'correction' $\{\mathscr{X}_{ij}^{[n_i]}\}_{j=1}^{n_i}$ defined as above to obtain RFS's with variances*

$$\sigma^2_{\mathscr{X}_1^{[n_1]}} = \left(1+\frac{a_{n_1}}{\sqrt{n_1}}\right)\sigma^2_{\mathscr{X}_1}, \text{ and } \sigma^2_{\mathscr{X}_i^{[n_i]}} = \sigma^2_{\mathscr{X}_i} \text{ for } i \in \{2,\ldots,k\}$$

where $a_{n_1} \in (0,\infty)$ converges to ∞ and $a_{n_1}/\sqrt{n_1} \to 0$ as $n_1 \to \infty$.

Then, if $\sup_{x_i \in (\mathscr{X}_i)_0} \|x_i\| \in L^4(\Omega,\mathscr{A},P)$, $n_i/N \to p_i \in (0,1)$ as $n_i \to \infty$ and the asymptotic procedure in Section 3 with significance level β is applied to *the sequence of the corrected samples* $\{\mathscr{X}_{i1}^{[n_i]},\ldots,\mathscr{X}_{in_i}^{[n_i]}\}_{n_i}$ *for* $i\in\{1,\ldots,k\}$*, we have that*

$$\lim_{n_i\to\infty} P(T^* > t_{(1-\beta)}) = 1.$$

Proof. First of all, the denominator of T^* (denoted by D^*) satisfies that

$$\begin{aligned} D^* &= \left(1+\frac{a_{n_1}}{\sqrt{n_1}}\right)\frac{1}{n_1}\sum_{j=1}^{n_1}\left[\left(D_\theta^\varphi\left(\mathscr{X}_{1j},\overline{\mathscr{X}_{1\cdot}}\right)\right)^2-\widehat{\sigma}^2_{\mathscr{X}_1^{[n_1]}}\right]^2 \\ &+\sum_{i=2}^{k}\frac{1}{n_i}\sum_{j=1}^{n_i}\left[\left(D_\theta^\varphi\left(\mathscr{X}_{ij},\overline{\mathscr{X}_{i\cdot}}\right)\right)^2-\widehat{\sigma}^2_{\mathscr{X}_i}\right]^2. \end{aligned}$$

In [9] it is shown that if $\sup_{x_i\in(\mathscr{X}_i)_0}\|x_i\|\in L^4(\Omega,\mathscr{A},P)$, then

$$\frac{1}{n_i}\sum_{j=1}^{n_i}\left[\left(D_\theta^\varphi\left(\mathscr{X}_{ij},\overline{\mathscr{X}_{i\cdot}}\right)\right)^2-\widehat{\sigma}^2_{\mathscr{X}_i}\right]^2 \xrightarrow{a.s.} \sigma^2_{(D_\theta^\varphi(\mathscr{X}_i,E(\mathscr{X}_i)))^2}$$

for $i\in\{2,\ldots,k\}$. Moreover, $\left(1+\dfrac{a_{n_1}}{\sqrt{n_1}}\right)\xrightarrow{n_1\to\infty} 1$. Therefore,

$$D^* \xrightarrow{a.s.} D = \sum_{i=1}^{k}\sigma^2_{(D_\theta^\varphi(\mathscr{X}_i,E(\mathscr{X}_i)))^2}.$$

The numerator of T^* (denoted by M^*) can be decomposed as

$$M^* = n_1\left[\widehat{\sigma}^2_{\mathscr{X}_1^{[n_1]}}-\frac{1}{N}\sum_{i=1}^{k}n_i\widehat{\sigma}^2_{\mathscr{X}_i^{[n_i]}}\right]^2+\sum_{i=2}^{k}n_i\left[\widehat{\sigma}^2_{\mathscr{X}_i^{[n_i]}}-\frac{1}{N}\sum_{l=1}^{k}n_l\widehat{\sigma}^2_{\mathscr{X}_l^{[n_l]}}\right]^2.$$

We can expand the first term of M^*, M_1^*, as follows:

$$\begin{aligned} M_1^* &= \left[\sqrt{n_1}\big(\widehat{\sigma}^2_{\mathscr{X}_1}-\frac{1}{N}\sum_{i=1}^{k}n_i\widehat{\sigma}^2_{\mathscr{X}_i}\big)\right]^2 && (4.2)\\ &+\left[a_{n_1}\left(\widehat{\sigma}^2_{\mathscr{X}_1}-\frac{n_1}{N}\widehat{\sigma}^2_{\mathscr{X}_1}\right)\right]^2 && (4.3)\\ &+2n_1\left(\widehat{\sigma}^2_{\mathscr{X}_1}-\frac{1}{N}\sum_{i=1}^{k}n_i\widehat{\sigma}^2_{\mathscr{X}_i}\right)\left(\frac{a_{n_1}}{\sqrt{n_1}}\left(\widehat{\sigma}^2_{\mathscr{X}_1}-\frac{n_1}{N}\widehat{\sigma}^2_{\mathscr{X}_1}\right)\right) && (4.4) \end{aligned}$$

The second term of M^*, M_2^*, has the following decomposition:

$$\begin{aligned} M_2^* &= \sum_{i=2}^{k}n_i\left[\widehat{\sigma}^2_{\mathscr{X}_i}-\frac{1}{N}\sum_{l=1}^{k}n_l\widehat{\sigma}^2_{\mathscr{X}_l}\right]^2 && (4.5)\\ &+(N-n_1)\left(\frac{n_1a^2_{n_1}}{N^2}\right)\widehat{\sigma}^4_{\mathscr{X}_1} && (4.6)\\ &-2\frac{\sqrt{n_1}a_{n_1}}{N}\widehat{\sigma}^2_{\mathscr{X}_1}\sum_{i=2}^{k}n_i\left[\widehat{\sigma}^2_{\mathscr{X}_i}-\frac{1}{N}\sum_{l=1}^{k}n_l\widehat{\sigma}^2_{\mathscr{X}_l}\right] && (4.7) \end{aligned}$$

Firstly, (4.2) is a function of the variables $\{\xi_i\}_{i=1}^k$ given by

$$f_n(\xi_1,\ldots,\xi_k) = \left(\xi_1 - \sum_{i=1}^k \sqrt{\frac{n_1}{N}}\sqrt{\frac{n_i}{N}}\xi_l\right)^2$$

Taking into account that $n_i/N \to p_i \in (0,1)$ as $n_i \to \infty$ for all $i \in \{1,\ldots,k\}$, this function converges uniformly to the function defined as

$$f(y_1,\ldots,y_k) = \left(y_1 - \sum_{i=1}^k \sqrt{p_1 p_i} y_i\right)^2,$$

where the vector $(y_1,\ldots,y_k)^T$ is distributed as the one in Section 3. Analogously, (4.5) converges uniformly to

$$f(y_1,\ldots,y_k) = \sum_{i=2}^k \left(y_i - \sum_{l=1}^k \sqrt{p_i p_l} y_l\right)^2.$$

Since $a_{n_1} \to \infty$ and $\left(\widehat{\sigma}^2_{\mathscr{X}_1} - \frac{n_1}{N}\widehat{\sigma}^2_{\mathscr{X}_1}\right)^2 \xrightarrow{a.s} (1-p_1)^2\sigma^4_{\mathscr{X}_1}$, we conclude that the term (4.3) divided by $(D^* a_{n_1})$ converges to ∞ as $n_i/N \overset{n_i\to\infty}{\longrightarrow} p_i$ for $i \in \{1,\ldots,k\}$. In addition, by taking into account the preceding arguments, it is satisfied that (4.4) divided by $(D^* a_{n_1})$ converges in law to

$$2(1-p_1)\sigma^2_{\mathscr{X}_1}(y_1 - \sum_{i=1}^k \sqrt{p_1 p_i} y_i)^2.$$

On the other hand, since $\left(\frac{n_1}{N}\right)\left(\frac{N-n_1}{N}\right)^2 \widehat{\sigma}^4_{\mathscr{X}_1} \xrightarrow{a.s} p_1 \sigma^4_{\mathscr{X}_1} \sum_{i=2}^k p_i$ and the sequence $a_{n_1} \to \infty$, then the term (4.6) divided by $(D^* a_{n_1})$ converges to ∞ as $n_i/N \overset{n_i\to\infty}{\longrightarrow} p_i$ for $i \in \{1,\ldots,k\}$. Moreover, from the former results it is easy to see that (4.7) divided by $(D^* a_{n_1})$ converges in law to

$$\left[-2p_1 \sum_{i=2}^k p_i\big(y_i - \sum_{l=1}^k \sqrt{p_i p_l} y_l\big)\right].$$

Finally, from the preceding expressions, we have that

$$T^*/a_{n_1} \xrightarrow{P} \infty$$

as $n_i/N \overset{n_i\to\infty}{\longrightarrow} p_i$ for $i \in \{1,\ldots,k\}$ and, as a result,

$$\lim_{n_i\to\infty} P(T^* > t_{(1-\alpha)}) = 1. \qquad \square$$

Remark 1. Theorem 1 indicates that for any sequence $\{a_{n_1}\}$ such that $a_{n_1} \in (0,\infty)$, $a_{n_1} \to \infty$ and $a_{n_1}/\sqrt{n_1} \to 0$ as $n_1 \to \infty$, then $\sigma^2_{\mathscr{X}_1^{[n_1]}} \longrightarrow \sigma^2_{\mathscr{X}_1}$ pointwise, and the proposed asymptotic procedure rejects H_0 with probability 1 in the limit.

5 Simulation Studies

To analyze the consistency of the Test (3.1) for RFS's, some simulations of the power function have been carried out by using local alternatives. Given $a \in (0,\infty)$, 3 triangular RFS's have been considered, namely,

- $\mathscr{X}_1^a \equiv \sqrt{a/6.44} \cdot T(l_1,c_1,r_1)$ s.t. $l_1 \equiv \chi_3^2$, $c_1 \equiv \mathscr{N}(1,2)$ and $r_1 \equiv \chi_8^2$,
- $\mathscr{X}_2 \equiv T(l_2,c_2,r_2)$ s.t. $l_2 \equiv \chi_8^2$, $c_2 \equiv \mathscr{N}(0,2)$ and $r_2 \equiv \chi_3^2$,
- $\mathscr{X}_3 \equiv T(l_3,c_3,r_3)$ s.t. $l_3 \equiv \chi_5^2$, $c_3 \equiv \mathscr{N}(-1,2)$ and $r_3 \equiv \chi_6^2$,

These random elements verify that

$$\sigma^2_{\mathscr{X}_1^a} = a \text{ for all } a \in (0,\infty) \quad \text{and} \quad \sigma^2_{\mathscr{X}_2} = \sigma^2_{\mathscr{X}_3} = 6.44.$$

It should be noted that when $a = 6.44$, then the null hypothesis $H_0 = \sigma^2_{\mathscr{X}_1^a} = \sigma^2_{\mathscr{X}_2} = \sigma^2_{\mathscr{X}_3}$ is verified. The power function at the level $\alpha = .05$ as a function of a is shown in Figure 1. The involved metric D_θ^φ was chosen so that $\theta = 1/3$ and φ=Lebesgue measure. Three simple random samples of size 100 from $\mathscr{X}_1^a$, $\mathscr{X}_2$ and $\mathscr{X}_3$ respectively have been drawn for different values of a ranging in $(0,17]$, and 10,000 simulations of the asymptotic testing procedure have been carried out. As a result, Figure 1 shows that the power is close to 1 as far as the alternative hypothesis is from the null one.

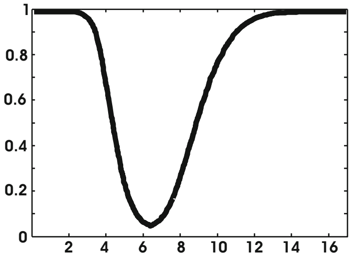


Fig. 1 Power of the test $H_0 : \sigma^2_{\mathscr{X}_1} = \sigma^2_{\mathscr{X}_2} = \sigma^2_{\mathscr{X}_3}$

6 Concluding Remarks

The power of the homoscedasticity test for k RFSŠs has been theoretically and empirically analyzed by using local alternatives. The results show that the asymptotic procedure is consistent.

One immediate problem to be discussed is that of analyzing the power under local alternatives of the bootstrap testing procedure proposed in [10]. It would be also interesting to develop sensitivity analyses concerning the influence of choice of the value θ and the measure φ.

Acknowledgements. The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02, by the Principality of Asturias Project IB09-042C2, by the NFU/LILA-CATA project and by the COST Action IC0702.

References

1. Aumann, R.J.: Integrals of set-valued functions. J. Math. Anal. Appl. 12, 1–12 (1965)
2. Colubi, A., Domínguez-Menchero, J.S., López-Díaz, M., Ralescu, R.: A $D_E[0,1]$ representation of random upper semicontinuous functions. Proc. Amer. Math. Soc. 130, 3237–3242 (2002)
3. Fréchet, M.: Les éléments aléatoires de nature quelconque dan un espace distancié. Ann. Inst. Henri Poincaré 10, 215–310 (1948)
4. Klement, E., Puri, M.L., Ralescu, D.A.: Limit theorems for fuzzy random variables. Proc. Roy. Soc. London. Ser. A 1832, 171–182 (1986)
5. Körner, R., Näther, W.: On the variance of random fuzzy variables. In: Bertoluzza, C., Gil, M.A., Ralescu, D.A. (eds.) Statistical Modeling, Analysis and Management of Fuzzy Data, pp. 22–39. Physica-Verlag, Heidelberg (2002)
6. Levene, H.: Robust Tests for Equality of Variances. In: Olkin, I. (ed.) Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling, pp. 278–292. Standford University Press, Stanford (1960)
7. Lubiano, M.A., Alonso, C., Gil, M.A.: Statistical inferences on the S-mean squared dispersion of a fuzzy random variable. In: Proceedings of the Joint EUROFUSE-SIC 1999 International Conference, EUROFUSE-SIC 1999, Budapest, Hungary, pp. 532–537 (1999)
8. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math Anal. Appl. 114, 409–422 (1986)
9. Ramos-Guajardo, A.B., Colubi, A., González-Rodríguez, G., Gil, M.A.: One-sample tests for a generalized Fréchet variance of a fuzzy random variable. Metrika 71(2), 185–202 (2009)
10. Ramos-Guajardo, A.B., Lubiano, M.A.: K-sample homoscedasticity tests for random fuzzy sets (submitted for publication, 2010)
11. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inform. Sci. 179, 3964–3972 (2009)
12. Zadeh, L.A.: 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 (1975)

Periodic Generalized-Differentiable Solutions to Fuzzy Differential Equations

Rosana Rodríguez-López

Abstract. We study the existence of solution to a class of periodic boundary value problems for first-order fuzzy differential equations under generalized differentiability.

We allow the coefficient of the linear equation to change its sign an arbitrary number of times in the interval of interest, extending some previous results.

Keywords: Fuzzy differential equations, Generalized-differentiability, Periodic boundary value problems.

1 Introduction

The study of the existence of solution to fuzzy differential equations is based on the establishment of the concept of differentiability of a fuzzy function. The differentiability in the sense of Hukuhara [15] is one of the first approaches, and it allows to study the existence and uniqueness of solution [6]. However, one important drawback of this approach is the nondecreasing character of the diameter of the level sets of the solutions, so that the introduction of impulses [16] is a procedure to obtain periodic solutions to first-order fuzzy differential equations under Hukuhara differentiability. In this reference, a periodic boundary value problem related to an impulsive fuzzy differential equation is considered, and the monotone iterative technique is used to approximate the extremal solutions in a fixed fuzzy functional interval. We also cite references [4, 5, 6, 7, 9, 11, 12, 14] for the foundations of fuzzy sets, fuzzy differential equations and fuzzy functional differential equations, other

Rosana Rodríguez-López
Departamento de Análisis Matemático, Facultad de Matemáticas,
Universidad de Santiago de Compostela, Spain
e-mail: rosana.rodriguez.lopez@usc.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 545–552.
springerlink.com

approaches in the analysis of uncertain systems, and some periodic boundary value problems for second-order fuzzy differential equations. For strongly generalized differentiability, we refer to [1] and [2, 3, 8, 13, 19].

In [19], the authors solve interval differential equations by proving a characterization by ODEs and using a numerical procedure, which is applied to some initial value problems for fuzzy linear differential equations whose solutions are obtained by combination of two types of derivatives using a switching point. In [10], the authors use generalized differentiability and establish sufficient conditions for the existence of solution to boundary value problems for first-order differential equations of the type

$$\begin{cases} y'(t) = a(t)y(t) + b(t), \, t \in J, \\ y(0) = y(T), \end{cases}$$

where $J = [0,T]$, and $a : [0,T] \to \mathbb{R}$, $b : [0,T] \to \mathbb{R}_F$ are continuous functions. In the results provided, the sign of function a is allowed to change only once in the interval J.

We study a more general problem, proving the existence of periodic solutions for a class of fuzzy differential equations under strongly generalized differentiability, providing weaker hypotheses for the coefficients of the equation. Under this approach, we illustrate the capability of fuzzy differential equations to modeling periodic phenomena.

1.1 Preliminaries

We consider the space of fuzzy intervals $\mathbb{R}_F$ as the class of elements $u : \mathbb{R} \to [0,1]$ such that

(i) u is normal, i.e., there exists $s_0 \in \mathbb{R}$ such that $u(s_0) = 1$,
(ii) u is fuzzy-convex, that is, $u(ts + (1-t)r) \geq \min\{u(s), u(r)\}$, $\forall t \in [0,1]$, and $s, r \in \mathbb{R}$,
(iii) u is upper semicontinuous on $\mathbb{R}$,
(iv) $cl\{s \in \mathbb{R} \mid u(s) > 0\}$ is compact, where cl denotes the closure of a subset.

For each $0 < \alpha \leq 1$, we denote the level set $[u]^\alpha = \{s \in \mathbb{R} \mid u(s) \geq \alpha\}$ and $[u]^0 = cl\{s \in \mathbb{R} \mid u(s) > 0\}$, which is a non-empty compact interval for all $0 \leq \alpha \leq 1$ and every $u \in \mathbb{R}_F$.

We also denote $[u]^\alpha = [\underline{u}^\alpha, \overline{u}^\alpha]$ and $diam[u]^\alpha = \overline{u}^\alpha - \underline{u}^\alpha$. Functions $\underline{u}$ and $\overline{u}$ are the lower and upper branches of u, respectively.

The addition and multiplication by an scalar in $\mathbb{R}_F$ are defined levelsetwise and the metric structure is given by the distance $D : \mathbb{R}_F \times \mathbb{R}_F \to \mathbb{R}_+ \cup \{0\}$, defined as

$$D(u,v) = \sup_{\alpha \in [0,1]} \max\{|\underline{u}^\alpha - \underline{v}^\alpha|, |\overline{u}^\alpha - \overline{v}^\alpha|\}, \text{ for } u, v \in \mathbb{R}_F. \tag{1}$$

With this metric, $\mathbb{R}_F$ is a complete metric space. We also define the difference of Hukuhara of fuzzy intervals.

Definition 1. *Given $x,y \in \mathbb{R}_F$, if there exists $z \in \mathbb{R}_F$ with $x = y + z$, we say that z is the H-difference of x,y, denoted by $x \ominus y$.*

For the differentiability of fuzzy-valued functions, we consider the concept of generalized differentiability [1, 2].

2 Existence of Periodic Solutions

We consider the periodic boundary value problem

$$\begin{cases} y'(t) = a(t)y(t) + b(t), \ t \in J, \\ y(0) = y(T), \end{cases} \tag{2}$$

where $J = [0,T]$, and $a : [0,T] \to \mathbb{R}$, $b : [0,T] \to \mathbb{R}_F$ are continuous functions. This problem is connected with problem (6) in [3] and the equation studied in [10].

In the interval $J = [0,T]$, we consider a finite sequence of real numbers $\delta_k \in (0,T)$, $k = 1,2,\ldots,m$, in such a way that $0 = \delta_0 < \delta_1 < \cdots < \delta_m < \delta_{m+1} = T$.

We define the concept of solution to problem (2), by defining the space of functions where the solutions lie, as follows.

Definition 2. *Let $J = [0,T]$ be a real interval and consider the sequence of real numbers $\delta_k \in (0,T)$, $k = 1,2,\ldots,m$, satisfying that*

$$0 = \delta_0 < \delta_1 < \cdots < \delta_m < \delta_{m+1} = T.$$

We define the space $\mathscr{F}_{\{\delta_k\}} = \mathscr{F}_{\{\delta_1,\ldots,\delta_m\}}$, consisting on the functions $u \in C(J,\mathbb{R}_F)$ which are differentiable in the sense of generalized differentiability on $J \setminus \{\delta_1,\ldots,\ \delta_m\}$ and such that there exist the one-sided limits $u'(\delta_k^-)$ and $u'(\delta_k^+)$ in the sense of generalized differentiability, for every $k = 1,2,\ldots,m$.

Definition 3. *A solution of (2) is a function in the space $\mathscr{F}_{\{\delta_k\}} = \mathscr{F}_{\{\delta_1,\ldots,\delta_m\}}$ satisfying the conditions in (2).*

For an arbitrary number of terms of the sequence $\{\delta_k\}$, we consider that the continuous real function $a : [0,T] \to \mathbb{R}$ is such that

$$a > 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an even number with } k \leq m,$$

and

$$a < 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an odd number with } k \leq m.$$

It is obvious that $a(\delta_k) = 0$, for every $k = 1,\ldots,m$.

Theorem 1. *Suppose that $J=[0,T]$, $a:[0,T]\to\mathbb{R}$, $b:[0,T]\to\mathbb{R}_F$ are continuous functions satisfying*

$$a>0 \text{ on } (\delta_k,\delta_{k+1}), \text{ for } k \text{ an even number with } k\le m,$$

$$a<0 \text{ on } (\delta_k,\delta_{k+1}), \text{ for } k \text{ an odd number with } k\le m.$$

Suppose also that, for every $\alpha\in[0,1]$, and every odd number j with $\delta_j<T$ ($1\le j\le m$),

$$\sum_{l=0}^{j}(-1)^l\int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\,ds\ge 0, \tag{3}$$

and

$$\int_0^T a(u)\,du<0. \tag{4}$$

Then there exists a solution to problem (2) in the space $\mathscr{F}_{\{\delta_1,\dots,\delta_m\}}$. Furthermore, this solution u is (i)-differentiable on $\bigcup_{k\ even}(\delta_k,\delta_{k+1})$ and (ii)-differentiable on $\bigcup_{k\ odd}(\delta_k,\delta_{k+1})$.

Proof. The proof is based on the definition of an operator which maps a certain initial condition y_0 into a fuzzy number equal to the value at T of a solution to the equation (in the sense of Definition 3) starting at y_0 which is obtained piecewise by using (i)-differentiable and (ii)-differentiable solutions according to the sign of the coefficient a on each interval, where the expressions of the solutions given in [3] play an important role. Indeed, for each initial condition $y_0\in\mathbb{R}_F$, we consider the solution $y\in C(J,\mathbb{R}_F)$ to the linear fuzzy differential equation $y'(t)=a(t)y(t)+b(t)$, $t\in J$, given recursively by

$$y(t)=\begin{cases} e^{\int_{\delta_j}^t a(r)\,dr}\left(y(\delta_j)+\int_{\delta_j}^t b(s)\,e^{-\int_{\delta_j}^s a(r)\,dr}\,ds\right), \\ \qquad\qquad\qquad\text{for } t\in(\delta_j,\delta_{j+1}], \text{ and } j \text{ even}, \\ e^{\int_{\delta_j}^t a(r)\,dr}\left(y(\delta_j)\ominus\int_{\delta_j}^t (-b(s))\,e^{-\int_{\delta_j}^s a(r)\,dr}\,ds\right), \\ \qquad\qquad\qquad\text{for } t\in(\delta_j,\delta_{j+1}], \text{ and } j \text{ odd}, \end{cases}$$

where $y(\delta_0)=y(0)=y_0$. Next, we define the operator

$$\mathscr{G}:\mathbb{R}_F\longrightarrow\mathbb{R}_F,$$

given by $\mathscr{G}(y_0)=y(T)$, for $y_0\in\mathbb{R}_F$.

The integral condition (4) provides the contractive character of the operator $\mathscr{G}$ which has, in virtue of the contractive mapping theorem, a unique fixed point, which is a periodic solution to the fuzzy differential equation. The remaining conditions are imposed in order to define the operator properly, in what concerns the expression of the solution on the intervals of (ii)-differentiability, that is, to prove the existence of the Hukuhara differences

$$y(\delta_j) \ominus \int_{\delta_j}^{t} (-b(s))\, e^{-\int_{\delta_j}^{s} a(r)\,dr}\, ds,$$

for every $t \in (\delta_j, \delta_{j+1}]$, where j is odd. In our procedure, it is also important to control the value of the diameter of the level sets of the piecewise solution at the conjunction points δ_j. For the full details of the proof of this result, we refer to [18]. □

Corollary 1. *In Theorem 1, taking $m = 3$, that is, $0 = \delta_0 < \delta_1 < \delta_2 < \delta_3 < \delta_4 = T$ and $a : [0,T] \to \mathbb{R}$ a continuous real function satisfying that*

$$a > 0 \text{ on } (0, \delta_1) \cup (\delta_2, \delta_3),$$

$$a < 0 \text{ on } (\delta_1, \delta_2) \cup (\delta_3, T).$$

we just have to check condition (3) for $j = 1$ and $j = 3$, that is, we derive conditions

$$\int_{\delta_1}^{\delta_2} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\, ds \le \int_0^{\delta_1} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\, ds \tag{5}$$

and

$$\sum_{j=1,3} \int_{\delta_j}^{\delta_{j+1}} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\, ds$$
$$\le \sum_{j=0,2} \int_{\delta_j}^{\delta_{j+1}} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\, ds, \tag{6}$$

for every $\alpha \in [0,1]$. These hypotheses, joint to condition (4) which provides the contractive character of the operator, allow to affirm the existence of a solution to problem (2) which is (i)-differentiable on $(0, \delta_1) \cup (\delta_2, \delta_3)$ and (ii)-differentiable on $(\delta_1, \delta_2) \cup (\delta_3, T)$.

Corollary 2. *For $m = 2$, $0 = \delta_0 < \delta_1 < \delta_2 < \delta_3 = T$ and $a : [0,T] \to \mathbb{R}$ a continuous real function satisfying that*

$$a > 0 \text{ on } (0, \delta_1) \cup (\delta_2, T), \qquad a < 0 \text{ on } (\delta_1, \delta_2),$$

conditions (4) and (5) provide the existence of a solution to problem (2) which is (i)-differentiable on $(0, \delta_1) \cup (\delta_2, T)$ and (ii)-differentiable on (δ_1, δ_2).

Corollary 3. *For the case $m = 1$, if $a > 0$ on $(0, \delta)$, and $a < 0$ on (δ, T), then Theorem 1 is reduced to the results in [10].*

Next, we consider the case where the continuous function $a : [0,T] \to \mathbb{R}$ is such that

$$a < 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an even number with } k \le m,$$

$$a > 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an odd number with } k \le m.$$

Again $a(\delta_k) = 0$, for every $k = 1, \dots, m$.

Theorem 2. *Suppose that $J=[0,T]$, and $a:[0,T]\to\mathbb{R}$, $b:[0,T]\to\mathbb{R}_F$ are continuous functions satisfying*

$$a<0 \text{ on } (\delta_k,\delta_{k+1}), \text{ for } k \text{ an even number with } k\le m,$$

$$a>0 \text{ on } (\delta_k,\delta_{k+1}), \text{ for } k \text{ an odd number with } k\le m.$$

Suppose also that, for every $\alpha\in[0,1]$, and every even number j satisfying that $j\in\{0,\ldots,m\}$, the following inequality holds

$$e^{\int_0^T a(u)\,du}\sum_{l=j+1}^{m}(-1)^{l+1}\int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\,ds$$

$$\ge \sum_{l=0}^{j}(-1)^{l}\int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^{\alpha}\, e^{-\int_0^s a(u)\,du}\,ds, \qquad (7)$$

and

$$\int_0^T a(u)\,du<0.$$

Then there exists a solution to problem (2) in the space $\mathscr{F}_{\{\delta_1,\ldots,\delta_m\}}$. Furthermore, this solution u is (i)-differentiable on $\bigcup_{k \text{ odd}}(\delta_k,\delta_{k+1})$ and (ii)-differentiable on $\bigcup_{k \text{ even}}(\delta_k,\delta_{k+1})$.

Proof. For the proof of this result, we consider, for a fixed initial condition $y_0\in\mathbb{R}_F$, the solution $y\in C(J,\mathbb{R}_F)$ to the linear fuzzy differential equation $y'(t)=a(t)y(t)+b(t)$, $t\in J$, given recursively by

$$y(t)=\begin{cases} e^{\int_{\delta_j}^t a(r)\,dr}\left(y(\delta_j)\ominus\int_{\delta_j}^t(-b(s))\,e^{-\int_{\delta_j}^s a(r)\,dr}\,ds\right), \\ \qquad\qquad\qquad \text{for } t\in(\delta_j,\delta_{j+1}], \text{ and } j \text{ even}, \\ e^{\int_{\delta_j}^t a(r)\,dr}\left(y(\delta_j)+\int_{\delta_j}^t b(s)\,e^{-\int_{\delta_j}^s a(r)\,dr}\,ds\right), \\ \qquad\qquad\qquad \text{for } t\in(\delta_j,\delta_{j+1}], \text{ and } j \text{ odd}, \end{cases}$$

where $y(\delta_0)=y(0)=y_0$. We define the operator

$$\widetilde{\mathscr{G}}:C\longrightarrow C,$$

given by $\widetilde{\mathscr{G}}(y_0)=y(T)$, for $y_0\in C$, where

$$C=\Big\{y_0\in\mathbb{R}_F : diam[y_0]^{\alpha}\ge\sum_{l=0}^{j}(-1)^l\int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^{\alpha}\,e^{-\int_0^s a(u)\,du}\,ds,$$
$$\text{for every } \alpha\in[0,1], \text{ and every even number } j \text{ with } 0\le j\le m\Big\}.$$

The choice of C is made in order to ensure that the Hukuhara differences in the expression of y exist. The hypothesis (4) can be used again to check the contractive character of the operator $\widetilde{\mathscr{G}}$ so that there exists a unique fixed point of $\widetilde{\mathscr{G}}$, which is a solution to problem (2). On the other hand, condition (7) is important for a good definition of $\widetilde{\mathscr{G}}$. For the complete proof of this result, we refer again to [18]. □

The previous result extends some results given in [10] for $a < 0$ on $(0, \delta)$ and $a > 0$ on (δ, T).

Remark 1. In [18], it is also proved that in the same context

$$a < 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an even number with } k \leq m,$$

$$a > 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an odd number with } k \leq m,$$

we can also deduce the existence of solution to the periodic boundary value problem (2) in the set

$$\widehat{C} = \left\{ y_0 \in \mathbb{R}_F : diam[y_0]^\alpha \geq \int_0^{\delta_1} diam[b(s)]^\alpha e^{-\int_0^s a(u)\,du} ds, \forall \alpha \in [0,1] \right\}, \quad (8)$$

just by assuming (4) and the corresponding sufficient conditions which are established in order to guarantee that operator $\widetilde{\mathscr{G}}$ is well-defined and maps $\widehat{C}$ into itself. See [18] for details.

Finally, similarly to [10], it is possible to study conditions to guarantee that solutions are crisp at the points where the type of differentibility changes from an interval of (ii)-differentiability to another interval of (i)-differentiability, that is:

- Points δ_j where j is an even number with $2 \leq j \leq m$, in the case
$$a > 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an even number with } k \leq m,$$
$$a < 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an odd number with } k \leq m.$$
- Points δ_j where j is an odd number $1 \leq j \leq m$, in the case
$$a < 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an even number with } k \leq m,$$
$$a > 0 \text{ on } (\delta_k, \delta_{k+1}), \text{ for } k \text{ an odd number with } k \leq m.$$

In this last case, as established in [18], the condition to be imposed is

$$\begin{aligned} &e^{\int_0^T a(u)\,du} \sum_{l=j+1}^{m} (-1)^{l+1} \int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^\alpha e^{-\int_0^s a(u)\,du} ds \\ &= \sum_{l=0}^{j} (-1)^l \int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^\alpha e^{-\int_0^s a(u)\,du} ds, \; \forall \alpha \in [0,1], \end{aligned} \quad (9)$$

which has to be fulfilled for every even number j with $0 \leq j < m$. If m is an even number, we also have to impose (see [18]) the additional hypothesis

$$\sum_{l=0}^{m} (-1)^l \int_{\delta_l}^{\delta_{l+1}} diam[b(s)]^\alpha e^{-\int_0^s a(u)\,du} ds \leq 0. \quad (10)$$

References

1. Bede, B., Gal, S.G.: Almost periodic fuzzy-number-valued functions. Fuzzy Sets Syst. 147(3), 385–403 (2004)
2. Bede, B., Gal, S.G.: Generalizations of the differentiability of fuzzy-number-valued functions with applications to fuzzy differential equations. Fuzzy Sets Syst. 151(3), 581–599 (2005)
3. Bede, B., Rudas, I.J., Bencsik, A.L.: First order linear fuzzy differential equations under generalized differentiability. Inform. Sci. 177(7), 1648–1662 (2007)
4. Diamond, P., Kloeden, P.: Metric Spaces of Fuzzy Sets. World Scientific, Singapore (1994)
5. Dubois, D., Prade, H.: Towards fuzzy differential calculus part 3: Differentiation. Fuzzy Sets Syst. 8(3), 225–233 (1982)
6. Kaleva, O.: Fuzzy differential equations. Fuzzy Sets Syst. 24(3), 301–317 (1987)
7. Kaleva, O.: A note on fuzzy differential equations. Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods 64(5), 895–900 (2006)
8. Khastan, A., Bahrami, F., Ivaz, K.: New Results on Multiple Solutions for Nth-Order Fuzzy Differential Equations under Generalized Differentiability. Bound. Value Probl. Article ID 395714 (2009)
9. Khastan, A., Nieto, J.J.: A boundary value problem for second order fuzzy differential equations. Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods 72(9-10), 3583–3593 (2010)
10. Khastan, A., Nieto, J.J., Rodríguez-López, R.: Periodic boundary value problems for first-order differential equations with uncertainty under generalized differentiability (Preprint, 2010)
11. Hüllermeier, E.: An Approach to Modelling and Simulation of Uncertain Dynamical Systems. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 5(2), 117–138 (1997)
12. Lupulescu, V.: On a class of fuzzy functional differential equations. Fuzzy Sets Syst. 160(11), 1547–1562 (2009)
13. Nieto, J.J., Khastan, A., Ivaz, K.: Numerical solution of fuzzy differential equations under generalized differentiability. Nonlinear Anal., Hybrid Syst. 3, 700–707 (2009)
14. Nieto, J.J., Rodríguez-López, R., Franco, D.: Linear first-order fuzzy differential equations. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 14(6), 687–709 (2006)
15. Puri, M.L., Ralescu, D.A.: Differentials of fuzzy functions. J. Math. Anal. Appl. 91, 552–558 (1983)
16. Rodríguez-López, R.: Periodic boundary value problems for impulsive fuzzy differential equations. Fuzzy Sets Syst. 159(11), 1384–1409 (2008)
17. Rodríguez-López, R.: On Boundary Value Problems for Fuzzy Differential Equations. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Handling Variability and Imprecision. Advances in Soft Computing, vol. 48, pp. 218–225. Springer, Heidelberg (2008)
18. Rodríguez-López, R.: On the existence of periodic solutions to fuzzy linear differential equations (Preprint, 2010)
19. Stefanini, L., Bede, B.: Generalized Hukuhara differentiability of interval-valued functions and interval differential equations. Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods 71(3-4), 1311–1328 (2009)

The Selection of the Shrinkage Region in Small Area Estimation

Cristina Rueda and José A. Menéndez

Abstract. In this article we consider general mixed models to derive small area estimators. The fixed part of the models links the area parameters to the auxiliary variables using a shrinkage region. We show how the selection of the shrinkage region depends on two main factors: the inter-area variation and the correlation coefficient of the auxiliaries with the response.

Keywords: James-Stein estimator, Order restricted inference, Mixed models, Fay-Herriot model.

1 Introduction

Typically, a Fay-Herriot model assumes a $D \times p$ matrix (D being the number of areas) of auxiliary variables $\mathbf{x}$ related to the D-dimensional parameter of interest $\overline{Y}$ by a linear model $\overline{Y_d} = x_d'\beta + u_d$, $d = 1,...,D$. Moreover, it is also assumed that the direct estimates verify $y_d = \overline{Y_d} + e_d$, where e_d is the survey error. An extreme case, when the auxiliary is constant, gives the James-Stein estimator.

In this paper, we propose a general approach to small area estimation that includes the James-Stein and Fay-Herriot methodologies as particular cases and also other estimators obtained by relaxing the linearity assumption about the relationship between the auxiliaries and the response.

To motivate the definition of the proposed models, we will first comment on the role of auxiliary variables in the small area estimation problem. A general assumption in this context is that some kind of shrinkage would provide more precise estimators in the presence of high sample variability. The simplest

Cristina Rueda and José A. Menéndez
Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, 47005 Valladolid, Spain
e-mail: crueda@eio.uva.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 553–560.
springerlink.com

way of achieving this shrinkage is by using the James-Stein estimator based on shrinkage towards the mean, which therefore implies the use of the a priori idea of equality among the areas. Mathematically, this means that the shrinkage region is of dimension 1. In many applications, this seems to be too strong an assumption and the availability of auxiliary information is used to define other (higher dimensional) shrinkage regions. The Fay-Herriot methodology uses p auxiliary variables to define linear shrinkage subspaces of dimension p. Using only the knowledge of the monotonicity, the shrinkage regions can be defined in several ways, mathematically represented by convex cones in $\Re^D$, with dimensions going from 2 to D. In a similar way, other shrinkage regions arise if a more complex shape relationship, such as a convex relation, is assumed.

We will show that the mathematical formulation is the same for the different alternatives. Estimators for area means prediction are derived by combining Order Restricted Inference (ORI) and mixed models standard approaches. We analyze, using simulated experiments, two important factors determining the selection of the shrinkage region: the small-area variation and the correlation coefficient. A bootstrap approach is proposed to select the best alternative and to provide estimators of the mean square prediction error and confidence intervals for the means.

First, we introduce the general model and the estimators for the small areas and for the variance of the random effect in Section 2. The bootstrap approach is also introduced in this section. In Section 3, several simulation results are presented and conclusions about the selection of the shrinkage region are derived. Finally, in Section 4, the baseball data set is revisited and it is used to illustrate the questions introduced in the paper.

2 Isotonic Area-Level Models

Let $\overline{Y}_d = \mu_d$, $d = 1,...D$, be the parameters of interest that we will assume, for simplicity, are the area means. For each area, we have a direct estimator y_d and the information on p auxiliary variables $\mathbf{x}_d$. Consider the general restricted mixed model that is given by the two-level model:

Level 1: Sampling model $y_d/\mu_d \rightsquigarrow N(\mu_d, \sigma_d^2)$, $d = 1,...,D$.

Level 2: Linking model $\mu_d \rightsquigarrow N\left(\theta_d, \sigma_u^2\right)$, $d = 1,...,D$. $\theta = (\theta_1,...\ \theta_d) \in \mathbf{C}(\mathbf{x})$.

Level 2 links the true small area means μ_d to the auxiliary variable by using $\mathbf{C}(\mathbf{x})$, a region in $\Re^D$ that defines the relationship between $\mathbf{x}$ and $\overline{\mathbf{Y}}$. As usual in this framework, we assume σ_d^2 is known.

An important particular case is the Fay-Herriot model where $\mathbf{C}(\mathbf{x}) = \mathbf{L}(\mathbf{x}) = \left\{\theta \in \Re^D/\theta = \alpha + \beta'\mathbf{x}\right\}$, is the linear subspace that generates $\mathbf{x}$. Besides, when $\mathbf{C}(\mathbf{x}) = \left\{\theta \in \Re^D/\theta_i = \theta_{i+1}\right\}$ the James-Stein estimator arises (Rao [7]).

In this paper we consider the following isotonic models:

The $C-SimpleOrder$ model: define, from the auxiliary information, a new variable $\mathbf{z} = g(\mathbf{x})$ and assume that the responses are ordered according to

$\mathbf{z}$ ($z_i = z_{(i)}$). Now, let $\mathbf{C}(\mathbf{x}) = \mathbf{C}_0(\mathbf{z})$ be the cone that defines the order induced by $\mathbf{z}$. As the areas are ordered according to $\mathbf{z}$ we have that: $\mathbf{C}_0(\mathbf{z})$ $=\left\{\theta \in \Re^D / \theta_i \leq \theta_{i+1}\right\}$, which is usually known as the simple order cone. This latter cone represents the a priori idea that $\overline{\mathbf{Y}}$ increases with $\mathbf{z}$, and then that the auxiliary information is used to derive an order between the area parameters: $\theta_1 \leq ... \leq \theta_D$.

The $C-Additive$ model: define $\mathbf{C}(\mathbf{x}) = \mathbf{C}_A(\mathbf{x}) = \mathbf{C}_0(x_1) + ... + \mathbf{C}_0(x_p)$ where $\mathbf{C}_0(x_i)$ is the simple order cone that generates the auxiliary x_i. In this case the shrinkage region increases with respect to that defined by $\mathbf{C}_0(z)$.

The $C-Intersection$ model: consider the intersection cone $\mathbf{C}(\mathbf{x}) = \mathbf{C}_I(\mathbf{x}) =$ $\mathbf{C}_0(x_1) \cap ... \cap \mathbf{C}_0(x_p)$, which causes a reduction of dimensionality of the shrinkage region compared with the other alternatives.

Although we will only consider the models with shrinkage regions determine by the cones defined above in this paper, alternative models with more complex shrinkage regions could be defined, depending on the application at hand. For instance, a model with a linear component plus an isotonic component, a model with isotonic components defined by the levels of a categorical predictor, or models with shrinkage regions that define other shape restrictions different from the monotone one, etc...

2.1 The ORI Estimators

Some mathematics (see [9]) give, for the isotonic models defined above, the empirical maximum likelihood predictor for the area means (ORI predictors). The ORI predictors have a similar expression to the Fay-Herriot predictors as follows,

$$\widehat{\mu_d}^C = \left(1 - \frac{\widetilde{\sigma_u}^{2,C}}{\sigma_d^2 + \widetilde{\sigma_u}^{2,C}}\right)\widehat{\theta_d}^C + \frac{\widetilde{\sigma_u}^{2,C}}{\sigma_d^2 + \widetilde{\sigma_u}^{2,C}} y_d, \quad d = 1,2,....,D \qquad (1)$$

where $\widehat{\theta}^C = \underset{\theta \in \mathbf{C}(\mathbf{x})}{arg\ min} \sum_{d=1}^{D} \frac{(y_d - \theta_d)^2}{\sigma_d^2 + \sigma_u^2}$ is the projection of y onto $\mathbf{C}(\mathbf{x})$ (see [8]), and $\widetilde{\sigma_u}^{2,C}$ is an estimator of the random effect variance σ_u^2 that depends on $\widehat{\theta}^C$ and is obtained using an iterative algorithm ([9]). In the simple case of $\sigma_d = \sigma$, $d = 1,...,D$, the estimator is given by: $\widetilde{\sigma_u}^{2,C} = max(0, \widehat{\sigma_u}^{2,C})$ where,

$$\widehat{\sigma_u^2}^C = \frac{\sum_{d=1}^{D} \left(y_d - \widehat{\theta}_d^C\right)^2}{D - l} - \sigma^2$$

where $l = dim(\mathbf{L}), \mathbf{L}$ the linear subspace such as $P(\mathbf{y}|\mathbf{L}) = P(\mathbf{y}|\mathbf{C})$. In particular, for the isotonic models defined above:

$\mathbf{C}-SimpleOrder$: $\widehat{\theta}^C = P(\mathbf{y}|\mathbf{C}_0(\mathbf{z}))$, $l = \dim \mathbf{L}$.

$\mathbf{C}-Additive$: $\widehat{\theta}^C$ is the additive regression: $\widehat{\theta}^C = \widehat{\theta}_1 + \ldots + \widehat{\theta}_p, \widehat{\theta}_i \in \mathbf{L_i}$, $l = \dim(\mathbf{L_1} + \ldots + \mathbf{L_p})$. ([6]).

$\mathbf{C}-Intersection$: $\widehat{\theta}^C = P(\mathbf{y}|\mathbf{C}_I(\mathbf{x}))$, $l = \dim \mathbf{L}$.

To compare the performance of the estimators we use the mean squared prediction error (MSPE) : $E(\widehat{\mu} - \mu)^2$.

2.2 The Bootstrap Approach

We follow a similar parametric bootstrap approach as the one in Chatterjee et al. ([2]) and Hall and Maiti ([5]).

Consider $\widehat{\theta}^C = P(\mathbf{y}/\mathbf{C}(\mathbf{x}))$ and calculate $\widetilde{\sigma}_u^{2,C}$ using the definitions above, with σ^2 known. A bootstrap sample is given by $y_d^* = \widehat{\theta}_d^C + u_d^* + e_d^*$, where u_d^* and e_d^* are values generated from independent $N(0, \widetilde{\sigma}_u^{2,C})$ and $N(0, \sigma_d^2)$, respectively. Now, obtaining $\widehat{\theta}^{C*}$ and $\widetilde{\sigma}_u^{2,C*}$ as before, but applied to $\mathbf{y}^*$, we get the bootstrap empirical predictor $\widehat{\mu}^{C*}$ using Eq 1. By selecting B bootstrap samples, the bootstrap estimator for the MSPE in the area d is defined by $\widehat{MSPE}_{d,B}^{C} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\mu}_{d,b}^{C*} - \widehat{\theta}_d^C - u_d^*)$. Also, we define $\widehat{MSPE}_{B}^{C} = \sum_{d=1}^{18} \widehat{MSPE}_{d,B}^{C}$.

3 Analyzing Simulated Data

To evaluate the behavior of the ORI estimators defined in Section 2 we have conducted two simulation experiments. In the first one, we generate an artificial auxiliary variable. For the second one, we have selected two auxiliaries from the Australian farm data ([1]) to generate the data. In all the scenarios, for simplicity we consider the equal variance cases, $\sigma_d^2 = \sigma^2 = 1$, and $\sigma_u^2 = 1$, and we simulate three different models for standardized θ and three different values of $\|\theta\|$. Scenarios with high values of $\|\theta\|$ correspond to high inter-area variation and represent situations where the area means are far from the total mean. In these cases, the synthetic estimator, the sample mean, is expected to perform badly as an estimator for the individual areas, it is also expected that the James-Stein approach will not be a good choice. These scenarios represent the situations where more sophisticated estimation methods, which use auxiliary information, should be used because they serve a purpose.

Besides the James-Stein, $\widehat{\mu}^{JS}$, and the Fay-Herriot estimator, $\widehat{\mu}^{FH}$, we have also calculated two ORI estimators, the estimator that uses the additive cone, $\widehat{\mu}^{C_A}$, and the estimator that uses the intersection cone, $\widehat{\mu}^{C_I}$. For the first experiment, with only one auxiliary, these latter two cones are equal and we refer to it as C.

3.1 One Auxiliary

In order to simplify the experiment, we get $(x_1,...,x_{30})=(1,...,30)$, for which: $\mathbf{C}=\left\{\theta\in\Re^{30}/\theta_1\leq...\leq\theta_{30}\right\}$ and $\mathbf{L}(\mathbf{x})=\{\theta\in\Re^{30}/\theta_d-\theta_{d-1}=\theta_{d+1}-\theta_d\}$. We also select different standardized θ values verifying: $\theta=\theta^{\mathbf{0}}\in\mathbf{L}(\mathbf{x})$ (the Fay-Herriot model is true), $\theta=\theta^{\mathbf{c1}},\theta=\theta^{\mathbf{c2}}\in\mathbf{C}-\ \mathbf{L}(\mathbf{x})$ (the isotonic model is true but the Fay-Herriot model is not true) and $\theta=\theta^{\mathbf{x}}\notin\mathbf{C}$ (neither the Fay-Herriot nor the isotonic model are true). We compare, in Table 1, the MSE of $\widetilde{\sigma_u}^{2,C}$ and $\widetilde{\sigma_u}^{2,FH}$, and the MSPE of $\widehat{\mu}^C$ with $\widehat{\mu}^{FH}$ and $\widehat{\mu}^{JS}$. For each scenario, Table 1 also gives the correlation, $\rho(\theta,x)$.

Table 1 MSE for $\widetilde{\sigma_u}^{2,C}$ and $\widetilde{\sigma_u}^{2,FH}$ and MSPE for $\widehat{\mu}^C$, $\widehat{\mu}^{FH}$ and $\widehat{\mu}^{JS}$

			MSE		MSPE		
θ	$\lVert\theta\rVert$	$\rho(\theta,x)$	$\widetilde{\sigma_u}^{2,C}$	$\widetilde{\sigma_u}^{2,FH}$	$\widehat{\mu}^C$	$\widehat{\mu}^{FH}$	$\widehat{\mu}^{JS}$
θ^0	low	1.00	0.299	**0.278**	0.611	0.574	**0.566**
θ^0	moderate	1.00	0.412	**0.278**	0.664	**0.574**	0.916
θ^0	high	1.00	0.493	**0.278**	0.693	**0.574**	0.967
θ^{c1}	low	0.95	0.300	**0.278**	0.611	0.574	**0.566**
θ^{c1}	moderate	0.95	**0.369**	0.963	**0.662**	0.694	0.917
θ^{c1}	high	0.95	**0.404**	4.862	**0.684**	0.800	0.968
θ^{c2}	low	0.77	0.297	**0.278**	0.611	0.576	**0.565**
θ^{c2}	moderate	0.77	**0.347**	12.372	**0.679**	0.842	0.913
θ^{c2}	high	0.77	**0.399**	89.006	**0.721**	0.926	0.965
θ^x	low	0.61	0.298	**0.279**	0.612	0.577	**0.564**
θ^x	moderate	0.61	**2.1350**	26.732	**0.799**	0.881	0.912
θ^x	high	0.61	**16.088**	231.205	**0.902**	0.949	0.965

3.2 Two Auxiliaries

In this experiment we simulate the model: $y_d=\theta_d+u_d+e_d$, $d=1,2,\ldots,29$, $\theta_d=f_1(x_d)+f_2(x_d)$.

We define three scenarios from different functions f_1 and f_2. $S1:f_1(x)=\log(x_1)$, $f_2(x)=\log(x_2)$; $S2:f_1(x)=x_1^{1/2}$, $f_2(x)=x_2^5$; $S3:f_1(x)=-\sin(x_1/3)$, $f_2(x)=sin(x_1/4+x_2/4)$.

We include, in Table 2, the MSPE of $\widehat{\mu}^M$, $M=C_A,C_I,FH,JS$. For each scenario, Table 2 also gives the correlations $\rho(\theta,x)$.

3.3 Conclusions

Table 1 shows that the MSE is clearly smaller for $\widetilde{\sigma_u}^{2,FH}$ than for $\widetilde{\sigma_u}^{2,C}$, only for θ^0. Also, from the figures in Tables 1 and 2, it is clear that, in most cases,

Table 2 MSPE for $\widehat{\mu}^M, M = C_A, C_I, FH, JS$

Scenario	$\lVert\theta\rVert$	$\rho(\theta,\mathbf{x})$	$\widehat{\mu}^{C_A}$	$\widehat{\mu}^{C_I}$	$\widehat{\mu}^{FH}$	$\widehat{\mu}^{JS}$
$S1$	low	0.93	0.684	0.724	**0.651**	0.709
$S1$	moderate	0.93	**0.743**	0.921	0.839	0.924
$S1$	high	0.93	**0.780**	0.970	0.927	0.931
$S2$	low	0.74	0.690	**0.654**	0.708	0.707
$S2$	moderate	0.74	**0.737**	0.835	0.917	0.926
$S2$	high	0.74	**0.761**	0.924	0.970	0.973
$S3$	low	0.43	0.739	**0.677**	0.734	0.711
$S3$	moderate	0.43	0.914	**0.877**	0.937	0.931
$S3$	high	0.43	0.972	**0.953**	0.980	0.977

the ORI estimators for the area means have smaller MSPE than the FH or the JS estimators.

Specifically, for scenarios where the inter-area variability is low, the best estimators are those associated with the more restrictive shrinkage regions: $\widehat{\mu}^{JS}$ and $\widehat{\mu}^{C_I}$. For scenarios with moderate or high inter-area variability, Fay-Herriot is clearly better only when $\rho(\theta,x) = 1$. However, in scenarios corresponding to $\rho(\theta,x) < 1$, the ORI estimators are better; $\widehat{\mu}^{C_I}$ being the best choice when $\rho(\theta,x) < 0.5$, and $\widehat{\mu}^{C_A}$ the one with the smallest MSPE values, when $0.5 < \rho(\theta,x) < 1$.

4 The Baseball Data Set Revisited

We now revisit the baseball data example given in Efron and Morris ([3]) and used by many other authors.

For the player d $(d = 1,...,18)$, let p_d and π_d be the batting average for the first 45 'at-bat' and the true season batting average of the 1970 season respectively. Consider, also used by other authors, the arc-sine transformation: $y_d = \sqrt{n}\arcsin(2p_d - 1)$, $\mu_d = \sqrt{n}\arcsin(2\pi_d - 1)$. We use as an auxiliary information, $\mathbf{x}$, the previous 'at-bat', and consider the players ordered using $\mathbf{x}$. Assume the following model: $y_d \sim N(\mu_d, 1)$, $\mu_d \sim N\left(\theta_d, \sigma_u^2\right)$, $\theta_d = f(\mathbf{x})$.

In order to find a plausible θ_d, we fit a polynomial regression of μ against $\mathbf{x}$. $\sigma_u^2 = 0.5$ is also obtained from this fitted model.

We have simulated the above model (100 iterations). The results of the MSPE for the different estimators of μ appear in Table 3. Figures in the first column show that the James-Stein estimator, $\widehat{\mu}^{JS}$, is the best and also that the auxiliary information based estimators, $\widehat{\mu}^{C}$ and $\widehat{\mu}^{FH}$, perform very well, comparing the MSPE values with those of the direct estimator. This is due to the relatively low inter-area variance. The second column of Table 3 compares the estimators using only the observed $\mathbf{y}$, instead of several repetitions of the

Table 3 Baseball example: MSE for estimators with different methods

Method	model MSE	Observed data MSE
ORI	9.60	5.81
Fay − Herriot	8.64	5.44
James − Stein	8.53	4.45
Direct	17.91	13.71

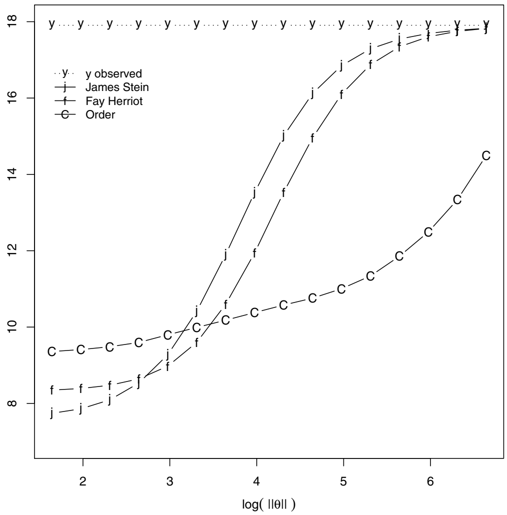


Fig. 1 *Baseball example: MSPE of the predictors* $\widehat{\mu}^{FH}$, $\widehat{\mu}^{JS}$, $\widehat{\mu}^{C}$ *and* $\mathbf{y}$.

model; the conclusions are similar to those in column 1. The good behavior of the James-Stein estimator shown in Table 3 implies that the auxiliary information is not useful in this particular case.

Figure 1 displays the MSPE of the new estimator $\widehat{\mu}^{C}$ along with that of $\widehat{\mu}^{FH}$, $\widehat{\mu}^{JS}$ and the direct estimator $\mathbf{y}$ for different scenarios, using the model above for different values of $\boldsymbol{\theta} = \lambda\boldsymbol{\theta}^0$, $log(\lambda)$ ranging from -1 to 4. From Fig. 1, we conclude that when the inter-area variance increases, $\widehat{\mu}^{JS}$ is no longer the best estimator and that in these cases, the new estimator, $\widehat{\mu}^{C}$, is the one with the best behavior and the estimator that uses the auxiliary information more efficiently.

The bootstrap approach introduced in Section 2.2 was used with this data. In order to validate the bootstrap estimation, we have included a summary, in Table 4, of the results from 100 simulated samples, $B = 1000$, and different values of $\|\boldsymbol{\theta}\|$. The mean values of $\widehat{MSPE_B}$ for the three estimators replicate the values in Fig. 1, demonstrating the good performance of the proposed

Table 4 Baseball example: Mean values of bootstrap estimates, $\widehat{MSPE}_B$, from 100 simulated samples under different values of $\|\boldsymbol{\theta}\|$.

	$log(\|\boldsymbol{\theta}\|)$				
Method	2	3	4	5	6
ORI	8.47	8.66	9.74	10.20	11.87
Fay − Herriot	7.06	7.50	10.88	16.05	17.65
James − Stein	7.10	8.56	13.71	17.14	17.84
Direct	17.99	18.00	17.99	18.00	17.97

bootstrap approach, which could be adopted to select the best estimator in a given scenario and to estimate the *MSPEs*.

Acknowledgements. This research was partially supported by Spanish DGES (grant MTM 2009-11161).

References

1. Chambers, R., Tzavidis, N.: M-quantile models for small area estimation. Biometrika 93(2), 255–268 (2006)
2. Chatterjee, S., Lahiri, P., Li, H.: Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. Ann. Statist. 36(3), 1221–1245 (2008)
3. Efron, B., Morris, C.N.: Data analysis using Stein's estimator and its generalizations. J. Amer. Statistical Assoc. 70, 311–319 (1975)
4. Fay, R.E., Herriot, R.A.: Estimates of income for small places: an application of James-Stein procedures to census data. J. Amer. Statistical Assoc. 74, 341–353 (1979)
5. Hall, P., Maiti, T.: On parametric bootstrap methods for small area prediction. J. Roy. Statist. Soc. Ser. B 68(2), 221–238 (2006)
6. Mammen, E., Yu, K.: Additive isotone regression. IMS Lecture Notes–Monograph Series Asymptotics: Particles, Processes and Inverse Problems 55, 179–195 (2007)
7. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)
8. Robertson, T., Wright, F.T., Dykstra, R.L.: Order Restricted Statistical Inference. Wiley, Chichester (1988)
9. Rueda, C., Menéndez, J.A., Gómez, F.: Small area estimators based on restricted mixed models (submitted for publication, 2010) doi:10.1007/s117489-010-0186- 2

Set-Valued Stochastic Processes and Sets of Probability Measures Induced by Stochastic Differential Equations with Random Set Parameters

Bernhard Schmelzer

Abstract. We consider stochastic differential equations depending on parameters whose uncertainty is modeled by random compact sets. Several approaches are discussed how to construct set-valued processes from the solutions. The induced lower and upper probabilities are compared to a set of probability measures constructed from the distributions of the solutions and the selections of the random set.

Keywords: Stochastic differential equation, Random set, Parameter uncertainty, Set-valued stochastic process, Parameterized probability measures.

1 Introduction

We consider stochastic differential equations (SDEs) whose initial value and coefficients depend on random set parameters. More precisely, we study solutions of SDEs of the integral form

$$x_{t,a} = x_{t_0,a} + \int_{t_0}^{t} f(s,a,x_{s,a})\,\mathrm{d}s + \int_{t_0}^{t} G(s,a,x_{s,a})\,\mathrm{d}b_s \tag{1}$$

where time t ranges within some finite interval $[t_0,\bar{t}]$, the (deterministic) parameter a takes values in some set $\mathbb{A} \subseteq \mathbb{R}^p$ and b denotes an m-dimensional Brownian motion (Wiener process) on a probability space $(\Omega_b, \Sigma_b, P_b)$. The initial value x_{t_0} and the coefficients f and G are maps of the form

$$\begin{aligned} x_{t_0} &: \mathbb{A} \times \Omega_b \to \mathbb{R}^d, & (a,\omega_b) &\mapsto x_{t_0,a}(\omega_b), \\ f &: [t_0,\bar{t}] \times \mathbb{A} \times \mathbb{R}^d \to \mathbb{R}^d, & (t,a,x) &\mapsto f(t,a,x), \\ G &: [t_0,\bar{t}] \times \mathbb{A} \times \mathbb{R}^d \to \mathbb{R}^{d\times m}, & (t,a,x) &\mapsto G(t,a,x). \end{aligned}$$

Bernhard Schmelzer
Unit for Engineering Mathematics, University of Innsbruck,
A-6020 Innsbruck, Austria
e-mail: bernhard.schmelzer@uibk.ac.at

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 561–568.
springerlink.com

If for each a the conditions for the existence and uniqueness of a solution (see [9, Th. 1, 3]) are fulfilled this leads to a family of stochastic processes which can be summarized by the following map

$$x : [t_0, \bar{t}] \times \mathbb{A} \times \Omega_b \to \mathbb{R}^d, (t, a, \omega_b) \mapsto x_{t,a}(\omega_b). \tag{2}$$

In [18, Prop. 2, 3] conditions have been given under which the map x is a continuous and thus measurable stochastic process, that is, for each $\omega_b \in \Omega_b$ the sample path $x(\omega_b)$ is continuous on $[t_0, \bar{t}] \times \mathbb{A}$ and x is measurable with respect to the product σ-algebra $\mathscr{B}([t_0, \bar{t}]) \otimes \mathscr{B}(\mathbb{A}) \otimes \Sigma_b$. (Note that $\mathscr{B}(\mathbb{A}) = \mathscr{B}(\mathbb{R}^p)|_{\mathbb{A}}$.) Throughout the paper we will assume that x is continuous and measurable. As a model for the parameter uncertainty of a we use a random compact set

$$A : \Omega_{\mathbb{A}} \to \mathscr{K}'(\mathbb{A}), \omega_{\mathbb{A}} \mapsto A(\omega_{\mathbb{A}})$$

on the probability space $(\Omega_{\mathbb{A}}, \Sigma_{\mathbb{A}}, P_{\mathbb{A}})$ where $\mathscr{K}'(\mathbb{A})$ denotes the set of non-empty compact subsets of $\mathbb{A}$. Since $\mathscr{K}'(\mathbb{A}) \subseteq \mathscr{K}'(\mathbb{R}^p)$ (i.e. A is also a random compact set in $\mathbb{R}^p$) this means that for each Borel set $B \in \mathscr{B}(\mathbb{R}^p)$ its upper inverse defined by $A^-(B) = \{\omega_{\mathbb{A}} : A(\omega_{\mathbb{A}}) \cap B \neq \emptyset\}$ lies in $\Sigma_{\mathbb{A}}$ ([17]). It is a well-known fact ([2, 11, 17, 16]) that A has measurable selections, that is, there exist measurable functions $\alpha : \Omega_{\mathbb{A}} \to \mathbb{A}$ such that $\alpha(\omega_{\mathbb{A}}) \in A(\omega_{\mathbb{A}})$ for all $\omega_{\mathbb{A}} \in \Omega_{\mathbb{A}}$. Denoting by $\mathscr{S}(A)$ the set of all measurable selections of A one can even find a Castaing representation of A, that is, a sequence $\{\alpha_n\}_{n \in \mathbb{N}} \subseteq \mathscr{S}(A)$ such that $\mathrm{cl}(\{\alpha_n(\omega\mathbb{A})\}_{n \in \mathbb{N}}) = A(\omega_{\mathbb{A}})$ for all $\omega_{\mathbb{A}} \in \Omega_{\mathbb{A}}$.

Let (Ω, Σ, P) denote the product space $(\Omega_{\mathbb{A}} \times \Omega_b, \Sigma_{\mathbb{A}} \otimes \Sigma_b, P_{\mathbb{A}} \otimes P_b)$. The basic idea in [18] was to define a set-valued map

$$X : (t, \omega) \mapsto X_t(\omega) = \{x_{t,a}(\omega_b) : a \in A(\omega_{\mathbb{A}})\} \tag{3}$$

by merging at each time and for each sample path of the Brownian motion all values of the solutions. Furthermore, for $\alpha \in \mathscr{S}(A)$ one can consider

$$\xi^{\alpha} : [t_0, \bar{t}] \times \Omega \to \mathbb{R}^d, (t, \omega) \mapsto x_{t, \alpha(\omega_{\mathbb{A}})}(\omega_b) \tag{4}$$

which is a measurable and continuous stochastic process. We cite [18, Prop. 4] which is proved by using the selections ξ^{α}.

Proposition 1. *The map X is a continuous and measurable set-valued stochastic process whose values are non-empty compact subsets of $\mathbb{R}^d$. More precisely, $X^-(G) \in \Sigma$ for each open set $G \subseteq \mathbb{R}^d$ and $X^-(B) \in \Sigma^0$ for each Borel set $B \in \mathscr{B}(\mathbb{R}^d)$ where Σ^0 denotes the completion of Σ with respect to P. For each $\omega \in \Omega$ the sample function $X(\omega)$ is continuous with respect to the Hausdorff-metric on $\mathscr{K}'(\mathbb{R}^d)$. Furthermore, for a Castaing representation $\{\alpha_n\}_{n \in \mathbb{N}}$ of A the processes $\{\xi^{\alpha_n}\}_{n \in \mathbb{N}}$ form a Castaing representation of X and for each $t \in [t_0, \bar{t}]$ the family $\{\xi_t^{\alpha_n}\}_{n \in \mathbb{N}}$ forms a Castaing representation of $X_t = X_t(\cdot)$.*

As an example let us consider the trivial SDE

$$\mathrm{d}x_{t,a} = a\,\mathrm{d}b_t, \quad x_0 = 0 \tag{5}$$

whose solution is clearly $x_{t,a} = ab_t$. Let $\mathbb{A} = [1,2]$ and $A = \mathbb{A}$ be deterministic ($\Omega_{\mathbb{A}}$ exists of only one element). Then Equation (3) ($\Omega_{\mathbb{A}}$ can be omitted) implies

$$X : (t, \omega_b) \mapsto \{x_{t,a}(\omega_b) : a \in \mathbb{A}\} = \begin{cases} [b_t(\omega_b), 2b_t(\omega_b)] & \text{if } b_t(\omega_b) > 0, \\ \{0\} & \text{if } b_t(\omega_b) = 0, \\ [2b_t(\omega_b), b_t(\omega_b)] & \text{if } b_t(\omega_b) < 0. \end{cases} \tag{6}$$

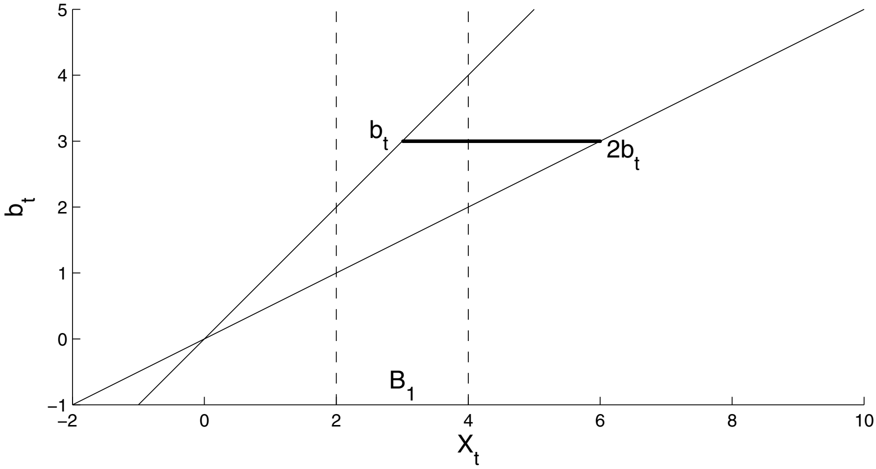


Fig. 1 Bounds of the focal sets X_t (thin solid lines), together with a special focal set (thick solid line) and the interval B_1 (dotted bounds).

From Figure 1 one can see that the lower and upper probability ([5]) of $B_1 = (2,4)$, i.e. the probabilities of $\{\omega_b : X_t(\omega_b) \subseteq B_1\}$ and $X_t^-(B_1)$, are computed as:

$$\underline{P}(B_1) = P_b(\{\omega_b : X_t(\omega_b) \subseteq B_1\}) = P_b(\{\omega_b : 2 < b_t(\omega_b) < 2\}),$$
$$\overline{P}(B_1) = P_b(\{\omega_b : X_t(\omega_b) \cap B_1 \neq \emptyset\}) = P_b(\{\omega_b : 1 < b_t(\omega_b) < 4\}).$$

Note that b_t is normally distributed with mean zero and variance t which leads to the values $\underline{P}(B_1) = 0$ and $\overline{P}(B_1) = 0.286$ at time $t = 4$. Since in this simple example the distributions of the random variables $x_{t,a}$ are known one might think of computing for each a the probability of $x_{t,a}$ lying in B_1 and taking the smallest and the greatest values instead of $\underline{P}$ and $\overline{P}$. If we do so we obtain as extreme values 0.136 and 0.161 which are much tighter bounds than $\underline{P}(B_1)$ and $\overline{P}(B_1)$.

The purpose of this paper is to find out the reason for these quite different results. In Section 2 alternative definitions of set-valued processes induced by (1) are proposed and compared to (3). In Section 3, an approach is presented how to use the distributions of the solution processes to compute probability bounds for certain events. Furthermore, the discrepancy between the two approaches is clarified theoretically and illustrated by continuing the above example.

2 Alternative Definitions of Set-Valued Processes

From Equation (4) one can see that for each $\alpha \in \mathscr{S}(A)$ the process ξ^{α} consists of sample functions that are sample paths of solutions of the parameterized SDE (1) although the corresponding value of the parameter varies with ω. But note that the set-valued process X encloses as selections many other stochastic processes that are not directly related to solutions or their sample paths. Instead of x it is thus reasonable to consider the map

$$\tilde{x} : \mathbb{A} \times \Omega_b \to \mathscr{C}([t_0,\bar{t}]), (a,\omega_b) \mapsto x_{\cdot,a}(\omega_b) \tag{7}$$

where $\mathscr{C}([t_0,\bar{t}])$ together with $\|\cdot\|_\infty$ (defined by $\|f\|_\infty = \sup_{t\in[t_0,\bar{t}]} \|f(t)\|$) denotes the separable Banach space of continuous functions from $[t_0,\bar{t}]$ to $\mathbb{R}^d$. Hence, the map $\tilde{x}$ assigns to each parameter value a and each ω_b the whole solution path. We assume that $\tilde{x}$ consists of continuous sample functions and is thus $\mathscr{B}(\mathbb{A}) \otimes \Sigma_b$-measurable (which is a consequence of the continuity of x assumed in the introduction). We can now proceed as in the foregoing section which means that we define a set-valued map by

$$\tilde{X} : \omega \mapsto \{x_{\cdot,a}(\omega_b) : a \in A(\omega_{\mathbb{A}})\} \tag{8}$$

and that for $\alpha \in \mathscr{S}(A)$ we consider the $\mathscr{C}([t_0,\bar{t}])$-valued random variable

$$\tilde{\xi}^{\alpha} : \Omega \to \mathscr{C}([t_0,\bar{t}]), \omega \mapsto x_{\cdot,\alpha(\omega_{\mathbb{A}})}(\omega_b).$$

We can state the following proposition which is proved in a similar manner as Proposition 1.

Proposition 2. *The map $\tilde{X}$ is a random set on the completed probability space (Ω,Σ^0,P^0) whose values are non-empty compact subsets of $\mathscr{C}([t_0,\bar{t}])$. More precisely, $\tilde{X}^-(G) \in \Sigma$ for each open set $G \subseteq \mathscr{C}([t_0,\bar{t}])$ and $\tilde{X}^-(B) \in \Sigma^0$ for each Borel set $B \in \mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$. For a Castaing representation $\{\alpha_n\}_{n\in\mathbb{N}}$ of A the processes $\{\tilde{\xi}^{\alpha_n}\}_{n\in\mathbb{N}}$ form a Castaing representation of $\tilde{X}$.*

Note that every selection of $\tilde{X}$ consists of sample functions of solutions of the SDE (1), indeed let $\tilde{\xi} \in \mathscr{S}(\tilde{X})$ which means that for all $\omega \in \Omega$ we have $\tilde{\xi}(\omega) \in \tilde{X}(\omega) = \{x_{\cdot,a}(\omega_b) : a \in A(\omega_{\mathbb{A}})\}$. Hence, for each $\omega \in \Omega$ there exists an $a \in A(\omega_{\mathbb{A}})$ such that $\tilde{\xi}(\omega) = x_{\cdot,a}(\omega_b)$. Thus, $\tilde{X}$ seems to be conceptually more appropriate then X.

To establish the relation between $\tilde{X}$ and X we consider the evaluation map $\pi_t : \mathscr{C}([t_0,\bar{t}]) \to \mathbb{R}^d, f \mapsto f(t)$ that assigns for fixed $t \in [t_0,\bar{t}]$ to each continuous function its value at time t. Obviously, π_t is continuous. Applying π_t to the random set $\tilde{X}$ leads to the random set X_t which means that in view of the example given in the introduction $\tilde{X}$ leads to the same results as X:

$$\pi_t(\tilde{X}(\omega)) = \{\pi_t(x_{\cdot,a}(\omega_b)) : a \in A(\omega_{\mathbb{A}})\} = \{x_{t,a}(\omega_b) : a \in A(\omega_{\mathbb{A}})\} = X_t(\omega)$$

We will now present a third possibility for defining a set-valued process which has also been used in [12, 13, 14] and is based on the following theorem ([10, Theorem 3.1]).

Theorem 1. *Let $(\mathbb{M},\mathscr{M},\mu)$ be a complete σ-finite measure space, let $\mathbb{E}$ be a separable Banach space and let S be a non-empty closed subset of $L^p(\mathbb{M};\mathbb{E}) = \{f : \mathbb{M} \to \mathbb{E} : \int_{\mathbb{M}} \|f\|^p \mathrm{d}\mu < \infty\}$, $1 \le p < \infty$. Then there is a random closed set Y on $\mathbb{M}$ such that S equals the set $\mathscr{S}^p(Y)$ of p-integrable selections of Y if and only if S is decomposable which means that for each finite partition $\{C_1,\dots,C_n\} \subseteq \mathscr{M}$ of $\mathbb{M}$ and $\{f_1,\dots,f_n\} \subseteq S$ it holds that $\sum_{i=1}^n f_i \mathbb{1}_{C_i} \in S$.*

In [18, Prop. 6] it has been proved that under certain conditions $\xi^\alpha \in \mathscr{S}^2(X)$ for any bounded $\alpha \in \mathscr{S}(A)$. (These conditions are also assumed in [18, Prop. 2] implying the continuity of x.) We assume that these conditions are fulfilled and consider $S_0 = \{\xi^\alpha : \alpha \in \mathscr{S}(A) \text{ bounded}\} \subseteq \mathscr{S}^2(X)$. By $\mathsf{decl}(S_0)$ we denote the smallest closed and decomposable subset of $L^2([t_0,\bar{t}] \times \Omega; \mathbb{R}^d)$ containing S_0 (see [12, 13, 14]). One can infer that there exists a random closed set $\hat{X}$ on the completion of the product space $([t_0,\bar{t}] \times \Omega, \mathscr{B}([t_0,\bar{t}]) \otimes \Sigma^0, \lambda \otimes P^0)$ such that $\mathsf{decl}(S_0) = \mathscr{S}^2(\hat{X})$ (λ denotes the Lebesgue measure on $\mathscr{B}([t_0,\bar{t}])$). The next proposition shows that X has almost surely the same values as $\hat{X}$ which can be interpreted as the smallest set-valued process such that S_0 is a set of square-integrable selections.

Proposition 3. *For almost all $(t,\omega) \in [t_0,\bar{t}] \times \Omega$ it holds that $X_t(\omega) = \hat{X}_t(\omega)$.*

Proof. As argued above it holds that $S_0 \subseteq \mathscr{S}^2(X)$ and consequently (by the above theorem) $\mathscr{S}^2(\hat{X}) = \mathsf{decl}(S_0) \subseteq \mathscr{S}^2(X)$ which implies that $\hat{X}_t(\omega) \subseteq X_t(\omega)$ for almost all (t,ω) because of [10, Lemma 1.1]. From the same lemma we follow that there is a Castaing representation $\{\alpha_n\}_{n\in\mathbb{N}}$ consisting of bounded selections of A. Since $\xi^{\alpha_n} \in S_0 \subseteq \mathsf{decl}(S_0) = \mathscr{S}^2(\hat{X})$ one can conclude that by [10, Lemma 1.1] $\xi_t^{\alpha_n}(\omega) \in \hat{X}_t(\omega)$ almost surely. Since $X_t(\omega) = \mathrm{cl}(\{\xi_t^{\alpha_n}(\omega)\}_{n\in\mathbb{N}})$ by Proposition 1 and $\hat{X}_t(\omega)$ is closed one obtains $X_t(\omega) \subseteq \hat{X}_t(\omega)$ for almost all (t,ω). □

3 Measures Induced by Parameterized Stochastic Processes

In this section we will mainly consider the map $\tilde{x}$ defined by Equation (7) and the corresponding distributions on $\mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$, more precisely, for $a \in \mathbb{A}$ we define

$$p_a : \mathscr{B}(\mathscr{C}([t_0,\bar{t}])) \to [0,1], B \mapsto P_b \circ \tilde{x}_a^{-1}(B) = P_b(\{\omega_b : x_{\cdot,a}(\omega_b) \in B\}).$$

Again, we use the random compact set A to model the parameter uncertainty of a which induces different sets of probability measures on $\mathscr{B}(\mathbb{A})$ (see [7, 16, 15, 3]), for example the set $\mathfrak{S}(A) = \{P_{\mathbb{A}} \circ \alpha^{-1} : \alpha \in \mathscr{S}(A)\}$ of distributions of measurable selections of A. Following [7, 8] we consider

$$\mathfrak{P} = \left\{ \int_{\mathbb{A}} p_a \,\mathrm{d}\nu(a) : \nu \in \mathfrak{S}(A) \right\} \tag{9}$$

which consists of probability measures on $\mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$ under the assumption that the map $a \mapsto p_a(B)$ is measurable for any $B \in \mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$. In view of [9, Th. I.7.3] and [1, Th. 2.1] it turns out that this measurability holds under weaker conditions than those ([18, Prop. 2]) implying the continuity of x or $\tilde{x}$.

From [16, Th. 8, 14] or [3, Th. 1, Prop. 3] one can deduce the following proposition concerning the set $\mathfrak{P}(B)$ of values of $\mathfrak{P}$ for $B \in \mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$.

Proposition 4. *Let $B \in \mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$ such that $a \mapsto p_a(B)$ is measurable. Then the following two equalities hold*

$$\int_{\mathbb{A}} p_a(B)\,\mathrm{d}\underline{P}_{\mathbb{A}}(a) = \inf \mathfrak{P}(B), \quad \int_{\mathbb{A}} p_a(B)\,\mathrm{d}\overline{P}_{\mathbb{A}}(a) = \sup \mathfrak{P}(B)$$

where the left-hand integrals are the Choquet-integrals ([4, 6]) with respect to the lower and upper probability induced by the random set A. If the probability space $(\Omega_{\mathbb{A}}, \Sigma_{\mathbb{A}}, P_{\mathbb{A}})$ is non-atomic then the set $\mathfrak{P}(B)$ is convex.

We will now show that the values of $\mathfrak{P}$ are bounded by the lower and the upper probability induced by the random compact set $\tilde{X}$.

Proposition 5. *Let $B \in \mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$ such that $a \mapsto p_a(B)$ is measurable and assume that $\tilde{x}(\omega_b)$ continuously depends on a for each $\omega_b \in \Omega_b$. Then*

$$\underline{P}(B) \leq \inf \mathfrak{P}(B) \leq \sup \mathfrak{P}(B) \leq \overline{P}(B)$$

where $\underline{P}$ and $\overline{P}$ are the lower and upper probabilities induced by $\tilde{X}$.

Proof. From Prop. 2 we can infer that for an open set $B \subseteq \mathscr{C}([t_0,\bar{t}])$ it holds that $\tilde{X}^-(B) \in \Sigma = \Sigma_{\mathbb{A}} \otimes \Sigma_b$. Hence, for any $\nu = P_{\mathbb{A}} \circ \alpha^{-1} \in \mathfrak{S}(A)$ $(\alpha \in \mathscr{S}(A))$

$$\int_{\mathbb{A}} p_a(B)\,\mathrm{d}\nu(a) = \int_{\Omega_{\mathbb{A}}} P_b(\{\omega_b : \xi^{\alpha}(\omega_{\mathbb{A}}, \omega_b) \in B\})\,\mathrm{d}P_{\mathbb{A}}(\omega_{\mathbb{A}}) \leq$$
$$\leq \int_{\Omega_{\mathbb{A}}} P_b(\{\omega_b : \tilde{X}(\omega_{\mathbb{A}}, \omega_b) \cap B \neq \emptyset\})\,\mathrm{d}P_{\mathbb{A}}(\omega_{\mathbb{A}}) = P_{\mathbb{A}} \otimes P_b(\tilde{X}^-(B)) = \overline{P}(B)$$

since $\tilde{x}_{\alpha(\omega_{\mathbb{A}})}(\omega_b) = \xi^{\alpha}(\omega) \in \tilde{X}(\omega)$ for all $\omega \in \Omega$. By using [3, Prop. 2] this inequality can be extended to $\mathscr{B}(\mathscr{C}([t_0,\bar{t}]))$. The inequality for $\underline{P}$ follows from the duality of $\overline{P}$ and $\underline{P}$. □

Assume that one can find an $\alpha \in \mathscr{S}(A)$ such that $P_b(\{\omega_b : \tilde{\xi}^{\alpha}(\omega_{\mathbb{A}}, \omega_b) \in B\}) = P_b(\{\omega_b : \tilde{X}(\omega_{\mathbb{A}}, \omega_b) \cap B \neq \emptyset\})$ for all $\omega_{\mathbb{A}}$. Then the above proof shows that the equality $\sup \mathfrak{P}(B) = \overline{P}(B)$ holds.

Let us now return to the example given in the introduction. Note that $\mathscr{S}(A) = \mathbb{A}$ and $\mathfrak{S}(A) = \{\delta_a : a \in \mathbb{A}\}$ is a set of point measures. As argued in Section 2 the set-valued process (6) is equivalent to the $\mathscr{K}'(\mathscr{C}([t_0, \bar{t}]))$-valued random set

$$\tilde{X} : \omega_b \to \{ab(\omega_b) : a \in \mathbb{A}\}$$

defined by Equation (8). Let us consider the event $B_2 = (2, \infty)$ which corresponds to $\tilde{B}_2 = \pi_4^{-1}(B_2) \in \mathscr{B}(\mathscr{C}([t_0, \bar{t}]))$. Then $\tilde{X}^{-}(\tilde{B}_2)$ is completely determined by the selection process $\tilde{\xi}^{\alpha} = 2b$ ($\alpha = a = 2$) representing the upper interval bound of $\tilde{X}$. Furthermore, $\{\tilde{X} \subseteq \tilde{B}_2\}$ is completely determined by the selection process $\tilde{\xi}^{\alpha} = b$ ($\alpha = a = 1$) representing the lower interval bound. Hence, we obtain $\overline{P}(\tilde{B}_2) = \sup \mathfrak{P}(\tilde{B}_2) = 0.31$ and $\underline{P}(\tilde{B}_2) = \inf \mathfrak{P}(\tilde{B}_2) = 0.16$. The reason for the different results in the case of B_1 is that the events $X_4^{-}(B_1)$ and $\{X_4 \subseteq B_1\}$ can only be expressed by using both the lower and the upper boundary process.

4 Summary and Conclusions

The aim of this work is to consider solution processes of stochastic differential equations depending on parameters whose uncertainty is modeled by random compact sets. Two different approaches have been presented how to estimate the probability of the solution taking values in a certain set. The first approach uses set-valued processes obtained from the parameterized solutions. After reviewing a pointwise construction two alternative definitions have been proposed. Although the latter might be conceptually more appropriate, it has turned out that the three constructions lead to the same random sets at any fixed time. In the second approach a set of probability measures has been constructed from the distributions of the solution processes and the selections of the parameter random set. It has been shown that probability bounds are not wider than in the first approach. A very simple example has demonstrated that both approaches can lead to quite different results.

There seem to be many applications, for example in earthquake engineering where stochastic processes (related to white noise) can be used to model ground accelerations. Random sets can serve as a robust model for the uncertainty of structural parameters. The reader is referred to [19] for an application of set-valued processes to a problem in earthquake engineering.

References

1. Billingsley, P.: Probability and measure. Wiley, New York (1986)
2. Castaing, C., Valadier, M.: Convex analysis and measurable multifunctions. Springer, Berlin (1977)

3. Castaldo, A., Marinacci, M.: Random correspondences as bundles of random variables. In: De Cooman, G., Fine, T., Seidenfeld, T. (eds.) Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications, ISIPTA 2001, Ithaca, NY, USA, pp. 77–82. Shaker Publishing, Maastricht (2001)
4. Choquet, G.: Theory of capacities. Ann. Inst. Fourier, Grenoble 5(1953-1954), 131–295 (1955)
5. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Statist. 38, 325–339 (1967)
6. Denneberg, D.: Non-additive measure and integral. Kluwer, Dordrecht (1994)
7. Fetz, T., Oberguggenberger, M.: Propagation of uncertainty through multivariate functions in the framework of sets of probability measures. Reliab. Eng. Syst. Saf. 85, 73–88 (2004)
8. Fetz, T.: Multiparameter models: Probability distributions parameterized by random sets. In: De Cooman, G., Vejnarova, J., Zaffalon, M. (eds.) Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2007, Prague, Czech Republic, pp. 183–192. Action M Agency for SIPTA (2007)
9. Gikhman, I.I., Skorokhod, A.V.: Stochastic differential equations. Springer, New York (1972)
10. Hiai, F., Umegaki, H.: Integrals, conditional expectations, and martingales of multivalued functions. J. Multivariate Anal. 7, 149–182 (1977)
11. Himmelberg, C.J.: Measurable relations. Fund. Math. 87, 53–72 (1975)
12. Li, S., Ren, A.: Representation theorems, set-valued and fuzzy set-valued Itô integral. Fuzzy Sets Syst. 158, 949–962 (2007)
13. Li, J., Li, S.: Set-Valued Stochastic Lebesgue Integral and Representation Theorems. Int. J. Comput. Intell. Syst. 1, 177–187 (2008)
14. Li, J., Li, S.: Itô Type Set-Valued Stochastic Differential Equation. J. Uncertain Syst. 3, 52–63 (2009)
15. Miranda, E., Couso, I., Gil, P.: Random sets as imprecise random variables, J. Math. Anal. Appl. 307, 32–47 (2005)
16. Miranda, E., Couso, I., Gil, P.: Approximations of upper and lower probabilities by measurable selections. Inform. Sci. 180, 1407–1417 (2010)
17. Molchanov, I.: Theory of random sets. Springer, London (2005)
18. Schmelzer, B.: On solutions of stochastic differential equations with parameters modelled by random sets. Internat. J. Approx. Reason. (accepted for publication, 2010)
19. Schmelzer, B., Oberguggenberger, M., Adam, C.: Efficiency of Tuned Mass Dampers with Uncertain Parameters on the Performance of Structures under Stochastic Excitation. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability (accepted for publication, 2010)

Coupled Brownian Motion

Carlo Sempi

Abstract. We present a way of considering a stochastic process $\{B_t : t \geq 0\}$ with values in $\mathbb{R}^2$ such that each component is a Brownian motion. The distribution function of B_t, for each t, is obtained as the copula of the distribution functions of the components. In this way a "coupled Brownian motion" is obtained. The (one–dimensional) Brownian motion is the example of a stochastic process that (a) is a Markov process, (b) is a martingale in continuous time, and (c) is a Gaussian process. It will be seen that while the coupled Brownian motion is still a Markov process and a martingale, it is not in general a Gaussian process.

Keywords: Copulas, Brownian motion, Markov processes, Martingales, Gaussian processes.

1 Introduction and Definitions

In a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ let $\{B_t^{(1)} : t \geq 0\}$ and $\{B_t^{(2)} : t \geq 0\}$ be two Brownian motions (=BM's) and consider, for every $t \geq 0$, the random vector

$$B_t := \left(B_t^{(1)}, B_t^{(2)}\right). \tag{1}$$

Then $\{B_t : t \geq 0\}$ defines a stochastic process with values in $\mathbb{R}^2$. The literature deals mainly with the independent case, viz. $B_t^{(1)}$ and $B_t^{(2)}$ are independent for every $t \geq 0$; this is usually called the *two–dimensional* BM (see, for instance, [3, 5]). However, we think it useful, both for its own sake and in view of potential applications, to introduce a more general multidimensional BM.

Carlo Sempi
Dipartimento di Matematica "Ennio De Giorgi", Università del Salento,
73100 Lecce, Italy
e-mail: carlo.sempi@unisalento.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 569–574.
springerlink.com

For every $t \geq 0$, let $F_t^{(1)}$ and $F_t^{(2)}$ be the (right–continuous) distribution functions (=d.f.'s) of $B_t^{(1)}$ and $B_t^{(2)}$, respectively; thus, for every $x \in \mathbb{R}$,

$$F_t^{(j)}(x) = \mathbb{P}\left(B_t^{(j)} \leq x\right) \qquad (j = 1,2).$$

Actually, For every $t \geq 0$, $F_t^{(1)}(x) = F_t^{(2)}(x) = \Phi(x)$, where Φ is the d.f. of the standard normal distribution $N(0,1)$.

In order to describe the bivariate d.f. of B_t, the concept of *copula* will be needed. There is now an extensive literature on copulas; we shall quote here only the original papers ([7, 8]), and books that deal with this topic ([1, 2, 4, 6]).

A bivariate copula C is the restriction to the unit square $\mathbb{I}^2$ ($\mathbb{I} = [0,1]$) of a two–dimensional d.f. that concentrates all the probability mass on $\mathbb{I}^2$ and which has uniform marginals on $\mathbb{I}$. Equivalently a bivariate copula C is a function $C : \mathbb{I}^2 \to \mathbb{I}$ such that

(a) satisfies the boundary conditions

$$\forall\, t \in \mathbb{I} \qquad C(t,0) = C(0,t) = 0, \qquad C(t,1) = C(1,t) = t; \tag{2}$$

(b) is 2–increasing, namely, for all u, u', v and v' in $\mathbb{I}$, with $u \leq u'$ and $v \leq v'$,

$$C(u',v') - C(u,v') - C(u',v) + C(u,v) \geq 0. \tag{3}$$

The importance of copulas stems from Sklar's theorem, which is here stated in a form adapted to the aims of the present paper.

Theorem 1 (Sklar). *For every two–dimensional d.f. H with marginals F_1 and F_2, there exists a copula C such that, for all x and y in $\mathbb{R}$,*

$$H(x,y) = C\left(F_1(x), F_2(y)\right). \tag{4}$$

If F_1 and F_2 are both continuous, then the copula C is unique.

For every $t \geq 0$, we shall consider a bivariate copula C_t, which depends on t, to be the copula of the random pair $(B_t^{(1)}, B_t^{(2)})$. Then the d.f. $H_t : \mathbb{R}^2 \to \mathbb{I}$ of the random pair B_t, is given, for all x and y in $\mathbb{R}$, by

$$H_t(x,y) = C_t\left(F_t^{(1)}(x), F_t^{(2)}(y)\right). \tag{5}$$

The BM's $B_t^{(1)}$ and $B_t^{(2)}$ may be assumed to be continuous (we shall always make this assumption), and, in fact, absolutely continuous since both $B_t^{(1)}$ and $B_t^{(2)}$ are normally distributed. Therefore, their respective d.f.'s $F_t^{(1)}$ and $F_t^{(2)}$ are also absolutely continuous for every $t \geq 0$, and, as a consequence, the copula C_t, for every $t \geq 0$, is uniquely determined.

Through an abuse of notation we shall write

$$B_t := C_t\left(B_t^{(1)}, B_t^{(2)}\right). \tag{6}$$

Notice that, in principle, a different copula is allowed for every $t \geq 0$. The process $\{B_t : t \geq 0\}$ will be called the 2*–dimensional coupled Brownian motion.*

The traditional two–dimensional BM is included in the picture we have presented so far; in order to recover it, it suffices to choose the independence copula $\Pi_2(u,v) := uv$ $((u,v) \in \mathbb{I}^2)$ and set $C_t = \Pi_2$ in (5) for every $t \geq 0$, so as to obtain

$$H_t(x,y) = F_t^{(1)}(x)\, F_t^{(2)}(y) \qquad ((x,y) \in \mathbb{R}^2).$$

A possible extension will have to be considered in future developments. Since there is no reason why one should limit oneself to dimension 2, it should also be possible to consider also the d–dimensional case, namely, given d BM's $(B_t^{(1)}, B_t^{(2)}, \dots, B_t^{(d)})$ on $(\Omega, \mathscr{F}, \mathbb{P})$, let $B_t : \Omega \to \mathbb{R}^d$ be, for every $t \geq 0$, the random vector defined by

$$B_t := \left(B_t^{(1)}, B_t^{(2)}, \dots, B_t^{(d)}\right). \tag{7}$$

Thus, by recourse to d–dimensional copulas, briefly d–copulas, one can write

$$B_t := C_t\left(B_t^{(1)}, B_t^{(2)}, \dots, B_t^{(d)}\right). \tag{8}$$

In this paper, we shall limit ourselves to studying the model (6).

The (one–dimensional) BM is the example of a stochastic process that has three properties

- it a Markov process;
- it is a martingale in continuous time;
- it is a Gaussian process.

These three possible aspects of a coupled BM will be examined in the following sections.

2 The Markov Property

Since the Markov property for a d–dimensional process $\{X_t : t \geq 0\}$ disregards the dependence relationship of its components at every $t \geq 0$, but is solely concerned with the dependence structure of the random vector X_t at different times, the traditional proof for the ordinary (independent) BM (see, for instance, [3, §2.5.12]) holds for the coupled BM $\{B_t := C_t(B_t^{(1)}, B_t^{(2)}) : t \geq 0\}$. Therefore, we have

Theorem 2. *A coupled Brownian motion* $\{B_t := C_t(B_t^{(1)}, B_t^{(2)}) : t \geq 0\}$ *is a Markov process.*

3 The Coupled Brownian Motion Is a Martingale

One can prove the following result, already announced by the title of this section.

Theorem 3. *The coupled Brownian motion* $\{B_t := C_t(B_t^{(1)}, B_t^{(2)}) : t \geq 0\}$ *is a martingale.*

Proof. Let $(\mathscr{F}_t)_{t\geq 0}$ be the natural filtration: for every $t \geq 0$, $\mathscr{F}_t$ is the smallest σ–algebra with respect to which both $B_t^{(1)}$ and $B_t^{(2)}$ are measurable, and let $\mathbb{E}_s := \mathbb{E}(\cdot \mid \mathscr{F}_s)$ denote the conditional expectation with respect to $\mathscr{F}_s$. Let s and t be such that $0 \leq s < t$. Then, since each component $B_t^{(j)}$ $(j = 1,2)$ of B_t is a martingale, one has, for every set $A \in \mathscr{F}_s$,

$$\begin{aligned}\int_A \mathbb{E}_s\left(C_t\left(B_t^{(1)}, B_t^{(2)}\right)\right) d\,\mathbb{P} &= \int_A C_t\left(B_t^{(1)}, B_t^{(2)}\right) d\,\mathbb{P} = \int_A B_t\, d\,\mathbb{P} \\ &= \left(\int_A B_t^{(1)}\, d\,\mathbb{P}, \int_A B_t^{(2)}\, d\,\mathbb{P}\right) = \left(\int_A \mathbb{E}_s\left(B_t^{(1)}\right) d\,\mathbb{P}, \int_A \mathbb{E}_s\left(B_t^{(2)}\right) d\,\mathbb{P}\right) \\ &= \left(\int_A B_s^{(1)}\, d\,\mathbb{P}, \int_A B_s^{(2)}\, d\,\mathbb{P}\right) = \int_A B_s\, d\,\mathbb{P} = \int_A C_s\left(B_s^{(1)}, B_s^{(2)}\right) d\,\mathbb{P},\end{aligned}$$

which proves the martingale property

$$\mathbb{E}_s\left(C_t\left(B_t^{(1)}, B_t^{(2)}\right)\right) = C_s\left(B_s^{(1)}, B_s^{(2)}\right) \qquad a.s.,$$

or, equivalently,

$$\mathbb{E}_s(B_t) = B_s \qquad a.s.,$$

which proves the assertion. □

4 Is a Coupled Brownian Motion a Gaussian Process?

One has first to state what is meant by the expression "Gaussian process" when a stochastic process with values in $\mathbb{R}^2$ is considered. We shall adopt the following definition.

Definition 1. *A stochastic process* $\{X_t : t \geq 0\}$ *with values in* $\mathbb{R}^d$ *is said to be* Gaussian *if, for* $n \in \mathbb{N}$*, and for every choice of* n *times* $0 \leq t_1 < t_2 < \cdots < t_n$*, the random vector* $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ *has a* $(d \times n)$*–dimensional normal distribution.*

As the following examples will make it clear, the answer to the question of the title is an emphatic *No*.

Example 1. Let the copula C_t coincide, for every $t \geq 0$, with M_2, i.e., $M_2(u,v) = \min\{u,v\}$, u and v in $\mathbb{I}$. Then

$$\begin{aligned} H_t(x,y) &= \frac{1}{\sqrt{2\pi t}} \min\left\{ \int_{-\infty}^{x} \exp\{-v^2/(2t)\}\, du, \int_{-\infty}^{y} \exp\{-u^2/(2t)\}\, dv \right\} \\ &= \Phi\left(\frac{\min\{x,y\}}{\sqrt{t}}\right). \end{aligned}$$

A simple calculation shows that

$$\frac{\partial^2 H_t(x,y)}{\partial x \partial y} = 0 \qquad a.e. \tag{9}$$

with respect to the Lebesgue measure λ_2, so that H_t is not even absolutely continuous.

Example 2. If the copula C_t is given, for every $t \geq 0$, by W_2, where

$$W_2(u,v) := \max\{u+v-1, 0\},$$

then the d.f. H_t of B_t is given by

$$H_t(x,y) = \max\left\{\Phi\left(\frac{x}{\sqrt{t}}\right) + \Phi\left(\frac{y}{\sqrt{t}}\right) - 1, 0\right\},$$

which again leads, after simple calculations, to (9), so that, again, B_t is not even absolutely continuous.

It will have been noticed that in both the previous examples the copula considered was singular; thus it is hardly surprising that the coupled BM B_t turns out not to be absolutely continuous, and, *a fortiori*, not a Gaussian vector.

The two previous examples represent extreme cases; in fact, since the d.f.'s involved are continuous, the copula of two random variables is M_2 if, and only if, they are comonotone, namely, each of them is an increasing function of the other, while their copula is W_2 if, and only if, they are countermonotone, namely, each of them is a decreasing function of the other. In this sense both examples are the opposite of the independent case, which is characterized by the copula Π_2.

We recall that a copula can be either absolutely continuous or singular or, again, a mixture of the two types (see, e.g., [4]). In general, if the copula C is singular, namely the d.f. of a probability measure concentrated on a subset of zero Lebesgue measure λ_2 in the unit square $\mathbb{I}^2$, then also B_t is singular.

Now let the copula C_t be absolutely continuous with density c_t; a simple calculation shows that B_t is absolutely continuous and that its density is given by

$$h_t(x,y) = \frac{1}{2\pi t} \exp\left(-\frac{x^2+y^2}{2t}\right) c_t\left(\Phi\left(\frac{x}{\sqrt{t}}\right), \Phi\left(\frac{y}{\sqrt{t}}\right)\right) \qquad a.e..$$

As a consequence, B_t has a normal law if, and only if, $c_t(u,v) = 1$ for almost all u and v in $\mathbb{I}$; together with the boundary conditions (2), this implies $C_t(u,v) = uv = \Pi_2(u,v)$. One has thus proved the following result

Theorem 4. *In a coupled Brownian motion*

$$\left\{B_t = C_t\left(B_t^{(1)}, B_t^{(2)}\right) : t \geq 0\right\},$$

B_t has a normal law if, and only if, $C_t = \Pi_2$, viz., if, and only if, its components $B_t^{(1)}$ and $B_t^{(2)}$ are independent.

References

1. Durante, F., Sempi, C.: Copula theory: an introduction. In: Jaworski, P., Durante, F., Härdle, W., Rychlik, T. (eds.) Workshop on Copula Theory and its Applications. Springer, Heidelberg (2010)
2. Jaworski, P., Durante, F., Härdle, W., Rychlik, T.: Workshop on Copula Theory and its Applications. Springer, Heidelberg (2010)
3. Karatzas, I., Shreve, S.E.: Brownian motion and stochastic calculus, 2nd edn. Springer, Berlin (1991)
4. Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer, New York (2006)
5. Rogers, L.C.G., Williams, D.: Diffusions, Markov processes and martingales. Foundations, 2nd edn., vol. 1. Cambridge University Press, Cambridge (2000)
6. Schweizer, B., Sklar, A.: Probabilistic Metric Spaces. North–Holland, New York (1983); Reprinted, Dover, Mineola NY, (2005)
7. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges, Publ. Inst. Statist. Univ. Paris 8, 229–231 (1959)
8. Sklar, A.: Random variables, joint distribution functions and copulas. Kybernetika 9, 449–460 (1973)

The Median of a Random Interval

Beatriz Sinova, María Rosa Casals, Ana Colubi, and María Ángeles Gil

Abstract. In dealing with real-valued random variables, the median of the distribution is the 'central tendency' summary measure associated with its 'middle position'. When available random elements are interval-valued, the lack of a universal ranking of values makes it impossible to formalize the extension of the concept of median as a middle-position summary measure. Nevertheless, the use of a generalized L^1 Hausdorff-type metric for interval data enables to formalize the median of a random interval as the central-tendency interval(s) minimizing the mean distance with respect to the random set values, by following the alternate equivalent way to introduce the median in the real-valued case. The expression for the median(s) is obtained, and main properties are analyzed. A short discussion is made on the main different features in contrast to the real-valued case.

Keywords: Generalized Hausdorff metric, Median, Random interval.

1 Introduction

Interval data in connection with random experiments usually come either from the observation/measurement of an intrinsically interval-valued random attribute (say fluctuations, ranges, etc.), from an uncertain measurement or from a grouping of real-valued data in accordance with a given list of intervals (like often happens with age or income groups).

The statistical analysis of interval data, and especially the inferential developments, requires an appropriate formalization within the probabilistic

Beatriz Sinova, María Rosa Casals, Ana Colubi, and María Ángeles Gil
Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,
33007 Oviedo, Spain
e-mail: sinovabeatriz.uo,rmcasals,colubi,magil@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 575–583.
springerlink.com

setting. In this respect, compact convex random sets represent a suitable tool to handle the random mechanisms producing interval data.

In the literature, one can find several statistical studies devoted to interval data, most of them being based on descriptive techniques and approaches. Concerning the random set approach to deal with interval data some developments can be found, for instance, in López-García et al. [9], Gil et al. [4], [5], González-Rodríguez et al. [6], Montenegro et al. [11], Blanco [2], Sinova et al. [12].

In most of the studies carried out using the random set approach, the central tendency measure to be used is the mean, which extends the mathematical expectation of a real-valued random variable. It is formalized by the Aumann-type integral, is supported by the Strong Laws of Large Numbers and satisfies Fréchet's principle w.r.t. a generalized L^2-type metric.

As for the real-valued case, the mean of a random interval takes into account all the possible values, and it is usually very much influenced by 'small' or 'high' values, so that for very large one-sided skews (like it happens with income ranges) it is hardly the most suitable representative of the central tendency. This reason motivates to consider an extension of the median of random variables to alternately summarize the central tendency of a random interval. Usually, the median(s) is introduced to be the middle position value once data have been sorted from 'smallest' to 'largest'. However, there is no universal ranking between interval data, so one cannot extend this formalization to the interval-valued case but in a very restrictive way.

The median of the distribution of a real-valued random variable can also be introduced as the value minimizing the mean Euclidean distance w.r.t. variable values. This approach will be now followed to formalize the extension of the median to random intervals. Furthermore, the extension can be made by involving appropriate L^1 metrics, and this is to be made in Section 3, after recalling in Section 2 some preliminaries on interval data and random intervals. In Section 4 we will analyze the main properties of the extended median and the differences in contrast to the real-valued case. Finally, some concluding remarks will be commented in Section 5.

2 Preliminaries on Interval Data and Random Intervals

Interval arithmetic (in particular, the sum and the product by a real number) is a particular case of set arithmetic, and is stated as follows:

- If $K, K' \in \mathscr{K}_c(\mathbb{R})$ = class of the nonempty compact intervals, the *sum of K and K'* is defined as the Minkowski sum of K and K', i.e., as the interval
$$K + K' = \left[\inf K + \inf K', \sup K + \sup K'\right].$$
- If $K \in \mathscr{K}_c(\mathbb{R})$ and γ is a real number, the *product of K by the scalar* γ is defined as the interval in $\mathscr{K}_c(\mathbb{R})$ such that

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

It should be pointed out that $(\mathscr{K}_c(\mathbb{R}), +, \cdot)$ has not a linear (but a conical) structure.

In previous statistical developments with interval data using the random set approach, the particularization of the Bertoluzza et al. [1] metric has shown to be very useful and satisfying convenient properties, especially in connection with least squares approaches and other statistical developments. This metric has been alternatively defined (see Gil et al. [4],) as a generalized L^2-distance allowing us to weight the influence of the 'location' of interval values (represented by the corresponding mid-points or centers, mid) in contrast to the influence of the 'imprecision' of the values (represented by their spread or radius, spr). More concretely, given $\theta \in (0, +\infty)$, the d_θ*-metric* is defined for $K, K' \in \mathscr{K}_c(\mathbb{R})$ so that

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

d_θ is an L^2-type metric on $\mathscr{K}_c(\mathbb{R})$, and $(\mathscr{K}_c(\mathbb{R}), d_\theta)$ is a separable metric space.

The notion of random interval, as a model for a random mechanism producing interval data, can be introduced in several equivalent ways. The random set-based way makes use of the well-known Hausdorff metric on $\mathscr{K}_c(\mathbb{R})$, which by following a similar equivalence to the one stated between Bertoluzza et al.'s metric and d_θ by Gil et al. [4], and by considering some properties of absolute values, can be expressed for the interval-valued case so that for $K, K' \in \mathscr{K}_c(\mathbb{R})$ it is given by

$$d_H(K, K') = \left|\text{mid} K - \text{mid} K'\right| + \left|\text{spr} K - \text{spr} K'\right|$$

(see, for instance Chavent et al. [3], Trutschnig et al. [13]). Thus, given a probability space $(\Omega, \mathscr{A}, P)$ a mapping $X : \Omega \to \mathscr{K}_c(\mathbb{R})$ is said to be a *random interval* (for short RI) associated with it, if it is a compact convex random set (that is, it is a Borel measurable mapping w.r.t. $\mathscr{A}$ and the Borel σ-field generated by the topology induced by d_H. Analogously, X is an RI if, and only if, the real-valued functions $\inf X : \Omega \to \mathbb{R}, \sup X : \Omega \to \mathbb{R}$ (with $\inf X \leq \sup X$) are real-valued random variables, which is equivalent to say that $\text{mid} X : \Omega \to \mathbb{R}, \text{spr} X : \Omega \to [0, \infty)$ are real-valued random variables. The Borel measurability of RIs guarantees that one can properly refer to concepts like the *distribution induced by an RI*, the *stochastic independence of RIs*, and so on, which are crucial for inferential developments.

The mean value of an RI X is given by the Aumann expectation, so that the mid-point of the mean equals the expected value of $\text{mid} X$ and the spread of the mean equals the expected value of $\text{spr} X$. This (interval-valued) mean satisfies the usual properties of linearity and it is the Fréchet expectation w.r.t. d_θ (that is, it is the unique interval value minimizing over $K \in \mathscr{K}_c(\mathbb{R})$ the expected squared distance $E\left[(d_\theta(X,K))^2\right]$, which corroborates its adequacy as a *central tendency measure*). It is also coherent with the usual interval arithmetic.

3 The Median of an RI

Aiming to weigh the influence of the 'location' in contrast to the influence of the 'imprecision' as for d_θ (since 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), we can state

Definition 1. *Given* $\theta \in (0,+\infty)$*, the mapping* $d_{H,\theta} : \mathcal{K}_c(\mathbb{R}) \times \mathcal{K}_c(\mathbb{R}) \to [0,+\infty)$ *such that for any* $K, K' \in \mathcal{K}_c(\mathbb{R})$

$$d_{H,\theta}(K,K') = \left|\mathrm{mid}\,K - \mathrm{mid}\,K'\right| + \theta \cdot \left|\mathrm{spr}\,K - \mathrm{spr}\,K'\right|$$

will be called the **generalized Hausdorff metric** *on* $\mathcal{K}_c(\mathbb{R})$*.*

$d_{H,\theta}$ is an L^1-type metric on $\mathcal{K}_c(\mathbb{R})$, and $\left(\mathcal{K}_c(\mathbb{R}), d_{H,\theta}\right)$ is a separable metric space.

If one now looks for the interval value minimizing the expected distance

$$E\left[d_{H,\theta}(X,K)\right] = E\left[|\mathrm{mid}\,X - \mathrm{mid}\,K|\right] + \theta \cdot E\left[\left|\mathrm{spr}\,K - \mathrm{spr}\,K'\right|\right],$$

over $K \in \mathcal{K}_c(\mathbb{R})$, then one can look for minimizing $E\left[\left|\mathrm{mid}\,X - \mathrm{mid}\,K\right|\right]$ over $\mathrm{mid}\,K \in \mathbb{R}$ and $E\left[\left|\mathrm{spr}\,X - \mathrm{spr}\,K\right|\right]$ over $\mathrm{spr}\,K \in [0,\infty)$ separately. Therefore, irrespectively of the value of θ, we can define the median of an RI as the interval value(s) minimizing $E\left[d_{H,\theta}(X,K)\right]$ (which corroborates its adequacy as a *central tendency measure*), that is,

Definition 2. *Given a probability space* $(\Omega, \mathcal{A}, P)$ *and an RI* X *associated with it, the* **median** *of the distribution of* X*,* $\mathrm{Me}[X]$ *is(are) the nonempty compact interval(s) such that*

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

As for the real-valued case, the median of an RI can be either defined by a unique interval or not, depending on the medians of $\mathrm{mid}\,X$ and $\mathrm{spr}\,X$ being both defined in a unique way or not. The definition above coincides with the descriptive idea given recently by Irpino and Verde [7].

The next two real-life examples illustrate the computation of the median of an RI in two different situations. In the first one, the RI corresponds to an intrinsically interval-valued random element obtained by observing the daily fluctuation of a certain real-valued variable in a sample. The second one deals with the usual way in which data are grouped either to present them properly (because of having been measured in a large population or sample) or to ensure 'statistical confidentiality'.

Example 1. The data in Table 1 have been supplied in 1998 by the Nephrology Unit of the Hospital Valle del Nalón in Langreo (Asturias, Spain). The associated RI is the “range of the pulse rate over a day”, X, observed over a sample of 59 patients (suffering different types of illness) from a population of 3000 who are hospitalized per year in a given area.

Values of X are obtained from several registers of the Pulse rate of each patient measured at different moments (usually 60 to 70) over a concrete day. Rate pulse data are often collected by taking into account simply the fluctuation during a day (actually, some devices used for this purpose only record these extreme values along a day). In these cases, the knowledge of the whole registers for a day and the associated variation could distort the information on the characteristic which is considered to be relevant in some clinical studies: the range/fluctuation. In this way, X is intrinsically interval-valued; irrespectively of whether it is or not based on a real-valued characteristic, the attribute of interest or observable is interval-valued.

Table 1 Data on the ranges of pulse rate (X)

58-90	64-107	54-78	52-78	56-133	75-124
47-68	54-84	53-103	55-84	37-75	58-99
32-114	47-95	47-86	61-101	61-94	59-78
61-110	56-90	70-132	65-92	44-110	55-89
62-89	44-108	63-115	38-66	46-83	55-80
63-119	63-109	47-86	48-73	52-98	70-105
51-95	62-95	56-103	59-98	56-84	40-80
49-78	48-107	71-121	59-87	54-92	56-97
43-67	26-109	68-91	49-82	53-120	37-86
55-102	61-108	62-100	48-77	49-88	

By computing the empirical distribution functions of the mids and the spreads of the intervals, we get that

$$F_{\mathrm{mid}X}(73) = \frac{29}{59} < .5 < \frac{30}{59} = F_{\mathrm{mid}X}(74),$$
$$F_{\mathrm{spr}X}(19) = \frac{29}{59} < .5 < \frac{30}{59} = F_{\mathrm{spr}X}(19.5),$$

whence $\mathrm{Me}(\mathrm{mid}X) = 74$ and $\mathrm{Me}(\mathrm{spr}X) = 19.5$ and, hence, the median of X corresponds to

$$\mathrm{Me}[X] = [74 - 19.5, 74 + 19.5] = [54.5, 93.5].$$

Example 2. Table 2 has been constructed on the basis of data supplied in `http://politicalcalculations.blogspot.com/2007/06/comparing-us-distribu-tion-of-income-by.html`.

This table collects US income ranges X in US dollars in 2005. This situation corresponds to grouping data (either for reasons of presentation or for statistical confidentiality) from a very large population (close to 177 millions

Table 2 Number of US income earners by income range in 2005

Income range	# of income earners
1-2499	11547934
2500-4999	10907381
5000-7499	10368655
7500-9999	9907187
10000-12499	9498515
12500-14999	9122469
15000-17499	8761899
17500-19999	8406086
20000-22499	8045078
22500-24999	7673230
25000-27499	7287988
27500-29999	6889501
30000-32499	6480073
32500-34999	6066183
35000-37499	5645333
37500-39999	5229308
40000-42499	4821088
42500-44999	4425150
45000-47499	4045250
47500-49999	3684309
50000-52499	3344410
52500-54999	3026840
55000-57499	2732199
57500-59999	2460499
60000-62499	2211300
62500-64999	1983806
65000-67499	1776968
67500-69999	1589578
70000-72499	1420325
72500-74999	1267858
75000-77499	1130828
77500-79999	1007912
80000-82499	897842
82500-84999	799420
85000-87499	711518
87500-89999	633097
90000-92499	563091
92500-94999	500926

of income earners). Although there is an underlying real-valued characteristic the usually observable attribute is interval-valued.

As it often happens in this case, the width of intervals (an hence their spread) is degenerate, so that the median for the spreads is trivially computed. On the other hand, by accumulating the absolute frequencies of the intervals, we get that the cumulative distribution function of mid$X$ satisfies that

$$F_{\mathrm{mid}X}(\mathrm{mid}\,[20000,22499]) = .489 < .5 < .533 = F_{\mathrm{mid}X}(\mathrm{mid}\,[22500,24999]),$$

whence the median of X corresponds to

$$\mathrm{Me}[X] = [22500,24999].$$

4 Properties of the Median of an RI

An immediate difference in contrast to the real-valued case is that the median of an RI as defined in Section 3 should not be necessarily an interval value the RI takes on. Thus, in Example 1 Me$[X]$ does not coincide with any of the (sample) interval data, whereas in Example 2 it does.

On the other hand, and as for the real-valued case, Me$[X]$ is not always a unique interval value, since the uniqueness of Me(mid$X$) and Me(spr$X$) does not hold necessarily (but in case of using some conventions). In this way, if $\mathrm{Me}(\mathrm{mid}X)$ and $\mathrm{Me}(\mathrm{spr}X)$ are any median of $\mathrm{mid}X$ and $\mathrm{spr}X$, respectively, then the interval

$$[\mathrm{Me}(\mathrm{mid}X) - \mathrm{Me}(\mathrm{spr}X), \mathrm{Me}(\mathrm{mid}X) + \mathrm{Me}(\mathrm{spr}X)]$$

is a median of the distribution of X.

The median of an RI preserves most of the elementary operational properties for real-valued random variables, namely,

Proposition 1. *Suppose that X is an RI associated with a probability space. Then,*

i) if the distribution of X is degenerate at an interval value $K \in \mathscr{K}_c(\mathbb{R})$, $\mathrm{Me}[X] = K$;
ii) whatever $K \in \mathscr{K}_c(\mathbb{R})$ and $\gamma \in \mathbb{R}$ may be, $\mathrm{Me}[\gamma \cdot X + K] = \gamma \cdot \mathrm{Me}[X] + K$.

Furthermore, although there is no universal ranking between elements in $\mathscr{K}_c(\mathbb{R})$, there is at least a partial ordering which is fully coherent with $\mathrm{Me}[X]$ in Definition 2 as a *measure of middle position* in accordance with such an ordering. The ordering is the one stated in Ishibuchi and Tanaka [8], and stating that $K \leq_{CW} K'$ if, and only if, $\mathrm{mid}K \leq \mathrm{mid}K'$ and $\mathrm{spr}K \geq \mathrm{spr}K'$, i.e., an interval value is considered to be *CW-larger* than another iff its location is greater and its imprecision lower than those for the second one. One can then prove that

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

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

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,* $\mathrm{mid}\mathit{Me}[X]$ *can be any value in* $\left[\mathrm{mid}X(\omega_{n/2}), \mathrm{mid}X(\omega_{(n/2)+1})\right]$, *whereas* $\mathrm{spr}\mathit{Me}[X]$ *can be any value in* $\left[\mathrm{spr}X(\omega_{(n/2)+1}), \mathrm{spr}(\omega_{n/2})\right]$.

Since this study is an introductory one, deeper statistical developments (especially those related to robustness, comparison with the mean, asymptotic relative efficiency, and so on) will be left for future analysis. However, what one can easily derive is the strong consistency of the sample median in estimating the population one under very mild conditions. Thus,

Proposition 3. *Suppose that X is an RI associated with a probability space $(\Omega, \mathscr{A}, P)$ 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*

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

5 Concluding Remarks

This introductory study presents a generalized Hausdorff-type metric, and the median of the distribution of an RI is defined as a central tendency summary measure minimizing the mean distance w.r.t. the RI values. The median defined in this way preserves relevant properties of the real-valued case and agrees with the middle-position approach for a known partial ordering between intervals.

A future direction to be considered is the one related to the relationship between the mean and the median of RIs for symmetric and skewed distributions of RIs. In this respect, the robustness of the median should be discussed in depth.

Furthermore, it should be interesting to extend the notion to compact convex random sets of higher dimension and to random fuzzy sets, by extending the L^1 metric in a way similar to what has been followed in Trutschnig et al. [13].

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, the Principality of Asturias Grants IB09-042C1 and IB09-042C2, and the COST Action IC0702. Their financial support is gratefully acknowledged.

References

1. Bertoluzza, C., Corral, N., Salas, A.: On a new class of distances between fuzzy numbers. Math. Soft Comput. 2, 71–84 (1995)
2. Blanco, A.: Análisis estadístico de un nuevo modelo de regresión lineal flexible para intervalos aleatorios. PhD Thesis, University of Oviedo (2009)
3. Chavent, M., De Carvalho, F., Lechevallier, Y., Verde, R.: Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. Rev. Stat. Appl. 4, 5–29 (2003)
4. Gil, M.A., Lubiano, M.A., Montenegro, M., López-García, M.T.: Least squares fitting of an affine function and strength of association for interval data. Metrika 56, 97–111 (2002)
5. Gil, M.A., González-Rodríguez, G., Colubi, A., Montenegro, M.: Testing linear independence in linear models with interval data. Comput. Statist. Data Anal. 51, 3002–3015 (2007)
6. González-Rodríguez, G., Blanco, A., Corral, N., Colubi, A.: Least squares estimation of linear regression models for convex compact random sets. Adv. Data Anal. Clas. 1, 67–81 (2007)
7. Irpino, A., Verde, R.: Dynamic clustering of interval data using a Wasserstein-based distance. Pattern Recog. Lett. 29, 1648–1658 (2008)
8. Ishibuchi, H., Tanaka, H.: Multiobjective programming in optimization of the interval objective function. Europ. J. Oper. Res. 48, 219–225 (1990)
9. López-García, H., López-Díaz, M., Gil, M.A.: Interval-valued quantification of the inequality associated with a random set. Statist. & Probab. Lett. 46, 149–159 (2000)

10. Montenegro, M.: Estadística con datos imprecisos basada en una métrica generalizada. PhD Thesis. University of Oviedo (2003)
11. Montenegro, M., Casals, M.R., Colubi, A., Gil, M.A.: Testing 'two-sided' hypothesis about the mean of an random interval. In: Dubois, D., Lubiano, M.A., Prade, H., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Handling Variability and Imprecision. Advances in Soft Computing, vol. 48, pp. 133–139. Springer, Heidelberg (2008)
12. Sinova, B., Colubi, A., Gil, M.A., González-Rodríguez, G.: Interval arithmetic-based simple linear regression between interval data: empirical sensitivity analysis on the choice of the metric (submitted for publication, 2010)
13. Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread Inform. Sci. 179, 3964–3972 (2009)

The Use of Sets of Stochastic Operators to Constructing Imprecise Probabilities

Damjan Škulj

Abstract. A new approach to constructing sets of probabilities is presented. We use sets of stochastic operators that represent rules that preserve desirability of gambles. We also provide a set of criteria that allow constructing imprecise probability models consistent with the desirability preserving rules. The model is more general than the standard imprecise probability models using lower and upper previsions. The greater generality means that credal sets and therefore lower previsions can be understood as a special case. Some results on extensions of such models are also provided that generalise the corresponding results from the theory of lower and upper previsions.

Keywords: Imprecise probabilities, Lower previsions, Stochastic operators, Credal sets.

1 Introduction

Lack of precision in probability distributions is modelled in a very convenient way by the use of lower an upper previsions [5], which turn out to coincide with the lower and upper expectations with respect to sets of finitely additive probabilities. Another equivalent way to arrive to the same implications is through sets of desirable gambles, i.e. those random gambles that a decision maker is willing to accept.

Additional assumptions, such as risk aversion, may sometimes be imposed on the sets of desirable gambles. Thus, a risk averse decision maker is supposed to always accept a gamble with the same expectation and that is less risky than a desirable gamble. It has been shown previously in literature that such additional assumptions imply interesting properties to the corresponding sets of probabilities.

Damjan Škulj
Faculty of Social Sciences, University of Ljubljana, SI-1000 Ljubljana, Slovenia
e-mail: damjan.skulj@fdv.uni-lj.si

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 585–592.
springerlink.com

The aim of this paper is to build a unified model that allows imposing additional requirements to the models of imprecise probabilities. Very different types of such requirements can be combined. The result is an imprecise probability model that satisfies all given requirements.

The paper has the following structure. In the next section we repeat the most basic elements of imprecise probabilities. In Section 3 we describe our model and show that it is able to describe the classical imprecise probability models. In Section 4 we give two basic results on extension which generalise known results from the theory of imprecise probabilities.

Although the results are mostly general enough to be valid in more general probability spaces, to avoid technical difficulties and to be able to use simplified notations, we assume all probability spaces to be finite.

2 Elements of Imprecise Probability Theory

The theory of imprecise probabilities, as proposed by Walley [5], is based on subjective Bayesian interpretation of probability. Thus a set $\mathscr{L}$ of *gambles* is assumed with a subset $\mathscr{D}$ of *desirable gambles*. A desirable gamble is a gamble that a decision maker is willing to accept.

Mathematically, the set of gambles is represented by a set of bounded random variables on a set $\mathscr{X}$. A gamble is denoted as $X\colon \mathscr{X} \to \mathbb{R}$. The set of desirable gambles in particular contains all gambles with sure gains, i. e. X such that $\inf_{x\in\mathscr{X}} X(x) \geq 0$ and no sure losses, i.e. gambles where $\sup_{x\in\mathscr{X}} X(x) < 0$.

Moreover, a pair of real values $\underline{P}(X)$ and $\overline{P}(X)$ is assumed for each gamble X, called the *lower* and the *upper prevision* respectively, and interpreted as the buying and selling price. They are defined using the set of desirable gambles as follows. Given a gamble X its lower prevision $\underline{P}(X)$ is defined as the supremum of the set $\{\mu; X-\mu \in \mathscr{D}\}$. The upper prevision can be obtained as $\overline{P}(X) = -\underline{P}(-X)$. Therefore it is enough to focus on lower previsions only.

In general $\underline{P}(X) \leq \overline{P}(X)$ holds, however, if equality holds for every $X \in \mathscr{L}$ then we are talking about a *linear prevision*. In the case of a finite space every linear prevision can be represented by a probability mass function p so that $P(X) = E_p(X) = \sum_{x\in X} p(x)X(x)$.

A lower prevision $\underline{P}$ on a linear space of gambles $\mathscr{K}$ which may be a proper subspace of $\mathscr{L}$ is called *coherent* whenever it satisfies the following axioms:

(P1) $\underline{P}(X) \geq \inf_{\mathscr{X}} X$ for all $X \in \mathscr{K}$ (accepting sure gains);
(P2) $\underline{P}(\lambda X) = \lambda \underline{P}(X)$ for any $X \in \mathscr{K}$ and $\lambda > 0$ (positive homogeneity);
(P3) $\underline{P}(X+Y) \geq \underline{P}(X) + \underline{P}(Y)$ for all $X, Y \in \mathscr{K}$ (superlinearity).

One of the consequences of coherence is

$$\underline{P}(X) = \min_{P \geq \underline{P}} P(X), \tag{1}$$

where the minimum is taken over the set of all linear previsions dominating $\underline{P}$. The set $\mathscr{M}(\underline{P}) = \{P; P \geq \underline{P}\}$ of linear previsions or, equivalently, finitely additive probabilities, is called the *credal set* of $\underline{P}$.

In the finite case, a probability on the set $\mathscr{X}$ can be represented via its probability mass function, which is a real map on $\mathscr{X}$. Thus, the set Λ of all probabilities on $\mathscr{X}$ can be viewed as a subset of $\mathscr{L}$. Let p be a probability mass function on $\mathscr{X}$. Then the expectation operator with respect to p maps a gamble X to $E_pX = \sum_{x \in X} p(x)X(x)$. In the case of finite spaces we may write $E_pX = pX$ if p is assumed as a row and X as a column vector. It follows from the Riesz representation theorem that every positive functional on $\mathscr{L}$ with the norm 1 can be represented as an expectation with respect to some probability mass function.

3 Sets of Linear Stochastic Operators

Except for convexity no additional structure is required in general for sets of desirable gambles or credal sets. However, very often there are additional requirements for these sets that have to be considered. In this section we will show that such requirements can often be described in terms of (linear) stochastic operators that preserve desirability. Thus given a desirable gamble $X \in \mathscr{D}$ and an operator T preserving desirability, we assume that the gamble TX is desirable as well. We will show that sets of such operators allow a more general description of imprecise probabilities and additionally, they may be used to force additional structure to sets of desirable gambles and the corresponding credal sets.

Definition 1. *A positive linear operator $T \in B(\mathscr{L})$ is called* stochastic *whenever $T1_{\mathscr{X}} = 1_{\mathscr{X}}$.*

We denote by $S(\mathscr{K})$ the set of all stochastic operators on a linear space of gambles, where we assume that $\mathscr{K}$ contains all constant gambles.

Clearly, the adjoint of any stochastic operator maps the set of probability density functions into itself. Since all spaces considered here are assumed to be finite, we may consider all operators to be given in the form of matrices. We will therefore write pT to denote the action of the adjoint operator on a probability mass function. For set operations, we will write $\mathscr{M}\mathscr{T} := \{pT; p \in \mathscr{M}, T \in \mathscr{T}\}$, and similarly for the actions of sets of operators on sets of gambles.

Very often in decision theory risk aversion with respect to some reference probability distribution p is assumed. One of the interpretations of risk aversion is that a conditional expectation $E_p(X|\mathscr{B})$, where $\mathscr{B}$ is an algebra of sets, should be considered more desirable than the gamble itself. The operator $E(\cdot|\mathscr{B})$ is clearly a stochastic operator. It has been shown in literature (see e.g. [3, 6]) that if the set of desirable gambles is closed under conditional expectations with respect to some reference probability measure p then the corresponding credal set is closed for the set of adjoint operators which in this case coincide with the Jeffrey's conditioning rule (see [3]).

Another class of desirability preserving transformations that partially overlaps with stochastic operators has been studied by Walley [5, Section 3.5] and de Cooman and Miranda [1]. They consider imprecise probability models that are invariant for sets of transformations of the set $\mathscr{X}$. The induced transformations on the set of gambles are not necessarily stochastic operators; however, in the case where permutations of elements of $\mathscr{X}$ are considered, the induced operators are *permutation operators* that can be described as follows. Let π be a permutation of elements in $\mathscr{X}$. Then the operator T_π such that $T_\pi X(x) = X(\pi^{-1}(x))$ is a permutation operator. If a decision maker has a symmetric information about the probabilities of occurrence of elements in $\mathscr{X}$ then the gamble $T_\pi X$ should be desirable whenever X is desirable, and therefore permutation operators can as well be used sometimes as desirability preserving operators.

The following definitions give the most important consistency requirements between the set $\mathscr{T}$ and the corresponding set of desirable gambles $\mathscr{D}$ and credal set $\mathscr{M}$.

Definition 2. *A credal set $\mathscr{M}$*

(i) is consistent *with $\mathscr{T}$ iff, for every $X \in \mathscr{K}$ and for every $T \in \mathscr{T}$, $\underline{E}_{\mathscr{M}}TX \geq \underline{E}_{\mathscr{M}}X$;*
(ii) dominates *$\mathscr{T}$ iff, for every $X \in \mathscr{K}$, $\min_{T\in\mathscr{T}} \underline{E}_{\mathscr{M}}TX \leq \underline{E}_{\mathscr{M}}X$;*
(iii) is generated *by $\mathscr{T}$ iff it is both consistent with $\mathscr{T}$ and dominates it.*

Consistency, (i), clearly implies that for every desirable gamble X the gamble TX is desirable too, while dominance, (ii), implies that for every undesirable gamble X there exists at least one $T \in \mathscr{T}$ such that TX is undesirable. The connection with the usual requirements for credal sets will become clear later in this section where we connect the representation with operators with the standard models.

Proposition 1. *Let $\mathscr{T} \subseteq S(\mathscr{L})$ and $\mathscr{M}$ be a set of probabilities. Then $\mathscr{M}$*

(i) is consistent with $\mathscr{T}$ iff $\mathscr{M}\mathscr{T} \subseteq \mathscr{M}$;
(ii) dominates $\mathscr{T}$ iff $\mathscr{M}\mathscr{T} \supseteq \mathscr{M}$;
(iii) is generated by $\mathscr{T}$ iff $\mathscr{M}\mathscr{T} = \mathscr{M}$.

Proof. All parts are easy consequences of the definitions and the fact that $\underline{E}_{\mathscr{M}}TX = \underline{E}_{\mathscr{M}T}X$ and $\min_{T\in\mathscr{T}} \underline{E}_{\mathscr{M}}TX = \underline{E}_{\mathscr{M}\mathscr{T}}X$ respectively.

The sets with the property (iii) of the last proposition are known from the study of imprecise Markov chains where they are called *invariant sets of probabilities*. The following proposition is implied by a general result about invariant sets of probabilities (see e.g. [4]).

Proposition 2. *Let $\mathscr{T}$ be a set of operators. Then there always exists the largest set of probabilities generated by $\mathscr{T}$. If additionally for some $T \in \mathscr{T}$*

exists the limit $\lim_{n\to\infty} pT^n = q$ *and is independent of* p *then there exists the smallest closed set of probabilities generated by* $\mathscr{T}$.

The largest set generated by $\mathscr{T}$ turns out to be

$$\mathscr{M} = \bigcap_{n\in\mathbb{N}} \mathscr{M}_n, \tag{2}$$

where $\{\mathscr{M}_n\}$ is the sequence of sets where $\mathscr{M}_0 = \Lambda$ is the set of all probabilities and $\mathscr{M}_{n+1} = \mathscr{M}_n\mathscr{T}$. The largest set generated by $\mathscr{T}$ is always non-empty and corresponds to the most conservative lower prevision that is consistent with and dominates $\mathscr{T}$, which is usually the object of interest when studying extensions of lower previsions. Therefore, we denote the largest set generated by a set of operators $\mathscr{T}$ with $\mathscr{M}(\mathscr{T})$.

Proposition 3. *The maximal set of probabilities* $\mathscr{M}$ *generated by* $\mathscr{T}$ *is exactly the maximal set that dominates* $\mathscr{T}$.

Proof. It is easy to see that a maximal set (which could not be unique) dominating $\mathscr{T}$ must be generated by $\mathscr{T}$, and since $\mathscr{M}$ contains all sets generated by $\mathscr{T}$, it follows that it must be the unique largest set dominating $\mathscr{T}$.

Now we demonstrate that every lower prevision can equivalently be described with sets of operators. Let $\underline{P}$ be a lower prevision on a linear subspace of gambles $\mathscr{K}$. Then we can define the following set of operators

$$\mathscr{T}_{\underline{P}} = \{T_P | P \text{ is a linear prevision on } \mathscr{K}, P \geq \underline{P}\}, \tag{3}$$

where $T_P X = P(X)1_{\mathscr{X}}$.

The classical model is therefore only a special case of the generalised model where every operator maps the set of gambles into constant gambles. The decision whether a gamble is desirable or not is then obvious since there is a clear cut about desirability of constant gambles, the non-negative gambles are desirable while the negative ones are not. In the general case the gambles TX are not constant and therefore there is no such an obvious criterion about their desirability; however, the consistency requirements allow assigning, not necessarily in unique ways, sets of desirable gambles consistent with the more general sets of operators.

4 Extensions

A set of operators $\mathscr{T}$ may not be defined on the space of all gambles $\mathscr{L}$ but rather on a proper subspace $\mathscr{K}$. An *extension* $\bar{\mathscr{T}}$ is a set of operators on a larger domain, say $\mathscr{H} \geq \mathscr{K}$ such that $\mathscr{T} = \{\bar{T}|_{\mathscr{K}}; \bar{T} \in \bar{\mathscr{T}}\}$. In this section we study such extensions from the point of view of the lower expectation operators corresponding to the generated sets of probabilities. We give two basic results. The first one shows that extensions do not affect the behaviour on the set $\mathscr{K}$, and the second is a generalisation of the *marginal extension theorem.*

4.1 The Basic Extension Theorem

In this subsection we prove that given a set of operators $\mathscr{T}$ on the domain $\mathscr{K}$ and its arbitrary extension $\bar{\mathscr{T}}$ to a larger domain, say $\mathscr{H}$, then the lower expectation operators corresponding to $\mathscr{M}(\mathscr{T})$ and $\mathscr{M}(\bar{\mathscr{T}})$ coincide on $\mathscr{K}$. To prove this we first provide an equivalent way to arrive to the most conservative lower prevision generated by $\mathscr{T}$. The alternative way makes use of the smallest set of desirable gambles generated by $\mathscr{T}$. It is clear that smaller sets of desirable gambles correspond to more conservative lower previsions and therefore larger sets of probabilities.

Let $\mathscr{T} \subseteq S(\mathscr{K})$ be a set of operators. The smallest set of desirable gambles is constructed using the following monotone sequence of sets of gambles. Let $\mathscr{D}_0 = \{X \in \mathscr{K}; X \geq 0\}$ and $\mathscr{D}_n$ be the largest set of gambles in $\mathscr{K}$ such that $\mathscr{T}\mathscr{D}_n \subseteq \mathscr{D}_{n-1}$. Since $\mathscr{T}\mathscr{D}_0 \subseteq \mathscr{D}_0$ we have that, for all $n > 0$, $\mathscr{D}_{n-1} \subseteq \mathscr{D}_n$ (see similar constructions in e.g. [4]). Monotonicity of the sequence allows the construction of the set $\mathscr{D}_{\mathscr{K}} = \mathscr{D}_\infty := \bigcup_{n\in\mathbb{N}} \mathscr{D}_n$. The following holds.

Proposition 4. *The set $\mathscr{D}_{\mathscr{K}}$ is the smallest set of desirable gambles in $\mathscr{K}$ generated by $\mathscr{T}$, i.e. corresponding to a credal set generated by $\mathscr{T}$.*

To form a set of desirable gambles on the whole space $\mathscr{L}$ of course we need to take also all gambles that dominate those in $\mathscr{D}_{\mathscr{K}}$: $\bar{\mathscr{D}}_{\mathscr{K}} = \{Y \in \mathscr{L}; \exists X \in \mathscr{K} : X \leq Y\}$, which is a convex set that contains all gambles $Y \geq 0$.

Proof. Clearly the set $\mathscr{D}_{\mathscr{K}}$ is convex and contains all non-negative gambles in $\mathscr{K}$ which are the minimal requirements for a set of desirable gambles. We only need to prove that all its members are necessarily desirable. Let $\mathscr{M}$ be a set of probabilities generated by $\mathscr{T}$. Take any $X \in \mathscr{D}_{\mathscr{K}}$. Then there is some $r \in \mathbb{N}$ such that, for every possible sequence $T_1, \ldots, T_r \in \mathscr{T}, T_1 \ldots T_r X \geq 0$. The fact that $\mathscr{M}$ is generated by $\mathscr{T}$ now implies that $\min_{T_1\in\mathscr{T}} \underline{E}_{\mathscr{M}} T_1 \ldots T_r X = \underline{E}_{\mathscr{M}} T_2 \ldots T_r X$. A sequential application of this argument r times gives that $0 \leq \min_{T_i\in\mathscr{T}} \underline{E}_{\mathscr{M}} T_1 \ldots T_r X = \underline{E}_{\mathscr{M}} X$, which implies that X must be desirable.

The extension theorem follows.

Theorem 1. *Let $\mathscr{T} \subseteq S(\mathscr{K})$ be given and $\bar{\mathscr{T}} \subseteq S(\mathscr{H})$, where $\mathscr{K} \leq \mathscr{H}$ an extension. Let $\underline{E} = \underline{E}_{\mathscr{M}(\mathscr{T})}$ and $\underline{\tilde{E}} = \underline{E}_{\mathscr{M}(\bar{\mathscr{T}})}$. Then we have that $\underline{\tilde{E}}|_{\mathscr{K}} = \underline{E}|_{\mathscr{K}}$.*

Proof. The claim will clearly follow from the fact that $\mathscr{D}_{\mathscr{H}} \cap \mathscr{K} = \mathscr{D}_{\mathscr{K}}$. We will prove this by inductively proving that every set $\mathscr{D}_n$, corresponding to $\mathscr{T}$, coincides with $\mathscr{D}'_n \cap \mathscr{K}$ where $\mathscr{D}'_n$ corresponds to $\bar{\mathscr{T}}$. In the case $n = 0$ this holds by definition. Now suppose that $\mathscr{D}_{n-1} = \mathscr{D}'_{n-1} \cap \mathscr{K}$. We have $\bar{\mathscr{T}}(\mathscr{D}_n) \subseteq \mathscr{D}_{n-1} = \mathscr{D}'_{n-1} \cap \mathscr{K} \subseteq \mathscr{D}'_{n-1}$. Thus $\mathscr{D}_n \subseteq \mathscr{D}'_n \cap \mathscr{K}$. Next we have that $\mathscr{T}(\mathscr{D}'_n \cap \mathscr{K}) = \bar{\mathscr{T}}(\mathscr{D}'_n \cap \mathscr{K}) \subseteq \mathscr{D}'_{n-1} \cap \mathscr{K} = \mathscr{D}_{n-1}$, and therefore $\mathscr{D}'_n \cap \mathscr{K} \subseteq \mathscr{D}_n$. Together this implies that $\mathscr{D}_n = \mathscr{D}'_n \cap \mathscr{K}$.

The following proposition is an immediate consequence of definitions.

Proposition 5. *Let $\mathscr{T}$ and $\bar{\mathscr{T}}$ be as in Theorem 1. Then any set $\mathscr{M}$ that dominates $\bar{\mathscr{T}}$ also dominates $\mathscr{T}$.*

4.2 Generalised Marginal Extensions

Another important extension in the theory of lower previsions is the *marginal extension* (see [5, 2]). It is applied in the situation where a lower prevision $\underline{P}$ is defined on the set $\mathscr{L}(\mathscr{B})$ of $\mathscr{B}$-measurable gambles and additionally we have a conditional lower prevision $\underline{P}(\cdot|\mathscr{B})$ that maps a set of gambles $\mathscr{H} \supseteq \mathscr{L}(\mathscr{B})$ to $\mathscr{L}(\mathscr{B})$ so that $\underline{P}(X|\mathscr{B}) = X$ for all $X \in \mathscr{L}(\mathscr{B})$ and that satisfies certain coherence requirements. The marginal extension theorem essentially says that there exists the smallest lower prevision $\underline{E}$ on the set $\mathscr{L}$ that coincides with $\underline{P}(\underline{P}(\cdot|\mathscr{B}))$ on $\mathscr{H}$ and with the natural extension of $\underline{P}$ on $\mathscr{L}(\mathscr{B})$.

This situation as well can naturally be generalised to sets of operators. Thus we will assume a set $\mathscr{T} \subseteq S(\mathscr{K})$ where $\mathscr{K} \leq \mathscr{L}$. Further we have a set of projections $\mathscr{R}$ from $\mathscr{H} \geq \mathscr{K}$ to $\mathscr{K}$. That is, every $R \in \mathscr{R}$ is a positive linear operator $\mathscr{H} \to \mathscr{K}$ such that $RX = X$ for every $X \in \mathscr{K}$. Under the above assumptions, the set $\mathscr{T}\mathscr{R} \subseteq S(\mathscr{H})$ can be formed that maps $\mathscr{H}$ to $\mathscr{K}$ and coincides with $\mathscr{T}$ on $\mathscr{K}$. The following theorem shows that $\mathscr{M}(\mathscr{T}\mathscr{R})$ satisfies the properties of marginal extension.

Theorem 2. *Let $\mathscr{K} \leq \mathscr{H} \leq \mathscr{L}$ be linear spaces of gambles and let $\mathscr{T}$ and $\mathscr{R}$ have the properties described above. Then we have:*

(i) A set of probabilities $\mathscr{M}$ dominates $\mathscr{T}\mathscr{R}$ iff it dominates $\mathscr{T}$ and $\mathscr{R}$. Thus, $\mathscr{M}(\mathscr{T}\mathscr{R})$ is the largest credal set that dominates both $\mathscr{T}$ and $\mathscr{R}$.
(ii) $\underline{E}_{\mathscr{M}(\mathscr{T})}X = \underline{E}_{\mathscr{M}(\mathscr{T}\mathscr{R})}X$ for every $X \in \mathscr{K}$.

Proof. Theorem 1 implies (ii), since $\mathscr{T}\mathscr{R}$ is an extension of $\mathscr{T}$ from $\mathscr{K}$ to $\mathscr{H}$. Now let $\mathscr{M}$ dominate $\mathscr{T}\mathscr{R}$. Then it must dominate $\mathscr{T}$, by Proposition 5. Moreover, any set $\mathscr{M}$ that dominates $\mathscr{T}$ dominates $\mathscr{T}\mathscr{R}$ iff it dominates $\mathscr{R}$. To see this, take any $X \in \mathscr{L}$. We have that $\min_{\mathscr{T}\mathscr{R}} \underline{E}_{\mathscr{M}} TRX = \min_{\mathscr{R}} \min_{\mathscr{T}} \underline{E}_{\mathscr{M}} TRX = \min_{\mathscr{R}} \underline{E}_{\mathscr{M}} RX$, which implies our claim.

5 Concluding Remarks

The use of sets of stochastic operators has been shown to be a convenient new way how imprecise probabilities can be constructed. For finite probability spaces we have shown that the classical model with lower probabilities can naturally be considered as a special case of the new more general model.

There are of course many tasks that still wait to be accomplished. The assumption of finiteness of the probability spaces that allowed us to use simplified forms of some concepts, such as the marginal extension, does not seem crucial in most of the theory. So transition to more general spaces seems to be a very reasonable next step.

Another interesting question is the following. We have shown that if the domain for a set of operators $\mathscr{T}$ is not the whole space $\mathscr{L}$, this does not present great difficulties, due to the natural extension. However, it would be interesting to have a simple method of the extension of such a set of operators to a set $\bar{\mathscr{T}}$ defined on the whole space $\mathscr{L}$ such that $\mathscr{M}(\mathscr{T})$ and $\mathscr{M}(\bar{\mathscr{T}})$ would coincide.

References

1. de Cooman, G., Miranda, E.: Symmetry of models versus models of symmetry. In: Harper, W., Wheeler, G. (eds.) Probability and Inference: Essays in Honor of Henry E. Kyburg, Jr., pp. 67–149. Kings College Publications, London (2007)
2. Miranda, E., de Cooman, G.: Marginal extension in the theory of coherent lower previsions. Internat. J. Approx. Reason. 46(1), 188–225 (2007), doi:10.1016/j.ijar.2006.12.009
3. Škulj, D.: Jeffrey's conditioning rule in neighbourhood models. Internat. J. Approx. Reason. 42(3), 192–211 (2006), doi:10.1016/j.ijar.2005.11.002
4. Škulj, D.: Discrete time Markov chains with interval probabilities. Internat. J. Approx. Reason. 50(8), 1314–1329 (2009), doi:10.1016/j.ijar.2009.06.007
5. Walley, P.: Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London (1991)
6. Werner, J.: Risk aversion for multiple-prior and variational preferences. University of Minnesota (submitted for publication, 2010)

Prediction of Future Order Statistics from the Uniform Distribution

K.S. Sultan and S.A. Alshami

Abstract. In this paper, we use the r smallest Type-II censored order statistics $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{r:n}$ from the uniform distribution to predict the upper bound for the remaining $n-r$ observations. We use a certain statistic based both classical and Bayesian approaches. In order to show the efficiency of the proposed techniques, we point out some numerical illustrations.

Keywords: Percentage points, Probability coverage, Simulation.

1 Introduction

The prediction of future observations of order statistics have many applications in the real life studies such as, the biological studies, life testing and quality control problems. Prediction problems come up naturally in several real life situations, for example, the prediction of rain fall extremes, highest water level of the seas and temperatures. Also, some other applications involving data from weather, sport and economics.

Prediction of order statistics have been investigated by many authors, for example, Lawless [5] and Lingappaiah [6] have considered the case of exponential distribution for fixed sample size. Specifically, they have predicted the j-th order observation in a sample of size n based on observing the first i ordered observations $(j > i)$. Wright and Singh [15] and Adatia and Chan [2] have considered the problem of predicting the future order statistics from Weibull distribution. Nelson and Schmee[9] have investigated the prediction limits for the last failure time of a sample from lognormal distribution. Lu [7] has predicted intervals of an ordered observation from one-parameter

K.S. Sultan and S.A. Alshami
Department of Statistics and Operations Research, College of Science,
King Saud University, Riyadh 11451, Saudi Arabia
e-mail: ksultan@ksu.edu.sa,alshami99@gmail.com

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 593–602.
springerlink.com

exponential based of multiple type-II censoring samples. Wu and Lu [16] have predicted intervals for an ordered observation from the logistic distribution based on censored samples. Balasooriya [4] and Ogunyemi and Nelson [10] have considered the prediction of future distribution order statistics based on Type-II censored samples from gamma distribution. Abd Ellah and Sultan [1] and Sultan and Abd Ellah [13] have generalized the results by Lawless [5] and Lingappaiah [6] to predict future order statistics from the exponential distribution. Uniform distribution has many applications in real life problem. Goodness-of-fit test of the uniform distribution is considered by many authors, see for example, Samuel-Cahn [12], Marries and Szynal [8] and Steele and Chaseling [14].

Let $X_{1:n} \leq X_{2:n}, \cdots \leq X_{r:n}$ be the available Type-II censored order statistics from uniform distribution $U(0,1)$. In this paper, we follow similar approach of Abd Ellah and Sultan [1] and Sultan and Abd Ellah [13] to develop the classical and Bayesian prediction of the upper bound of the future $n-r$ observations from $U(0,1)$. This technique can be used to construct the upper bound of the future order statistics from some other continuous distributions with explicit cumulative distribution functions. In Section 2, we discuss the classical approach while in Section 3, we discuses the Bayesian approach. In Section 4, we give some numerical illustrations and application. Finally, in Section 5, we draw some conclusions.

2 Classical Approach

In this section, we propose the following statistic to develop the predictive function of the future order statistics from the uniform distribution

$$W = X_{j:n} - X_{i:n}, \quad 1 \leq i < j \leq n, \tag{1}$$

where $X_{i:n}$ and $X_{j:n}$ represent the i-th and j-th order statistics from uniform distribution $U(0,1)$.

The following lemma presents the probability density function (pdf) and the corresponding cumulative distribution function (cdf) of the proposed statistic W.

Lemma 1. *Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ be order statistics from uniform distribution $U(0,1)$, then the statistic W defined in (1) is distributed as $\beta(j-i, n-j+i+1)$ and the corresponding cdf is given by (see Arnold, Balakrishnan and Nagaraja [3])*

$$F_W(w) = IB_w(j-i, n-j+i+1), \tag{2}$$

where $IB_W(p,q)$ is the incomplete beta function defined by

$$IB_w(p,q) = \frac{1}{\beta(p,q)} \int_0^w u^{p-1}(1-u)^{q-1} du, \quad 0 < u < 1. \tag{3}$$

From Lemma 1, the $(1-\alpha)100\%$ predictive upper bound for the future order statistic $X_{j:n}$ is given by

$$Pr(X_{j:n} \le x_{i:n} + w) = 1 - \alpha, \tag{4}$$

where w represents the $(1-\alpha)100\%$ percentage points of $F_W(w)$ given in (2) and calculated in Table 1 by solving the nonlinear equation

$$F_W(w) = 1 - \alpha. \tag{5}$$

As examination of the entries in Table 1, we note that the upper percentage points of W increase as the confidence level and the difference $j-i$ increase.

Also, from Table 1, we see that when i increases for a given value of j and a given confidence level, we get better (sharper) upper bounds which is expected since we increase i more information is obtained. In addition, when the pair (i,j) increases and the confidence level increases, the upper bound of W increases. In order to show how to use Table 1, we give the following example.

Example 1. In this example we generate the first r order statistics from $U(0,1)$ (say $r=8$) when $n=10$ as follows: 0.072, 0.140, 0.162, 0.192, 0.234, 0.366, 0.466, 0.536. By using the 8−th order statistic and Table 1, then the 90% upper bound of the 9−th order statistic is obtained as: $U_{8:10} + w = 0.536 + 0.2057 = 0.7417$.

To examine the efficiency of our technique, the probability coverage of the predictive confidence intervals are simulated when $n = 10$ based on 10,000 repetitions through Monte Carlo simulation as given in Table 2. Some other table for different sample sizes are available with the authors upon request.

From Table 2, we note that the simulated probability coverage is quite close to the corresponding confidence levels for the cases $90\%, 95\%, 97.5\%$ and 99%. Also the simulated probability coverage increases when the confidence level increases for any pair (i,j).

Table 1 The upper percentage points of the upper bound of W.

i	j	90%	95%	97.5%	99%	i	j	90%	95%	97.5%	99%
5	6	0.2057	0.2589	0.3085	0.3690	7	8	0.2057	0.2589	0.3085	0.3690
5	7	0.3368	0.3942	0.4450	0.5043	7	9	0.3368	0.3942	0.4450	0.5043
5	8	0.4496	0.5069	0.5561	0.6117	7	10	0.4496	0.5069	0.5561	0.6117
5	9	0.5517	0.6066	0.6524	0.7029	8	9	0.2057	0.2589	0.3085	0.3690
5	10	0.6458	0.6965	0.7376	0.7817	8	10	0.3368	0.3942	0.4450	0.5043
6	7	0.2057	0.2589	0.3085	0.3690	9	10	0.2057	0.2589	0.3085	0.3690
6	8	0.3368	0.3942	0.4450	0.5043						
6	9	0.4496	0.5069	0.5561	0.6117						
6	10	0.5517	0.6066	0.6524	0.7029						

Table 2 The Probability coverage of the upper bound of W.

i	j	90%	95%	97.5%	99%	i	j	90%	95%	97.5%	99%
5	6	0.9017	0.9491	0.9746	0.9893	7	8	0.8959	0.9481	0.9729	0.9889
5	7	0.9032	0.9473	0.9778	0.9906	7	9	0.8997	0.9494	0.9726	0.9921
5	8	0.9046	0.9499	0.9740	0.9892	7	10	0.9058	0.9528	0.9740	0.9916
5	9	0.9022	0.9477	0.9749	0.9895	8	9	0.8982	0.9474	0.9760	0.9898
5	10	0.8966	0.9463	0.9725	0.9906	8	10	0.9012	0.9474	0.9739	0.9890
6	7	0.9021	0.9518	0.9760	0.9894	9	10	0.8978	0.9527	0.9724	0.9894
6	8	0.8999	0.9473	0.9743	0.9897						
6	9	0.9029	0.9520	0.9757	0.9900						
6	10	0.8992	0.9514	0.9756	0.9893						

Example 2 (Prediction from the exponential distribution). Let $t_{1:n} \leq t_{2:n} \leq \cdots \leq t_{r:n}$ denote the first r order observations (Type-II censoring) from the standard exponential distribution, with pdf $f(t) = \exp\{-t\},\ 0 \leq t < \infty$. By using the inverse transform $x = F^{-1}(u) = -\log(1-u)$, where u is a random variate follows $U(0,1)$, we can obtain the predictive upper bounds of the exponential distribution through the following steps:

1. Generate the first r ordered observations from the exponential distribution say: $t_{1:n} \leq t_{2:n} \leq \cdots \leq t_{r:n}$
2. Calculate the corresponding first r ordered observations from $U(0,1)$ as $U_{i:n} = 1 - \exp\{-t_{i:n}\}, \quad i = 1,2,\ldots r$
3. Predict the upper bound of the future order statistics from $U(0,1)$ as explained before, say: $U_{r+1:n}, U_{r+2:n} \ldots U_{n:n}$.
4. Then use $x = F^{-1}(u) = -\log(1-u)$ to calculate the corresponding predicted upper bounds $t_{r+1:n}, t_{r+2:n}, \ldots t_{n:n}$ from the exponential distribution.

In this case, we generate the first 5 order statistics from the standard exponential distribution when $n = 10$ as follows: 0.0167, 0.0902, 0.1110, 0.2716, 0.3117 and the corresponding uniform observations are: 0.0166, 0.0862, 0.1051, 0.2379, 0.2678. Then the 90% upper bound of the 6-th to 10-th ordered observations from $U(0,1)$ can be obtained based on statistic W from Table 1. Next, use the the inverse transformation method to get the predicted upper bound from the the exponential distribution as displayed in the following table:

j	6	7	8	9	10
$U_{j:n}$	0.4735	0.6046	0.7174	0.8195	0. 9136
$t_{j:n}$	0.6415	0.9278	1.2637	1.7119	2.4485

3 Bayesian Approach

In this section, we derive the exact Bayesian predictive function of the upper bound of the future order statistics from the uniform distribution.

Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{i:n}$ be the first i order statistics from the uniform distribution $U(0,\theta)$ and $X_{i+1:n} \le X_{i+2:n} \le \cdots \le X_{n:n}$ be the remaining $(n-i)$ order statistics.

The predictive density function of $y = X_{j:n},\ j = i+1, i+2, \ldots, n$ given $X = (X_{1:n}, X_{2:n}, \ldots X_{i:n})$ can be written as

$$h(y \mid X) = \int f(y \mid \theta)\pi(\theta \mid X)d\theta, \tag{6}$$

where $f(y \mid \theta)$ is the conditional pdf of the future observation y and $\pi(\theta \mid X)$ is the posterior pdf. The following theorem gives the Bayesian predictive function of the upper bound of the future order statistics from $U(0,\theta)$ based on the statistics W given in (1).

Theorem 1. *Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{i:n}$ be the smallest i Type-II order statistics from the uniform distribution $U(0,\theta)$. Considering the prior $g(\theta) = 1,\ 0 < \theta < 1$, then the predictive density function of W defined in (1) can be obtained from (6) as*

$$h(w \mid x) = \begin{cases} C_{i,j:n} \sum_{r=0}^{n-j} \sum_{k=0}^{n-i} \sum_{d=0}^{n-j+i} (-1)^{d+k+r} \binom{n-j}{r} \binom{n-i}{k} \binom{n-j+i}{d} \\ \times [(\frac{1}{x_i})^{d+j-1} - x_i^k] \frac{w^{d+j-i-1}}{(i+r)(a)(d+j+k-1)},\ 0 < w < x_i, \\ \\ C_{i,j:n} \sum_{r=0}^{n-j} \sum_{k=0}^{n-i} \sum_{d=0}^{n-j+i} (-1)^{d+k+r} \binom{n-j}{r} \binom{n-i}{k} \binom{n-j+i}{d} \\ \times \frac{x_i^k}{(i+r)(a)(d+j+k-1)} [(\frac{1}{w})^{i+k} - w^{j+d-i-1}],\ x_i < w < 1, \end{cases} \tag{7}$$

where $C_{i,j:n} = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}$ and a is given by

$$a = \sum_{m=0}^{n-i} \frac{(-1)^m \binom{n-i}{m} (x_i)^m [(\frac{1}{x_i})^{i+m-1} - 1]}{i+m-1}. \tag{8}$$

Proof. Let $W_3 = X_{j:n} - X_{i:n}$ and $V_3 = X_{i:n}$, then the joint pdf of W_3 and V_3 is given by

$$f_{i,j:n}(v_3, w_3) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \sum_{r=0}^{n-j} (-1)^r \binom{n-j}{r} {w_3}^{j-i-1} v_3^{i+r-1}$$
$$\times (\frac{1}{\theta})^{j+r} (1 - \frac{w_3}{\theta})^{n-j-r}, \quad 0 < v_3 < \theta - w_3, \quad 0 < w_3 < \theta,$$

hence, the marginal pdf of W_3 is obtained as

$$f(w_3 \mid \theta) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\sum_{r=0}^{n-j}(-1)^r\binom{n-j}{r}\frac{\theta^{i-j}}{i+r} \times w_3^{j-i-1}(1-\frac{w_3}{\theta})^{n-j+i},\ 0 < w_3 < \theta. \tag{9}$$

Using the prior density $g(\theta) = 1, 0 < \theta < 1$, and the likelihood function

$$L(x \mid \theta) = \frac{n!}{(n-i)!}(\theta - x_i)^{n-i}(\frac{1}{\theta})^n,\ 0 < \theta < 1,$$

then posterior distribution based on W_3 is given by

$$\pi(\theta \mid X) \propto L(x \mid \theta)g(\theta)\ = \frac{(\theta - x_i)^{n-i}(\frac{1}{\theta})^n}{a},\ \ x_i < \theta < 1, \tag{10}$$

where a is defined as in defined in (8).

From (6), (9) and (10), we have the predictive function of W_3 is given by

$$\begin{aligned} h(w_3 \mid X) &= \int_{\theta=t}^{1} f(w_3 \mid \theta)\pi(\theta \mid X)d\theta,\ t = max(x_i, w_3), \\ &= \int_{\theta=t}^{1} C_{i,j:n}\sum_{r=0}^{n-j}(-1)^r\binom{n-j}{r}w_3^{j-i-1}[\frac{1}{a(i+r)}]\theta^{i-j}[1-\frac{w_3}{\theta}]^{n-j+i} \\ &\times\ (\theta - x_i)^{n-i}(\frac{1}{\theta})^n d\theta. \end{aligned}$$

Expanding $(\theta - x_i)^{n-i}$ and $(1-\frac{w_3}{\theta})^{n-j+i}$ binomially and simplify we get (7). □

Lemma 2. *The corresponding cdf of the predictive pdf in Theorem 1 is given by*

$$H(w \mid x) = \begin{cases} \frac{C_{i,j:n}}{a}\sum_{r=0}^{n-j}\sum_{k=0}^{n-i}\sum_{d=0}^{n-j+i}(-1)^{d+k+r}\binom{n-j}{r}\binom{n-i}{k}\binom{n-j+i}{d} \\ \times\frac{x_i^k}{(i+r)(j+d-i)}\frac{[(\frac{1}{x_i})^{d+j+k-1}-1]}{d+j+k-1}w^{j+d-i},\ 0 < w < x_i, \\ \\ \frac{C_{i,j:n}}{a}\sum_{r=0}^{n-j}\sum_{k=0}^{n-i}\sum_{d=0}^{n-j+i}(-1)^{d+k+r}\binom{n-j}{r}\binom{n-i}{k}\binom{n-j+i}{d} \\ \times\frac{x_i^k}{(i+r)(d+j+k-1)}\Bigg[(\frac{1}{k+i-1})[(\frac{1}{w})^{k+i-1}-(\frac{1}{x_i})^{k+i-1}] \\ -(\frac{w^{j-i+d}}{j-i+d}-\frac{x_i^{j-i+d}}{j-i+d})+((\frac{1}{x_i})^{i-1}-x_i^{d+j+k-i})\Bigg],\ x_i < w < 1, \end{cases} \tag{11}$$

where a is defined in (8) *and* $C_{i,j:n} = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}$.

Then the Bayesian percentage points of the upper bound of the future order statistics can be calculated in view of (5) when $n = 10$ as given in Table 3. The conclusions of Table 3 are the same as those of Table 1.

Table 3 The upper percentage points of the upper bound of W based on Bayesian approach..

i	j	90%	95%	97.5%	99%	i	j	90%	95%	97.5%	99%
i	j	90%	95%	97.5%	99%	i	j	90%	95%	97.5%	99%
5	6	0.1219	0.1585	0.1952	0.2435	7	8	0.1518	0.1932	0.2330	0.2832
5	7	0.2059	0.2513	0.2951	0.3504	7	9	0.2512	0.2985	0.3423	0.3960
5	8	0.2821	0.3341	0.3827	0.4422	7	10	0.3385	0.3891	0.4350	0.4900
5	9	0.3549	0.4121	0.4643	0.5261	8	9	0.1533	0.1946	0.2341	0.2836
5	10	0.4258	0.4872	0.5417	0.6044	8	10	0.2530	0.2996	0.3426	0.3949
6	7	0.1432	0.1833	0.2223	0.2722	9	10	0.1402	0.1779	0.2138	0.2592
6	8	0.2383	0.2853	0.3293	0.3838						
6	9	0.3227	0.3740	0.4211	0.4778						
6	10	0.4017	0.4562	0.5050	0.5623						

Example 3. In this example, we generate 7 order statistics from $U(0,1)$ when $n = 10$ as follows: $0.072, 0.140, 0.162, 0.192, 0.234, 0.366, 0.466$. To predict, for example, the 90% upper bound of the 9-th order statistic by using the 7-th order statistics, we have

$$\text{The 90\% upper bound of } U_{9:10} = U_{7:10} + w = 0.466 + 0.2512 = 0.7172.$$

In order to examine the efficiency of our technique in this case, the probability coverage of the predictive confidence intervals can be calculated as given in Table 4.

From Table 4, we see that the empirical probability coverage in this case of 80% and even lower for a theoretical of 90%. This probability coverage becomes more efficient for large n.

Table 4 The Probability coverage of the upper bound of W based on Bayesian approach.

i	j	90%	95%	97.5%	99%	i	j	90%	95%	97.5%	99%
5	6	0.7285	0.8250	0.8886	0.9388	7	8	0.8100	0.8842	0.9334	0.9668
5	7	0.6500	0.7679	0.8471	0.9207	7	9	0.7634	0.8510	0.9044	0.9510
5	8	0.5753	0.7100	0.8114	0.9010	7	10	0.7112	0.8163	0.8829	0.9398
5	9	0.5080	0.6555	0.7733	0.8774	8	9	0.8062	0.8866	0.9316	0.9658
5	10	0.4338	0.5973	0.7274	0.8490	8	10	0.7575	0.8489	0.9107	0.9531
6	7	0.7981	0.8774	0.9230	0.9604	9	10	0.7764	0.8579	0.9053	0.9489
6	8	0.7379	0.8353	0.8956	0.9460						
6	9	0.6862	0.7938	0.8664	0.9277						
6	10	0.6270	0.7521	0.8387	0.9114						

4 Application

The following data given by Proschan [11] that are the times between successive failure of air conditioning equipment in a Boeing 720 airplane, arranged in increasing order of magnitude. The first nine of the data points are: 12, 21, 26, 29, 29, 48, 57, 59, 70. Assuming that the data follows an exponential distribution $\tau e^{-\tau t}$, then upon using the percentage points of W in Tables 1 and 3, we predict the 90% upper bounds for the 10-th observation based on the first 9 order observations as follows:

1. Calculate the maximum likelihood estimate of τ as:

$$\hat{\tau} = \frac{r}{\sum_{i=1}^{r} t_i + (n-r)t_r} = \frac{9}{351+70} = 0.021377672.$$

2. Calculate the corresponding 9-th uniform observation as:

$$U_{9:10} = 1 - e^{-0.021377672 \times 70} = 0.776073.$$

3. From Table 3, we have the 90% upper bound of the 10-th uniform observation is $0.776073 + 0.14020 = 0.916273$. Similarly from Table 1, we have the 90% upper bound of the 10-th uniform observation is $0.776073 + 0.2057 = 0.981773$.
4. The 90% classical and Bayesian prediction of the upper bound of the 10-th order statistic from the exponential distribution is given, respectively, by

$$\frac{-\ln(1-0.981772)}{\hat{\tau}} = 187 \text{ and } \frac{-\ln(1-0.916273)}{\hat{\tau}} = 116.$$

If we assume the true value of the 10-th observation lies in the middle of the intervals (70,187) and (70, 116) for the classical and Bayesian techniques, respectively. Then the classical and Bayesian estimates of the 10-th observation are 128.5 and 93, respectively.

From the above results, we recommend the Bayesian technique for this data since the box plot identifies the value 128.5 as an outlier. Also, the standard error of the mean in the Bayesian case is 8.04 while it is 10.7 in the classical case.

5 Conclusions

In this paper, some techniques for predicting the upper bound of the future order statistics from the uniform distributions are developed. Both Classical and Bayesian approaches are used to develop the upper bound of the future order statistics by using the smallest Type-II censoring observations. To show the performance of the developed techniques, Some numerical illustrations via Monte Carlo simulations are used in different situations. Also, some examples and application are discussed. In conclusion, we summarize the following:

1. The classical and Bayesian predictive upper bounds for the future order statistics from uniform distribution are obtained and used to develop the corresponding prediction of upper bound of the future order statistics from the exponential distribution.
2. The Bayesian approach gives better results than the classical approach in the sense of the average predictive width which is observed to be narrower in the Bayesian approach than that of the classical one.
3. The empirical probability coverages for the classical approach are closer to the theoretical confidence levels than the corresponding values for the Bayesian approach. It is observed, however, that the probability coverage values for the two approaches give very close results for large n. Adapting reasonable prior distrusting could leads to better efficiency in the Bayesian setup.

Acknowledgements. The authors would like to thank the referees for their helpful comments, which improved the presentation of the paper.

References

1. Abd Ellah, A.H., Sultan, K.S.: Exact Bayesian Prediction of Exponential Lifetime Based on Fixed and Random Sample Sizes. Quality Technology & Quantitative Management 2, 161–175 (2005)
2. Adatia, A., Chan, L.K.: Robust Procedures for Estimating the Scale Parameter and Predicting Future Order Statistics of the Weibull Distribution. IEEE Trans. Reliability R-31(5), 491–498 (1982)
3. Arnold, B.C., Balakrishnan, N., Nagaraja, H.N.: A First Course in Order Statistics. John Wiley & Sons, New York (1992)
4. Balasooriya, U.: A Comparison of the Prediction of Future Order Statistics for the 2-Parameter Gamma Distribution. IEEE Trans. Reliability R-36(5), 591–594 (1987)
5. Lawless, J.F.: A prediction problem concerning samples from the exponential distribution, with application in life testing. Technometrics 13, 725–730 (1971)
6. Lingappaiah, G.: Prediction in exponential life testing. Canad. J. Statist. 1, 113–117 (1973)
7. Lu, H.L.: Prediction Intervals of an Ordered Observation from One-Parameter Exponential Distribution Based on Multiple Type II Censored Samples. J. Chinese Institute of Industrial Engineers 21(5), 494–503 (2004)
8. Morris, K.W., Szynal, D.: A Goodness-of-Fit Test for the Uniform Distribution Based on a Characterization. J. Math. Sci. 106, 2719–2724 (2001)
9. Nelson, W., Schmee, J.: Prediction Limits for the Last Failure Time of a (Log) Normal Sample from Early Failures. IEEE Trans. Reliability R-30(5), 461–465 (1981)
10. Ogunyemi, O.T., Nelson, P.I.: Prediction of Gamma failure times. IEEE Trans. Reliability R-46(3), 400–405 (1997)
11. Proschan, F.: Theoretical explanation of observed decreasing failure rate. Technometrics 5, 375–383 (1963)
12. Samuel-Cahn, E.: Two Kinds of Repeated Significance Tests, and Their Application for the Uniform Distribution. Comm. Statist. Simulation Comput. 3(5), 419–431 (1974)

13. Sultan, K.S., Abd Ellah, A.H.: Exact prediction intervals for exponential lifetime based on random sample size. Int. J. Comput. Math. 83(12), 867–878 (2006)
14. Steele, M., Chaseling, J.: Powers of Discrete Goodness-of-Fit Test Statistics for a Uniform Null Against a Selection of Alternative Distributions. Comm. Statist. Simulation Comput. 35, 1067–1075 (2006)
15. Wright, W.P., Singh, N.: A Prediction Interval in Life Testing: Weibull Distribution. IEEE Trans. Reliability R-30(5), 466–467 (1981)
16. Wu, T.H., Lu, H.L.: Prediction intervals for an ordered observation from the logistic distribution based on censored samples. J. Stat. Comput. Simul. 77(5), 389–405 (2007)

Balance Sheet Approach to Agent-Based Computational Economics: The EURACE Project

Andrea Teglio, Marco Raberto, and Silvano Cincotti

Abstract. Handling carefully monetary and real flows, given by agents' behaviors and interactions, is a key requirement when dealing with complex economic models populated by a high number of agents. The paper shows how the stock-flows consistency issue has been faced in the EURACE model, by considering a dynamic balance sheet approach for modeling and validation purposes.

Keywords: Agent-based computational economics, Balance sheet approach, EURACE project.

1 Introduction

This paper proposes a rigorous approach to agent-based economic modelling, which stresses the importance of agents' balance sheets consistency. This approach has been adopted while designing a highly complex economic model, like the EURACE model, and has proved to be a very useful theoretical tool.

The EURACE simulator is the outcome of a three years project, started in September 2006 with the aim to realize an agent-based macroeconomic design and simulation platform. It is a fully-specified agent-based model of a

Andrea Teglio
Departament d'Economia, Universitat Jaume I, Castellón de la Plana, Spain
e-mail: teglio@eco.uji.es

Marco Raberto
School of Science and Engineering, Reykjavik University, 101 Reykjavik, Iceland
e-mail: raberto@ru.is

Silvano Cincotti
DIBE-CINEF, Università di Genova, 16145 Genova, Italy
e-mail: silvano.cincotti@unige.it

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 603–610.
springerlink.com

complete economy that integrates many different sectors and markets, whose main features are described in Section 2.

The balance sheet approach is alternative to the mainstream paradigm, which is based on the inter-temporal optimization of welfare by individual agents. It introduces a new methodology for studying how institutions (firms, banks, governments and households) create flows of income, expenditure and production together with stocks of assets (including money) and liabilities, thereby determining how the whole economy evolves through time. Indeed, our crucial assumption is that the EURACE model, as any realistic representation of a monetary economy, must be grounded in a fully articulated system of income and flow of funds accounts.

The overall philosophy of the EURACE Project is part of the research program on a Generative Social Science [7, 8], which seeks to explain socio-economic phenomena by constructing artificial societies that generate possible explanations from the bottom-up. The field of agent-based computational economics (ACE) has been characterized by a great deal of development in recent years (see [15] for a recent survey), but the EURACE project has probably been the first successful effort to build a complete economy that integrates all the main markets and economic mechanisms. As a matter of fact, in the last decade there have been many studies regarding finance in the ACE field (see [12] for a review), while others have focused on labour and goods market [14, 16, 6] or industrial organization [11]. However, only a few partial attempts have been made to model a multiple-market economy as a whole [1, 2, 13, 5]. In this respect, the EURACE simulator is certainly more complete, incorporating many crucial connections between real economy and financial markets.

Therefore, the EURACE agent-based framework provides a powerful computational facility where experiments concerning policy design issues can be performed. It offers a realistic environment, characterized by non-clearing markets and bounded rational agents, well suited for studying the out-of equilibrium transitory dynamics of the economy caused by policy parameter changes.

2 The Simulator

The EURACE model represents a fully integrated macroeconomy, consisting of three economic spheres: the real sphere (consumption goods, investment goods, and labour market), the financial sphere (credit and financial markets), and the public sector (Government, Central Bank and Eurostat).

The implementation of the EURACE platform is based on innovative technological instruments, in order to be scalable to a large number of agents. In particular, the software framework is based on FLAME (Flexible Large-scale Agent-based Modelling Environment), a logical communicating extended finite state machine theory (X-machine) which gives the agents more power to enable writing of complex models for large complex systems. The agents are modeled as X-machines allowing them to communicate among each other by broadcasting messages according to a specific model design. This information is automatically read by the FLAME framework and generates a simulation program for efficient execution in a parallel computing environment [10, 3].

FLAME uses X-machines to represent all agents acting in the system. Each agent possesses a set of internal states, transitions functions operating between states, internal memory and a language for broadcasting messages among agents.

Given the complexity of the underlying technological framework and given the considerable extension of the EURACE model, it is not possible to present within this paper an exhaustive explanation of the economic modelling choices, together with a related mathematical or algorithmic description. We will limit the presentation to a general qualitative explanation of the key features of the model.

Full details about the EURACE implementation can be found in [9]. Moreover, when needed, we will cite specific EURACE deliverables. Some general information on EURACE can be found in [4].

We resume in the following some of the main distinguishing features of EURACE.

- Closure: EURACE is one of the very rare fully-specified agent-based models of a complete economy. EURACE is dynamically complete, that is, it specifies all real and financial stocks and flows, allowing us to aggregate upward from the micro-specifications to the macroeconomic variables of interest.
- Encompassing real and financial markets populated by economic interacting agents.
- Wide use of empirically documented behavioural rules.
- Different levels of time and space granularity. It is possible to investigate the impact of real-life granularity on the economic outcomes, and to analyse the consequences of a modification of this granularity.
- Treatment of time: asynchronous decision-making across different agents.
- Explicit spatial structure, allowing to take into account not only regional and land-use aspects, but also more generally the fact that all human activities are localized in geographical space.
- Evolving social network structure linking the different agents.
- Very large number of agents, possibly allowing to discover emerging phenomena and/or rare events that would not occur with a smaller population.
- Use and development of innovative software frameworks, code parallelization in order to employ super-computers, allowing very large-scale simulations.
- Calibration on European economic data and focus on European policy analysis.
- Use of a balance sheet approach as a modelling paradigm.

In the next section we describe the last feature of this list, stressing the importance of using a balance sheet approach as a modelling paradigm for agent-based computational economics.

3 The Balance Sheet Approach

In the Eurace model, a double-entry balance sheet with a detailed account of all monetary and real assets as well as monetary liabilities is defined for each agent. Monetary and real flows, given by agents' behaviors and interactions, determine the period by period balance sheet dynamics. A stock-flow model is then created and used to check that all monetary and real flows are accounted for, and that all changes to stock variables are consistent with these flows. This provides us with a solid and economically well-founded methodology to test the consistency of the model.

In order to explain our approach, let us consider the balance sheets of the different agents of the model.

Household's balance sheet is reported in Table 1. Its financial wealth is given by

$$W = M^h + \sum_{f \in \{firms\}} n_f^h p_f + \sum_{b \in \{banks\}} n_b^h p_b + \sum_{g \in \{governments\}} n_g^h p_g$$

where p_f, p_b are daily prices of equity shares issued by firm f and bank b, respectively; while p_g is the daily price of the bond issued by government g.

Table 1 Household (H): balance sheet overview

Assets	**Liabilities**
M^h: liquidity deposited at a given *bank* n_g^h: government bonds holdings n_f^h, n_b^h: equity shares holdings of firm f and bank b	(none)

Table 2 Firm (f): balance sheet overview

Assets	**Liabilities**
M^f: liquidity deposited at a given *bank*	D_b^f: debts to *banks*
I_m^f: inventories at *malls*	E^f: equity
K^f: physical capital	

Firm's balance sheet is shown in Table 2. M^f and I^f_m are updated daily following firms' cash flows and sales, while K^f and D^f_b are updated updated monthly. The equity E^f is also updated monthly according to the following rule:

$$E^f = M^f + p_C \sum_{m \in \{malls\}} I^f_m + p_K K^f - \sum_{b \in \{banks\}} D^f_b$$

where p_C is the average price level of consumption goods and p_K is the price of capital goods.

Table 3 reports the balance sheet of the bank. M^b_h, M^b_f, L^b_f are updated daily following the private sector deposits changes and the credit market outcomes. M^b and E^b are updated daily following banks cash flows and keeping into account the balance constraint:

$$M^b = D^b + \sum_{h \in \{households\}} M^b_h + \sum_{f \in \{firms\}} M^b_f + E^b - \sum_{f \in \{firms\}} L^b_f$$

If M^b becomes negative, D^b is increased to set $M^b = 0$. If both M^b and D^b are positive, D^b is partially or totally repaid.

Table 3 Bank (b): balance sheet overview

Assets	**Liabilities**
M^b: liquidity (reserves) deposited at the *central bank*	D^b: standing facility (debts to the *central bank*)
L^b_f: loans to firms	M^b_h: households' deposits at the bank
	M^b_f: firms' deposits at the bank
	E^b: equity

In order to understand the functioning of money creation, circulation and destruction in EURACE, we first need to explain the outlay of bank's balance sheet.

Let's start with the money creation issue: four channels of money formation are open. The first, and most important one, activates when banks grant loans to firms, and new money (M1) appears in the form of firm's increased payment account (and, thus, increased deposits). The second channel operates when the central bank is financing commercial banks through lending of last resort, and money creation (Fiat money) translates in augmented bank's reserves. Government Bond issuing constitutes the third channel: it is at work whenever the quantitative easing (QE) feature is active, allowing the CB to buy government bonds in the financial market. Finally, the fourth and last channel is represented by bailouts of commercial banks by the CB.

Table 4 Government (g): balance sheet overview

Assets	**Liabilities**
M^g: liquidity deposited at the central bank	D^g: standing facility with the central bank
	n^g: number of outstanding bonds

So far we have dealt with money creation, let us now comment money circulation and destruction. Since there is no currency, that is no money is present outside the banking system, when agents (firms, households or Government) use their liquid assets to settle in favor of other agents, money should simply flow from payer's bank account to taker's bank account, obviously keeping itself constant (such cash movements are accounted at the end of the day, when agents communicate to banks all their payments). On the contrary, whenever a debt is repaid, money stock has to decrease accordingly. For technical details and a more exhaustive discussion on these issues, see [9].

Finally, the balance sheets of the government and of the central bank are reported in Tables 4 and 5, respectively.

The government budget is composed by taxes on corporate profits, household labor and capital income, as revenues, and unemployment benefits, transfer and subsidies, as expenses.

Since the Central Bank is not allowed to make a profit, its revenues from government bonds and bank advances are distributed to the government in the form of a dividend. In case of multiple governments, the total dividend payment is equally divided among the different governments.

These modelling hypothesis lead to the definition of a precise "EURACE time invariant" feature, consisting in a fundamental macroeconomic accounting identity:

$$\underbrace{\Delta\left(\sum_h M^h + \sum_f M^f\right)}_{\text{private sector deposits}} + \underbrace{\Delta\left(\sum_b E^b\right)}_{\text{banks equity}} + \underbrace{\Delta\left(\sum_g M^g + M^c\right)}_{\text{public sector deposits}} =$$

$$\underbrace{\Delta\left(M^c + \sum_b L_b^c + \sum_g L_g^c\right)}_{\text{fiat money}} + \underbrace{\Delta\left(\sum_b \sum_f L_f^b\right)}_{\text{credit money}}$$

This accounting identity ensures the coherence of the aggregate stock-flow in the EURACE model. For policy considerations, it is clearly important to consider the monetary endowment of agents in the private sector, i.e.,

$$\sum_h M^h + \sum_f M^f + \sum_b E^b$$

Table 5 Central Bank (c): balance sheet overview

Assets	**Liabilities**
n_g^c: Government bonds (QE)	M^c: fiat money due to QE
M^c: liquidity	M_g^c: Governments liquidity
L_b^c: loans to banks	M_b^c: banks reserves
L_g^c: loans to governments	E^c: equity

A higher monetary endowment due, e.g., to a loose fiscal policy and QE, leads to a higher nominal demand. Depending on the behavior of prices, the higher nominal demand could translate into a higher real demand.

4 Conclusions

This paper describes how the EURACE modeling approach is based on balance sheets. This approach guarantees several advantages in the agent-based economics computational framework. From the point of view of model validation, considering aggregate balance sheets allows to monitor stock flows, checking the presence of conceptual errors. Indeed, the agent-based approach is characterized by modelling the behavior of the single agent, independently from the aggregate behavior, which is achieved by the balance sheet which depends on the interaction of many different agents and therefore represents the aggregate vision. Moreover, features that are not included in single agents and that emerge from the bottom-up can be detected and analyzed by looking at aggregate balance sheets. Network relations, for instance, can be described by the balance sheets structure. Finally, the recent crises showed how shock propagation and financial fragility depend on balance sheets. As a conclusion, we think that the balance sheet approach should be the standard approach for every agent-based economic model.

Acknowledgements. This work was carried out in conjunction with the EURACE project (EU IST FP6 STREP grant: 035086) which is a collaboration lead by S. Cincotti (Università di Genova), H Dawid (Universitaet Bielefeld), C. Deissenberg (Université de la Mediterranee), K. Erkan (TUBITAK-UEKAE National Research Institute of Electronics and Cryptology), M. Gallegati (Universitţa Politecnica delle Marche), Holcombe (University of Sheffield), M. Marchesi (Universitţa di Cagliari), C. Greenough (Science and Technology Facilities Council Rutherford Appleton Laboratory).

References

1. Basu, N., Pryor, R., Quint, T.: ASPEN: a microsimulation model of the economy. Comput. Econ. 12(3), 223–241 (1998)
2. Bruun, C.: Agent-based Keynesian economics: simulating a monetary production system bottom-up, University of Aalborg, Denmark (1999)

3. Coakley, S., Kiran, M.: EURACE Report D1.1: X-Agent framework and software environment for agent-based models in economics. Department of Computer Science, University of Sheffield, UK (2007)
4. Deissenberg, C., van der Hoog, S., Dawid, H.: EURACE: A Massively Parallel Agent-based Model of the European Economy. Appl. Math. Comput. 204, 541–552 (2008)
5. Dosi, G., Fagiolo, G., Roventini, A.: Schumpeter Meeting Keynes: A Policy-Friendly Model of Endogenous Growth and Business Cycles. J. Econ. Dynam. Control (in press, 2010)
6. Dosi, G., Fagiolo, G., Roventini, A.: The Microfoundations of Business Cycles: An Evolutionary, Multi-Agent Model. J. Evol. Econ. 18(3-4), 413–432 (2008)
7. Epstein, J.M.: Agent-Based Computational Models And Generative Social Science. Complexity 4(5), 41–60 (1999)
8. Epstein, J.M., Axtell, R.L.: Growing Artificial Societies: Social Science from the Bottom Up (Complex Adaptive Systems). MIT Press, Cambridge (1996)
9. EURACE Final Activity Report (2009), http://www.eurace.org/
10. Holcombe, M., Coakley, S., Smallwood, R.: A General Framework for Agent-based Modelling of Complex Systems. EURACE Working paper WP1.1, Department of Computer Science, University of Sheffield, UK (2006)
11. Kutschinski, E., Uthmann, T., Polani, D.: Learning Competitive Pricing Strategies by Multi-Agent Reinforcement Learning. J. Econometrics 27, 2207–2218 (2001)
12. LeBaron, B.D.: Agent-based computational finance. In: Tesfatsion, L., Judd, K. (eds.) Handbook of Computational Economics, North Holland, Amsterdam (2006)
13. Sallans, B., Pfister, A., Karatzoglou, A., Dorffner, G.: Simulations and validation of an integrated markets model. J. Artif. Soc. Social Simul. 6(4) (2003), http://jasss.soc.surrey.ac.uk/6/4/2.html
14. Tassier, T.: Emerging small-world referral networks in evolutionary labor markets. IEEE Trans. Evol. Comput. 5(5), 482–492 (2001)
15. Tesfatsion, L., Judd, K.: Agent-Based Computational Economics. North Holland, Amsterdam (2006)
16. Tesfatsion, L.: Structure, behaviour, and market power in an evolutionary labour market with adaptive search. J. Econom. Dynam. Control 25, 419–457 (2001)

Connections between Statistical Depth Functions and Fuzzy Sets

Pedro Terán

Abstract. We show that two probabilistic interpretations of fuzzy sets via random sets and large deviation principles have a common feature: they regard the fuzzy set as a depth function of a random object. Conversely, some depth functions in the literature can be regarded as the fuzzy sets of central points of appropriately chosen random sets.

1 Statistical Depth Functions

Depth functions try to order, in a center-outward sense, the points in a data set or the whole $\mathbf{R}^d$ with respect to a given probability distribution. That allows one to extend concepts from univariate statistics to the multivariate setting. A general reference on depth functions is [5], see also [8]. Applications can be found in statistical problems like classification, outlier detection, exploratory analysis, measures of multivariate scatter, skewness and kurtosis, and so on.

Zuo and Serfling [15] distilled the following four *desirable* properties from various particular cases in the literature: (a) Depth is maximal at a center of symmetry of the distribution, if the latter exists. (b) Depth decreases along any ray departing from the center. (c) Depth vanishes as the distance to the center goes to infinity. (d) The depth function should be affinely equivariant (so that conclusions do not depend on the chosen coordinate system).

However, one should be warned that abundant examples exist of depth functions failing one or another of those properties (e.g. L^p-depth, defined in that very paper by Zuo and Serfling themselves).

Pedro Terán
Escuela de Ingeniería Técnica Industrial, Departamento de Estadística
e Investigación Operativa y Didáctica de la Matemática, Universidad de Oviedo,
E-33071 Gijón, Spain
e-mail: teranpedro@uniovi.es

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 611–618.
springerlink.com

Since depth functions are $[0,1]$-valued, superficially they look quite like a fuzzy set measuring the degree of 'deepness' of each point of $\mathbf{R}^d$ in the probability distribution. Our aim is to pursue, at a more rigorous level, the connections between fuzzy sets and depth functions.

2 Fuzzy Sets and Statistical Depth Functions

Let us show that the two main probability-theoretical interpretations of fuzzy sets can be interpreted as depth functions. Yet a third view of fuzzy sets as depth functions generated by random objects appears when the given fuzzy set is obtained as the expectation of a fuzzy random variable.

2.1 Fuzzy Sets as Coverage Functions of Random Sets

The first interpretation, dating back to the late seventies and variously attributed to Goodman and Nguyen, is the coverage function of a random set. If X is a random set, its *coverage function* is the function given by

$$p_X(x) = P(x \in X),$$

which generalizes the probability mass function of a random variable.

For any random set, p_X is $[0,1]$-valued and so can be regarded as a fuzzy set. If X is almost surely closed, then p_X is upper semicontinuous. Conversely, for any upper semicontinuous fuzzy set A there is a canonical random set $L_A : \alpha \in [0,1] \mapsto A_\alpha$ (mapping each α to the α-cut of A) whose coverage function is that fuzzy set.

Is p_X a depth function? Properties (a) through (d) of depth functions are devised for random variables and it is not too easy to translate them literally to general random sets. For instance, the sense of properties like (b) and (c) depends implicitly on the fact that single points are bounded and convex, which no longer holds true for arbitrary sets. One possibility is to show that p_X arises as a particular case of a general method for constructing depth functions. A second possibility is to restrict our analysis to random sets which preserve some special properties of points.

Cascos and López–Díaz [2] presented an abstract approach to depth functions for random variables, extended in [14] in a way which allows one to define depth for random sets and other imprecise probability models. In the latter, an *integral depth function* is defined as follows: given a specified family of functions $\mathscr{F}$, the integral depth of a point x in a random set X is given by

$$d_{\mathscr{F}}(x;X) = \sup\{\alpha \in (0,1] \mid \forall f \in \mathscr{F}, f(x) \le \sup_{Q \le \alpha^{-1} P \text{ for some } P \in \mathrm{Sel}(X)} \int f \mathrm{d}Q\},$$

where $\mathrm{Sel}(X)$ is the set of all selectionable distributions of X.

Proposition 1. *Let X be a random closed set. Then, p_X is an integral depth function with respect to the family $\mathscr{F}$ of all indicator functions of singletons.*

Proof. By definition,

$$d_{\mathscr{F}}(x;X) = \sup\{\alpha \in (0,1] \mid \forall a \in \mathbf{R}^d, I_{\{a\}}(x) \le \sup_{Q \le \alpha^{-1}P, P \in \mathrm{Sel}(X)} \int I_{\{a\}} \mathrm{d}Q\}$$

$$= \sup\{\alpha \in (0,1] \mid \sup_{Q \le \alpha^{-1}P, P \in \mathrm{Sel}(X)} Q(\{x\}) = 1\}.$$

For any α satisfying the condition in the right-hand side, one finds probability distributions Q_n and P_n under the following restrictions:

$$Q_n \le \alpha^{-1}P_n, \quad P_n \in \mathrm{Sel}(X), \quad Q_n(\{x\}) \to 1.$$

But then

$$p_X(x) = P(x \in X) \ge \sup_n P_n(\{x\}) \ge \alpha \sup_n Q_n(\{x\}) = \alpha.$$

Since $d_{\mathscr{F}}(x;X)$ is the supremum of all those α, we have $d_{\mathscr{F}}(x;X) \le p_X(x)$. To show the converse inequality, we need to check the identity

$$\sup_{Q \le p_X(x)^{-1}P, P \in \mathrm{Sel}(X)} Q(\{x\}) = 1.$$

The only distribution Q where the supremum might be reached is the Dirac distribution δ_x at x. To show that

$$\delta_x \le p_X(x)^{-1}P \text{ for some } P \in \mathrm{Sel}(X),$$

we just need to find $P \in \mathrm{Sel}(X)$ such that $p_X(x) \le P(\{x\})$. The Measurable Selection Theorem yields a selection ξ of X. Define $\eta = x \cdot I_{x \in X} + \xi \cdot I_{x \notin X}$. The mapping η is still a selection of X, but $P(\eta = x) = P(x \in X) = p_X(x)$, so it suffices to take P to be the distribution of η. That concludes the proof. □

For the second, direct approach we assume the random set is almost surely convex. That corresponds to convex fuzzy sets.

Proposition 2. *Let X be a random closed convex set. Then, p_X is a depth function satisfying properties (a), (b), and (d). If X is almost surely bounded, then property (c) holds as well.*

Proof. Proof of (a): Assume that X is centrally symmetric with respect to some point x, namely $x + u \in X$ if and only $x - u \in X$. Since X is almost surely non-empty, for any fixed $\omega \in \Omega$ there is some u such that $x + u \in X(\omega)$. Taking into account the symmetry, $x - u \in X(\omega)$ as well, so

$$x = .5(x+u) + .5(x-u) \in X(\omega)$$

by the convexity of X. Since ω was arbitrary, we have

$$p_X(x) = P(x \in X) = 1$$

so indeed x maximizes p_X.

Proof of (b): Let x be the center of symmetry in part (a). We need to show

$$p_X(x+tu) \geq p_X(x+u)$$

for any arbitrary $t > 1, u \in \mathbf{R}^d$. As shown above, $x \in X$ almost surely. If $x+tu \in X$ then, by the convexity of X, also

$$x+u = (1-t^{-1})x + t^{-1}(x+tu) \in X.$$

Thus, $\{x+tu \in X\} \subset \{x+u \in X\}$.

Proof of (c): Under the boundedness assumption, the random variable $\|X\| = \sup_{y \in X} |y|$ is almost surely finite. For any sequence $\{y_m\}_m$ with $|y_m| \to \infty$,

$$p_X(y_m) = P(y_m \in X) \leq P(\|X\| \geq |y_m|) \to 0$$

since

$$\bigcap_m \{\|X\| \geq |y_m|\} = \emptyset.$$

Proof of (d): Let A be a regular matrix of size d, and let $b \in \mathbf{R}^d$. Since the transformation $\phi : x \mapsto Ax+b$ is bijective,

$$p_{\phi(X)}(\phi(x)) = P(\phi(x) \in \phi(X)) = P(x \in X) = p_X(x). \qquad \square$$

2.2 Fuzzy Sets as Large Deviation Limits

The second interpretation is due to Nguyen and Bouchon-Meunier [10] and relies on the so-called 'idempotent probability' approach to large deviation problems in probability theory [11]. For a given probability distribution, take a sequence of i.i.d. random vectors $\{\xi_n\}_n$. The weak law of large numbers gives conditions under which the sample average $S_n = n^{-1}\sum_{i=1}^n \xi_i$ converges in probability. The question is how likely it is that S_n remains far from its limit as n progresses (hence the name 'large deviations'). Under appropriate assumptions on ξ, that probability decreases exponentially. Moreover, there exists a non-negative, lower semicontinuous rate function I_ξ such that

$$\liminf_n n^{-1} \log P(S_n \in A) \geq -\inf_{x \in \operatorname{int} A} I_\xi(x)$$

and

$$\limsup_n n^{-1} \log P(S_n \in A) \leq -\inf_{x \in \operatorname{cl} A} I_\xi(x),$$

namely ξ satisfies a large deviation principle. It must be noted that the function $J_\xi = \exp(-I_\xi)$ is $[0,1]$-valued and it is very natural to interpret it as the possibility distribution of a possibility measure $\mathit{\Pi}_\xi$. Indeed, it follows easily from the inequalities above that

$$\liminf_n P(S_n \in G)^{1/n} \geq \mathit{\Pi}_\xi(G)$$

and

$$\limsup_n P(S_n \in F)^{1/n} \leq \mathit{\Pi}_\xi(F)$$

for any open G and closed F, whence $\mathit{\Pi}_\xi$ is in a sense, very close to convergence in distribution, the limit of the set functions $(P_{S_n})^{1/n}$ (see [9, 11]). Therefore, this probabilistic semantics of possibility measures and distributions yields a probabilistic interpretation of the fuzzy set J_ξ.

Proposition 3. *Let ξ be an integrable random vector satisfying the large deviation principle. Then, J_ξ is a depth function in the sense of properties (a) through (d) above, taking its maximum value 1 at the expectation of ξ.*

Proof. Proof of (a): The expectation is the center of symmetry of the random vector ξ for some definitions of symmetry, e.g. when $\xi - x$ and $x - \xi$ are identically distributed for some x.

Therefore, it suffices to show that indeed the expectation $E\xi$ maximizes J_ξ. Note that, for each $\varepsilon > 0$,

$$\sup J_\xi(E\xi + \varepsilon B) \geq \liminf_n P(S_n \in E\xi + \varepsilon B)^{1/n}$$

$$= \liminf_n P(|S_n - E\xi| \leq \varepsilon)^{1/n} \geq \liminf_n P(|S_n - E\xi| \leq \varepsilon) = 1.$$

Taking a decreasing sequence $\varepsilon_n \searrow 0$, we find a sequence $x_n \to E\xi$ with $J_\xi(x_n) \to 1$. Since J_ξ is upper semicontinuous,

$$J_\xi(E\xi) \geq \limsup_n J_\xi(x_n) \to 1.$$

Since J_ξ is $[0,1]$-valued, clearly $E\xi$ maximizes J_ξ.

Proof of (b): Let us show that J_ξ is quasiconcave. Since

$$\{x \mid J_\xi(x) \geq \alpha\} = \{x \mid e^{-I_\xi(x)} \geq \alpha\} = \{x \mid I_\xi(x) \leq -\ln\alpha\},$$

but the latter level sets are convex since the function I_ξ is convex [3, Theorem 2.2.30 and Lemma 2.2.31].

Now, using the quasiconcavity, for any $t > 1$,

$$J_\xi(E\xi + \lambda) = J_\xi\left((1 - t^{-1})E\xi + t^{-1}(E\xi + t\lambda)\right) \geq \min\{J_\xi(E\xi), J_\xi(E\xi + t\lambda)\}$$

$$= \min\{1, J_\xi(E\xi + t\lambda)\} = J_\xi(E\xi + t\lambda).$$

Proof of (c): Since I_ξ is a good rate function, the level sets $\{x \mid I_\xi \leq t\}$ are compact [3, Theorem 2.2.30 and Lemma 2.2.31]. Reasoning like before, the α-cuts of J_ξ are compact as well.

Let $\{y_m\}_m$ be a sequence such that $|y_m| \to \infty$. Clearly, for each $\varepsilon \in (0,1]$ we eventually have $y_m \notin (J_\xi)_\varepsilon$, and so $J_\xi(y_m) \leq \varepsilon$.

Proof of (d): Let A be a regular matrix of size d, and let $b \in \mathbf{R}^d$. Since the transformation $\phi : x \mapsto Ax + b$ is continuous and bijective, the equivariance property follows from a direct application of the contraction principle [3, Theorem 4.2.1]. □

The conclusion is that both probabilistic interpretations are very different (based on random sets vs. limit theorems) but both interpret fuzzy sets as depth functions of certain random objects.

3 Statistical Depth Functions and Fuzzy Sets

It is also possible to take the opposite viewpoint: instead of 'explaining' fuzzy sets as depth functions, to explain some depth functions as special fuzzy sets.

Recently [12, 13] we introduced a gradual notion of centrality for location estimation, based on defining a *fuzzy set of central points*. Given a specified family of fuzzy events $\mathscr{A}$, the degree of centrality of a point x in a family of distribution $\mathscr{P}$ is defined to be

$$C(x) = \sup\{\alpha \in (0,1] \mid \forall A \in \mathscr{A}, \sup_{P \in \mathscr{P}} P(A) \geq \alpha A(x)\}.$$

We begin by showing that coverage functions of random sets are special cases of fuzzy sets of central points.

Proposition 4. *Let X be a random closed set. Then, p_X is the fuzzy set of central points of the family of all selectionable distributions of X with respect to the family $\mathscr{A}$ of all indicator functions of singletons.*

Proof. Taking into account Proposition 1, it suffices to prove

$$d_{\mathscr{F}}(x;X) = C(x)$$

for the choice $\mathscr{F} = \mathscr{A}$. Observe that both definitions are similar, though not exactly the same. While the definition of C involves the condition

$$\sup_{P \in \mathrm{Sel}(X)} P(A) \geq \alpha A(x), \ \forall A \in \mathscr{A},$$

that of $d_{\mathscr{F}}$ translates into

$$\sup_{Q \leq \alpha^{-1} P \text{ for some } P \in \mathrm{Sel}(X)} P(A) \geq A(x), \ \forall A \in \mathscr{A}.$$

Replacing A by the indicator functions of all singletons, we must compare conditions

$$\sup_{P \in \mathrm{Sel}(X)} P(\{x\}) \geq \alpha \qquad (1)$$

and

$$\sup_{Q \leq \alpha^{-1} P, P \in \mathrm{Sel}(X)} Q(\{x\}) \geq 1 \qquad (2)$$

Take $x \in \mathbf{R}^d$ such that (1) holds. There exist $P_n \in \mathrm{Sel}(X)$ such that $P_n(\{x\}) \geq (1-n^{-1})\alpha$. For each n, let ξ_n be a selection of X having law P_n, and define a random variable η_n which equals x with probability $\min\{\alpha^{-1}P_n(\{x\}),1\}$ and equals ξ_n otherwise. Let Q_n be the law of η_n. Then,

$$Q_n \leq \alpha^{-1}P_n, \quad Q_n(\{x\}) \geq 1-n^{-1}, \quad P_n \in \mathrm{Sel}(X),$$

so

$$\sup_{Q \leq \alpha^{-1}P, P \in \mathrm{Sel}(X)} Q(\{x\}) \geq \sup_n Q_n(\{x\}) = 1.$$

Conversely, if (2) holds, there exist appropriate distributions Q_n, P_n with $Q_n \leq \alpha^{-1}P_n$ and $Q_n(\{x\}) \to 1$. Thus,

$$P_n(\{x\}) \geq \alpha Q_n(\{x\}) \to \alpha$$

and

$$\sup_{P \in \mathrm{Sel}(X)} P(\{x\}) \geq \sup_n P_n(\{x\}) \geq \alpha.$$

□

As a consequence, some statistical depth functions in the literature can easily be shown to be fuzzy sets of central points.

Let $\{x_1,\ldots,x_n\}$ be a data sample. Convex hull peeling [4, 1] proceeds by taking the convex hull C_1 of the sample and finding its vertices V_1. Those vertices are 'peeled' away and we go on iteratively defining $C_2 = \mathrm{co}(C_1 \setminus V_1)$, V_2 its set of vertices and so on. The depth of a point x in the sample is then defined to be proportional to the number of data points that need to be peeled away so as to leave x out of the convex hull of the remaining data, namely

$$d_{CHP}(x) = \sum_i \frac{\mathrm{card}\,(V_i)}{n} \cdot I_{\{x \in C_i\}}.$$

The simplicial depth [6] of a point x in the distribution of a random vector ξ is defined to be the probability that x is in the simplex generated by $d+1$ independent observations of ξ.

The majority depth [7] of x is defined to be the probability that x is in the *major side* of d independent observations of ξ, that is, the P-largest halfspace having $\xi_1,\ldots,\xi_d$ in its boundary.

Proposition 5. *Let ξ be a random vector. For appropriate choices of $\mathscr{A}$ and $\mathscr{P}$, the corresponding fuzzy set of central points is:*

(1)The convex hull peeling depth function.
(2)The simplicial depth function.
(3)The majority depth function.

Proof. Proof of (1): The depth function is the coverage function of the random set which takes on values C_i with probabilities $n^{-1}\mathrm{card}\,(V_i)$.

Proof of (2): The simplicial depth function is the coverage function of the random set $\mathrm{co}\{\xi_1,\dots,\xi_{d+1}\}$, where the ξ_i are $d+1$ independent copies of ξ.

Proof of (3): The majority depth function is the coverage function of the random set $\mathsf{MajorSide}[\xi_1,\dots,\xi_d]$, where the ξ_i are d independent copies of ξ. □

Acknowledgements. Research has been partially funded by Spain's *Ministerio de Ciencia e Innovación* (MTM2008-01519) and Aragón's *Gobierno* (PM2007-034).

References

1. Barnett, V.: The ordering of multivariate data (with discussion). J. Royal Statist. Soc. A 139, 318–352 (1976)
2. Cascos, I., López-Díaz, M.: Integral trimmed regions. J. Multivariate Anal. 96, 404–424 (2005)
3. Dembo, A., Zeitouni, O.: Large deviations. Techniques and applications, 2nd edn. Springer, New York (1998)
4. Huber, P.J.: Robust statistics: a review. Ann. Math. Statist. 43, 1041–1067 (1972)
5. Liu, R., Serfling, R., Souvaine, D.: Data depth: robust multivariate analysis, computational geometry and applications. Amer. Math. Soc., Providence (2006)
6. Liu, R.Y.: On a notion of data depth based on random simplices. Ann. Statist. 18, 405–414 (1990)
7. Liu, R.Y., Singh, K.: A quality index based on data depth and multivariate rank tests. J. Amer. Statist. Assoc. 88, 252–260 (1993)
8. Mosler, K.: Multivariate dispersion, central regions and depth. In: The lift zonoid approach. Lect. Notes Statist, vol. 165. Springer, Berlin (2002)
9. Norberg, T.: Random capacities and their distributions. Probab. Theory Related Fields 73, 281–297 (1986)
10. Nguyen, H.T., Bouchon-Meunier, B.: Random sets and large deviations principle as a foundation for possibility measures. Soft Computing 8, 61–70 (2003)
11. Puhalski, A.: Large deviations and idempotent probabilities. Chapman and Hall/CRC, Boca Raton (2001)
12. Terán, P.: A new bridge between fuzzy sets and statistics. In: Proceedings of the Joint 2009 Int. Fuzzy Systems Assoc. World Congress and 2009 Eur. Soc. Fuzzy Logic and Technology Conference, IFSA-EUSFLAT, Lisbon, Portugal, pp. 1887–1891 (2009)
13. Terán, P.: Centrality as a gradual notion: a new bridge between Fuzzy Sets and Statistics (submitted for publication, 2010)
14. Terán, P.: Integral central regions and integral depth (Unpublished manuscript)
15. Zuo, Y., Serfling, R.: General notions of statistical depth function. Ann. Statist. 28, 461–482 (2000)

An Alternative Approach to Evidential Network Construction

Jiřina Vejnarová

Abstract. We present an alternative approach to belief network construction based on operator of composition of basic assignments. We show that belief networks constructed in this way have similar structural properties to Bayesian networks in contrary to previously proposed directed evidential networks by Ben Yaghlane at al.

1 Introduction

Bayesian networks are at present the most popular representative of so-called graphical Markov models. Therefore it is not surprising that some attempts to construct an analogy of Bayesian networks have also been made in other frameworks as e.g. in possibility theory [4] or evidence theory [3].

In this paper we bring an alternative to [3], which does not seem to us to be satisfactory, as graphical tools well-known from Bayesian networks are used in different sense. Our approach is based on previously introduced operator of composition for basic assignments [7, 6]. The evidential network is reconstructed from the resulting compositional model. We concentrate ourselves to structural properties of the network, the problem of definition of conditional beliefs is not solved here.

The paper is organized as follows. After a brief summary of basic notions from evidence theory (Section 2), in Section 3 we recall the definition of the operator of composition (and its basic properties) and in Section 4 after recalling perfect sequences of basic assignments we present an algorithm for transformation of a perfect sequence into an evidential network. We also demonstrate, through a simple example, in which sense our approach is superior to the previous one [3].

Jiřina Vejnarová
Institute of Information Theory and Automation of the ASCR,
182 08 Prague, Czech Republic
e-mail: vejnar@utia.cas.cz

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 619–626.
springerlink.com

2 Basic Notions

In this section we will briefly recall basic concepts from evidence theory [9] concerning sets, set functions and (conditional) independence.

2.1 Set Projections and Joins

For an index set $N = \{1,2,\dots,n\}$ let $\{X_i\}_{i\in N}$ be a system of variables, each X_i having its values in a finite set $\mathbf{X}_i$. In this paper we will deal with *multidimensional frame of discernment* $\mathbf{X}_N = \mathbf{X}_1 \times \mathbf{X}_2 \times \ldots \times \mathbf{X}_n$, and its *subframes* (for $K \subseteq N$) $\mathbf{X}_K = \times_{i\in K}\mathbf{X}_i$. When dealing with groups of variables on these subframes, X_K will denote a group of variables $\{X_i\}_{i\in K}$ throughout the paper.

A *projection* of $x = (x_1,x_2,\dots,x_n) \in \mathbf{X}_N$ into $\mathbf{X}_K$ will be denoted $x^{\downarrow K}$, i.e., for $K = \{i_1,i_2,\dots,i_k\}$

$$x^{\downarrow K} = (x_{i_1},x_{i_2},\dots,x_{i_k}) \in \mathbf{X}_K.$$

Analogously, for $M \subset K \subseteq N$ and $A \subset \mathbf{X}_K$, $A^{\downarrow M}$ will denote a *projection* of A into $\mathbf{X}_M$:[1]

$$A^{\downarrow M} = \{y \in \mathbf{X}_M \mid \exists x \in A : y = x^{\downarrow M}\}.$$

In addition to the projection, in this text we will also need an opposite operation, which will be called a join. By a *join*[2] of two sets $A \subseteq \mathbf{X}_K$ and $B \subseteq \mathbf{X}_L$ $(K,L \subseteq N)$ we will understand a set

$$A \bowtie B = \{x \in \mathbf{X}_{K\cup L} : x^{\downarrow K} \in A \;\;\&\;\; x^{\downarrow L} \in B\}.$$

Let us note that for any $C \subseteq \mathbf{X}_{K\cup L}$ naturally $C \subseteq C^{\downarrow K} \bowtie C^{\downarrow L}$, but generally $C \neq C^{\downarrow K} \bowtie C^{\downarrow L}$.

2.2 Set Functions

In evidence theory [9] (or Dempster-Shafer theory) two measures are used to model the uncertainty: belief and plausibility measures (the latter one will not be used in this paper). Both of them can be defined with the help of another set function called a *basic (probability* or *belief) assignment* m on $\mathbf{X}_N$, i.e.,

$$m : \mathscr{P}(\mathbf{X}_N) \longrightarrow [0,1],$$

where $\mathscr{P}(\mathbf{X}_N)$ is power set of $\mathbf{X}_N$ and $\sum_{A\subseteq\mathbf{X}_N} m(A) = 1$. Furthermore, we assume that $m(\emptyset) = 0$. A set $A \in \mathscr{P}(\mathbf{X}_N)$ is a *focal element* if $m(A) > 0$.

Belief measure is defined for any $A \subseteq \mathbf{X}_N$ by the equality

$$Bel(A) = \sum_{B\subseteq A} m(B). \tag{1}$$

[1] Let us remark that we do not exclude situations when $M = \emptyset$. In this case $A^{\downarrow\emptyset} = \emptyset$.

[2] This term and notation are taken from the theory of relational databases [1].

For a basic assignment m on $\mathbf{X}_K$ and $M \subset K$, a *marginal basic assignment* of m on $\mathbf{X}_M$ is defined (for each $A \subseteq \mathbf{X}_M$):

$$m^{\downarrow M}(A) = \sum_{B \subseteq \mathbf{X}_K : B^{\downarrow M} = A} m(B).$$

Having two basic assignments m_1 and m_2 on $\mathbf{X}_K$ and $\mathbf{X}_L$, respectively ($K, L \subseteq N$), we say that these assignments are *projective* if

$$m_1^{\downarrow K \cap L} = m_2^{\downarrow K \cap L},$$

which occurs if and only if there exists a basic assignment m on $\mathbf{X}_{K \cup L}$ such that both m_1 and m_2 are marginal assignments of m. Let us note that according to the convention $m^{\downarrow \emptyset} \equiv 1$ for arbitrary basic assignment m, m_1 and m_2 are projective whenever $K \cap L = \emptyset$.

2.3 Independence

When constructing graphical models in any framework, (conditional) independence concept plays an important role. In evidence theory the most common notion of independence is that of random set independence [5]: Let m be a basic assignment on $\mathbf{X}_N$ and $K, L \subset N$ be disjoint. We say that groups of variables X_K and X_L are *independent with respect to basic assignment m* (in notation $K \perp\!\!\!\perp L\ [m]$) if

$$m^{\downarrow K \cup L}(A) = m^{\downarrow K}(A^{\downarrow K}) \cdot m^{\downarrow L}(A^{\downarrow L})$$

for all $A \subseteq \mathbf{X}_{K \cup L}$ for which $A = A^{\downarrow K} \times A^{\downarrow L}$, and $m(A) = 0$ otherwise.

This notion can be generalized in various ways [10, 2, 11]; the concept of conditional non-interactivity $X_K \perp_m X_L | X_M$ from [2], based on conjunction combination rule, is used for construction of directed evidential networks in [3]. In this paper we will use the concept introduced in [11, 6], as we consider it more suitable (the arguments can be found in [11]).

Definition 1. Let m be a basic assignment on $\mathbf{X}_N$ and $K, L, M \subset N$ be disjoint, $K \neq \emptyset \neq L$. We say that groups of variables X_K and X_L are *conditionally independent given X_M with respect to m* (and denote it by $K \perp\!\!\!\perp L | M\ [m]$), if the equality

$$m^{\downarrow K \cup L \cup M}(A) \cdot m^{\downarrow M}(A^{\downarrow M}) = m^{\downarrow K \cup M}(A^{\downarrow K \cup M}) \cdot m^{\downarrow L \cup M}(A^{\downarrow L \cup M}) \qquad (2)$$

holds for any $A \subseteq \mathbf{X}_{K \cup L \cup M}$ such that $A = A^{\downarrow K \cup M} \bowtie A^{\downarrow L \cup M}$, and $m(A) = 0$ otherwise.

It has been proven in [11] that this conditional independence concept satisfies so-called semi-graphoid properties taken as reasonable to be valid for any conditional independence concept (see e.g. [8]).

3 Operator of Composition and Its Basic Properties

Operator of composition of basic assignments was introduced in [7] in the following way.

Definition 2. *For two arbitrary basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ a* composition *$m_1 \triangleright m_2$ is defined for all $C \subseteq \mathbf{X}_{K\cup L}$ by one of the following expressions:*

[a] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) > 0$ and $C = C^{\downarrow K} \bowtie C^{\downarrow L}$ then

$$(m_1 \triangleright m_2)(C) = \frac{m_1(C^{\downarrow K}) \cdot m_2(C^{\downarrow L})}{m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L})};$$

[b] if $m_2^{\downarrow K\cap L}(C^{\downarrow K\cap L}) = 0$ and $C = C^{\downarrow K} \times \mathbf{X}_{L\setminus K}$ then

$$(m_1 \triangleright m_2)(C) = m_1(C^{\downarrow K});$$

[c] in all other cases

$$(m_1 \triangleright m_2)(C) = 0.$$

Its basic properties are contained in the following lemma proven in [7].

Lemma 1. *For arbitrary two basic assignments m_1 on $\mathbf{X}_K$ and m_2 on $\mathbf{X}_L$ the following properties hold true:*

(i) $m_1 \triangleright m_2$ *is a basic assignment on* $\mathbf{X}_{K\cup L}$*,*
(ii) $(m_1 \triangleright m_2)^{\downarrow K} = m_1$*,*
(iii) $m_1 \triangleright m_2 = m_2 \triangleright m_1 \iff m_1^{\downarrow K\cap L} = m_2^{\downarrow K\cap L}$*.*

From these basic properties one can see that operator of composition is not commutative in general, but it preserves first marginal (in case of projective basic assignments both of them). In both these aspects it differs from conjunctive combination rule. Furthermore, operator of composition is not associative and therefore its iterative applications must be made carefully, as we will see in the next section.

A lot of other properties possessed by the operator of composition can be found in [6, 7], nevertheless here we will confine ourselves to the following theorem (proven in [6]) expressing the relationship between conditional independence and operator of composition.

Theorem 1. *Let m be a joint basic assignment on $\mathbf{X}_M$, $K, L \subseteq M$. Then $(K \setminus L) \perp\!\!\!\perp (L \setminus K) | (K \cap L)$ $[m]$ if and only if*

$$m^{\downarrow K\cup L}(A) = (m^{\downarrow K} \triangleright m^{\downarrow L})(A)$$

for any $A \subseteq \mathbf{X}_{K\cup L}$.

4 Belief Network Generated by a Perfect Sequence

Now, let us consider a system of low-dimensional basic assignments m_1, m_2, …, m_n defined on $\mathbf{X}_{K_1}, \mathbf{X}_{K_2}, \dots, \mathbf{X}_{K_n}$, respectively. Composing them together by multiple application of the operator of composition, one gets multidimensional basic assignments on $\mathbf{X}_{K_1 \cup K_2 \cup \ldots \cup K_n}$. However, since we know that the operator of composition is neither commutative nor associative, we have to properly specify what "composing them together" means.

To avoid using too many parentheses let us make the following convention. Whenever we put down the expression $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$ we will understand that the operator of composition is performed successively from left to right:[3]

$$m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n = (\ldots((m_1 \triangleright m_2) \triangleright m_3) \triangleright \ldots) \triangleright m_n. \tag{3}$$

Therefore, multidimensional model (3) is specified by an ordered sequence of low-dimensional basic assignments — a *generating sequence* $m_1, m_2, \dots, m_n$.

4.1 Perfect Sequences

From the point of view of artificial intelligence models used to represent knowledge in a specific area of interest, a special role is played by the so-called *perfect sequences*, i.e., generating sequences $m_1, m_2, \dots, m_n$, for which

$$\begin{aligned} m_1 \triangleright m_2 &= m_2 \triangleright m_1, \\ m_1 \triangleright m_2 \triangleright m_3 &= m_3 \triangleright (m_1 \triangleright m_2), \\ &\vdots \\ m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n &= m_n \triangleright (m_1 \triangleright \ldots \triangleright m_{n-1}). \end{aligned}$$

The property explaining why we call these sequences "perfect" is expressed by the following assertion proven in [6].

Theorem 2. *A generating sequence $m_1, m_2, \dots, m_n$ is perfect if and only if all $m_1, m_2, \dots, m_n$ are marginal assignments of the multidimensional assignment $m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n$:*

$$(m_1 \triangleright m_2 \triangleright \ldots \triangleright m_n)^{\downarrow K_j} = m_j,$$

for all $j = 1, \dots, n$.

4.2 Reconstruction of a Belief Network

Having a perfect sequence $m_1, m_2, \dots, m_n$ (m_ℓ being the basic assignment of X_{K_ℓ}), we first order all the variables for which at least one of the basic assignments m_ℓ is defined in such a way that first we order (in an arbitrary

[3] Naturally, if we want to change the ordering in which the operators are to be performed we will do so using parentheses.

way) variables for which m_1 is defined, then variables from m_2 which are not contained in m_1, etc.[4] Finally we have

$$\{X_1, X_2, X_3, \ldots, X_k\} = \{X_i\}_{i \in K_1 \cup \ldots \cup K_n}.$$

Then we get a graph of the constructed belief network in the following way:

1. the nodes are all the variables $X_1, X_2, X_3, \ldots, X_k$;
2. there is an edge $(X_i \rightarrow X_j)$ if there exists a basic assignment m_ℓ such that both $i, j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ and either $i \in K_1 \cup \ldots \cup K_{\ell-1}$ or $i < j$.

Evidently, for each j the requirement $j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$ is met exactly for one $\ell \in \{1, \ldots, n\}$. It means that all the parents of node X_j must be from the respective set $\{X_i\}_{i \in K_\ell}$ and therefore the necessary conditional belief function $Bel(X_j | X_{pa(j)})$ can easily be computed from basic assignment m_ℓ via (1) and some (not yet specified) conditioning rule. As far as we know, the use of a conditioning rule is still not fixed in evidence theory, and therefore we leave this question open for the present.

It is also evident, that if both i and j are in the same basic assignment and not in previous ones, then the direction of the arc depends only on the ordering of the variables. This might lead to different independences, nevertheless, the following theorem sets forth that any of them is induced by the perfect sequence.

Theorem 3. *For a belief network defined by the above procedure the following independence statements are satisfied for any $j = 2, \ldots k$:*

$$\{j\} \perp\!\!\!\perp (\{i < j\} \setminus pa(j)) \,|\, pa(j). \tag{4}$$

Proof. Let $j \in K_\ell$, $j \notin K_1 \cup \ldots \cup K_{\ell-1}$. Due to the fact that

$$m_1 \triangleright m_2 \triangleright \ldots \triangleright m_{\ell-1} \triangleright m_\ell = (\cdots (m_1 \triangleright m_2) \triangleright \cdots \triangleright m_{\ell-1}) \triangleright m_\ell$$

and Theorem 1 we have that

$$K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) \perp\!\!\!\perp (K_1 \cup \ldots \cup K_{\ell-1}) \setminus K_\ell \,|\, K_\ell \cap (K_1 \cup \ldots \cup K_{\ell-1}). \tag{5}$$

It is evident that $(K_1 \cup \ldots \cup K_{\ell-1}) \setminus K_\ell = \{i < j\} \setminus pa(j)$, let us denote it by L. Now, there are two possibilities: either $K_\ell \cap (K_1 \cup \ldots \cup K_{\ell-1}) = pa(j)$ (if j does not have any parents appearing first in K_ℓ) or $K_\ell \cap (K_1 \cup \ldots \cup K_{\ell-1}) \subsetneq pa(j)$ (otherwise).

In the first case either $K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) = \{j\}$ and we immediately obtain (4), or $K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) \supsetneq \{j\}$ and (4) follows from (5) due to $K \cup M \perp\!\!\!\perp L | I\ [m] \;\Rightarrow\; K \perp\!\!\!\perp L | I\ [m]$ (following for any mutually disjoint sets I, K, L, M from semi-graphoid properties), where $K = \{j\}, M = K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) \setminus \{j\}$ and $I = K_\ell \cap (K_1 \cup \ldots \cup K_{\ell-1}) = pa(j)$.

In the latter case, we start by application of the implication $K \cup M \perp\!\!\!\perp L | I\ [m] \;\Rightarrow\; K \perp\!\!\!\perp L | M \cup I\ [m]$, whose validity for any mutually disjoint sets

[4] Let us note that variables $X_1, X_2, \ldots, X_k$ may be ordered arbitrarily, nevertheless, for the above ordering proof of Theorem 3 is simpler than in the general case.

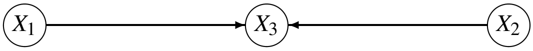
I,K,L,M follows again from semi-graphoid properties, to $K = K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) \setminus \{j\} \setminus pa(j)$, $M = K_\ell \setminus (K_1 \cup \ldots \cup K_{\ell-1}) \cap pa(j)$ and $I = K_\ell \cap (K_1 \cup \ldots \cup K_{\ell-1})$. As $M \cup I = \{i < j\} \setminus pa(j)$ we can then proceed analogous to previous paragraph to obtain (4). □

Let us note that it is different than in the case of directed evidential networks with conditional belief functions introduced in [3], where is no distinction between conditionally and unconditionally independent variables, as the following simple example suggests.

Example 1. Let us consider a sequence of basic assignments m_1, m_2 and m_3, defined on $\mathbf{X}_1, \mathbf{X}_2$ and $\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3$. This sequence need not be perfect, in general, but it is perfect iff

$$m_3^{\downarrow\{1,2\}}(x_1,x_2) = m_1(x_1) \cdot m_2(x_2).$$

This perfect sequence induces independence statements $1 \perp\!\!\!\perp 2$, but generally not $1 \perp\!\!\!\perp 2|3$. Using the above-presented algorithm, we can easily obtain the following graph expressing the same independence statements.



On the other hand, in [3] the same situation is described by $Bel(X_1)$, $Bel(X_2)$, $Bel(X_3|X_1)$ and $Bel(X_3|X_2)$ and the joint belief function is computed using conjunctive combination rule. Therefore, in the resulting model $X_1 \perp_m X_2|X_3$, which corresponds rather to so-called pseudobayesian networks than to Bayesian ones.

5 Conclusions

We introduced an alternative approach to evidential network construction to that presented in [3]. The evidential network is constructed from so-called perfect sequences of basic assignments through a simple transformation algorithm. We proved that the independence relations in the resulting models are analogous to those valid in Bayesian networks, while it does not hold for models introduced in [3]. Due to the limited extent of the paper we are not able to bring more detailed comparison of these two approaches, but we believe that Theorem 3 and Example 1 give the basic idea. Nevertheless, still one substantial problem should be solved — the choice of a proper conditioning rule compatible with (conditional) independence concept used in our models. It will be one of the main goals of our future research.

Acknowledgements. The work of the author was supported by the grant GA ČR 201/09/1891, by the grant GA AV ČR A100750603 and by the grant MŠMT 2C06019.

References

1. Beeri, C., Fagin, R., Maier, D., Yannakakis, M.: On the desirability of acyclic database schemes. J. Assoc. Comput. Mach. 30, 479–513 (1983)
2. Ben Yaghlane, B., Smets, P., Mellouli, K.: Belief functions independence: II. the conditional case. Internat. J. Approx. Reason. 31, 31–75 (2002)
3. Ben Yaghlane, B., Smets, P., Mellouli, K.: Directed evidential networks with conditional belief functions. In: Nielsen, T.D., Zhang, N.L. (eds.) ECSQARU 2003. LNCS (LNAI), vol. 2711, pp. 291–305. Springer, Heidelberg (2003)
4. Benferhat, S., Dubois, D., Gracia, L., Prade, H.: Directed possibilistic graphs and possibilistic logic. In: Bouchon-Meunier, B., Yager, R.R. (eds.) Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, IPMU 1998, Paris, France, pp. 1470–1477 (1998)
5. Couso, I., Moral, S., Walley, P.: Examples of independence for imprecise probabilities. In: de Cooman, G., Cozman, F.G., Moral, S., Walley, P. (eds.) Proceedings of the First International Symposium on Imprecise Probabilities and Their Applications ISIPTA 1999, Ghent, Belgium, pp. 121–130 (1999)
6. Jiroušek, R., Vejnarová, J.: Compositional models and conditional independence in evidence theory. Int. J. Approx. Reasoning (2010) doi:10.1016/j.ijar.2010.02.005
7. Jiroušek, R., Vejnarová, J., Daniel, M.: Compositional models for belief functions. In: de Cooman, G., Vejnarová, J., Zaffalon, M. (eds.) Proceedings of the Fifth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2007, Prague, Czech Republic, pp. 243–252 (2007)
8. Lauritzen, S.L.: Graphical models. Oxford University Press, Oxford (1996)
9. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
10. Shenoy, P.P.: Conditional independence in valuation-based systems. Int. J. Approx. Reasoning 10, 203–234 (1994)
11. Vejnarová, J.: On conditional independence in evidence theory. In: Augustin, T., Coolen, F.P.A., Moral, S., Troffaes, M.C.M. (eds.) Proceedings of the Sixth International Symposium on Imprecise Probability: Theories and Applications, ISIPTA 2009, Durham, UK, pp. 431–440 (2009)

Large Deviations of Random Sets and Random Upper Semicontinuous Functions

Xia Wang and Shoumei Li

Abstract. In this paper, we obtain large deviations (necessary and sufficient conditions) of random sets which take values of bounded closed convex sets on the underling separable Banach space with respect to the Hausdorff distance d_H. We also give necessary and sufficient conditions of large deviations for random upper semicontinuous functions whose values are of bounded closed convex levels on the underling separable Banach space in the sense of the uniform Hausdorff distance d_H^∞. The main tool is the work of Wu on the large deviations for empirical processes [16].

Keywords: Random sets, Random upper semicontinuous functions, Large deviations.

1 Introduction

The theory of large deviation principle (LDP) deals with the asymptotic estimation of probabilities of rare events and provides exponential bound on probability of such events. Some authors have discussed LDP on random sets and random upper semicontinuous functions. In 1999, Cerf [2] proved LDP for sums i.i.d. compact random sets in a separable type p Banach space with respect to the Hausdorff distance d_H, which is called Cramér type LDP. In 2006, Ter*á*n obtained Cramér type LDP of random upper semicontinuous functions whose level sets are compact [14], and Bolthausen type LDP of random upper semicontinuous functions whose level sets are compact convex [15] on a separable Banach space in the sense of the uniform Hausdorff distance d_H^∞. In 2009, Ogura and Setokuchi [12] proved a Cramér type LDP for

Xia Wang and Shoumei Li
Department of Applied Mathematics, Beijing University of Technology,
Beijing 100124, P.R. China
e-mail: wangxia@bjut.edu.cn,lisma@bjut.edu.cn

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 627–634.
springerlink.com

random upper semicontiunous functions on the underling separable Banach space with respect to the metric d_Q (see [12] for the notation) in a different method, which is weaker than the uniform Hausdorff distance d_H^∞. In 2010, Ogura, Li and Wang [11] also discuss LDP for random upper semicontinuous functions whose underlying space is d-dimensional Euclidean space $\mathbb{R}^d$ under various topologies for compact covex random sets and random upper semicontinuous functions. However, previous works in this direction were restricted to compact random sets and compact random upper semicontinuous functions. In this paper, we will obtain LDP for bounded closed convex random sets and related random upper semicontiunous functions. The results of LDP on a Banach space (not necessarily separable) related to this paper come from Wu's paper [16].

The paper is structured as follows. Section 2 will give some preliminaries about bounded closed convex random sets and random upper semicontinuous functions. In Section 3, we will give large deviations (necessary and sufficient conditions) of random sets which take values of bounded closed convex sets on the underling separable Banach space with respect to the Hausdorff distance d_H, and prove that of random upper semicontiunous functions whose values are of bounded closed convex levels on the underling separable Banach space in the sense of the uniform Hausdorff distance d_H^∞.

2 Preliminaries

Throughout this paper, we assume that $(\Omega, \mathscr{A}, P)$ is a complete probability space, $(\mathfrak{X}, \|\cdot\|_{\mathfrak{X}})$ is a real separable Banach space with its dual space $\mathfrak{X}^*$, which is separable with respect to usual norm $\|\cdot\|_{\mathfrak{X}^*}$. $\mathscr{K}(\mathfrak{X})$ is the family of all non-empty closed subsets of $\mathfrak{X}$, $\mathscr{K}_b(\mathfrak{X})$ (resp. $\mathscr{K}_{bc}(\mathfrak{X})$) is the family of all non-empty bounded closed (resp. bounded closed convex) subsets of $\mathfrak{X}$.

Let A and B be two non-empty subsets of $\mathfrak{X}$ and let $\lambda \in \mathbb{R}$, we can define addition and scalar multiplication by $A+B = cl\{a+b : a \in A,\ b \in B\}, \lambda A = \{\lambda a :\ a \in A\}$, where clA is the closure of set A taken in $\mathfrak{X}$. The Hausdorff distance on $\mathscr{K}_b(\mathfrak{X})$ is defined by

$$d_H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}\|a-b\|_{\mathfrak{X}}, \sup_{b\in B}\inf_{a\in A}\|a-b\|_{\mathfrak{X}}\right\}.$$

In particular, we denote $\|A\|_{\mathscr{K}} = d_H(\{0\},A) = \sup_{a\in A}\{\|a\|_{\mathfrak{X}}\}$. Then $(\mathscr{K}_b(\mathfrak{X}), d_H)$ is a complete metric space (see Li, Ogura and Kreinovich [7, p.5 Theorem 1.1.2]).

X is called bounded closed convex random sets, if it is a measurable mapping from the space $(\Omega, \mathscr{A}, P)$ to the space $(\mathscr{K}_{bc}(\mathfrak{X}), \mathfrak{B}(\mathscr{K}_{bc}(\mathfrak{X})))$, where $\mathfrak{B}(\mathscr{K}_{bc}(\mathfrak{X}))$ is the Borel σ-field of $\mathscr{K}_{bc}(\mathfrak{X})$ generated by the Hausdorff distance d_H. The expectation of X denoted by $E[X]$, is defined by $E[X] = cl\{\int_\Omega \xi dP : \xi \in S_X\}$, where $\int_\Omega \xi dP$ is the usual Bochner integral in $L^1[\Omega; \mathfrak{X}]$ (the family of integral

$\mathfrak{X}$-valued random variables), and $S_X = \{\xi \in L^1[\Omega;\mathfrak{X}] : \xi(\omega) \in X(\omega), a.e.\ P\}$. We call $E[X]$ Auman integral (see Auman [1]).

Let S^* be unit sphere of $\mathfrak{X}^*$ with strong topology whose related strong distance is denoted by d_s^*. Since we assume the dual space $\mathfrak{X}^*$ is a separable Banach space, the unite sphere S^* is also separable. Let $D_1 = \{x_1^*, x_2^*, \cdots\}$ be the countable dense subset in the unit sphere S^*. Denote by $C(S^*, d_s^*)$ be space of all continuous functions on S^* with the strong topology with the uniform norm $\|\cdot\|_{C(S^*)}(\|f\|_{C(S^*)} = \sup\{|f(x^*)| : x^* \in S^*\}$, for $f \in C(S^*, d_s^*)$, in fact $\|f\|_{C(S^*)} = \sup\{|f(x^*)| : x^* \in D_1\})$. We know that $C(S^*, d_s^*)$ is a Banach space, and in general it is not separable.

For each $A \in \mathscr{K}_{bc}(\mathfrak{X})$, we define its support function $s(A) :\ S^* \to \mathbb{R}$ as

$$s(A)(x^*) = \sup\{x^*(x) :\ x \in A\}, \quad x^* \in S^*.$$

The mapping $s :\ \mathscr{K}_{bc}(\mathfrak{X}) \to C(S^*, d_s^*)$ has the following properties: for any $A_1, A_2 \in \mathscr{K}_{bc}(\mathfrak{X})$ and $\lambda \in \mathbb{R}^+ = [0, \infty)$, (1) $s(A_1 + A_2) = s(A_1) + s(A_2)$, (2) $s(\lambda A_1) = \lambda s(A_1)$, (3) $d_H(A_1, A_2) = \|s(A_1) - s(A_2)\|_{C(S^*)}$. In fact, the mapping s is an isometric embedding of $(\mathscr{K}_{kc}(\mathfrak{X}), d_H)$ into a closed convex cone of the Banach space $(C(S^*, d_s^*),\ \|\cdot\|_{C(S^*)})$ (see Li, Ogura and Kreinovich [7, p.11 Theorem 1.1.12]).

In the following, we introduce the definition of a random upper semicontinuous function. Let $I = [0,1], I_{0+} = (0,1]$. Let $\mathscr{F}_b(\mathfrak{X})$ denote the family of all functions $u :\ \mathfrak{X} \to I$ satisfying the conditions: (1) the 1-level set $[u]_1 = \{x \in \mathfrak{X} :\ u(x) = 1\} \neq \emptyset$, (2) each u is upper semicontinuous, i.e. for each $\alpha \in I_{0+}$, the α level set $[u]_\alpha = \{x \in \mathfrak{X} :\ u(x) \geq \alpha\}$ is a closed subset of $\mathfrak{X}$, (3) the support set $[u]_0 = \mathrm{cl}\{x \in \mathfrak{X} :\ u(x) > 0\}$ is bounded.

Let $\mathscr{F}_{bc}(\mathfrak{X})$ (resp. $\mathscr{F}_c(\mathfrak{X})$) be the family of all bounded closed convex (resp. convex) upper semicontinuous functions. It is known that u is convex in the above sense if and only if, for any $\alpha \in I$, $u_\alpha \in \mathscr{F}_c(\mathfrak{X})$ (see Chen [3, Theorem 3.2.1]).

For any two upper semicontinuous functions u_1, u_2, define

$$(u_1 + u_2)(x) = \sup_{x_1 + x_2 = x} \min\{u_1(x_1), u_2(x_2)\} \quad \text{for any} \quad x \in \mathfrak{X}.$$

Similarly, for any upper semicontinuous function u and for any $\lambda \geq 0$ and $x \in \mathfrak{X}$, define

$$(\lambda u)(x) = \begin{cases} u(\dfrac{x}{\lambda}), & \text{if } \lambda \neq 0, \\ I_0(x), & \text{if } \lambda = 0, \end{cases}$$

where I_0 is the indicator function of 0. It is known that for any $\alpha \in [0,1], [u_1 + u_2]_\alpha = [u_1]_\alpha + [u_2]_\alpha, [\lambda u]_\alpha = \lambda [u]_\alpha$.

The following distance is the uniform Hausdorff distance which is extension of the Hausdorff distance d_H : for $u, v \in \mathscr{F}_b(\mathfrak{X})$, $d_H^\infty(u, v) = \sup_{\alpha \in I} d_H([u]_\alpha, [v]_\alpha)$, this distance is the strongest one considered in the literatures. The space $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$ is complete. We denote $\|u\|_{\mathscr{F}} = d_H^\infty(u, I_{\{0\}}) = \|u_0\|_{\mathscr{K}}$.

X is called a random upper semicontinuous function (or random fuzzy set or fuzzy set-valued random variable), if it is a measurable mapping $X:\ (\Omega, \mathscr{A}, P) \to (\mathscr{F}_{bc}(\mathfrak{X}), \mathfrak{B}(\mathscr{F}_{bc}(\mathfrak{X})))$ (where $\mathfrak{B}(\mathscr{F}_{bc}(\mathfrak{X}))$ is the Borel σ-field of $\mathscr{F}_{bc}(\mathfrak{X})$ generated by the uniform Hausdorff distance d_H^∞). It is well known that the level mappings $L_\alpha : U \mapsto [U]_\alpha (\alpha \in I)$ are continuous from the space $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$ to the space $(\mathscr{K}_{bc}(\mathfrak{X}), d_H)$, so if X is a random upper semicontinuous function, then $[X]_\alpha$ is a bounded closed convex random set for any $\alpha \in I$. The expectation of an $\mathscr{F}_{bc}(\mathfrak{X})$-valued random variable X, denoted by $E[X]$, is an element in $\mathscr{F}_{bc}(\mathfrak{X})$ such that for every $\alpha \in I$, $[E[X]]_\alpha = cl \int_\Omega [X]_\alpha dP = cl\{E\xi :\ \xi \in S_{[X]_\alpha}\}$.

Let $D(I, C(S^*, d_s^*)) = \{f :\ I \to C(S^*, d_s^*)$ is left continuous at I_{0+}, right continuous at 0 and bounded, and f has right limit in $(0,1)\}$. Then it is a Banach space with respect to the norm $\|f\|_D = \sup\limits_{\alpha \in I} \|f(\alpha)\|_{C(S^*)}$ (see Li Ogura and Nguyen [9, Lemma 3.1]), and it is not separable.

For any $u \in \mathscr{F}_{bc}(\mathfrak{X})$, the support process of u is defined to be the process

$$j(u)(\alpha, x^*) = s([u]_\alpha)(x^*) = \sup_{x \in [u]_\alpha} \{x^*(x)\}, \quad (\alpha, x^*) \in I \times S^*.$$

The mapping $j :\ \mathscr{F}_{bc}(\mathfrak{X}) \to D(I, C(S^*, d_s^*))$ has the following properties: (1) $j(u+v) = j(u) + j(v)$, for any $u, v \in \mathscr{F}_{bc}(\mathfrak{X})$, (2) $j(\lambda u) = \lambda j(u)$, $\lambda \geq 0$, for any $u \in \mathscr{F}_{bc}(\mathfrak{X})$, (3)$\|j(u) - j(v)\|_D = d_H^\infty(u, v)$, for any $u, v \in \mathscr{F}_{bc}(\mathfrak{X})$.

In fact, the mapping j is an isometrically embedding of $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$ into a closed convex cone of the Banach space $(D(I, C(S^*, d_s^*)), \|\cdot\|_D)$.

Now, we will introduce some notations that we need corresponding to Wu's paper [16]. Since $D_1 = \{x_1^*, x_2^*, \cdots\}$ is countable dense in the unit sphere S^*, $\widetilde{D_1} = \{\widetilde{x_1^*}, \widetilde{x_2^*}, \cdots\}$ is a subset of the unit ball of the dual space of $(C(S^*, d_s^*), \|\cdot\|_{C(S^*)})$, where $\widetilde{x_i^*}(f) = f(x_i^*)$, for any $i \in \mathbb{N}, f \in C(S^*, d_s^*)$. Let $\ell_\infty(\widetilde{D_1})$ be the space of all bounded real function on $\widetilde{D_1}$ with supnorm $\|F\|_{\ell_\infty(\widetilde{D_1})} = \sup\limits_{v \in \widetilde{D_1}} |F(v)|$. This is a nonseparable Banach space.

Denote $M(C(S^*, d_s^*), \|\cdot\|_{C(S^*)})$ be space of probability measures on $(C(S^*, d_s^*),\ \|\cdot\|_{C(S^*)})$. For every $\nu \in M(C(S^*, d_s^*), \|\cdot\|_{C(S^*)})$, we can define an element $\nu^{\widetilde{D_1}}$ in $\ell_\infty(\widetilde{D_1})$ as $\nu^{\widetilde{D_1}}(\widetilde{x_i^*}) = \nu(\widetilde{x_i^*}) = \int_{C(S^*, d_s^*)} \widetilde{x_i^*} d\nu$, for all $\widetilde{x_i^*} \in \widetilde{D_1}$. In particular, denote the mapping $g_1 :\ C(S^*, d_s^*) \to \ell_\infty(\widetilde{D_1})$ given by

$$g_1(f) = \delta_f^{\widetilde{D_1}}, \quad \delta_f^{\widetilde{D_1}}(\widetilde{x_i^*}) = \delta_f(\widetilde{x_i^*}) = \int_{C(S*, d_s^*)} \widetilde{x_i^*} d\delta_f = \widetilde{x_i^*}(f) = f(x_i^*),$$

for all $\widetilde{x_i^*} \in \widetilde{D_1}, \delta_f$ is the Dirac measure concentrated at f. In fact, the mapping g_1 is linear and isometric from the Banach space $C(S^*, d_s^*)$ to $\ell_\infty(\widetilde{D_1})$, i.e.

(1)$g_1(\alpha f+\beta h)=\alpha g_1(f)+\beta g_1(h)$ for any $f,h\in C(S^*,d_s^*),\alpha,\beta\in\mathbb{R}$,
(2)$\|f-h\|_{C(S^*)}=\sup\limits_{x^*\in S^*}|\delta_f^{\widetilde{D_1}}(\widetilde{x^*})-\delta_h^{\widetilde{D_1}}(\widetilde{x^*})|=\|g_1(f)-g_1(h)\|_{\ell_\infty(\widetilde{D_1})}$.

Let Q_0 be all rational numbers in the interval I, $D_2=Q_0\times D_1$, $\widetilde{D_2}=\{\widetilde{(\alpha,x^*)}:(\alpha,x^*)\in D_2\}$ is a subset of the unit ball of the dual space of $(D(I,C(S^*,d_s^*)),\|\cdot\|_D)$, where $\widetilde{(\alpha,x^*)}(f)=f(\alpha,x^*)$, for any $f\in D(I,C(S^*,d_s^*))$. Let $\ell_\infty(\widetilde{D_2})$ be the space of all bounded real function on $\widetilde{D_2}$ with supnorm $\|F\|_{\ell_\infty(\widetilde{D_2})}=\sup\limits_{\nu\in\widetilde{D_2}}|F(\nu)|$. This is a nonseparable Banach space.

Denote $M(D(I,C(S^*,d_s^*)),\|\cdot\|_D)$ be space of probability measures on $(D(I,C(S^*,d_s^*)),\|\cdot\|_D)$. For every $\nu\in M(D(I,C(S^*,d_s^*)),\|\cdot\|_D)$, we can also define an element $\nu^{\widetilde{D_2}}$ in $\ell_\infty(\widetilde{D_2})$ as $\nu^{\widetilde{D_2}}(\widetilde{(\alpha,x_i^*)})=\nu(\widetilde{(\alpha,x_i^*)})=\int_{D(I,C(S^*,d_s^*))}\widetilde{(\alpha,x^*)}d\nu$, for all $\widetilde{(\alpha,x_i^*)}\in\widetilde{D_2}$. In particular, we define another mapping $g_2:\ D(I,C(S^*,d_s^*))\to\ell_\infty(\widetilde{D_2})$ given by

$$g_2(f)=\delta_f^{\widetilde{D_2}},$$

$$\delta_f^{\widetilde{D_2}}(\widetilde{(\alpha,x^*)})=\delta_f(\widetilde{(\alpha,x^*)})=\int_{D(I,C(S^*,d_s^*))}\widetilde{(\alpha,x^*)}d\delta_f=\widetilde{(\alpha,x^*)}(f)=f(\alpha,x^*),$$

for all $\widetilde{(\alpha,x^*)},f\in\widetilde{D_2}$. In fact, the mapping g_2 is also linear and isometric from Banach space $D(I,C(S^*,d_s^*))$ to $\ell_\infty(\widetilde{D_2})$, i.e.

(1)$g_2(\alpha f+\beta h)=\alpha g_2(f)+\beta g_2(h)$ for any $f,h\in D(I,C(S^*,d_s^*))$, $\alpha,\beta\in\mathbb{R}$,
(2)$\|f-h\|_D=\sup_{\alpha\in I}\sup_{x^*\in S^*}|f(\alpha,x^*)-h(\alpha,x^*)|=\sup_{(\alpha,x^*)\in D_2}|f(\alpha,x^*)-h(\alpha,x^*)|=\sup_{\widetilde{(\alpha,x^*)}\in\widetilde{D_2}}|\delta_f^{\widetilde{D_2}}(\widetilde{(\alpha,x^*)})-\delta_h^{\widetilde{D_2}}(\widetilde{(\alpha,x^*)})|=\|g_2(f)-g_2(h)\|_{\ell_\infty(\widetilde{D_2})}$.

3 Main Results and Proofs

Before giving LDP for bounded closed convex random sets and random upper semicontinuous functions, we define rate functions and LDP following Dembo and Zeitouni [5]. Let $\mathscr{X}$ be a regular Hausdorff topological space.

Definition 1. *([5, p.4, Definition]) (1) A* rate function *is a lower semicontinuous mapping $I:\mathscr{X}\to[0,\infty]$.*

(2) A good rate function *is a rate function such that the level sets $\Psi_I(\alpha):=\{x:I(x)\le\alpha\}$ are compact subset of $\mathscr{X}$.*

Definition 2. *([5, p.5, Definition]) A family of probability measures $\{\mu_n:n\in\mathbb{N}\}$ on a measurable space $(\mathscr{X},\mathscr{B})$ where $\mathscr{B}$ is the Borel σ-algebra is said to satisfy the* LDP *with the rate function I if, for all closed set $U\subset\mathscr{X}$,*

$$\limsup_{n\to\infty}\frac{1}{n}\ln\mu_n(U)\le-\inf_{x\in U}I(x),$$

for all open set $V \subset \mathscr{X}$,

$$\liminf_{n\to\infty} \frac{1}{n} \ln \mu_n(V) \geq -\inf_{x\in V} I(x).$$

We first give LDP for $(\mathscr{K}_{bc}(\mathfrak{X}), d_H)$-valued *i.i.d.* random variables.

Theorem 1. *Let* $X, X_1, \ldots, X_n$ *be* $(\mathscr{K}_{bc}(\mathfrak{X}), d_H)$*-valued i.i.d. random variables such that* $E\left[e^{\lambda \|X\|_{\mathscr{K}}}\right] < \infty$ *for all* $\lambda > 0$*. The following conditions*

(a)The space $(\widetilde{D_1}, d_2^{(1)})$ *is totally bounded, where* $d_2^{(1)}(\widetilde{x_i^*}, \widetilde{x_j^*}) = (E[s(X)(x_i^*) - s(X)(x_j^*)]^2)^{1/2}$,

(b)$d_H\left(\frac{X_1+\cdots+X_n}{n}, E[X]\right) \xrightarrow{P} 0$,

are necessary and sufficient that LDP holds, i.e. for any open set $U \subset (\mathscr{K}_{bc}(\mathfrak{X}), d_H)$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left\{\left(\frac{X_1+\cdots+X_n}{n}\right) \in U\right\} \geq -\inf_{A\in U} h^{(1)}_{\ell_\infty(\widetilde{D_1})}(g_1(s(A))),$$

any for any closed set $V \subset (\mathscr{K}_{bc}(\mathfrak{X}), d_H)$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left\{\left(\frac{X_1+\cdots+X_n}{n}\right) \in V\right\} \leq -\inf_{A\in V} h^{(1)}_{\ell_\infty(\widetilde{D_1})}(g_1(s(A))),$$

where $h^{(1)}_{\ell_\infty(\widetilde{D_1})}(F)$ *is*

$$\inf\{h(\nu;\ P\circ s(X)^{-1}): \quad \nu \in M(C(S^*, d_s^*), \|\cdot\|_{C(S^*)}),\ \nu^{\widetilde{D_1}} = F \ \text{ on } \widetilde{D_1}\}$$

and $h(\nu;\ P\circ s(X)^{-1}) = \int_{C(S^*,d_s^*)} \frac{d\nu}{d(P\circ s(X)^{-1})} \log \frac{d\nu}{d(P\circ s(X)^{-1})} d(P\circ s(X)^{-1})$, *if* $\nu \ll d(P\circ s(X)^{-1})$*. Otherwise,* $h(\nu;\ P\circ s(X)^{-1}) = +\infty$*. In fact,* $h(\nu;\ P\circ s(X)^{-1})$ *is the relative entropy of* ν *with respect to* $P\circ s(X)^{-1}$.

Remark 1. We omit the proof of Theorem 1 because the key point and main idea of proof are included in the proof below of Theorem 2.

In the following, we then give LDP for $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$-valued *i.i.d.* random variables.

Theorem 2. *Let* $X, X_1, \ldots, X_n$ *be* $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$*-valued i.i.d. random variables such that* $E\left[e^{\lambda \|X\|_{\mathscr{F}}}\right] < \infty$ *for all* $\lambda > 0$*. The following conditions*

(a)The space $(\widetilde{D_2}, d_2^{(2)})$ *is totally bounded, where* $d_2^{(2)}(\widetilde{(\alpha, x_i^*)}, \widetilde{(\beta, x_j^*)}) = (E[j(X)(\alpha, x_i^*) - j(X)(\beta, x_j^*)]^2)^{1/2}$,

(b)$d_H^\infty\left(\frac{X_1+\cdots+X_n}{n}, E[X]\right) \xrightarrow{P} 0$,

are necessary and sufficient that LDP holds, i.e. for any open set $U \subset (\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$,

$$\liminf_{n\to\infty} \frac{1}{n} \log P\left\{\frac{X_1+\cdots+X_n}{n} \in U\right\} \geq -\inf_{A\in U} h^{(1)}_{\ell_\infty(\widetilde{D_2})}(g_2(j(A))), \tag{1}$$

any for any closed set $V \subset (\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left\{\frac{X_1+\cdots+X_n}{n} \in V\right\} \leq -\inf_{A\in V} h^{(1)}_{\ell_\infty(\widetilde{D_2})}(g_2(j(A))), \tag{2}$$

where $h^{(2)}_{\ell_\infty(\widetilde{D_2})}(F)$ *is*

$$\inf\{h(\nu; P\circ j(X)^{-1}) : \nu \in M(D(I, C(S^*, d_s^*)), \|\cdot\|_D), \nu^{\widetilde{D_2}} = F \, on \, \widetilde{D_2}\} \tag{3}$$

and $h(\nu;\ P\circ j(X)^{-1}) = \int_{D(I,C(S^*,d_s^*))} \frac{d\nu}{d(P\circ j(X)^{-1})} \log \frac{d\nu}{d(P\circ j(X)^{-1}}) d(P\circ j(X)^{-1})$, *if* $\nu \ll d(P\circ j(X)^{-1})$. *Otherwise,* $h(\nu;\ P\circ j(X)^{-1}) = +\infty$. *In fact* $h(\nu; P\circ j(X)^{-1})$ *is the relative entropy of* ν *with respect to* $P\circ j(X)^{-1}$.

Proof. Since $\{X, X_n : n \in \mathbb{N}\}$ are $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$-valued i.i.d. random variables, j is an isometrical mapping from $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^\infty)$ to the Banach space $(D(I, C(S^*, d_s^*)), \|\cdot\|_D)$, and g_2 is an also linear and isometric mapping from $(D(I, C(S^*, d_s^*)), \|\cdot\|_D)$ to the Banach space $(\ell_\infty(\widetilde{D_2}), \|\cdot\|_{\ell_\infty(\widetilde{D_2})})$, we have that $\{j(X), j(X_n) : n \in \mathbb{N}\}$ are $(D(I, C(S^*, d_s^*)), \|\cdot\|_D)$-valued i.i.d. random variables satisfying $E\left[e^{\lambda \|g_2(j(X))\|_{\ell_\infty(\widetilde{D_2})}}\right] = E\left[e^{\lambda\|X\|_{\mathscr{F}}}\right] < \infty$ for all $\lambda > 0$. And since $\|(\frac{1}{n}\sum_{i=1}^n \delta_{j(X_i)} - P\circ(j(X))^{-1})^{\widetilde{D_2}}\|_{\ell_\infty(\widetilde{D_2})} = d_H^\infty\Big(\frac{X_1+\cdots+X_n}{n}, E[X]\Big)$, we know that condition (b) is equivalent to the following condition (b'): $(\frac{1}{n}\sum_{i=1}^n \delta_{j(X_i)} - P\circ(j(X))^{-1})^{\widetilde{D_2}} \to 0$, in probability in $\ell_\infty(\widetilde{D_2})$. One hand, by Theorem 4 in Wu [16], the conditions (a) in Theorem 2 and condition (b') are necessary and sufficient that $\big\{P\circ\big(\Big(\frac{1}{n}\sum_{i=1}^n \delta_{j(X_i)})^{\widetilde{D_2}}\Big)^{-1}\big) : n \in \mathbb{N}\big\}$ as $n\to\infty$ satisfy the LDP in $(\ell_\infty(\widetilde{D_2}), \|\cdot\|_{\ell_\infty(\widetilde{D_2})})$ with speed $\frac{1}{n}$ and with the good rate function given in (3). Further the image under $g_2\circ j$, $g_2(j(\mathscr{F}_{bc}(\mathfrak{X})))$, is a closed subset of the Banach space $(\ell_\infty(\widetilde{D_2}), \|\cdot\|_{\ell_\infty(\widetilde{D_2})})$, and $P\Big(\frac{\sum_{i=1}^n g_2(j(X_i))}{n} \in g_2(j(\mathscr{F}_{bc}(\mathfrak{X})))\Big) = 1, \forall\, n\geq 1$, then in view of [5, Lemma 4.1.5], the proposition that $\big\{P\circ\big(\Big(\frac{1}{n}\sum_{i=1}^n \delta_{j(X_i)})^{\widetilde{D_2}}\Big)^{-1}\big) : n \in \mathbb{N}\big\}$ satisfy the LDP in $g_2(j(\mathscr{F}_{bc}(\mathfrak{X})))$ equipped with the topology induced by $(\ell_\infty(\widetilde{D_2}), \|\cdot\|_{\ell_\infty(\widetilde{D_2})})$ is necessary and sufficient that $\big\{P\circ\big(\Big(\frac{1}{n}\sum_{i=1}^n \delta_{j(X_i)})^{\widetilde{D_2}}\Big)^{-1}\big) : n \in \mathbb{N}\big\}$ satisfy the LDP in $(\ell_\infty(\widetilde{D_2}), \|\cdot\|_{\ell_\infty(\widetilde{D_2})})$. On the other hand, by virtue of good properties of the mapping j and g_2 and the contraction principle [5, p.126, Theorem 4.2.1], we know the fact that $\left\{P\circ(\left(\frac{\sum_{i=1}^{n}X_i}{n}\right)^{-1}) : n\in\mathbb{N}\right\}$ satisfies LDP in the space $(\mathscr{F}_{bc}(\mathfrak{X}), d_H^{\infty})$ is equivalent to the fact that $\left\{P\circ(\left((\frac{1}{n}\sum_{i=1}^{n}\delta_{j(X_i)})^{\widetilde{D_2}}\right)^{-1}) : n\in\mathbb{N}\right\}$ satisfy the LDP in $g_2(j(\mathscr{F}_{bc}(\mathfrak{X})))$. In view of those, so we complete the proof of Theorem 2. □

Acknowledgements. This research is partially supported by NSFC (No.10771010), Research Fund of Beijing Educational Committee, PHR(IHLB) and 111 Talent Project Fund of BJUT, P.R. China.

References

1. Auman, A.: Integrals of set valued functions. J. Math. Anal. Appl. 12, 1–12 (1965)
2. Cerf, R.: Large deviations for sums of i.i.d. random compact sets. Proc. Amer. Math. Soc. 127, 2431–2436 (1999)
3. Chen, Y.: Fuzzy systems and mathematics. Huazhong institute press of Science and Technology (1984) (in Chinese)
4. Colubi, A., López-Díaz, M., Domínguez-Menchero, J.S., Gil, M.A.: A genaralized strong law of large numbers. Probab. Theory Related Fields 114, 401–417 (1999)
5. Dembo, A., Zeitouni, O.: Large deviations techniques and applications, 2nd edn. Springer, New York (1998)
6. Deuschel, J.D., Strook, D.W.: Large deviations. Pure and Applied Mathematics, vol. 137. Academic Press, Boston (1989)
7. Li, S., Ogura, Y., Kreinovich, V.: Limit Theorems and Applications of Set-valued and Fuzzy-valued Random Variables. Kluwer Academic Publishers, Dordrecht (2002)
8. Li, S., Ogura, Y., Proske, F.N., Puri, M.L.: Central limit theorems for generalized set-valued random variables. J. Math. Anal. Appl. 285, 250–263 (2003)
9. Li, S., Ogura, Y., Nguyen, H.T.: Gaussian processes and martingales for fuzzy valued random variables with continuous parameter. Inform. Sci. 133, 7–21 (2001)
10. Molchanov, I.S.: On strong laws of large numbers for random upper semicontinuous functins. J. Math. Anal. Appl. 235, 349–355 (1999)
11. Ogura, Y., Li, S., Wang, X.: Large and moderate deviations of random upper semicontinuous functions. Stoch. Anal. Appl. 28, 350–376 (2010)
12. Ogura, Y., Setokuchi, T.: Large deviations for random upper semicontinuous functions. Tohoku Math. J. 61, 213–223 (2009)
13. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 406–422 (1986)
14. Teran, P.: A large deviation principle for random upper semicontimuous functions. Proc. Amer. Math. Soc. 134, 571–580 (2006)
15. Teran, P.: On Borel measurability and large deviations for fuzzy random variables. Fuzzy Sets Syst. 157, 2558–2568 (2006)
16. Wu, L.M.: Large deviations, moderate deviations and LIL for empirical processes. Ann. Probab. 22, 17–27 (1994)

A Note of Proposed Privacy Measures in Randomized Response Models

Hong Zhimin, Yan Zaizai, and Wei Lidong

Abstract. Randomized response (say, RR) techniques on survey are used for collecting data on sensitive issues while trying to protect the respondents' privacy. The degree of confidentiality will clearly determine whether or not respondents choose to cooperate. There have been many proposals for privacy measures with very different implications for an optimal model design. These derived measures of protection privacy involves both conditional probabilities of being perceived as belonging to the sensitive group A under given an answer "yes" or "no", denoted as $P(A \mid yes)$ and $P(A \mid no)$. This motivates us to evaluate the proposed measures of protection privacy. This article shows that the most of the proposed measures of protection privacy are unified.

Keywords: Privacy of respondent, Randomized response, Efficiency.

1 Introduction

When conducting personal interview surveys on sensitive or highly personal questions, such as tax evasion, drug use, and sexual behavior, refusal to respond or untruthful responses are a major problem. Warner [6] introduced a technique known as "randomized response" to obtain information on π_A, the proportion of people with the sensitive characteristic A, while protecting the privacy of the respondent. Since that time, the problem has been

Hong Zhimin and Yan Zaizai
Department of Mathematics, Science College of Inner Mongolia University of Technology, Hohhot, Inner Mongolia, 010051, P.R. China
e-mail: zz.yan@163.com

Wei Lidong
Bureau of Statistics, Hohhot, Inner Mongolia, 010000, P.R. China

C. Borgelt et al. (Eds.): Combining Soft Computing & Stats. Methods, AISC 77, pp. 635–642.
springerlink.com

reconsidered with new techniques being proposed and studied. A major concern in randomized response techniques is how to produce good estimators of the population proportion of people having a stigmatizing characteristic while at the same time the privacy of participants is protected.

Except when stated otherwise, we assume truthful reporting in randomized queries. In RR surveys, although a respondent is not asked to divulge his or her true standing in respect to a sensitive characteristic, the person does, nevertheless, run certain risks of disclosures. It is possible that among the respondents there may be intelligent and knowledgeable people well equipped intellectually to analyze and weigh the hazards in giving out secrets. Naturally, they must be convinced that their privacy is well guarded before they will be persuaded to make available damaging and incriminating documents. So the degree of protection of privacy is an essential ingredient of RR theory and practice. In some RR models, a general phenomenon was observed that maintenance of privacy and efficient estimation with RR were in conflict. Hence, in RR surveys, a problem of optimally efficient estimation subject to practical constraints imposed by the requirement of protecting the privacy of a respondent. Several research workers ([1, 2, 3, 4, 5, 7, 8] among others) have discussed the relationship maintenance of privacy and efficient estimation with RR.

2 Statistical Analysis of Efficiency and Protection

Consider a population divided into complementary sensitive group A and non-sensitive group $\overline{A}$ with unknown proportions π_A and $1-\pi_A$, respectively. Considering a dichotomous response model, a typical response is R, which is "yes" (say, y) or "no"(say, n). The conditional probabilities that a response R comes from an individual of groups A and $\overline{A}$, respectively, are $P(R \mid A)$ and $P(R \mid \overline{A})$. These are quantities at the investigator's disposal.

The posterior probabilities that a respondent belongs to groups A and $\overline{A}$, when he or she reports R are $P(A \mid R)$ and $P(\overline{A} \mid R)$. These are the revealing probabilities. By Bayes' rule,

$$P(R \mid A) = \frac{P(R)P(A \mid R)}{P(A)}, \quad P(R \mid \overline{A}) = \frac{P(R)P(\overline{A} \mid R)}{P(\overline{A})},$$

$$\frac{P(R \mid A)}{P(R \mid \overline{A})} = \frac{1-P(A)}{P(A)} \cdot \frac{P(A \mid R)}{1-P(A \mid R)}.$$

Now, clearly, the probability of a "yes" response is

$$\lambda = P(A)P(y \mid A) + P(\overline{A})P(y \mid \overline{A}) = \pi_A\{P(y \mid A) - P(y \mid \overline{A})\} + P(y \mid \overline{A}) \qquad (1)$$

If a sample of n individuals is selected from the population according to simple random sampling with replacement(SRSWR), and $\hat{\lambda}$ is the sample proportion of "yes" replies, then an unbiased estimator of π_A is

$$\hat{\pi}_A = \frac{\hat{\lambda} - P(y\,|\,\overline{A})}{P(y\,|\,A) - P(y\,|\,\overline{A})} \tag{2}$$

which is defined if and only if

$$P(y\,|\,A) - P(y\,|\,\overline{A}) \neq 0 \tag{3}$$

holds.

Now suppose that the RR procedure is designed so that (3) is satisfied allowing unbiased estimation of π_A even at the cost of privacy, which has to be sacrificed to some extent. It is easy to work out the variance formula for $\hat{\pi}_A$ as

$$V(\hat{\pi}_A) = \frac{\lambda(1-\lambda)}{n\{P(y\,|\,A) - P(y\,|\,\overline{A})\}^2} = \frac{\pi_A^2(1-\pi_A)^2}{n(\pi_A - P(A\,|\,y))(P(A\,|\,n) - \pi_A)} \tag{4}$$

The response R is non-jeopardizing if and only if

$$P(A\,|\,y) = P(A\,|\,n) = P(A) = \pi_A \tag{5}$$

The difference of the conditional probabilities is

$$P(y\,|\,A) - P(y\,|\,\overline{A}) = \frac{(P(A\,|\,y) - \pi_A)(P(A\,|\,n) - \pi_A)}{\pi_A(1-\pi_A)(P(A\,|\,n) - P(A\,|\,y))} \tag{6}$$

Hence, by (3) and (6), for a defined $\hat{\pi}_A$, the expression (5) is violated.

Without loss of generality, we assume that $P(A\,|\,y)>\pi_A>P(A\,|\,n)$. It follows from (4) that

$$\frac{\partial V(\hat{\pi}_A)}{\partial P(A\,|\,y)} < 0, \quad \frac{\partial V(\hat{\pi}_A)}{\partial P(A\,|\,n)} > 0 \tag{7}$$

Hence, for the sake of efficiency, one needs a major level for $P(A\,|\,y)$ and a minor level for $P(A\,|\,n)$.

From practical considerations regarding protection of privacy, one can fix a maximal allowable level of $P(A\,|\,y)$ and a minimal allowable level of $P(A\,|\,n)$, say t_1 and t_2 in (0,1), respectively. Thus the problem now becomes one of constrained optimization about $P(A\,|\,y)$ and $P(A\,|\,n)$, that is, of minimizing $V(\hat{\pi}_A)$ subject to

$$\pi_A < P(A\,|\,y) \leq t_1, \quad t_2 \leq P(A\,|\,n) < \pi_A \tag{8}$$

Naturally, a relation (8) may be satisfied only for values of π_A in a subinterval of [0,1] , say $[\pi_1, \pi_2]$.

3 Proposed Privacy Measures

3.1 Leysieffer and Warner's (1976) Criterion for Protecting Privacy

Leysieffer and Warner [4] proposed the measures of jeopardy carried by R about A and $\overline{A}$, respectively. These measures are as follows:

$$g(R \mid A) = \frac{P(R \mid A)}{P(R \mid \overline{A})}, \quad g(R \mid \overline{A}) = \frac{1}{g(R \mid A)} \tag{9}$$

The response R is non-jeopardizing if and only if

$$g(R \mid A) = 1 \tag{10}$$

According to (3), an unbiased estimator of π_A is defined as (2) that is, (10) is violated. Assuming without loss of generality, that $P(y \mid A) > P(y \mid \overline{A})$, so that $g(y \mid A) > 1, g(n \mid \overline{A}) > 1$. Therefore, for the sake of efficiency, one needs as large magnitudes as possible for $g(y \mid A)$ and $g(n \mid \overline{A})$ and both above unity. Hence, from the practical point of view, regarding protection of privacy, one can fix some maximal allowable levels of $g(y \mid A)$ and $g(n \mid \overline{A})$, say k_1 and k_2, respectively. Thus the problem now becomes one of constrained optimization, that is, of minimizing $V(\hat{\pi}_A)$ subject to $1 < g(y \mid A) \leq k_1, 1 < g(n \mid \overline{A}) \leq k_2$. By (9), the jeopardy function is rewritten by, say,

$$g(y \mid A) = \frac{1-\pi_A}{\pi_A} \cdot \frac{P(A \mid y)}{1 - P(A \mid y)}, \quad g(n \mid \overline{A}) = \frac{\pi_A}{1-\pi_A} \cdot \frac{1 - P(A \mid n)}{P(A \mid n)}, \tag{11}$$

it follows from (11) that

$$\frac{\partial g(y \mid A)}{\partial P(A \mid y)} > 0, \ \frac{\partial g(n \mid \overline{A})}{\partial P(A \mid n)} < 0 \tag{12}$$

By (8) and (11), for every π_A in $[\pi_1, \pi_2]$,

$$g(y \mid A) \leq \frac{1-\pi_A}{\pi_A} \cdot \frac{t_1}{1-t_1}, \quad g(n \mid \overline{A}) \leq \frac{\pi_A}{1-\pi_A} \cdot \frac{1-t_2}{t_2}.$$

This implies that $g(y \mid A) \leq \frac{1-\pi_1}{\pi_1} \cdot \frac{t_1}{1-t_1}$ and $g(n \mid \overline{A}) \leq \frac{\pi_2}{1-\pi_2} \cdot \frac{1-t_2}{t_2}$.

Thus Leysieffer and Warner's (1976) criterion for protection of confidentiality effectively sets upper bounds for $g(y \mid A)$ and $g(n \mid \overline{A})$ and for minimizing $V(\hat{\pi}_A)$, subject to the foregoing restrictions (8), it is enough to take $P(A \mid y)$ and $P(A \mid n)$ at their maximal and minimal allowable levels.

3.2 Anderson's (1975) Criterion for Protecting Privacy

Anderson [1] called $P(A \mid R)$ and $P(\overline{A} \mid R)$ the two "risks of suspicion" corresponding to response R. He suggested restricting them such that

$$P(A \mid R) \leq \xi_1 < 1, \quad P(\overline{A} \mid R) \leq 1 - \xi_1 < 1, \tag{13}$$

respectively, if A and $\overline{A}$ are embarrassing, suitably fixing ξ_1 and ξ_2 in $(0,1)$. Evidently, (13) demands the revealing probability such that $\xi_1 \leq P(A \mid R) \leq \xi_2$. This implies that $P(A \mid y) \leq \xi_2$ and $\xi_1 \leq P(A \mid n)$.

Thus Anderson's (1975) criterion for protection of privacy subjects to the foregoing restrictions (8).

3.3 Lanke's (1976) Criterion for Protecting Privacy

Lanke [3] argued that "it is membership in Group A that people may want to hide, not membership in the complementary Group $\overline{A}$. Hence, it is 'yes' answers that may be embarrassing, and obviously a 'yes' answer is more embarrassing the less often the unrelated question has the answer 'yes'." This focus on the maximal "suspicion of belonging to A" led Lanke [3] to propose that the measure of protection privacy is

$$\varphi = \max\{P(A \mid y), P(A \mid n)\} \tag{14}$$

Assuming without loss of generality, that is $P(A \mid n) < \pi_A < P(A \mid y)$, so that Lanke's (1976) criterion is $\varphi = P(A \mid y)$. Following Lanke [3], the response R has less jeopardizing with respect to A if $\varphi = P(A \mid y)$ has less level. Hence, for the sake of efficiency, Lanke suggested restricting φ such that $\pi_A < P(A \mid y) \leq \eta$.

By the restrictions (8), it is enough to take $P(A \mid y)$ at a reasonable allowable level.

3.4 Fligner et al.'s (1977) Criterion for Protecting Privacy

Fligner et al. [2] took another criterion as the measures of a response jeopardy with respect to either A or $\overline{A}$. Let the expressions for $P(A \mid y)$ under the RR devices d_1 and d_2 be denoted by $P_{d_1}(A \mid y)$ and $P_{d_2}(A \mid y)$, respectively. Similarly, defined $P_{d_1}(A \mid n)$ and $P_{d_2}(A \mid n)$.

According to Fligner et al. [2], the two model afford equal treatment, in terms of protection of confidentiality if and only if

$$P_{d_1}(A \mid y) = P_{d_2}(A \mid y) \quad \text{and} \quad P_{d_1}(A \mid n) = P_{d_2}(A \mid n) \tag{15}$$

By (4) and (15), we obtain that $V_{d_1}(\hat{\pi}_A) = V_{d_2}(\hat{\pi}_A)$.

As a consequence, if both $P(A \mid y)$ and $P(A \mid n)$ are equal across models, two models are said to afford " equal treatment" to the respondent. It can be shown that the estimators have the same distribution and there is no reason to select one over the other.

3.5 Nayak's (1994) Criterion for Protecting Privacy

Nayak [5] formalized the commonly emphasized issues of Fligner et al. [2]. According to the discussion of Nayak [5], the posterior probabilities, $P(A \mid y)$, $P(A \mid n)$, $P(\overline{A} \mid y) = 1 - P(A \mid y)$ and $P(\overline{A} \mid n) = 1 - P(A \mid n)$, that are relevant for assessing respondents' protection. The efficiency increases as $V(\hat{\pi}_A)$ decreases, and the respondents' protection increases as the posterior probabilities $P(A \mid y)$ and $P(A \mid n)$ decrease.

Thus, Nayak has the following definitions:

A design d_1 is said to be better than another design d_2 if

$$P_{d_1}(A \mid y) \leq P_{d_2}(A \mid y)\ ,\ P_{d_1}(A \mid n) \leq P_{d_2}(A \mid n) \text{ and } V_{d_1}(\hat{\pi}_A) \leq V_{d_2}(\hat{\pi}_A), \qquad (16)$$

for all $\pi_A \in [0,1]$ at least one strict inequality holds for some π_A, where the subscripts d_1 and d_2 specify the design.

By (7), (8), and (16), Nayak's (1994) criterion shows that both efficiency and protection of pricacy increase as the posterior probability $P(A \mid n)$ decreases, which suggests that one should always take a minimal allowable level of $P(A \mid n)$ as t_2 for some admissible RR models. Nayak's (1994) criterion shows that efficiency and respondents' protection do not necessarily move in opposite directions.

3.6 Zaizai and Zankan's (2004) Criterion for Protecting Privacy

Zaizai and Zankan [7] considered the difference of two responses with respect to A as measure of protecting of privacy. Namely, $\mid P(A \mid y) - P(A \mid n) \mid$, say, Δ. The response R has less jeopardize as Δ is close to zero. Hence, for the sake of efficiency and protection of privacy, one can fix a maximal allowable level of Δ, say c.

Assuming without loss of generality, that is $P(A \mid n) < \pi_A < P(A \mid y)$, therefore, Zaizai and Zankan [7] took the measure of protection privacy that

$$\Delta = P(A \mid y) - P(A \mid n) \qquad (17)$$

It follows from (17) that $\dfrac{\partial \Delta}{\partial P(A \mid y)} = 1 > 0$ and $\dfrac{\partial \Delta}{\partial P(A \mid n)} = -1 < 0$. According to (8) and (17), that is $\Delta \leq t_1 - t_2$, namely, $c = t_1 - t_2$.

Thus Zaizai and Zankan's (2004) criterion for protection of confidentiality effectively sets upper bound for Δ and for minimizing $V(\hat{\pi}_A)$, subject to the foregoing restrictions ([8]).

3.7 Zhimin and Zaizai's (2008) Criterion for Protecting Privacy

Similar analysis on the problems of efficient estimation and protecting privacy was carried out by Zhimin and Zaizai [8], who reached parallel conclusions with Nayak's (1994) criterion. Zhimin and Zaizai's (2008) measures are that

$$\Delta(A) = \frac{P(y \mid A)}{P(n \mid A)}, \quad \Delta(\overline{A}) = \frac{P(n \mid \overline{A})}{P(y \mid \overline{A})} \tag{18}$$

This criterion measures the jeopardy carried by R about A and $\overline{A}$, respectively. The response R is non-jeopardizing if and only if $\Delta(A) = 1$ and $\Delta(\overline{A}) = 1$. But the existence of an unbiased estimator $\hat{\pi}_A$ necessarily makes a response jeopardy with respect to either A or $\overline{A}$.

Assuming without loss of generality, that $\Delta(A) \geq \Delta(\overline{A})$, $\Delta(A) \neq 1$ and $\Delta(\overline{A}) \neq 1$. Zhimin and Zaizai [8] defined $\Delta(\overline{A})$ as a measure of protecting of privacy.

Hence, for the sake of efficiency and protection of privacy, one can fix an allowable level of $\Delta(\overline{A})$, such that $\mid \Delta(\overline{A}) - 1 \mid$ is minimal.

By Bayes' rule,

$$\Delta(\overline{A}) = \frac{P(n \mid \overline{A})}{P(y \mid \overline{A})} = \frac{\pi_A - P(A \mid y)}{1 - P(A \mid y)} \cdot \frac{1 - P(A \mid n)}{P(A \mid n) - \pi_A}, \tag{19}$$

it follows from (19) that $\frac{\partial \Delta(\overline{A})}{\partial P(A|y)} > 0, \ \ \frac{\partial \Delta(\overline{A})}{\partial P(A|n)} > 0$. This implies that

$$\frac{\partial \mid \Delta(\overline{A}) - 1 \mid}{\partial P(A \mid y)} > 0, \quad \frac{\partial \mid \Delta(\overline{A}) - 1 \mid}{\partial P(A \mid n)} > 0 \quad (\Delta(\overline{A}) > 1) \tag{20}$$

or

$$\frac{\partial \mid \Delta(\overline{A}) - 1 \mid}{\partial P(A \mid y)} < 0, \quad \frac{\partial \mid \Delta(\overline{A}) - 1 \mid}{\partial P(A \mid n)} < 0 \quad (\Delta(\overline{A}) < 1) \tag{21}$$

By (7), (8), (20) and (21), note that both efficiency and respondents' protection increase as $P(A \mid n)$ decreases, which suggests that one should always take a minimal allowable level of $P(A \mid n)$ as t_2. This conclusion also disproves the common belief that in RR surveys efficiency and respondents' protection are always in conflict.

4 Summary

In this paper, we evaluated the difference of the proposed privacy measures. From practical considerations, regarding protection of privacy, the problem of all randomized response surveys becomes one of constrained optimization about $P(A \mid y)$ (a maximal allowable level of $P(A \mid y)$, say t_1) and $P(A \mid n)$ (a minimal allowable level of $P(A \mid n)$, say t_2), that is, of minimizing $V(\hat{\pi}_A)$ subject to $\pi_A < P(A \mid y) \leq t_1$, $t_2 \leq P(A \mid n) < \pi_A$. It also showed that the most of proposed measures of privacy are consistent with the conclusion that efficiency and respondents' protection are always in conflict, while Nayak's (1994) and Zhimin and Zaizai's (2008) criterions show that the maintenance of privacy and efficiency in RR surveys do not necessarily move in opposite directions.

Acknowledgements. This work was supported by National Natural Science Foundation of China (10761004), Talent Development Foundation of Inner Mongolia (2007), Higher School Science and Technology Foundation of Inner Mongolia (NJ10085) and Natural Science Foundation of Inner Mongolia (2009MS0107). The authors thank two referees for several suggestions, which have helped to improve the article.

References

1. Anderson, H.: Efficiency versus protection in the RR designs for estimating proportions. Tech. Rep. 9, University of Lund, Sweden (1975)
2. Fligner, M.A., Policello II, G.E., Singh, J.: A comparison of two randomized response survey method with consideration for the level of respondent protection. Commun. Statist. Theor. Meth. 6, 1511–1524 (1977)
3. Lanke, J.: On the degree of protection in randomized interviews. Internat. Statist. Rev. 44, 197–203 (1976)
4. Leysieffer, R.W., Warner, S.L.: Respondent jeopardy and optimal design in randomized response models. J. Amer. Statist. Assoc. 71, 649–656 (1976)
5. Nayak, T.K.: On randomized response surveys for estimating a proportion. Commun. Statist. Theor. Meth. 23(11), 3303–3321 (1994)
6. Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Assoc. 60, 63–69 (1965)
7. Zaizai, Y., Zankan, N.: A fair comparison of the randomized response strategies. Mathematica Scientia (in Chinese) 24, 362–368 (2004)
8. Zhimin, H., Zaizai, Y.: Efficiency comparison of randomized response strategies under the same level of respondent protection. Journal of Engineering Mathematics (in Chinese) 25(1), 97–102 (2008)

Index