Contributions to Statistics

Teresa A. Oliveira · Christos P. Kitsos
Amílcar Oliveira · Luís Grilo *Editors*

Recent Studies on Risk Analysis and Statistical Modeling

Springer

Contributions to Statistics

The series **Contributions to Statistics** contains publications in theoretical and applied statistics, including for example applications in medical statistics, biometrics, econometrics and computational statistics. These publications are primarily monographs and multiple author works containing new research results, but conference and congress reports are also considered.
Apart from the contribution to scientific progress presented, it is a notable characteristic of the series that publishing time is very short, permitting authors and editors to present their results without delay.

More information about this series at http://www.springer.com/series/2912

Teresa A. Oliveira • Christos P. Kitsos •
Amílcar Oliveira • Luís Grilo
Editors

Recent Studies on Risk Analysis and Statistical Modeling

Springer

Editors
Teresa A. Oliveira
Universidade Aberta
Lisboa, Portugal

Christos P. Kitsos
Department of Informatics
Technological Educational Institute of Athens
Egaleo, Greece

Amílcar Oliveira
Universidade Aberta
Lisboa, Portugal

Luís Grilo
Unidade Dept de Matematica e Fisica
Instituto Politécnico de Tomar
Tomar, Portugal

ISSN 1431-1968
Contributions to Statistics
ISBN 978-3-319-76604-1 ISBN 978-3-319-76605-8 (eBook)
https://doi.org/10.1007/978-3-319-76605-8

Library of Congress Control Number: 2018943717

Mathematics Subject Classification (2010): 62-XX

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Scientific progress depends on good methods, and in order to try to accomplish the developments in our days, it urges to explore and develop methodologies involving risk analysis and statistical modeling. Trying to minimize or even avoid risks and to have good ways to be prepared to future results based on real data observed in the past is in fact mandatory. With recent advances in these areas from theoretical, computational, and practical points of view, the problems analysis has become more complex, and yet there is a need for guidance to get into the more advanced literature. Most of this literature can be found in scientific journals and proceedings. Besides some books cover a few methods very well, most of them do not do it in a comprehensive way, mainly to the practitioners in these areas. From this point of view, our volume detaches the difference. This book tries to overcome that problem by covering an essential part of the quantitative approach in risk analysis, where statistical models and/or mathematical methods are linked with some phenomenon under investigation. Along the book, applications to real data are observed in several areas, like engineering, medicine, health sciences, education sciences, economy, finances, and industry.

At the same time modeling issues provide the methodology to gather a compact structure for the data. In the first stage of risk analysis, data were studied from the decision theory point of view. However, nowadays data analysis is expanded to medical and biological models, and moreover, it covers economical, industrial, environmental, and management problems. Very general definition could be that: risk analysis is the review of (estimation of) the risks associated with a particular event or action, resulting in another one. That is why in principle, risk management is the process of planning, organizing, leading, and controlling activities to minimize the adverse effects of accidental losses on the organization, such as a firm or an industrial unit. Similar is the definition of the environmental risk assessment (ERA): it aims at assessing the effects of stressors, often chemicals, on the local environment. But risk assessment is concerned with the determination of quantitative or qualitative estimate of the associated risk related to a well-defined situation and a recognized threat: thus, the hazard function is an essential "tool" for the threat evaluation.

The common target for risk assessment studies, either for toxicology/medicine or for biology/environmental, was the cancer risk assessment. Thus, interest was focused on the design of experiments and mixtures of experiments at the early stages, and later the study of tumor was through the "birth-death" stochastic process. Here the "risk" was the cell to be transformed to a tumor! The experiments of performing on rats were restricting from the "size of the experiment" for ethical and economic reasons. But such studies of risk cannot be applied on economic problems which need the estimation of the involved risk. Actuary mathematics is another approach to reduce the risk for insurance companies and others.

Our previous Springer volume in *Theory and Practice of Risk Assessment* reflects a first step to extend the applications of risk analysis, to obtain a broader area of research, rather than the one centered on biostatistics, as an extension (political) from game theory.

Needless to say, at the first stage of using risk methods, from a mathematical point of view relative risk was really a simple index, but so useful (distance) measure. We adopt this line of thought in this volume, thanks to Springer, and we include more areas of applying risk theory. Thus, we believe that we cement our point of view that risk analysis can be considered an independent branch of statistics, tackling areas of interest such as management, industry, and economics.

The problem of data is always at the first line of interest, not only if it exists or not. We must recall that in cases where the data set is small, less than 30 observations, or for big data sets, we need a special treatment of the data. In some cases (due to cost!) only very few observations can be obtained. We moved from the "data of status" that is statistics to analyzing data sets from a number of areas, when we were developing *Data Analysis* (thanks to John Tukey, who named the method) and, now dealing with big data sets and high-tech computers, we are moving to *Data Analytics (subset of Business Intelligence)*. But always the source of data we try to analyze is very essential and related to the risk we try to eliminate, minimize, or estimate.

It was essential to create biological data sets. Molecular biology through genomics, proteomics, and metabolomics increased our knowledge of biological processes, and several databases are now accessible through the Internet. Similar databases, not so easy accessible, were developed on cancer. Studies on risk analysis were therefore based on more reliable data sets, as far as cancer was concerned. Statistical modeling also tackles the data analysis problem. That is why this book is divided into two parts:

Part I. Risk Methodologies and Applications
Part II. Statistical Modeling and Risk Issues in Several Areas

The papers submitted to this volume were carefully reviewed by referees. The selected papers were placed appropriately and offer the readers the opportunity to look for a number of different approaches and a broad range of areas of application. We believe that this book will offer solutions to the existing problems, will provide the appropriate framework and background to real-life problems, and will cover a

gap that usually exists: some more time is needed for new theoretical results to be published in one book.

The book reflects contributions from invited experts, providing the reader with a comprehensive overview of the main areas by fostering links with several related disciplines, exploring computational issues, and presenting some research future trends. As this volume is multiauthor, multinational and covers different areas of applications, it offers a chance to the researchers working in different areas of having it as a reference book.

Lisboa, Portugal — Teresa A. Oliveira
Athens, Greece — Christos P. Kitsos
Lisboa, Portugal — Amílcar Oliveira
Tomar, Portugal — Luís Grilo

Introduction

Doubt is the beginning, not the end, of wisdom.

—George Iles, 1852–1942.

In fact, doubt raises the notion of hazard, promotes risk research and fosters new knowledge. This book tries to cover an essential part of the quantitative approach in Risk Analysis, where statistical models and/or mathematical methods are linked with some phenomenon under investigation. The general topic RISK is explored in order to understand, simulate, design and promote the analysis of real problems, fostering new challenges in several areas, such as Engineering, Medicine, Health Sciences, Education Sciences, Economy, Finances and Industry.

In an attempt to recognize the role that statistics and computation play in risk analysis, the International Committee on Risk Analysis of the International Statistical Institute, the ISI-CRA, decided to select a series of interesting papers in order to attempt as this book chapters, consisting of some of the most important and current methodologies under the Risk topic. With this book we aim to reinforce the bridge connecting theoretical topics and new methodologies to the practical applications, fostering a deep insight among the practitioners of several areas.

The book is presented into two main parts based on the subject matter covered:

Part I is devoted to *Risk Methodologies and Applications*

Part II is focused on *Statistical Modeling and Risk Issues in Several Areas*

Part I: Risk Methodologies and Applications

The papers in Part I mainly cover Risk theoretical issues and methodologies, with focus on applications in Health Sciences, Medicine, Economics, Finance, Engineering and on special issues in the main areas of Mathematics and Statistics.

The chapters are organized into sections based on the primary focus of the papers included. Some briefly description of the topics covered in these sections follows:

Section 1 The first section deals with Risk Analysis in Health Sciences and Medicine.

Chapter "Assessment of Maximum A Posteriori Image Estimation Algorithms for Reduced Acquisition Time Medical Positron Emission Tomography Data" considers a study to examine the effects of reduced radioactive dosage data collection on positron emissions tomography reconstruction reliability. The efficiency of various reconstruction methods is also investigated.

Chapter "On Mixed Cancer Risk Assessment" consider both mammary cancer and Wilms tumors, as two typical examples from oncology generating difficult multicriterial decision problems. The authors fit mixture models to box-counting fractal dimensions in order to better understand the variability, they explore the effect of chemotherapy and present a discussion on the shape analysis for Wilms tumors.

Section 2 The second section is devoted to Risk Analysis in Economics and Finance applications.

Chapter "Traditional Versus Alternative Risk Measures in Hedge Fund Investment Efficiency" deals with the Hedge funds which are financial institutions aiming at generating absolute rates of return, that is at realizing profits regardless of the market situation. Some measures of investment are explored and compared in a particular period.

Chapter "Estimating the Extremal Coefficient: A Simulation Comparison of Methods". Tail dependence is an important issue to evaluate risk and the multivariate extreme values theory is the most suitable to deal with the extremal dependence. The extremal coefficient measures the degree of dependence between the marginals of max-stable distributions, a natural class of models in this framework. The authors address the estimation of the extremal coefficient and a new estimator is compared through simulation with existing methods. An illustration with real data is also presented.

In *Chapter "On a Business Confidence Index and Its Data Analytics: A Chilean Case"* a methodology on a novel Chilean business confidence index is presented, which allows the description of some aspects of the market at a global level, as well as at industrial and sector levels of Chilean great brands. Some issues related to business intelligence, customer and business surveys, market variables and of the mentioned confidence index are discussed. Descriptive and Inferential results on this index are presented, as well as results on the competitiveness of the Chilean great brands.

In *Chapter "On the Application of Sample Coefficient of Variation for Managing Loan Portfolio Risks"* the application of Sample Coefficient of Variation for Managing Loan Portfolio Risk is presented. The authors obtain the lower and upper bounds for sample Coefficient Variation and study the possibility of using it for measuring the risk concentration in a loan portfolio. The capital adequacy and the single borrower limit are considered and some theoretical results are obtained.

Finally the implementation and illustration for this approach is presented using a real data set.

Section 3 The third section focuses on Risk Analysis considering Statistical Process Control (SPC) with Applications to Industrial and Environmental problems.

In *Chapter "Acceptance-Sampling Plans for Reducing the Risk Associated with Chemical Compounds"* a research study on acceptance—sampling plans for reducing the risk associated with chemical compounds is presented. The authors highlight the adequacy of the inflated Pareto distribution to model measurements obtained by chromatography, and define and evaluate acceptance-sampling plans under this distributional setup for lots of large dimension. Some technical results associated with the construction and evaluation of such sampling plans are provided as well as an algorithm for an easy implementation of the sampling plan that exhibits the best performance.

In *Chapter "Risk of Return Levels for Spatial Extreme Events"* a study on Risk of Return Levels for Spatial Extreme Events is presented and the authors illustrate it with an application, using real data on environmental issues.

In *Chapter "Nonparametric Individual Control Charts for Silica in Water"* non-parametric individual Control Charts for silica in water are presented. The authors present a comparison of the control limits obtained with different approaches and emphasize that the analysis with(out) outliers is very important for technicians, since the value of silica should be as small as possible.

Section 4 The fourth section focuses on Risk Analysis using Statistical and Mathematical methodologies. Chapters X, XI and XII are devoted to Risk issues on Extreme Theory and on Distribution Theory. In *Chapter "Revisiting Resampling Methods in the Extremal Index Estimation: Improving Risk Assessment"* the authors revisit resampling methods in the extremal index estimation, with the aim of improving risk assessment and, in *Chapter "Improving Asymptotically Unbiased Extreme Value Index Estimation"*, a research on improving Asymptotically Unbiased Extreme Value Index Estimation is presented. Moreover, in this section the Hazard Rate and Future Lifetime for Generalized Normal Distribution is explored and discussed in *Chapter "Hazard Rate and Future Lifetime for the Generalized Normal Distribution"*.

Part II: Statistical Modeling and Risk Issues in Several Areas

Section 1 The first section of Part II deals with Statistical Modeling and Applications in Time Series.

Chapter "Wavelet-Based Detection of Outliers in Poisson INAR(1) Time Series" deals with the Wavelet-based detection of outliers in Poisson INAR(1) time series. The authors give special emphasis to the problem of detecting outliers, additive or innovational, single, multiple or in patches, in count time series modelled by first-order Poisson integer-valued autoregressive, PoINAR(1), models. In order to

address this problem, two wavelet-based approaches that allow the identification of the time points of outlier occurrence are proposed and the effectiveness of the proposed methods is illustrated with synthetic as well as with an observed dataset.

Chapter "Surveillance in Discrete Time Series" is devoted to the problem of surveillance in Discrete Time Series. The principles for the construction of optimal alarm systems are discussed and their implementation is described. As there is no unifying approach to the modelling of all integer-valued time series, the focus of attention is on the class of observation-driven models and the implementation of the optimal alarm system is described in detail for a particular non-linear model in this class.

In *Chapter "On the Maxima of Integer Models Based on a New Thinning Operator"* the authors present their work on the maxima of integer models based on a new thinning operator. A non-negative integer-valued process is introduced and studied, referred as Y-INARMA(1,1), which is an extension of the well-known geometric ARMA(1,1) process.

Section 2 The second section of Part II is devoted to Statistical and Mathematical issues considering Risk and Modeling.

Chapter "Exact and Approximate Probabilities for the Null Distribution of Bartels Randomness Test", under the main topic of Distribution Theory, explores the exact and approximate probabilities for the null distribution of Bartels Randomness Test. A new approximation based on the Edgeworth series is presented for the null distribution of the Bartels randomness statistic, and the precision of this new approximation is discussed. Under this topic, *Chapter "Gamma-Series Representations for the Sum of Independent Gamma Random Variables and for the Product of Independent Beta Random Variables"* discusses the Gamma-series representations for the sum of independent gamma random variables and for the product of independent beta random variables. Still under the topic of Distribution Theory, in *Chapter "Likelihood Ratio Tests for Equality of Mean Vectors with Circular Covariance Matrices"* the likelihood ratio tests for equality of mean vectors with circular covariance matrices is presented. Numerical studies show the extreme closeness of these near-exact distributions to the exact distributions.

Chapter "Optimal Estimators in Mixed Models with Orthogonal Block Structures" is devoted to Mixed Models, namely the optimal estimators in such models with Orthogonal Block Structures are explored. Finally, in *Chapter "Constructing Random Klein Surfaces Without Boundary"*, algebraic curves are explored. Random oriented 3-regular graphs with big number of vertices provide surfaces of large genera and an experimental model to claim important geometrical properties of such surfaces. For instance the existence of a low bound for systoles when the genus tends to infinity. The construction of random Klein surfaces without boundary is then discussed.

Section 3 Modeling applications to Engineering and Economics.

Chapter "Performance Analysis of a GPS Equipment" considers the performance analysis of a GPS Equipment. The aim is to evaluate GPS SRP regarding accuracy, as the equivalent of a real time kinematic (RTK) network and to address

the practicality of using either a continuously operating reference stations (CORS) or a passive control point for providing accurate positioning control. *Chapter "Multivariate Generalized Birnbaum-Saunders Models Applied to Case Studies in Bio-Engineering and Industry"* presents a methodology based on multivariate generalized Birnbaum-Saunders models applied to case studies in bio-engineering and industry.

Chapter "Energy Prices Forecasting Using GLM" looks at energy prices forecasting using GLM. The aim of the problem is the short term forecast of hourly energy prices and an application is performed using real data proposed by the company EDP - Energy Solutions Operator, in the framework of ESGI 119th, European Study Group with Industry. The application was developed taking into account data concerning hourly electricity prices from 2008 to 2016 and these data were explored using different statistical software, namely IBM SPSS, Statistics, Matlab and R Statistical Software. In *Chapter "Pseudo Maximum Likelihood and Moments Estimators for Some Ergodic Diffusions"* the authors consider the pseudo maximum likelihood and moments estimators for some ergodic diffusions.

Section 4 Statistical modeling applications to Health Sciences, Education Sciences and Informatics.

In *Chapter "Statistical Modelling of Counts with a Simple Integer-Valued Bilinear Process"* the authors explore the Statistical Modelling of Counts with a Simple Integer-Valued Bilinear Process and present an empirical application to real epidemiological count data to attest for its practical applicability in data analysis. *Chapter "A Comparative Study of the Estimators for the Demand of Engineering Courses in Portugal"* presents a comparative study of the estimators with the purpose of modeling the demand of Engineering Courses in Portugal. The authors explore an application which is based on a data set that covers the results of the national contest from 1997 to 2015 provided by the Portuguese Ministry of Education and Science. Multivariate methodologies were performed in order to allow a better understanding of the students' allocation behavior. Finally in *Chapter "Statistical Methods for Word Association in Text Mining"* the authors present a research on statistical methods for word association in Text Mining. Some general techniques for text data mining, based on text retrieval models that can be applicable to any text in any natural language are explored.

In conclusion, papers selected for this volume are mostly focused on the development of methodologies to face problems related to the main proposed topics of Risk Analysis and Statistical Modeling. A considerable number of them establish a bridge between these two main topics, and illustrate the methodologies with applications using real data sets and exploring several computational approaches. Thus, the book covers applications in a broad range of areas and will be very useful not only to researchers and students but also for practitioners. Thanks to the straightforward nature of the book presentation, we believe that it will help a new generation of statisticians and practitioners to solve complex problems in risk analysis. Moreover, many models and methods used in Risk Analysis were developed recently and have yet to reach their largest possible audience. We assist to

the slip of results which are scattered in various journals and proceeding volumes. In that sense this book fills a gap in the market and it can easily serve as a textbook for a special topics courses both on risk analysis and statistical modeling. We sincerely hope that this book publication will enhance the spread of ideas that are currently trickling through the scientific literature.

All of the papers included in this volume were reviewed by two referees and by the editors.

We would like to extend our heartfelt thanks to all the reviewers who devoted their time to allow us to improve the quality of the submitted papers, and in turn the quality of the volume. At the same time, we express our sincere thanks to all the authors, not only for the submission of the papers, but also for their expeditious revisions and for incorporating the reviewers' suggestions.

Last but not least, the editors would like to express their heart-felt thanks and gratitude to SPRINGER for their help and support with this volume, particularly we are deeply grateful to Dr Eva Hiripi without whose valuable assistance we could never have realized this manuscript.

Lisboa, Portugal — Teresa A. Oliveira
Athens, Greece — Christos P. Kitsos
Lisboa, Portugal — Amílcar Oliveira
Tomar, Portugal — Luís Grilo

Contents

Part I
Risk Methodologies and Applications

Assessment of Maximum A Posteriori Image Estimation Algorithms for Reduced Acquisition Time Medical Positron Emission Tomography Data

Daniel Deidda, Robert G. Aykroyd, and Charalampos Tsoumpas

Abstract This study examines the effects of reduced radioactive dosage data collection on positron emission tomography reconstruction reliability and investigates the efficiency of various reconstruction methods. Also, it investigates properties of the reconstructed images under these circumstances and the limitations of the currently used algorithms. The methods are based on maximum likelihood and maximum a posteriori estimation, but no explicit solutions exist and hence iterative schemes are obtained using the expectation-maximisation and one-step-late methods, while greater efficiency is obtained by using an ordered-subset approach. Ten replicate real datasets, from the Hoffman brain phantom collected using a Siemens Biograph mMR scanner, are considered using standard deviation, bias and mean-squared error as quantitative output measures. The variability is very high when low prior parameter values are used but reduces substantially for higher values. However, in contrast, the bias is low for low parameter values and high for high parameter values. For individual reconstructions, low parameter values lead to detail being lost in the noise whereas high values produce unacceptable artefacts at the boundaries between different anatomical regions. Considering the mean-squared error, a balance between bias and variability, still identifies high prior parameter values as giving the best results, but this is in contradiction to visual inspection. These findings demonstrate that when it comes to low counts, variability and bias become significant and are visible in the images, but that improved reconstruction can be achieved by a careful choice of the prior parameter.

D. Deidda · C. Tsoumpas
Department of Biomedical Imaging Science, University of Leeds, Leeds, UK
e-mail: umdde@leeds.ac.uk; C.Tsoumpas@leeds.ac.uk

R. G. Aykroyd (✉)
Department of Statistics, University of Leeds, Leeds, UK
e-mail: r.g.aykroyd@leeds.ac.uk

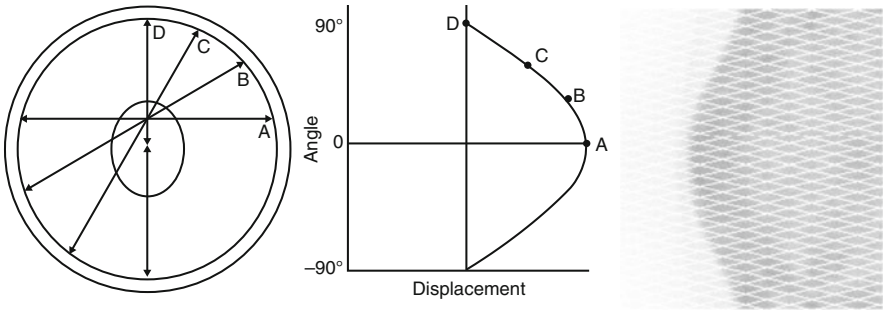
T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_1

1 Introduction

Positron emission tomography (PET) is a non-invasive imaging technique used in the clinical setting for routine diagnosis, dose delivery and treatment response evaluation. The use of medical imaging technologies is now commonplace in the clinical setting. Positron emission tomography leads the way in the detection of abnormalities such as cancer since, although it has low spatial resolution, it provides unrivalled functional information. As with all radiation-based methods, however, there is a risk of tissue damage which could lead to cancer at a later date. Hence there is a constant demand for decreases in radioactive dosages. Positrons are emitted by a radioactive-tracer, travelling a few millimetres before interacting with electrons. As soon as the interaction takes place a pair of photons, of identical energies, are emitted in opposite directions. The detection of the two photons allows a line-of-response to be defined which is characterised by an angle and the shortest distance between the line and the centre of the detector system—see Fig. 1a where four events are shown. When a large number of lines-of-response are plotted the resulting graph is half of a sine wave—hence the motivation for this type of graph being called a "sinogram". With complex objects the data will consist of a large number of overlapping sine waves [5]—see the real data example in Fig. 1b.

If this were a non-random system, then the data, Y say, would be a simple linear function of the radio-isotope concentrations, Λ say, and coefficients, C say, that is $Y = C\Lambda$. In principle, this can be inverted as $\Lambda = C^{-1}Y$, or the usual linear regression estimate $\Lambda = (C^T C)^{-1} C^T Y$. In a typical data collection system, however, there are about $256 \times 64^2 \approx 16$ million data values and a 3D reconstruction space of $256^3 \approx 16$ million voxels. This means that C is a 16 million by 16 million matrix with high multicollinearity. This is a highly ill-conditioned, and possibly ill-posed, big-data inverse problem which requires careful analysis.

Several iterative algorithms are currently used to estimate the radioactive-tracer distribution based upon the principles of maximum likelihood and maximum a pos-



a Diagram showing sinogram formulation **b** Example sinogram

Fig. 1 PET data acquisition: (**a**) events recorded as a line-of-response, and (**b**) total counts shown in a sinogram

teriori estimation. These have been widely studied using both simulated and clinical data, but less well for anatomically realistic measured phantom data, as is the theme here. This investigation will consider quantitative voxel-by-voxel output analysis, rather than the more common region-of-interest based approach, in terms of bias, standard deviation and mean squared error. A further novelty is the use of low count data and hence it will be the first to make the quantitative comparison of algorithm accuracy for such low count phantom data acquired with a PET-MR scanner.

This paper is organised as follows. A summary of the basic physical and statistical modelling is given in Sect. 2 with estimation discussed in Sect. 3. The data are described in Sect. 4 along with details of the investigation design. Numerical results are presented in Sect. 5 with discussion in Sect. 6.

2 Statistical Modelling for PET Data

Suppose that the 3D reconstruction space, $R \subset R^3$, is partitioned into small cubic voxels with the true radioactive-tracer concentration in voxel j labelled as λ_j giving the discretised unknown image $\Lambda = \{\lambda_j : j = 1, \ldots, M\}$. Note that this partitioning is arbitrary and hence can be tailored to the application or chosen for computational convenience. Let the data sinogram be denoted $Y = \{y_i : i = 1, \ldots, N\}$ which depends on the radioactive-tracer concentration through a Poisson model with

$$y_i \sim \text{Poisson}\left(\sum_{j=1}^{M} c_{ij}\lambda_j\right), \quad i = 1, \ldots, N, \tag{1}$$

where c_{ij} represents the, assumed known, probability that an event occurring in voxel j produces a coincidence in the i-th pair of detectors and takes into account attenuation and normalisation corrections. Before moving on it is worth noting that, for computational efficiency, the full summation over all N voxels is never performed. Instead, for each sinogram element the sum is only taken over voxels with non-negligible contribution, that is with c_{ij} above some threshold. The set of voxels with non-negligible contribution to y_i is denoted I_i and hence the model becomes

$$y_i \sim \text{Poisson}\left(\sum_{j \in I_i} c_{ij}\lambda_j\right), \quad i = 1, \ldots, N. \tag{2}$$

Based on this model the corresponding data log-likelihood, $l(Y, \Lambda) = \ln L(Y, \Lambda)$, is given by

$$l(Y, \Lambda) = \sum_{i \in N} \left[y_i \sum_{j \in I_i} \ln(c_{ij}\lambda_j) - \sum_{j \in I_i} (c_{ij}\lambda_j) - \ln(y_i!) \right] \tag{3}$$

and the maximum likelihood estimates of Λ given by

$$\widehat{\Lambda} = \max_{\Lambda} l(Y, \Lambda) \tag{4}$$

with the usual asymptotic approximations available for calculating a covariance matrix and construction of confidence intervals, etc. However, given that M is typically very large, e.g. $M = 16$ million, any direct solution of (4) is impractical. An alternative approach is to use an iterative method, such as the EM algorithm which is discussed in the next section.

Given that this is an inverse problem, stable solution of the maximum likelihood problem is unlikely, especially when the number of unknowns is large compared to the amount of data. One approach is to incorporate additional information into the maximisation step creating a penalised likelihood approach. This can equally be posed in a Bayesian setting with the penalty forming a prior distribution which is combined with the above likelihood to produce the posterior distribution and the maximum a posteriori (MAP) solution used as the reconstruction. Consider a prior distribution defined in terms of a Gibbs distribution

$$f(\Lambda) = \frac{1}{Z(\beta)} \exp\{-\beta V(\Lambda)\} \tag{5}$$

where $\beta > 0$ is a prior parameter and $V(\Lambda)$ is chosen so that f is large for values of Λ believed to be likely and small for implausible values. In particular, the following functional forms will be considered: (i) a Gaussian prior model on local differences

$$V_1(\Lambda) = \sum_{<j,k>} w_{jk}(\lambda_j - \lambda_k)^2/\lambda_j, \tag{6}$$

where $<j,k>$ denotes all voxel neighbours, $\{w_{jk}\}$ are positive constants that define a weight value for each neighbouring voxel (in general, 1 for first-order interactions between orthogonal nearest neighbours, and $1/\sqrt{2}$ for second-order diagonal interactions), and (ii) a Gaussian prior on local variability

$$V_2(\Lambda) = \sum_{j} (\lambda_j - \bar{\lambda}_j)^2/\bar{\lambda}_j, \tag{7}$$

were $\bar{\lambda}_j$ is the median of the values in the neighbourhood of voxel j. This is considered to be a more robust alternative allowing occasional sharp changes, or jumps, between neighbouring radioactive-tracer values. For further details of these prior distributions see, for example, [2] and [3].

The posterior distribution is then produced by combining the likelihood and the prior density using Bayes' Theorem resulting in the following log-posterior function (and ignoring constant terms)

$$p(\Lambda \mid Y) = \sum_{i \in N} \left[y_i \sum_{j \in I_i} \ln(c_{ij}\lambda_j) - \sum_{j \in I_i} (c_{ij}\lambda_j) - \beta V(\Lambda) \right], \tag{8}$$

with resulting definition of the maximum a posteriori (MAP) estimate as

$$\widehat{\Lambda} = \max_{\Lambda} p(Y, \Lambda). \tag{9}$$

Again, this cannot be solved easily and hence iterative algorithms can be used, one of which, based on the OSL algorithm [6], is described in the next section.

3 Maximum Likelihood and Maximum a Posteriori Estimation Using an EM Algorithm

In general, it is difficult to directly find the maximum of the log-likelihood in (3), recalling that there are millions of data points and millions of unknown parameters, and instead an iterative approach is used which considers single parameter updates one-by-one. The algorithms currently used in the clinical setting for PET are based on the approach originally proposed in [12], that is the maximum likelihood expectation maximisation (MLEM) algorithm, see also [4] and [10], which can be explained and derived as follows based on a "missing data" argument.

Suppose that instead of only Y being recorded, it had been possible to observe where all events originated, then this leaves a simple task of estimating the radioactive-tracer concentration. Hence a "complete dataset", $X = \{X_{ij} : i = 1, \ldots, N, j = 1, \ldots, M\}$, is considered where X_{ij} is defined as the number of photon pairs emitted from j and detected at i and is related to y_i by $y_i = \sum_{j \in I_i} X_{ij}$. The complete data log-likelihood is then:

$$l(X, \Lambda) = \sum_{i \in N} \left[X_{ij} \sum_{j \in I_i} \ln(c_{ij}\lambda_j) - \sum_{j \in I_i} (c_{ij}\lambda_j) - \ln(X_{ij}!) \right]. \tag{10}$$

In order to obtain the algorithm updating formula, the following two steps are necessary. In these n is the iteration number and $\widehat{\Lambda}^{(n)}$ is the estimated radioactive-tracer concentration at iteration n.

- E-STEP: During this step the algorithm estimates the conditional expectation of $l(X, \Lambda)$, $E(l(X, \Lambda)|Y, \widehat{\Lambda}^{(n)})$. The expected value for $l(X, \Lambda)$, given the measured data Y and $\widehat{\Lambda}^{(n)}$, is:

$$E(l(X, \Lambda) \mid Y, \widehat{\Lambda}^{(n)}) = \sum_{i,j} \left[\frac{c_{ij}\widehat{\lambda}_j^{(n)} y_i}{\sum_{k \in I_i} c_{ik}\widehat{\lambda}_k^{(n)}} \ln(c_{ij}\widehat{\lambda}_j^{(n)}) - c_{ij}\widehat{\lambda}_j^{(n)} \right] + \text{Const.} \tag{11}$$

  In the first iteration the image can consist of any non-negative solution, to assure the non-negativity constraint [10], though often an initial homogeneous image is used.

- M-STEP: In this step the algorithm finds the image that maximises the log-likelihood computed in the previous step by considering the partial derivatives:

$$\left.\frac{\partial E(l(X,\Lambda)\mid Y,\widehat{\Lambda}^{(n)})}{\partial\lambda_j}\right|_{\Lambda=\widehat{\Lambda}^{(n)}}=\sum_{i\in J_j}\left[\frac{c_{ij}\widehat{\lambda}_j^{(n)}y_i}{\sum_k c_{ik}\widehat{\lambda}_k^{(n)}}\lambda_j^{-1}-c_{ij}\right]=0, \tag{12}$$

  where J_j is the set of projections to which voxel j contributes. Dempster et al. [4] showed that Eq. (12) is equal to $\partial l(Y,\widehat{\Lambda}^{(n)})/\partial\lambda_j$. The resulting formula describes the MLEM algorithm:

$$\lambda_j^{(n+1)}=\frac{\widehat{\lambda}_j^{(n)}}{\sum_{i\in J_j}c_{ij}}\sum_{i\in J_j}\frac{c_{ij}y_i}{\sum_{k\in I_i}c_{ik}\widehat{\lambda}_k^{(n)}},\qquad j=1,\ldots,M. \tag{13}$$

  The resulting value for $\widehat{\Lambda}^{(n+1)}$ is then used in the E-STEP of the next iteration and the procedure is repeated until convergence is reached.

The MLEM algorithm is demonstrated to be a convergent algorithm and appropriately takes into account the random behaviour of the emission process. Nevertheless, it is computationally demanding and takes many iterations to converge. An accelerated version of MLEM was developed in [8] using ordered subsets (OS) of the data. The resulting OSEM converges in fewer iterations and is widely used in clinical practice because it is easily implemented and provides good images more quickly.

The OSEM algorithm provides acceleration of convergence, proportional to the number of subsets, by simply processing only the data within a subset at each sub-iteration. The data is organised in ordered subsets and the MLEM method is applied to each subset in turn. The reconstruction after each sub-iteration becomes the starting point for the following subset. In this way every iteration passes through every subset.

These subsets are usually chosen so that the projections within a subset corresponds to the projections of the image with down-sampled projection angles. The number of subsets has to be a divisor of the number of detector blocks in a ring (the Siemens mMR has 63 blocks per ring with 8×8 detectors per block and consequently, the choice could be one of 3, 9, 21 and 63). Following the same approach as for MLEM, the OSEM algorithm is obtained by substituting the sum over i by the sum over $s\in S_m$ in (13), where S_m is the chosen subset of detector pairs and $m=1,\ldots,M$, where M is the number of subsets:

$$\widehat{\lambda}_j^{(n,m+1)}=\frac{\widehat{\lambda}_j^{(n,m)}}{\sum_{s\in S_m}c_{sj}}\sum_{s\in S_m}\frac{c_{sj}y_s}{\sum_k c_{sk}\widehat{\lambda}_k^{(n,m)}},\qquad j=1,\ldots,M, \tag{14}$$

where $\widehat{\lambda}_j^{(n,m)}$ is the estimate of λ_j at sub-iteration m in the n-th full iteration. The resulting value for $\widehat{\Lambda}^{(n+1)}$ is then used in the next iteration and the procedure is

repeated until convergence is reached. The final solution then yields the maximum likelihood estimate, $\widehat{\Lambda}$, which will later be referred to as the MLE.

The OSEM method has been proven to converge quickly if the subset balance is respected [8]. However, with this method the MLEM noise artefacts are magnified after every iteration. For this reason, in the clinical practice it is often stopped at early iterations. Further, to remove the effects of noise in the reconstruction a Gaussian smoothing filter is applied as a post-processing step—later this will be referred to as the MLE+G.

To reduce the effects of noise in reconstruction in the final stages of the maximum likelihood algorithm, smoothing can be introduced through a prior distribution which then leads to the maximum a posteriori estimate. As with the log-likelihood, the log-posterior function in (9) cannot be maximised directly and hence an EM-based algorithm is again considered.

The OSMAPOSL algorithm is an extension of the OSEM algorithm which iteratively maximises the posterior distribution

$$l(\lambda \mid Y) = \sum_{i \in N} \left[y_i \sum_{j \in I_i} ln(c_{ij}\lambda_j) - \sum_{j \in I_i} (c_{ij}\lambda_j) - \beta V(\lambda) \right], \tag{15}$$

to produce the MAP estimate with the following updating equation:

$$\widehat{\lambda}_j^{(m+1)} = \frac{\widehat{\lambda}_j^{(m)}}{\left[\sum_{s \in S_m} c_{sj} - \beta \left. \frac{\partial V(\Lambda)}{\partial \lambda_k} \right|_{\Lambda = \widehat{\Lambda}^{(m)}} \right]} \sum_{s \in S_m} \frac{c_{sj} y_s}{\sum_k c_{sk} \widehat{\lambda}_k^{(m)}}, \qquad j = 1, \ldots, M. \tag{16}$$

This uses the same ordered-subset approach and hence has faster convergence than the original one-step-late algorithm [6]. The final solution then yields the maximum a posteriori estimate, $\widehat{\Lambda}$, which will later be referred to as the MAP. Further, the choice of prior function will be indicated as either V_1 or V_2 as defined in (6) and (7) respectively, and finally the value of the prior parameter will be given, for example $\beta = 100$, to give a full labelling such as MAP, V_1, $\beta = 100$.

4 Data Description and Assessment Criteria

The data used in this study were acquired with a 3D PET-MR system (Biograph mMR, Siemens Healthcare) by colleagues at the Institute of Nuclear Medicine of University College London Hospital. The Biograph mMR has 8 rings, each one divided in 63 blocks. The detector blocks have 8×8 LSO crystals, each $8 \times 8 \times 20\,\text{mm}^3$ in size. Experiments were carried out using the Hoffman 3D Brain Phantom [7], which can provide a realistic approximation of the radioactive-tracer distribution found in the normal brain. The phantom consists of a robust

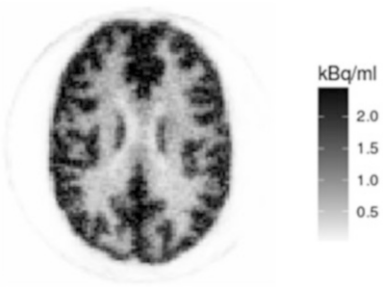


Fig. 2 MLEM image chosen as our high quality reference

plastic cylinder (diameter: 20.8 cm, height: 17.5 cm, fillable volume: ∼1.2 l) and 19 independent plates within the cylindrical phantom. It was filled with 60 MBq ^{18}F-fluorodeoxyglucose and the acquisition time was 3600 s giving a total number of events of about 10^9, which represents a standard for brain acquisitions. The numerical procedures have been developed within the Collaborative Computational Project in Positron Emission Tomography and Magnetic Resonance imaging (CCP-PET-MR)—see www.ccppetmr.ac.uk and make extensive use of STIR [13] for data correction (attenuation, scatter, normalisation and random matches) and reconstruction. The image size after the reconstruction is $289 \times 289 \times 127$ with voxel size $2.04 \times 2.04 \times 2.03\,\text{mm}^3$.

With real data the activity concentration is unknown and hence a gold standard is defined as a reference, using maximum likelihood estimation from the MLEM algorithm with the full 3600 s phantom data. The reconstructed image, as MLEM is globally convergent and because of the very high level of counts, is sufficient in terms of noise, bias and so on. The image, later denoted Λ^*, used as the "true" image was obtained after 126 iterations (see Fig. 2), to ensure convergence. Both "gray matter" and "white matter" voxel values can be distinguished well with substantial detail of the undulating and folded structure.

To ensure exactly equal experimental conditions, ten samples with roughly the same number of counts have been created, by sub-sampling from the 3600 s data, so as to mimic different acquisition times and to allow reconstruction assessment. The effective acquisition time was reduced from 3600 s to 36 s in order to simulate low count datasets. To analyse the low-count reconstructed images, different figures of merit have been chosen that are standard deviation (SD), Bias and Root Mean Squared Error (RMSE).

To define a general quantitative analysis design, suppose that K replicate datasets are available, $\{Y^k : k = 1, \ldots, K\}$ and that the corresponding results of the algorithm are estimated radioactive-tracer concentration images $\{\widehat{\Lambda}^k : k = 1, \ldots, K\}$. These can be used to define a mean image, $\bar{\Lambda} = \sum_{k=1}^{K} \widehat{\Lambda}^k / K$. Further,

recall that a "gold standard", Λ^* is available from the 3600 s dataset reconstructed using the MLEM algorithm. The SD, Bias and RMSE are then defined as follows:

$$\text{SD} = \sqrt{\frac{1}{K}\sum_{n=1}^{K}(\widehat{\Lambda}^k - \bar{\Lambda})^2}; \qquad \text{Bias} = \bar{\Lambda} - \Lambda^*; \qquad \text{RMSE} = \sqrt{\text{SD}^2 + \text{Bias}^2}. \tag{17}$$

The ideal reconstruction algorithm would produce low values for each of these measures indicating high reproducibility from the replicate datasets and lack of overall bias.

5 Experimental Results

A preliminary investigation was carried out to choose the number of sub-iterations in the ordered-subset algorithms, OSEM and OSMAP. To do so, two regions of interest (ROI) were chosen as representative of gray and white matter. Figure 3a, b shows the average value within each ROI as a function of sub-iteration from the OSEM reconstruction based on the 3600 s and 36 s datasets, respectively. This shows that at the 5-th iteration for 21 subsets (corresponding to 105 sub-iterations), the ROI values in both white and gray matter have stabilised. Although in the clinical setting, with 21 subsets, 63 sub-iteration are often used, here 105 is chosen as more appropriate for our datasets. In the rest of the analysis, 21 subsets and 5 complete iterations (105 sub-iterations) are used without further comment.

The results of the main statistical investigation with the 36 s dataset over the ten replicate datsets are represented in Figs. 4 and 5. These show a single dataset reconstruction (Individual), Mean, SD, Bias and RMSE images for six different

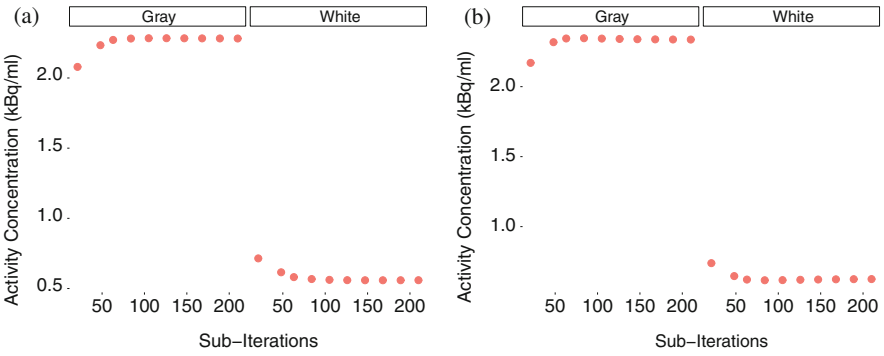


Fig. 3 Convergence of activity concentration values for white and gray matter using 21 subsets with OSEM. (**a**) 3600 s acquisition time. (**b**) 36 s acquisition time

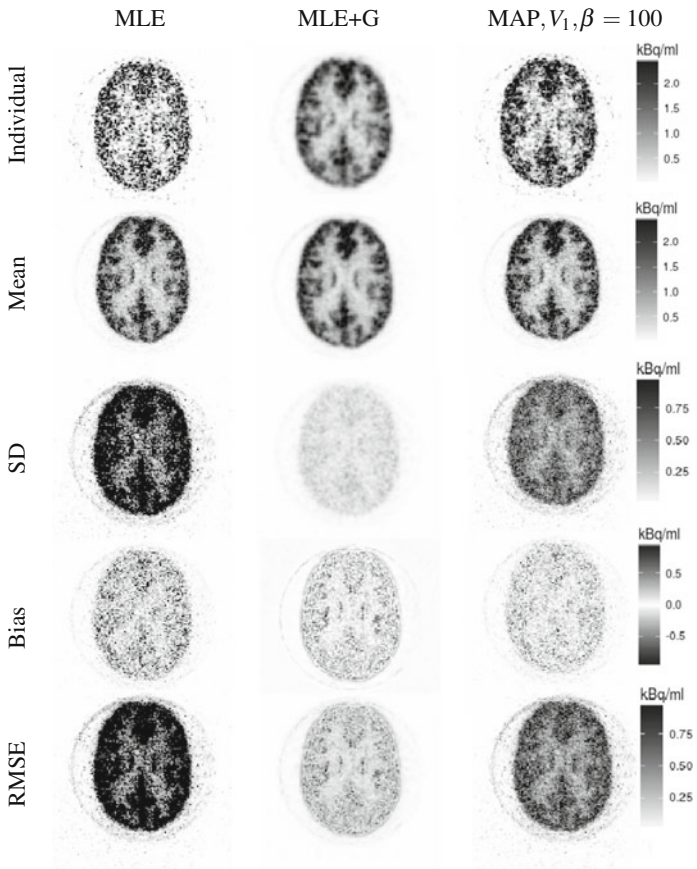


Fig. 4 Transverse view of images reconstructed with 21 subsets at the 105-th sub-iteration: 36 s acquisition

cases. Figure 4 shows the MLE from the OSEM algorithm, the corresponding image after the Gaussian smoothing post-processing, then the MAP estimate using the Gaussian prior on local differences, V_1, with $\beta = 100$. Figure 5 shows the remaining results for MAP estimates using the Gaussian prior on local differences, V_1, with $\beta = 1000$ and then MAP estimates using the Gaussian prior on local variability, V_2, with $\beta = 100$ and finally $\beta = 500$. The values of the prior parameters were chosen to give a range of reconstruction quality.

The SD images summarise the variability of the individual estimates over the ten samples, and show that the MLE and MAP estimates with small prior parameter have SD which is very high. Using a filter for the MLE to produce the MLE+G estimate, and a higher prior parameter for MAP estimation, helps to reduce the SD. Overall the Bias is higher in regions with the lowest activity concentration. In addition, too high a prior parameter leads to artefacts around the border between different anatomical features, this is the effect of over-smoothing.

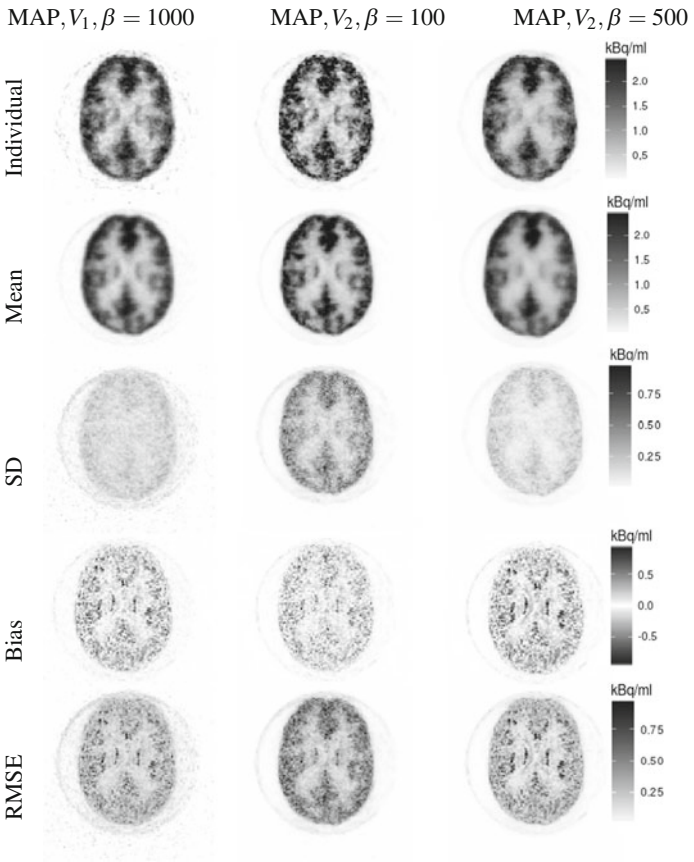


Fig. 5 Transverse view of images reconstructed with 21 subsets at the 105-th sub-iteration: 36 s acquisition

The RMSE images should take into account both Bias and SD, but here they show the same trend as the SD images. That is the variability between the ten samples swamps the Bias. With β higher than 500 and 1000 respectively for Gaussian prior on differences and Gaussian prior on variability, the Bias will become higher and more artefacts will be created. Table 1 shows the mean pixel values for two ROI, representing typical gray and white matter with each made up of three circles. Each number in the table is accompanied by the standard deviation over all the pixels within the ROI for each image in Figs. 4 and 5.

Table 1 Summary results, over ten samples, using three circular ROIs for each of gray and white matter

Gray matter								
	MLE		MAP with V_1			MAP with V_2		
	MLE	MLE+G	V_1, $\beta = 10$	V_1, $\beta = 100$	V_1, $\beta = 1000$	V_2, $\beta = 10$	V_2, $\beta = 100$	V_2, $\beta = 500$
Mean	2.47±0.59	2.39±0.12	2.46±0.51	2.47±0.26	2.39±0.15	2.49±0.47	2.49±0.21	2.42±0.16
SD	1.53±0.44	0.22±0.05	1.34±0.34	0.67±0.13	0.22±0.05	1.23±0.28	0.5±0.1	0.23±0.05
Bias	0.41±0.36	0.22±0.15	0.37±0.31	0.25±0.18	0.22±0.15	0.35±0.3	0.25±0.18	0.23±0.16
RMSE	1.61±0.48	0.33±0.12	1.42±0.37	0.74±0.14	0.33±0.12	1.3±0.31	0.59±0.13	0.34±0.12
White matter								
Mean	0.55±0.25	0.61±0.07	0.55±0.23	0.55±0.17	0.57±0.09	0.54±0.18	0.56±0.09	0.62±0.07
SD	0.57±0.35	0.14±0.03	0.54±0.29	0.38±0.13	0.14±0.03	0.41±0.16	0.17±0.04	0.08±0.02
Bias	0.16±0.13	0.14±0.08	0.15±0.12	0.11±0.08	0.1±0.07	0.11±0.08	0.09±0.06	0.13±0.09
RMSE	0.61±0.35	0.20±0.06	0.57±0.3	0.41±0.13	0.17±0.05	0.43±0.16	0.2±0.05	0.16±0.07

6 Discussion

The purpose of this work was to check the feasibility of image reconstruction when a short acquisition time or a low radioactive-tracer dosage is used and to compare the performance of various estimation methods in this situation. The study has used real data to assess how low-count conditions affect image reconstruction reliability, and in particular has compared different prior assumptions and different prior parameter values in terms of bias, standard deviation and mean squared error. The results show that good estimation can be achieved by a careful choice of the prior parameter. From a global perspective, low-count reconstructions show high noise and bias with all the methods showing the need for improvement. Moreover, it is clear that the convergence rate of OSEM is smaller in regions with lower pixel intensity; in fact, early-stopped OSEM images show a systematic bias in regions with lower activity concentration such as white matter and the background. In contrast, MAP methods with the right prior parameter value show better performance as low activity regions have less bias. This is due to the fact that they maximise the posterior distribution introducing prior information to remove the ill-conditioning. The maximum likelihood and MAP estimates with low prior parameter values have very high RMSE while it decreases with higher prior parameter values. Purely in terms of the RMSE, the best estimation occurs with the post-filtered maximum likelihood, and MAP estimation with the higher values of prior parameter. On visual inspection, however, these estimates are not completely acceptable as there is high bias across the boundary of grey and white matter, and further within a region it is overly smooth. These would make it challenging to recognise small abnormalities, such as cancers. Hence, a global goodness-fit measure which gives greater weight to bias, than does the RMSE, would be more suited to such medical investigations.

A key contribution of this study is to show the difficulty in the choice of a suitable prior parameter with low-count data. Even though the MAP estimates with the highest prior parameter values have low RMSE the images appear to be over-smoothed. In contrast, the maximum likelihood estimate with Gaussian smoothing, which is the preferred method in the clinic, produces results at least as well as any of MAP estimates. Hence, to achieve substantially better results with the MAP estimation methods will need careful choice of prior parameters. Recent studies, such as [1] and [9], have demonstrated that regularisation can significantly improve quantification and detectability compared to post-filtered maximum likelihood. The results of this investigation confirm what was suspected about low-count data, that is, under this special circumstance reconstruction is more greatly affected by bias and high levels of noise. Under this point of view, our results are in agreement with the results in [15]. The use of anatomical information from MR will result in the development of new hybrid reconstruction methods. This should help to preserve sharp contrast between adjacent anatomical features and avoid partial volume effects [11, 14].

This investigation and the procedures for iterative image reconstruction considered will be a useful guide for researchers who wish to study and extend image reconstruction and correction methods for PET data.

Conflict of Interest The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgements The authors are grateful to colleagues at University College London for use of the brain phantom data, and to the University of Leeds for financial support of Daniel Deidda through a University of Leeds University Scholarship. Part of this research is funded by the EPSRC Collaborative Computational Project (EP/P022200/1 & EP/M022587/1).

References

1. Ahn, S., Ross, S., Asma, E., Miao, J., Jin, X., Cheng, L., Wollenweber, S.D., Manjeshwar, R.M.: Quantitative comparison of OSEM and penalized likelihood image reconstruction using relative difference penalties for clinical PET. Phys. Med. Biol. **60**, 5733–5751 (2015)
2. Alenius, S., Ruotsalainen, U.: Generalization of median root prior reconstruction. IEEE Trans. Med. Imaging **21**(11), 1413–1420 (2002)
3. Bettinardi, V., Pagani, E., Gilardi, M., Alenius, S., Thielemans, K., Teras, M., Fazio, F.: Implementation and evaluation of a 3D one-step late reconstruction algorithm for 3D positron emission tomography brain studies using median root prior. Eur. J. Nucl. Med. Mol. Imaging **29**, 7–18 (2002)
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. (B). **39**, 1–38 (1977)
5. Fahey, F.: Data acquisition in PET imaging. J. Nucl. Med. Technol. **30**, 39–49 (2002)
6. Green, P.: Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Trans. Med. Imaging **9**, 84–93 (1990)
7. Hoffman, E., Cutler, P., Digby, W., Mazziotta, J.: 3-D phantom to simulate cerebral blood flow and metabolic images for PET. IEEE Trans. Nucl. Sci. **37**, 616–620 (1990)
8. Hudson, H., Larkin, R.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**, 601–609 (1994)
9. Karaoglanis, K., Polycarpou, I., Efthimiou, N., Tsoumpas, C.: Appropriately regularized OSEM can improve the reconstructed PET images of data with low count statistics. Hell. J. Nucl. Med. **18**, 140–145 (2015)
10. Lange, K., Carson, R.: EM reconstruction algorithms for emission and transmission tomography. J. Comput. Assist. Tomogr. **8**, 306–316 (1984)
11. Novosad, P., Reader, A.: MR-guided dynamic PET reconstruction with the kernel method and spectral temporal basis functions. Phys. Med. Biol. **61**, 46244645 (2016)
12. Shepp, L., Vardi, Y.: Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **1**, 113–122 (1982)
13. Thielemans, K., Tsoumpas, C., Mustafovic, S., Beisel, T., Aguiar, P., Dikaios, N., Jacobson, M.: STIR: software for tomographic image reconstruction release 2. Phys. Med. Biol. **57**, 867–883 (2012)
14. Vunckx, K., Atre, A., Baete, K., Reilhac, A., Deroose, C.M., Laere, K.V., Nuyts, J.: Evaluation of three MRI-based anatomical priors for quantitative PET brain imaging. IEEE Trans. Med. Imaging **31**, 599–612 (2012)
15. Walker, M., Asselin, M., Julyan, P., Feldmann, M., Talbot, P., Jones, T., Matthews, J.: Bias in iterative reconstruction of low-statistics PET data: benefits of a resolution mode. Phys. Med. Biol. **56**, 931–949 (2011)

Multifractal Analysis on Cancer Risk

Milan Stehlík, Philipp Hermann, Stefan Giebel, and Jens-Peter Schenk

Abstract Here we consider retroperitoneal tumors in childhood as examples from oncology generating difficult multicriterial decision problems. Inter-patient heterogeneity causes multifractal behavior of images for mammary cancer. Here we fit mixture models to box-counting fractal dimensions in order to better understand this variability. In this context the effect of chemotherapy is studied. The approach of Shape Analysis, proposed already in the work of Giebel (Bull Soc Sci Med Grand Duche Luxemb 1:121–130, 2008; Zur Anwendung der Formanalyse. Application of shape analysis. University of Luxembourg, Luxembourg, 2011), is used. This approach has considered a small number of cases and the test according to Ziezold (Biom J 3:491–510, 1994) is distribution free. Our method here is parametric.

1 Introduction

Most of the theories of tissue image analysis can be perspectively useful for better diagnostics of cancer. In this paper we address mainly multifractal phenomenon, observed in several recent studies (see, e.g., [9] and the references therein). In the last years, fractal and multifractal objects are largely used for modeling multiscale phenomena in several fields, including physics, geoscience, chemistry, and image processing. For theoretical background on multifractal approach to cancer, see [12].

M. Stehlík (✉)
Linz Institute of Technology (LIT) and Department of Applied Statistics, Johannes Kepler University, Linz, Austria

Institute of Statistics, Universidad de Valparaíso, Valparaíso, Chile

P. Hermann · S. Giebel
Department of Applied Statistics, Johannes Kepler University, Linz, Austria
e-mail: philipp.hermann@jku.at; sgiebel@gmx.de

J.-P. Schenk
Division of Pediatric Radiology, University Hospital of Heidelberg, Heidelberg, Germany
e-mail: Jens-peter.Schenk@med.uni-heidelberg.de

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_2

Data of landmarks are calculated according to the procedure, see [5]. As it was already discussed, this procedure is useful, if there are no landmarks in consequence of medical or theoretical aspects.

Shape analysis approach is applied on retroperitoneal tumors in childhood. Analysis is based on 2D images of magnetic resonance images with nephroblastomas (Wilms tumors). A platonic body C60 for 3D is constructed by using 2D images for 3D Shape Analysis. Additionally 3D Shape Analysis is performed for patients with neoadjuvant chemotherapy of the guidelines of the SIOP/GPOH study group for renal tumors in childhood.

2 Indication for Shape Analysis Approach for Wilms Tumors

Image recognition and classification of objects according to images is very important for medicine [1, 10, 11]. Important aspects when producing similarly processed images are, on the one hand, automated data entry, and on the other hand, its manageable evaluation. Mathematical procedures can support the applicants in their evaluation of magnetic resonance imaging, which is proposed in the example of nephroblastomas.

Nephroblastoma (Wilms tumor) is the most common tumor type in childhood and occurs in the majority of cases in the first decade of life [17, 18]. Genetic predisposition is suspected to increase the risk of nephroblastomas. The chance of curing cancer has been increased in the last decades depending on tumor stage and histological subtype. Tumor stage is dependent on primary imaging surgical and pathological findings. Modification of chemotherapy depends on multiple factors, e.g. tumor volume and volume regression under chemotherapy.

Nephroblastomas, limited on the kidneys, in early stages are a huge tumor mass, defined by a pseudocapsula. In higher stages this capsula is often penetrated from tumor tissue. The pseudocapsula helps to define the tumor shape and tumor tissue shows lower contrast enhancement than renal tissue, so tumor can be differentiated in MRI (magnetic resonance images). All the following conclusions refer to images and not to histological tumor extent. The differentiation of a nephroblastoma from other retroperitoneal tumors is complicated and multiple factors influence the radiological decision, e.g. tumor structure, tumor origin, displacement of neighboring anatomical structures, tumor volume, patients age [5], or presence of metastasis. Also the tumor shape is an aspect in the radiological decision.

In the radiological experience a huge round or oval mass is expected and a decision is performed empirically. The aspect of tumor shape should be studied on the basis of the theories of Ziezold [19] and the studies of Giebel et al. using 2D, 3D, and 4D Shape Analysis on medical data (see [5]).

Based on the data pool of [5] the statistical differences for shape of a nephroblastoma study group and a study group of non-Wilms tumors will be demonstrated using fractal analysis. All patients data in this paper are anonymized.

Magnetic resonance images deliver 2D images after being prescribed due to suspicion of Wilms tumor. Images of these screenings are basis for constructing a three-dimensional object of the renal tumor. Physicians have a huge interest to find markers for a good differentiation to avoid misclassifications. Especially with regard to automatic diagnosis in the next stage of future technologies all aspects of a tumor must be statistically analyzed. For this we can develop decision making techniques with probability of a tumor diagnosis.

Shape analysis allows to form more-dimensional objects with the aid of mathematical procedures, characterizing objects on the basis of key points, called landmarks. Standardization and centralization regarding size and position of the object allow comparing objects and differentiating between stages of tumors. In total 60 landmarks are taken as the cut-points between the surface of the tumor and the vector of the edge of the platonic body C60 [5, 6].

The size is eliminated by standardization, thus all objects are comparable from the point of statistics. We aim to differentiate tumors, not to detect them like in [14]. Therapy is organized with respect to therapy-optimizing studies of the Society of Pediatric Oncology and Hematology (SIOP). On the basis of radiological findings and reports indication of neoadjuvant chemotherapy is given. Since other tumors exist, we compare our nephroblastoma group with non-Wilms tumors, including tumors, which are difficult to differentiate [16]. See also [3, 7, 15].

2.1 Differentiation Between Different tumor Groups

Theoretical concepts of a group of objects like in anatomy enable to select landmarks describing these objects. Lack of theoretical concepts require other procedures to find landmarks. Here, 3D landmarks are obtained by constructing a three-dimensional object of the tumor, when every landmark consists of an x-, y-, and a z-coordinate. Then landmarks are taken as in Sect. 2. The data should be differentiated only on the basis of these three coordinates. This enables to compare the values of each of the coordinates between the two groups of tumors. The sample size ($n = 37$) comprised of 30 nephroblastomas and 7 non-Wilms tumors. The number of measured points in order to get the exact location of the landmarks varies to a greater extent (between 186 and 6638) due to the explorative approach based on geometric methods. Wilcoxon- and t-test yield evidence for *statistical significance of recognized differences* in descriptive statistics (mean, median) between the two

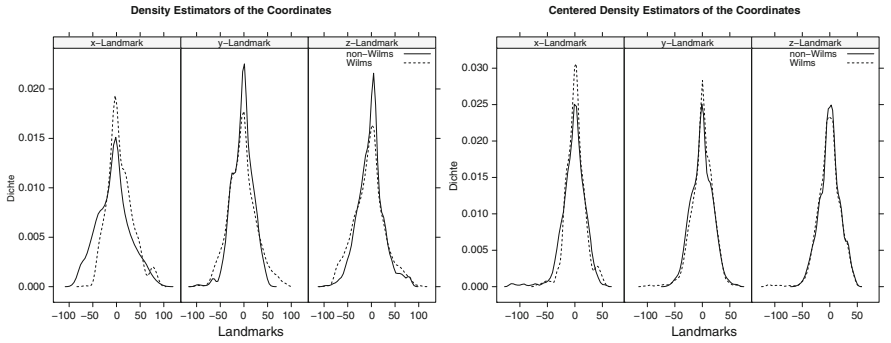


Fig. 1 Left: density estimators of the coordinates and groups; right: density estimators of the centered coordinates and groups

groups, for densities of the estimators of the coordinates, see Fig. 1. By graphical inspection of these figures (after centering of the non-representative small sample) we can see less variability in the group of non-Wilms tumors.

2.2 4D-Shape Analysis

By including the time after and before chemotherapy we have a 4D-approach to Shape Analysis. In ten patients a pre- and post-therapeutical MRI has been investigated in order to test the change in the landmarks before and after chemotherapy. This results to 20 observations. For analysis of the change of the tumor shape we suggest the term 4D Shape Analysis. For allowance of comparison of size of the object, each landmark was centered according to its center value in advance. A decrease in the mean of the standard deviation can be observed for y- and z-coordinate after therapy. Moreover, mean and median are close to zero. However, minima and maxima are smaller for all coordinates after therapy, which allows us for our sample to investigate the effects of the therapy. In order to have a graphical comparison of the data before and after chemotherapy, density estimators were given like in [8].

Figure 2 contains 3D-scatterplots in order to compare the standardized size of the object for three of the patients before and after therapy. Plots of the landmarks before therapy can be found in the first line and the corresponding plots after therapy are beneath. Apparently, by graphical inspection of the standardized volume of the object we can conclude that volume decreased for displayed patients 1 and 2 (all axes have same scale). This result is expectable from the point of medicine.

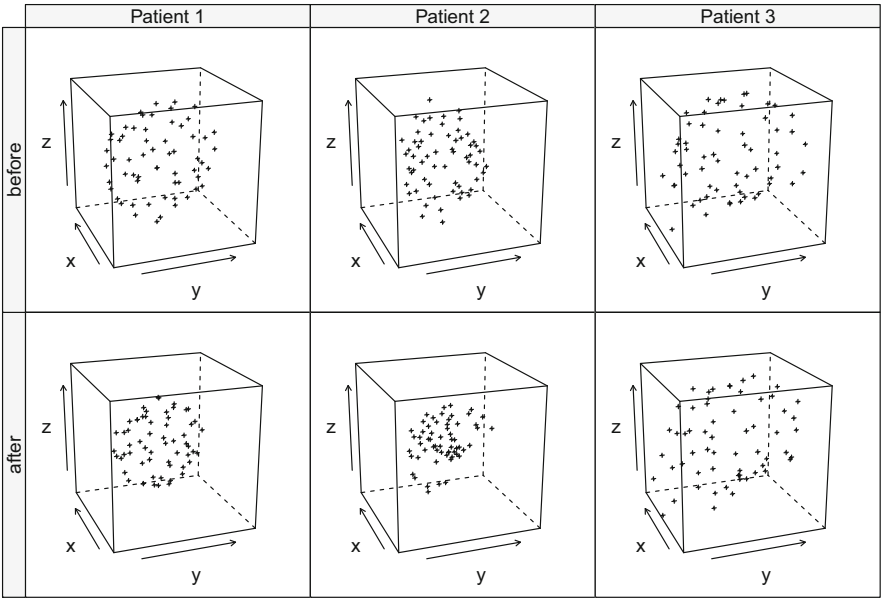


Fig. 2 Scatterplot for three patients before and after therapy

2.3 Mixed Distribution Estimation

Since our sample size is small in order to check for heterogeneity we use mixtures to model different modes in the data. One of the advantages of this approach is to recognize outliers in the sample. We estimate parameters of mixed normal distributions, obtained via the function `normalmixEM` of the package `mixtools` [2]. We use the following notation for later presentation of the results.

$$f(x) = \sum_{i=1}^{2} \lambda_i \, p_i(x), i = 1, 2, \tag{1}$$

where $p_i \sim N(\mu_i, \sigma_i^2), i = 1, 2$, and $\lambda_i, i = 1, 2$ being the weights. We will investigate the distributional behavior, i.e. affections on the distribution for every coordinate, for both, different groups of malignancy and the impact of chemotherapy.

2.3.1 4D Shape Analysis

We have conducted the same computations on different differential diagnosis, Wilms and non-Wilms tumors as for data on pre- and post-chemotherapy. Here, the data of ten patients, observed before and after chemotherapy is considered (see

Table 1 Mixture estimates of mean μ_i, standard deviation σ_i, and weighting parameter λ_i, $i = 1, 2$ for x-, y-, and z-coordinate

Coordinate	Status	μ_1	μ_2	σ_1	σ_2	λ_1	λ_2
x	Before	−19.65	20.46	17.52	20.10	0.49	0.51
	After	−10.26	21.95	18.93	12.68	0.81	0.19
	Before-centered	−6.41	33.41	24.22	8.83	0.84	0.16
	After-centered	−10.40	9.57	16.70	16.22	0.47	0.53
y	Before	−32.57	85.27	41.91	17.32	0.92	0.08
	After	−58.32	−10.12	27.75	18.47	0.54	0.46
	Before-centered	−15.32	20.97	24.17	20.70	0.61	0.39
	After-centered	0.04	−0.91	23.13	13.08	0.63	0.37
z	Before	−19.65	20.46	17.52	20.10	0.49	0.51
	After	−10.20	22.05	18.96	12.64	0.82	0.18
	Before-centered	−6.41	33.41	24.22	8.83	0.84	0.16
	After-centered	−0.60	21.40	19.12	4.30	0.96	0.04

Estimates are reported separately with respect to status of chemotherapy and whether original or centered (transformed) data is investigated

Sect. 2.2). The mixture distribution of the data is estimated via `normalmixEM` function and the corresponding estimated parameters are reported in Table 1. The first two columns represent the coordinate of interest, whereby for each coordinate four estimations were performed. These distinguish between the status of the chemotherapy (before/after) as well as between the constitution of the data (original/centered; note that original data does not have an extra label). These descriptors are followed by mean, standard deviation, and weighting parameters. As we can see in Table 1 at least one of both standard deviations of estimated mixture individual densities decreased after chemotherapy.

Figure 3 shows the estimated density estimates of the Gaussian mixture. Every plot contains the results for the x-, y-, and z-coordinate in black, red, and green, respectively. On the left-hand side the plots before therapy are provided with those containing the estimates after therapy on the right-hand side. The first row represents estimates based on original data and the second row those of the centered data. Plots in one column have the same scales for the sake of comparability. In the analysis of the coordinates the highest complexity relates to the coordinate z (see bimodal distribution of z coordinates (green color) in Fig. 3). This could be a result of the survey since in every MRI you can measure directly x and y coordinates, however z coordinate depends on thickness of the slides.

2.3.2 Checking of Consistency

The estimation procedure for mixed normal distribution leads to different results for every computation. The estimation of normal mixtures for small samples is not an easy task from the statistical perspective and it depends on starting values of

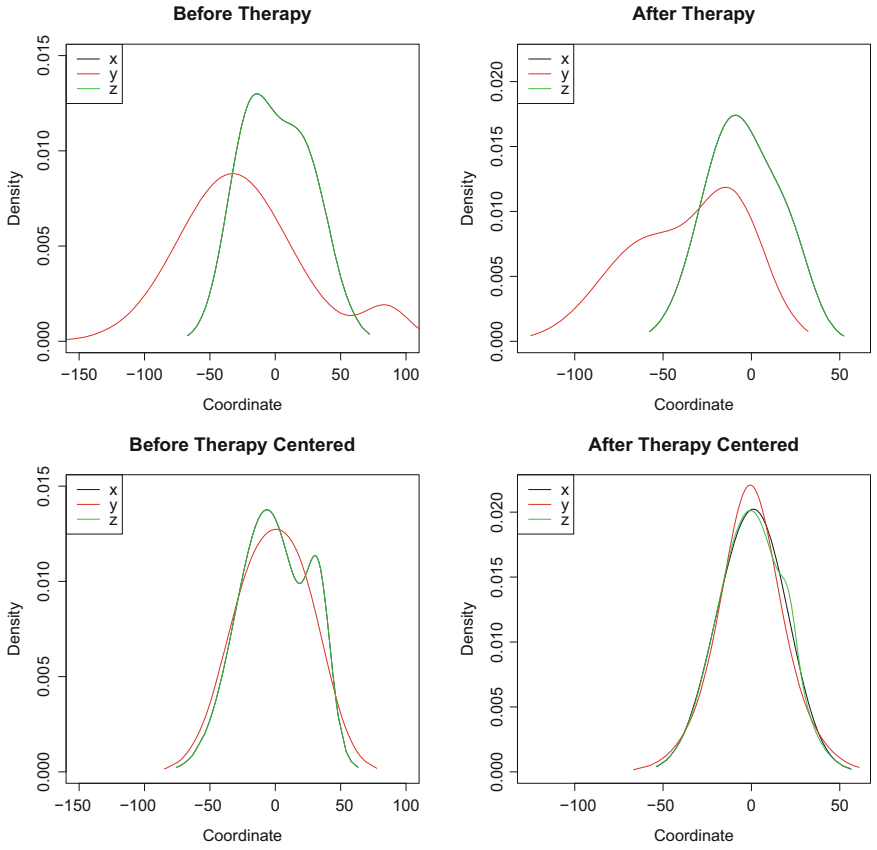


Fig. 3 Mixture density estimates for *x*-, *y*-, and *z*-coordinates before and after chemotherapy for original and centered data

parameters. Since we have not defined any starting values, one can expect that in cases of consistency, the same estimates (or slightly different) will be computed for the different (because randomly chosen by the program) starting values. Therefore, we have conducted a sensitivity study, where each computation is repeated 1000 times. Hence, the estimates for $\mu_1, \mu_2, \sigma_1, \sigma_2$ and λ_1, λ_2 are saved for every run and summarized by standard descriptive statistics in Table 2. These computations have been applied for both data sets, i.e. before and after therapy. We have obtained these values with statistics software *R* [13].

Table 2 shows the effects of the therapy on the obtained estimates in our small sample for $\mu_i, \sigma_i, \lambda_1, i = 1, 2$ in addition to the behavior of the estimates themselves. We only provide the estimated weight λ_1, because of $\lambda_2 = 1 - \lambda_1$. If the results of the parameters vary to a greater extent, it can be assumed that the estimates are not consistent. Since before therapy tumor tissue is larger, we can try to understand this that more than one distribution is needed to model the irregular

Table 2 Descriptive statistics of the estimates of the location, scale, and weighting parameter for the mixture distribution based on 1000 replicates of the sensitivity study

Coordinate	Parameter	Status	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	IQR	SD
x	μ_1	Before	−34.94	−34.75	−34.75	−34.21	−34.74	−4.91	0.01	3.83
		After	−42.75	−41.58	−41.58	−41.35	−41.58	−1.85	0.00	4.42
	μ_2	Before	−4.87	51.25	51.25	50.22	51.25	53.79	0.00	7.43
		After	−1.83	39.84	39.84	33.92	39.84	39.84	0.00	10.67
	σ_1	Before	22.55	23.82	23.82	24.36	23.82	52.39	0.00	3.88
		After	12.09	18.87	18.87	17.66	18.87	45.08	0.00	4.24
	σ_2	Before	36.44	36.55	36.55	36.85	36.55	52.39	0.00	2.16
		After	19.83	19.83	19.83	25.33	19.83	45.08	0.00	9.73
	λ_1	Before	0.00	0.35	0.35	0.34	0.35	0.49	0.00	0.03
		After	0.00	0.49	0.49	0.44	0.49	0.49	0.00	0.08
y	μ_1	Before	−32.57	−31.55	−31.55	−31.72	−31.55	−24.24	0.00	0.58
		After	−58.33	−58.33	−58.33	−58.28	−58.32	−54.72	0.01	0.42
	μ_2	Before	−18.20	−18.20	−18.20	1.69	−18.19	85.27	0.01	40.76
		After	−36.03	−10.13	−10.13	−10.49	−10.12	−10.12	0.01	3.05
	σ_1	Before	2.86	25.29	25.29	23.70	25.30	25.30	0.01	3.34
		After	0.42	18.47	18.48	18.22	18.48	18.48	0.01	2.12
	σ_2	Before	41.91	62.63	62.63	58.62	62.63	62.63	0.00	8.17
		After	27.74	27.74	27.74	27.83	27.75	34.04	0.01	0.74
	λ_1	Before	0.04	0.41	0.41	0.35	0.41	0.41	0.00	0.13
		After	0.01	0.46	0.46	0.45	0.46	0.46	0.00	0.05
z	μ_1	Before	−19.66	−19.66	−19.66	−18.59	−19.65	0.54	0.01	4.27
		After	−39.17	−10.51	−10.21	−10.84	−10.21	−4.25	0.30	3.14
	μ_2	Before	0.63	20.45	20.45	21.26	20.46	35.12	0.01	3.44
		After	−4.25	21.45	22.02	19.68	22.04	22.19	0.59	6.79
	σ_1	Before	0.82	17.52	17.52	16.94	17.52	27.52	0.00	2.43
		After	0.57	12.64	12.64	12.35	12.78	21.88	0.14	2.90
	σ_2	Before	20.10	20.10	20.11	20.50	20.11	27.58	0.01	1.60
		After	16.84	18.89	18.95	19.09	18.95	22.07	0.06	0.86
	λ_1	Before	0.00	0.49	0.49	0.47	0.49	0.49	0.00	0.10
		After	0.00	0.49	0.49	0.47	0.49	0.49	0.00	0.10

These statistics are reported for x-, y-, and z-coordinates as well as status of the chemotherapy (before/after)

tissue growth with the aid of the landmarks. However, after therapy, this chaos seems to be reduced and can occasionally be estimated adequately with the aid of a normal distribution.

Boxplots are presented in Figs. 4, 5 and 6 providing the estimates of the parameters of the mixture distribution. These figures build on the same pattern: leftmost group of boxplots contains the location parameters, $\mu_i, i = 1, 2$; middle representations provide information on the scale parameters, $\sigma_i, i = 1, 2$; right graphs contain the weightings, $\lambda_i, i = 1, 2$. The reason for presenting two different

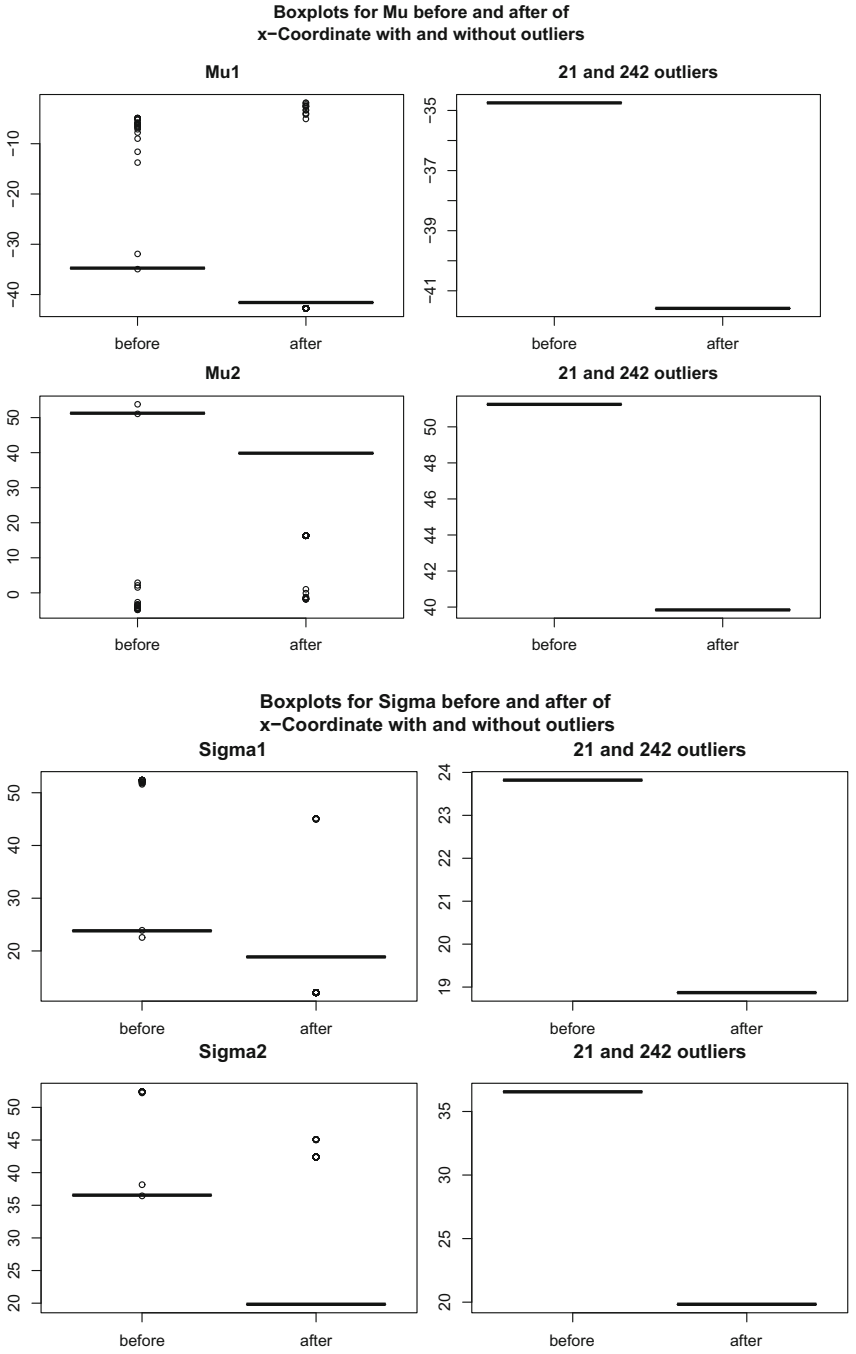


Fig. 4 Boxplots of the estimates of location μ_i, scale σ_i, and weighting parameters λ, $i = 1, 2$ for x-coordinate with and without candidates for outliers based on 1000 replicates

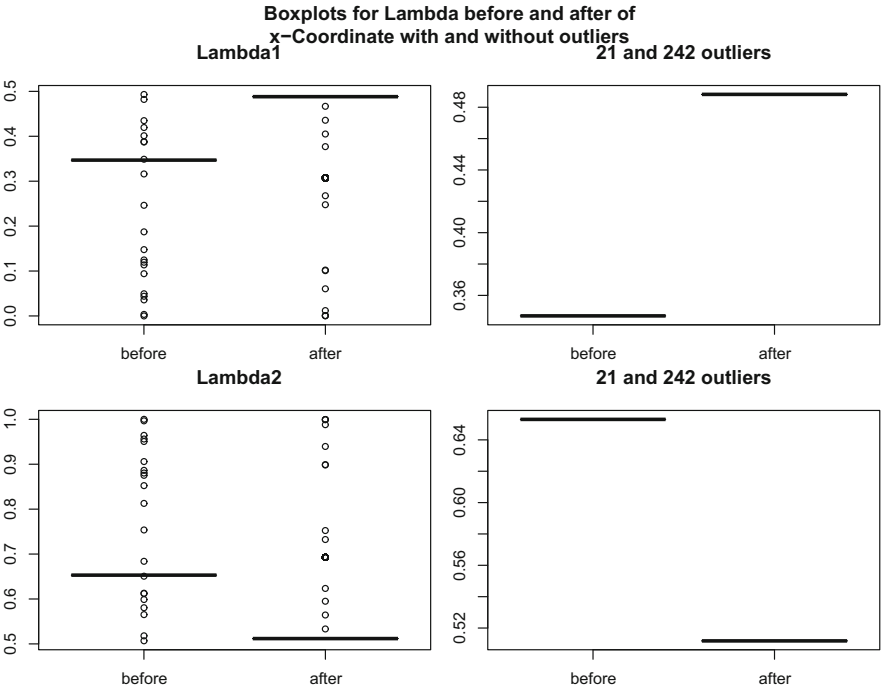


Fig. 4 (continued)

plots is the impact of "outliers," strongly affecting the comparability with respect to therapy status. Hereafter, outlier detection is based on the "boxplot-rule," which has more extreme values than $q_{0.25} - 1.5 \cdot \text{IQR}$ or $q_{0.75} + 1.5 \cdot \text{IQR}$. Note that IQR is the interquantile range as $q_{0.75} - q_{0.25}$ and q_α is the α-quantile. Moreover, the number of outliers of both groups is reported as the main caption for each plot. Hence, the information provided in the middle plots of Fig. 5 indicates that 195 outliers are detected for σ_1 before therapy and 14 candidates for outliers resulted for the after therapy data. These numbers show that there is a very strong variation present in the estimates. However, when comparing the results of the boxes there are different estimates for mean for every coordinate. In contrast to that, only for y- and z-coordinates the estimates of standard deviation and weightings differed, whereas x-coordinate shows overlapping boxes before and after therapy. Generally, the boxplots show that especially after therapy the parameters vary stronger, wherefore the possibility of mixture distribution of the landmarks after therapy has to be doubted. We do not claim this in general.

As we can see in Figs. 4, 5 and 6, even in the small non-representative sample of ten patients, statistically significant heterogeneity can be observed. This large

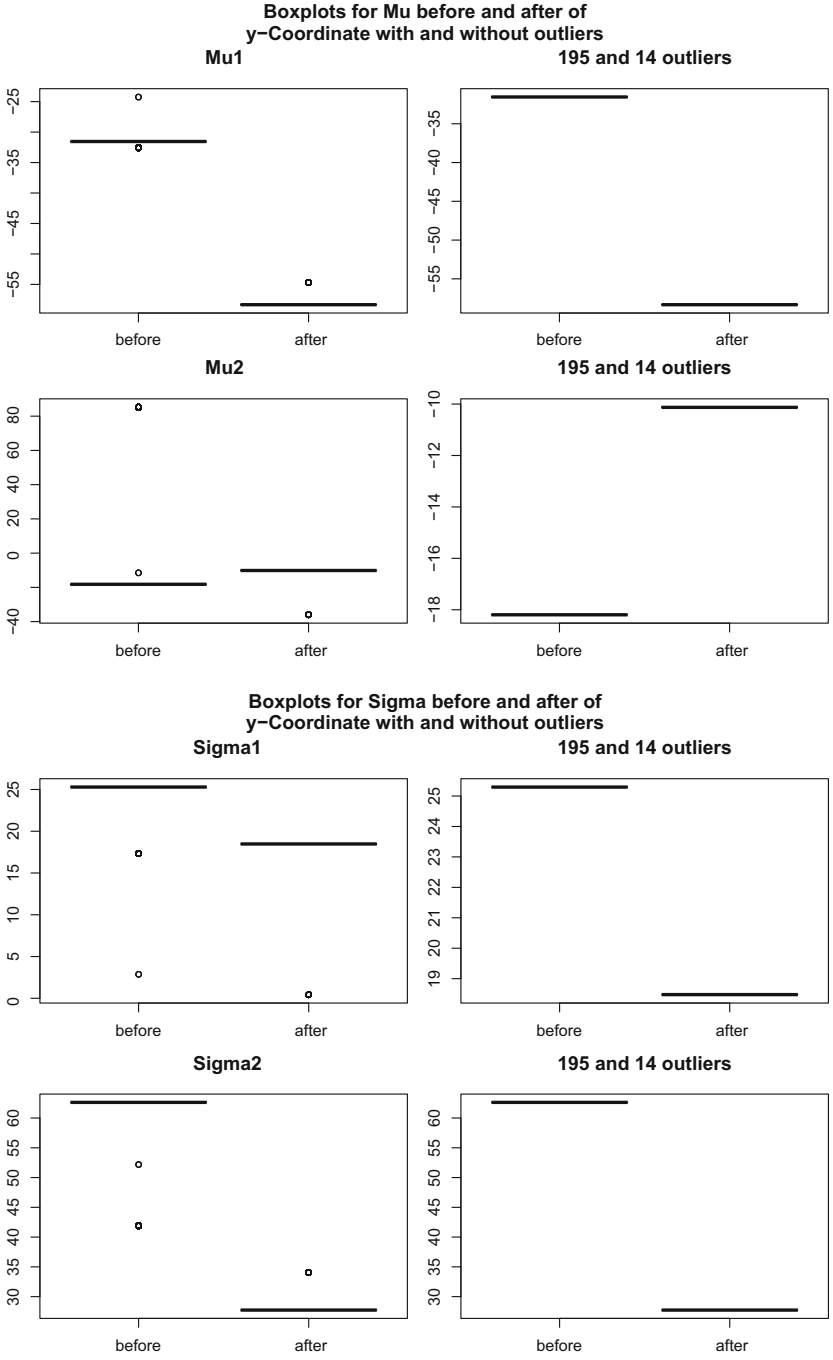


Fig. 5 Boxplots of the estimates of location μ_i, scale σ_i, and weighting parameters λ, $i = 1, 2$ for y-coordinate with and without candidates for outliers based on 1000 replicates

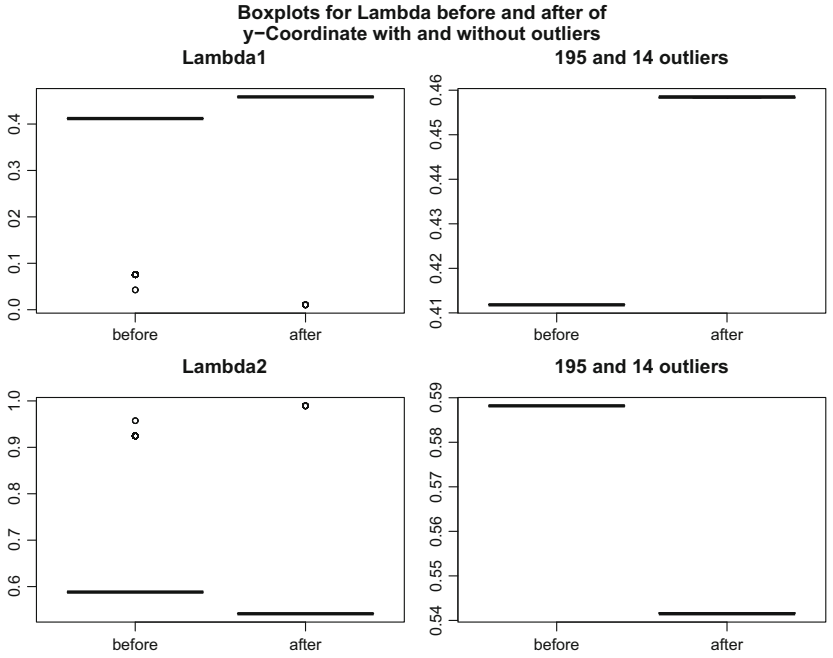


Fig. 5 (continued)

heterogeneity makes possible automatic decision procedure very complicated and thus an experienced oncologist and radiologist will always be necessary. From the statistical point of view we need also a bigger sample to prove this heterogeneity by means of statistics.

2.3.3 Estimation to Differentiate tumors

We have conducted the same computations on different diagnosis, Wilms and non-Wilms, as for the data on pre- and post-chemotherapy. Table 3 reports the estimated parameters for the mixture distributions differentiating between the groups and the constitution of the data.

As we can see in Table 3, after centering, there is a higher variance of normal bimodal mixture for non-Wilms in comparison with Wilms tumors. Centering is non-avoidable in order to stabilize the variance.

Figure 7 follows the same structure as Fig. 3 in terms of coordinates. Note that hereafter the different columns represent the groups Wilms (first column) and Non-Wilms (second column), differing between the original (first row) and the centered data (second row).

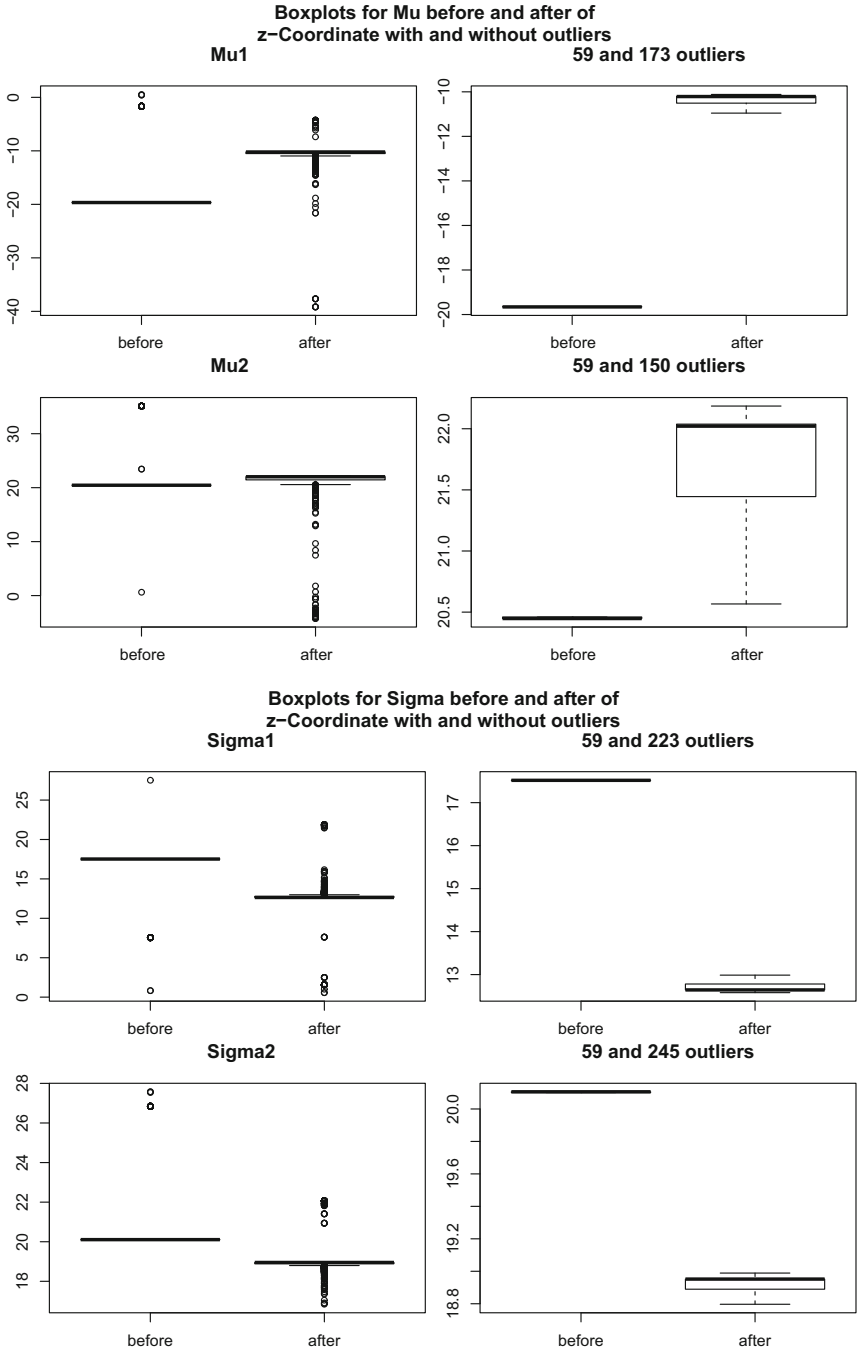


Fig. 6 Boxplots of the estimates of location μ_i, scale σ_i, and weighting parameters λ, $i = 1, 2$ for z-coordinate with and without candidates for outliers based on 1000 replicates

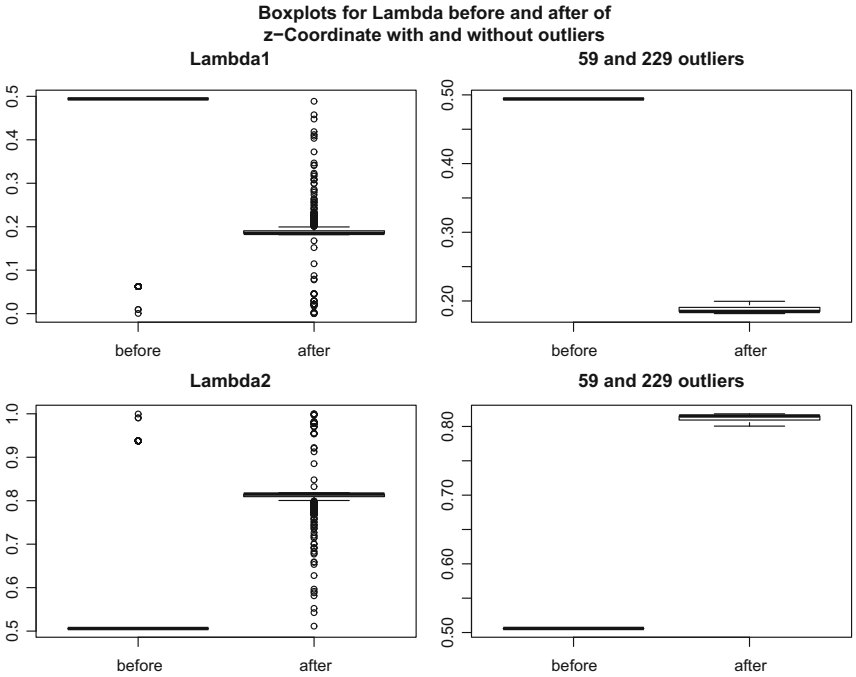


Fig. 6 (continued)

Table 3 Mixture estimates of mean μ_i, standard deviation σ_i, and weighting parameter λ_1, $i = 1, 2$ for x-, y-, and z-coordinate

Coordinate	Status	μ_1	μ_2	σ_1	σ_2	λ_1	λ_2
x	Non-Wilms	9.92	47.47	31.64	1.14	0.98	0.02
	Wilms	−24.54	50.09	27.95	19.40	0.78	0.22
	Non-Wilms-centered	2.93	−0.99	26.74	11.45	0.39	0.61
	Wilms-centered	0.09	−85.39	17.35	21.09	0.97	0.03
y	Non-Wilms	−43.12	1.47	15.84	18.54	0.06	0.94
	Wilms	−9.00	13.63	15.97	15.09	0.57	0.43
	Non-Wilms-centered	0.24	−0.79	22.37	8.23	0.67	0.33
	Wilms-centered	−16.20	5.15	13.80	15.66	0.28	0.72
z	Non-Wilms	−18.40	0.61	25.33	3.02	0.81	0.19
	Wilms	4.99	−2.33	3.00	28.94	0.14	0.86
	Non-Wilms-centered	0.31	−0.57	27.62	11.11	0.44	0.56
	Wilms-centered	1.47	−35.21	17.35	8.88	0.96	0.04

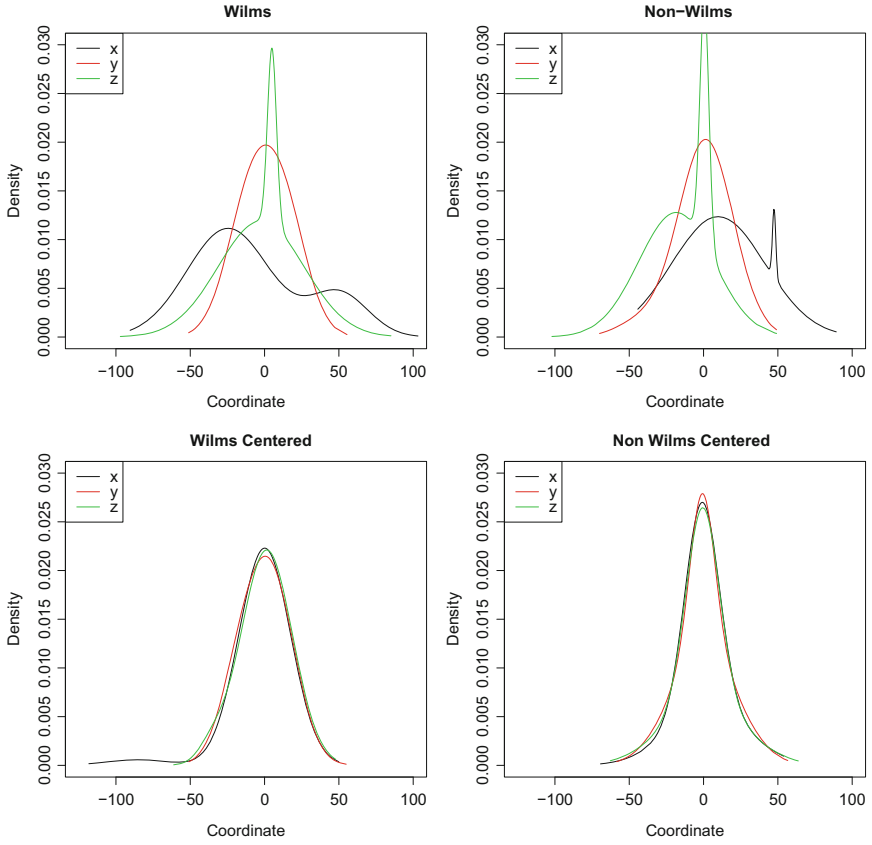


Fig. 7 Mixture density estimates for *x*-, *y*- and *z*-coordinate for original and centered data

3 Conclusion

Shape analysis and fractal analysis are tools in order to analyze cancer data. The standardly used mathematical procedure of constructing the platonic body C60 (see [4]) for the renal tumor allowed us to differentiate on the basis of distributions and means between the groups in our small sample volume with limited number of patients which cannot be representative for all nephroblastoma patients. Medical staff can be a support in the diagnosis decision process with this geometrical approach built on MR images. However in its current stage the analysis is preliminary. Limitations of the study are the primary approach with 2D images and a secondary reconstruction of 3D images. With primary 3D images in MRI a better reconstruction of tumor shape can be expected. Not all histological tumor risk groups are represented in this study and the Non-Wilms-tumor study group is a mixed group of retroperitoneal tumors including non-renal tumors, so

the statistical results are dependent on the composition of the study groups. All results are dependent on the primary radiological decision, where the tumor edge can be defined, so with higher image resolution in modern MRI scanners, it could be more difficult to define the tumor edge in single cases. To mark the tumor edge only early local tumor stages are suitable for this method because in late tumor stages tumor shape changes because of ruptures of the pseudocapsula and infiltrating tumor tissue. But even in the case of transversal images the Wilms tumors and the non-Wilms tumors showed a different statistical behavior in our results. Hence our results give a further development of the results of [6] and a possible support for decision making in a non-empiric way of a diagnosis in imaging when anatomical shape is a criterion for diagnosis.

In contrast to [6] minimization of variance by using a neural network is not necessary anymore. The test results in our original centered data using fractal analysis confirm already the differentiability of Wilms to non-Wilms tumors.

Acknowledgements Milan Stehlík acknowledges FONDECYT Regular No1151441 and LIT-2016-1-SEE-023 mODEC. This work was also supported by the Slovak Research and Development Agency under the contract No. SK-AT-2015-0019. Philipp Hermann was supported by ANR project DESIRE FWF I 833-N18.

References

1. Baish, J.W., Jain, R.K.: Fractals and cancer (2000). https://doi.org/10.1016/0306-9877(94)90163-5
2. Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S.: mixtools: An R package for analyzing finite. J. Stat. Softw. **32**(6), 1–29 (2009). https://doi.org/10.18637/jss.v032.i06
3. Furtwängler, R., Schenk, J.P., Reinhard, H., Leuschner, I., Rübe, C., von Schweinitz, D., Graf, N.: Nephroblastom - Wilms-tumor. Der Onkologe **11**(10), 1077–1089 (2005)
4. Giebel, S.: Statistical analysis of renal tumours with infants. Bull. Soc. Sci. Med. Grand Duche Luxemb. **1**, 121–130 (2008)
5. Giebel, S.: Zur Anwendung der Formanalyse (Application of Shape Analysis). University of Luxembourg, Luxembourg (2011)
6. Giebel, S., Schiltz, J., Graf, N., Nourkami, N., Leuschner, I., Schenk, J.P.: Application of shape analysis on 3D images-MRI of renal tumors. J. Iran. Stat. Soc. **11**(2), 131–146 (2012)
7. Günther, P., Schenk, J.P., Wunsch, R., Tröger, J., Waag, K.L.: Abdominal tumours in children: 3-D visualisation and surgical planning. Eur. J. Pediatr. Surg. **14**(5), 316–321 (2004). https://doi.org/10.1055/s-2004-821042
8. Hermann, P., Giebel, S.M., Schenk, J.P., Stehlík, M.: Dynamic shape analysis - before and after chemotherapy. In: Guillén, M., Juan, Á.A., Ramalhinho, H., Serra, I., Serrat, C. (eds.) Current Topics on Risk Analysis: ICRA 6 and RISK 2015 Conference, Barcelona, pp. 339–346 (2015)
9. Hermann, P., Mrkvička, T., Mattfeldt, T., Minárová, M., Helisová, K., Nicolis, O., Wartner, F., Stehlík, M.: Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process. Stat. Med. **34**(18), 2636–2661 (2015). https://doi.org/10.1002/sim.6497
10. Mattfeldt, T., Meschenmoser, D., Pantle, U., Schmidt, V.: Characterization of mammary gland tissue using joint estimators of Minkowski functionals. Image Anal. Stereol. **26**(1), 13–22 (2007). https://doi.org/10.5566/ias.v26.p13-22

11. Mrkvička, T., Mattfeldt, T.: Testing histological images of mammary tissues on compatibility with the boolean model of random sets. Image Anal. Stereol. **30**(1), 11–18 (2011). https://doi.org/10.5566/ias.v30.p11-18
12. Nicolis, O., Kiselák, J., Porro, F., Stehlík, M.: Multi-fractal cancer risk assessment. Stoch. Anal. Appl. **35**, 237–256 (2017). https://doi.org/10.1080/07362994.2016.1238766
13. R Core Team: R: A Language and Environment for Statistical Computing (2015). http://www.r-project.org/
14. Ratto, C., Sofo, L., Ippoliti, M., Merico, M., Doglietto, G.B., Crucitti, F.: Prognostic factors in colorectal cancer. Dis. Colon Rectum **41**(8), 1033–1049 (1998)
15. Schenk, J.P., Waag, K.L., Graf, N., Wunsch, R., Jourdan, C., Behnisch, W., Troger, J., Gunther, P.: 3D-visualization by MRI for surgical planning of Wilms tumors. RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin **176**(10), 1447–1452 (2004). https://doi.org/10.1055/s-2004-813398
16. Schenk, J.P., Graf, N., Günther, P., Ley, S., Göppl, M., Kulozik, A., Rohrschneider, W.K., Tröger, J.: Role of MRI in the management of patients with nephroblastoma. Eur. Radiol. **18**(4), 683–691 (2008). https://doi.org/10.1007/s00330-007-0826-4
17. Ward, E., DeSantis, C., Robbins, A., Kohler, B., Jemal, A.: Childhood and adolescent cancer statistics. CA Cancer J. Clin. **64**(2), 83–103 (2014). https://doi.org/10.3322/caac.21219
18. Wilms, M.: Die Mischgeschwülste der Niere. Verlag von Arthur Georgi, Leipzig (1889)
19. Ziezold, H.: Mean figures and mean shapes applied to biological figure and shape distributions in the plane. Biom. J. **36**, 491–510 (1994)

Traditional Versus Alternative Risk Measures in Hedge Fund Investment Efficiency

Izabela Pruchnicka-Grabias

Abstract The author presents results of the research conducted for hedge funds for the period of 1990–2014. They were divided into ten investment strategies and net asset values calculated for indexes created for them were used. Chosen alternative risk-return ratios (Calmar, Sterling and Burke ratio) were calculated and their values were compared with the Sharpe ratio for the same period of time. The main conclusion is that these alternative measures give different results from the traditional Sharpe ratio, that is hedge fund rankings made with these two kinds of measures are not the same. This in turn indicates that arguments of opponents of using traditional efficiency ratios in the hedge fund analysis may not be exaggerated.

1 Introduction

Hedge funds are financial institutions which aim at generating absolute rates of return, that is at realizing profits regardless of the market situation. They are the subject of a wide scientific discussion concerning the rates of return generated compared with other forms of investments. However, some scientists claim that although hedge funds really achieve attractive rates of return, at the same time, they generate high risk level. Another problem which has not been solved so far is what is the most adequate measure or group of measures for this type of risk (sometimes called the extreme risk). The most typical risk measure used both by hedge funds and investment funds when they present results of their portfolio management is the standard deviation and the risk-return measure is the Sharpe ratio. Opponents of using it say that it requires the assumption that the rates of return are normally distributed which doesn't have to be true. Assuming that it would be true, one can apply the whole range of alternative risk-return measures. Needless to say that they are more complex, require more advanced knowledge of

I. Pruchnicka-Grabias (✉)
Warsaw School of Economics, Institute of Banking, Warsaw, Poland
e-mail: ipruch@sgh.waw.pl

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_3

finance, mathematics, and statistics, and generate risk of making a mistake during calculations. It is hard to answer the question if the above named disadvantages are worth getting higher accuracy of results. It requires giving an answer to the question if the results achieved with the use of alternative measures are much different from those based on traditional ratios. If they are different, another question arises, if they are more adequate.

The author presents the results of the research conducted for hedge funds for the period of 1990–2014. They were divided into ten investment strategies and net asset values calculated for indexes created for them were used. The database used in the research comprises more than 2200 hedge funds, so it can be treated as a sufficient source. One hedge fund can use more than one strategy. Data were provided by Hedge Fund Research (www.hedgefundresearch.com) and generally they consider hedge funds from all over the world which report their results in American Dollars. However, the majority of hedge funds have their domiciles in the United States. The database prepared by Hedge Fund Research is one of the most acknowledged databases on hedge funds. The author did not divide hedge funds into strategies on her own but used classifications applied by the data provider which because of its high clarity, is often used in scientific research. Chosen alternative risk-return ratios (Calmar, Sterling, and Burke ratio) were calculated and their values were compared with the Sharpe ratio for the same period of time. As the minimum rate of return accepted by an investor, the risk-free interest rate was used. Mathematically it was reflected by the interest rate of 10Y American treasury bonds at the end of the research period, that is at the end of 2014 (2.16%). The main conclusion is that these two kinds of risk-return measures (traditional and alternative ones) give different results, that is hedge fund rankings made with them are not the same. This in turn indicates that arguments of opponents of using traditional efficiency ratios in the hedge fund analysis may not be exaggerated.

2 Literature Overview

The most spectacular example of the hedge fund extreme risk profile was the Long Term Capital Management bankruptcy in 1998. Many hedge fund investors achieved huge losses then. The hedge fund which existed for so many years suddenly was not able to operate any more. Many scientists blamed insufficient regulations of such entities created by governments [8]. Such a disaster has shown that hedge funds investments can be risky and that there is a need to monitor risk generated by these institutions with proper risk measures. Some others say that these were inadequate risk measures which are used to assess their results and look for some alternative ones [6, 9, 10, 13–15, 17, 20, 31]. Some measures of systematic risk made by them are also proposed in the literature [16]. Some authors stress that rates of return achieved by hedge funds are not normally distributed and in fact they are characterized by fat tails [18]. They find the fat tail risk is one of the most important factors influencing hedge fund performance. Cao et al. [4] pay attention

to the market liquidity which impacts hedge fund assets and question if hedge funds can make adjustments when it changes. Authors conclude that it is important to take the market liquidity into account while the process of decision making. Moreover, market liquidity makes that asset valuation is more difficult and thus is the assessment of hedge fund performance. Besides, it is not taken into consideration by the majority of models created by the theory of finance such as CAPM (Capital Asset Pricing Model) [29] or APT [26] (Arbitrage Pricing Theory) which means that it creates the model risk. Simultaneously, Sadka [27] reports that liquidity in the hedge fund world can be a good predictor of their prospective performance.

Another problem with hedge funds investment results assessment is that they start to report to databases when their returns are abnormally attractive and terminate to report when they are not so impressive anymore or when they suffer from capital outflows. The research shows that rates of return done by self-reporting funds are usually higher than those taken by non-reporting ones [1].

According to the Survey conducted by Ernst and Young [11], it is the asset growth which is the top priority for 57% of hedge fund managers, second priority for 20% of them, and third priority for 10% of managers. Talent management and operational efficiency are after that. It clearly shows that risk control is not the thing which they are worried about. Such an attitude towards investments creates not only high risk level to potential investors, but also high systemic risk which may result in financial system destabilization. This is why risk created by hedge funds should be monitored by financial market authorities.

Another aspect of hedge fund risk posed in the literature is the financial leverage used by these institutions. Duffie et al. [7] show that there is some optimal level of hedge fund leverage which should not be exceeded. McGuire et al. [23], Schneeweis et al. [28], AIMA [2] and FCA [12] show different financial and economic leverage definitions. No matter how it is understood, it is undoubtedly an important source of hedge fund risk.

There are many difficulties with analyzing hedge funds and monitoring risk generated by them because they use the whole variety of investment strategies. There are different classifications of them in the literature. For example, FCA [12] gives the following depicted in Table 1.

FCA [12] stresses that Long/Short Equity and Multi-strategy are the most popular strategies and they account for about 40% of the total number of hedge funds in their sample. Contrary to the above-mentioned classification, Tran [33] and Guizot [15] provide the following kinds of hedge fund investment strategies:

- Convertible Arbitrage—focused on investments in convertible bonds (long) and at the same time underlying stocks (short), looking for arbitrage opportunities aiming at generating profits higher that risk-free interest rate.
- Dedicated Short Bias—aimed at long and short positions in different securities, however with the majority of short ones. They concentrate on companies which are in a bad financial situation.
- Emerging Markets—targeting at investments in different securities, commodities, or currencies coming from so-called emerging economies.

Table 1 Hedge fund strategies classification

Strategies	Sub-strategies
Equity hedge	Long bias
	Long/Short
	Market neutral
	Short bias
Relative value	Convertible bond arbitrage
	Fixed income arbitrage
	Volatility arbitrage
Event driven	Distressed/Restructuring
	Equity special solutions
	Risk arbitrage/Merger acquisition
Credit	Asset backed lending
	Long/Short
Macro	Global macro
	Active trading
	Commodity
	Currency
Managed futures/CTA	Fundamental
	Quantitative
Multi-strategy	
Other	

Source: Financial Conduct Authority, Hedge Fund Survey, June 2015, pp. 14–15

- Long/Short Equity—taking long positions in stocks, as well as conducting short sale transactions without any assumption on the dominance of any of them. They can also use derivatives.
- Equity Market Neutral—constructing the risk-free stock portfolio which is not sensitive to chosen risk factors and at the same time trying to realize profits form chosen group of stocks.
- Fixed Income Arbitrage—the purpose is to take positions in fixed income securities and generate arbitrage profits at the low risk level. They use inefficiencies in their valuations. Interest rate swaps, futures, contracts, and mortgage securities can be also applied here.
- Event Driven—based on taking advantage of different market events like mergers, acquisitions, or restructuring.
- Global Macro—investing in various macroeconomics events. They often use high financial leverage.
- Managed Futures—actively managed accounts, with the use of commodity and currency futures contracts, stocks, or bonds high risk is generated and high rates of return are possible.
- Multistrategy—any strategy which can come into manager's mind on any market and in any securities.

The second classification is more transparent, so the author used it in the research.

3 Problems of Hedge Funds Efficiency Measurement

For the purpose of this study investment efficiency should be understood as the relation between the rate of return and risk. According to the [22] portfolio theory, investors choose higher rate of return at the same risk level and lower risk level if the rate of return is the same. A special situation arises if an investor is not punished for taking high risk level. It tempts them to generate high risk level because thanks to that high rate of return can be achieved. At first glance it seems to be an impossible situation, however it takes place in hedge funds investments in relation to their managers. The reason is that the system of their wages is based on commissions depending on the rates of return, not on the risk level taken by them. Besides, they are assessed on the basis of historical rates of return regardless of risk taken. Ethical matters are not taken into consideration. They are generally difficult to measure. Until high rates of return are generated by the manager, even high risk level is not a problem. The situation changes if the risk materializes. Investors start to withdraw their capital from such institutions. Hedge fund risk measurement is tricky because traditional risk measures are based on standard deviation which requires the assumption that rates of return are normally distributed. The research conducted in this field shows that they are not only abnormal, but they are usually far away from normality [24].

4 Basic Hedge Funds Statistics

Table 2 depicts basic statistics for all analyzed hedge fund strategies. Rates of return should be understood as the income generated on the invested capital in the analyzed period of time. The highest risk measured with standard deviation is generated by Short Bias strategy, then Emerging Markets and Equity Hedge. Short Bias strategy achieves the lowest average rate of return (0.01) at the same time, so according to the Sharpe ratio it is the least efficient strategy applied by hedge fund managers. The highest average rate of return in the examined period was done by Equity Hedge, however simultaneously the strategy generated one of the highest risk levels of all strategies. The least risky strategy is the Equity Market Neutral and simultaneously its average rate of return is not the lowest of all presented beneath. This suggests that it is quite efficient.

Rates of return fluctuations for hedge fund investment strategies can also be observed in Fig. 1.

Standard deviation is often used as a risk measure in the hedge fund industry. It is depicted in the following way:

$$\text{Standard Deviation} = \frac{\sum_{i=1}^{N}(r_i - r_i^d)^2}{N-1}$$

Table 2 Basic statistics on hedge funds rates of return

Strategy	Valid *N*	Mean	Minimum	Maximum	Standard deviation
Merger arbitrage	300	0.66	−6.46	3.12	1.14
Equity market neutral	300	0.54	−2.87	3.59	0.91
Short bias	300	0.01	−21.21	22.84	5.21
Emerging markets	300	0.99	−21.02	14.80	3.99
Equity hedge	300	1.00	−9.46	10.88	2.59
Event driven	300	0.90	−8.90	5.13	1.92
Macro	300	0.92	−6.40	7.88	2.12
Relative value	300	0.79	−8.03	5.72	1.23
Fixed income convertible arbitrage	300	0.68	−16.01	9.74	1.85
Multistrategy	300	0.67	−8.40	5.34	1.23

Source: Author's calculations

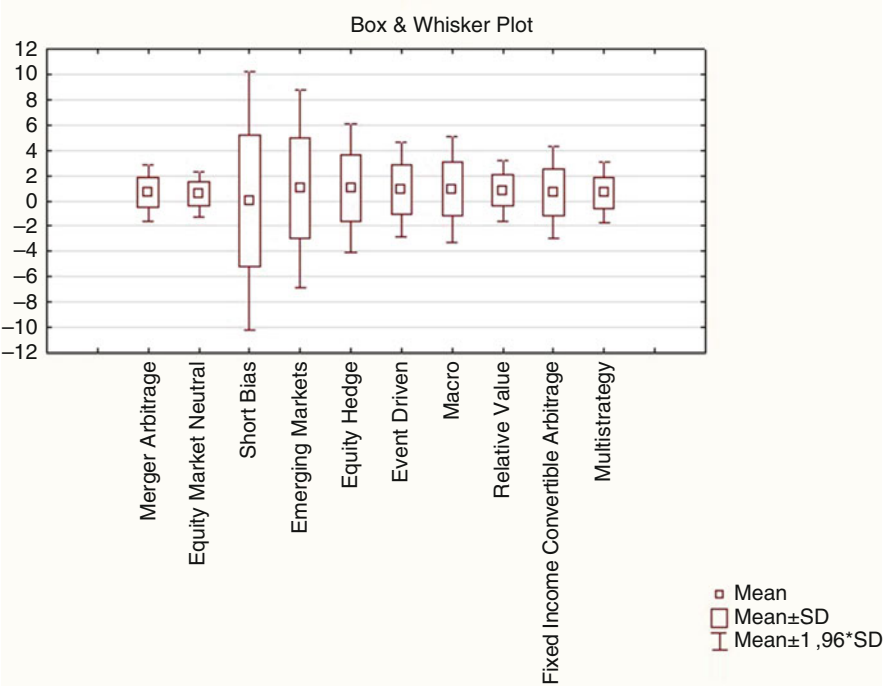


Fig. 1 Box and whisker plot for hedge fund strategies. Source: Author's study

where N is the number of observations of rates of return and r_i^d is the average value of the rate of return on the portfolio of hedge fund assets.

It is an important part of the Sharpe ratio, the most popular ratio to be used by hedge funds for their historical results [5]. It was created by W. Sharpe to be

used for measuring investment funds effectiveness. It is presented in the following way [30]:

$$\text{Sharpe Ratio} = \frac{r_i^d - r_f}{\sigma(r_i)}$$

where Sharpe Ratio is the investment result on the portfolio of hedge fund assets, $\sigma(r_i)$ is the standard deviation on rates of return on the portfolio of hedge fund assets, and r_f is the risk-free interest rate.

As far as the average rate of return (mean) is concerned, it is the highest for Equity Hedge and Emerging Markets strategies. Such values greatly show that strategies with higher risk levels let generate higher average rates of return. This in turn suggests that hedge fund managers should be paid not only for rates of return but also that their wages should depend on the risk which was necessary to be undertaken to achieve them. However, this poses the question, how risk should be measured. There are no doubts how rates of return can be interpreted, but there is not any risk measure which is widely acknowledged both by scientists and hedge fund practitioners. Standard deviation is the most popular one and quite easy to be calculated, but hedge funds generate risk similar to insurance companies. They may have good results for a very long time and suddenly because of some reason there appears one huge loss which decreases the capital dramatically. In the theory of finance this type of risk is called extreme risk. It has two features [19]:

- Low probability of appearance
- Huge loss if it appears anyway.

Stulz [32] reasonably compares the risk generated by hedge funds to the risk of a company which sells an earthquake insurance.

It is worth emphasizing that rates of return for all hedge fund strategies are correlated. Some of the Pearson correlation coefficients are low, some average, some high, however all of them are statistically significant at $p < 0.05$ (see Table 3). It confirms that hedge funds create high systemic risk.

5 Maximum Drawdown Measures and Their Definitions

Some scientists try to introduce alternative risk measures, however there are so many of them that nobody has proved if any is better than another one. In this paper maximum drawdown measures are being analyzed. To be exact, the author checks if their results differ from the results given by the traditional risk measure that is the Sharpe ratio. The typical maximum drawdown measures are the following [35–37]: Calmar ratio, Sterling and Burke ratio. The Calmar ratio is reflected by the following formula [9, 34]:

$$\text{CR} = \frac{r_i^d - r_f}{-\text{MD}_i}$$

Table 3 Correlation coefficients for different hedge fund strategies

Strategy	Merger arbitrage	Equity market neutral	Short bias	Emerging markets	Equity hedge	Event driven	Macro	Relative value	Fixed income convertible arbitrage	Multi-strategy
Merger arbitrage	1	0.35	−0.39	0.52	0.59	0.75	0.35	0.55	0.48	0.48
Equity market neutral	0.35	1	−0.14	0.27	0.50	0.41	0.33	0.41	0.32	0.38
Short bias	−0.39	−0.14	1	−0.57	−0.73	−0.62	−0.33	−0.39	−0.32	−0.44
Emerging markets	0.52	0.27	−0.57	1	0.74	0.76	0.56	0.62	0.53	0.67
Equity hedge	0.59	0.50	−0.73	0.74	1	0.84	0.56	0.68	0.60	0.67
Event driven	0.75	0.41	−0.62	0.76	0.84	1	0.51	0.76	0.66	0.76
Macro	0.35	0.33	−0.33	0.56	0.56	0.51	1	0.34	0.25	0.43
Relative value	0.55	0.41	−0.39	0.62	0.68	0.76	0.34	1	0.80	0.80
Fixed income convertible arbitrage	0.48	0.32	−0.32	0.53	0.60	0.66	0.25	0.80	1	0.78
Multistrategy	0.48	0.38	−0.44	0.67	0.67	0.76	0.43	0.80	0.78	1

Source: Author's calculations

where r_f is the risk-free interest rate, r_i^d is the average value of the rate of return on the portfolio of hedge fund assets, and MD_i is the lowest rate of return on hedge fund assets in the assumed period.

Thus Calmar ratio considers the worst scenario from the given period, which can be both an advantage and a disadvantage. It is safe but at the same time it is sensitive to some random values of low rates of return generated in the past and hardly probable to be repeated in the future. Similarly to the Sharpe ratio, the aim of the manager is to maximize it. Therefore, the optimum efficiency is when:

$$\mathrm{CR} \longrightarrow \max$$

Theoretically speaking, there is no maximum value for Calmar ratio. The same applies to Sterling or Burke ratio. All of them are ratios that measure the efficiency only compared with their other values for other investments, not on their own. One of the possibilities of making Calmar ratio less sensitive to some random loss is using the Starling ratio which considers the average value of a few maximum negative rates of return. The Sterling ratio can be depicted as [9, 21]:

$$\mathrm{SR} = \frac{r_i^d - r_f}{\frac{1}{N}\sum_{j=1}^{N}(-\mathrm{MD}_{ij})}$$

where N is the number of maximum negative rates of return.

A manager or an investor who wants to check the effectiveness of some hedge fund can choose an exact number of maximum negative rates taken into consideration depending on its preferences.

The Sterling ratio is also the higher the better. The optimal investment effectiveness is when:

$$\mathrm{SR} \longrightarrow \max$$

The third measure mentioned above is the Burke ratio which relates the excess rate of return to the square root of the powered sum of maximum negative rates of return generated in the researched period.

The mathematical formula for the Burke ratio can be presented as [3, 9]:

$$\mathrm{BR} = \frac{r_i^d - r_f}{\sqrt[2]{\sum_{j=1}^{N}\mathrm{MD}_{ij}^2}}$$

Similarly to the above presented ratios, the optimal investment effectiveness is assured when:

$$\mathrm{BR} \longrightarrow \max$$

6 Research Results of the Hedge Fund Risk Return Profile

Values of various risk return measures presented in Table 4 were applied to prepare hedge fund rankings depicted in Table 5. The risk-free interest rate used in the research is the interest rate of American 10Y treasury bonds at the end of the research period, that is from December 2014 (2.16%). If one compares the Sharpe ratio with alternative ratios, some differences can be seen. Only Short Bias strategy is always on the last position, no matter what measure is used. The biggest difference in the efficiency depending on if the traditional or alternative measure is used is seen for the Macro strategy. It is the sixth one in the case of the Sharpe ratio whereas it is the first one for the 5-period Sterling, 10-period Sterling, 5-period Burke, and the 10-period Burke ratio and the second one for the Calmar ratio. For example Merger Arbitrage is the second strategy for the Sharpe ratio whereas it is the sixth one for the Calmar, 5-period Sterling, and 5-period Burke ratio. For the 10-period Sterling it takes the fourth position and for the 10-period Burke the fifth one. As far as the Emerging Markets strategy is concerned, there is not a big difference between rankings made with the Sharpe ratio and alternative measures (1 up to 2 positions). Rankings for other strategies differ from each other by 1–4 positions (see Table 5) which is quite a lot. Generally, differences are rather important, so there raises the question of which ranking reflects the efficiency in the best way. Spearman's rank correlation coefficients (Table 6) between Sharpe ratio and alternative effectiveness measures are not significant for the majority of cases. It is only the 10-period Sterling ratio which is highly and significantly (at $p < 0.05$) correlated with the Sharpe ratio. At the same time correlation coefficients among alternative measures are high.

Data depicted in Appendix, Figs. 2 and 3, show that the majority of hedge fund strategies have negative skewness (apart from Short Bias and Macro) and high kurtosis. Thus, if these two central moments of the distribution are not considered

Table 4 Values of various risk return measures

Strategy	Sharpe	Calmar	Sterling5	Sterling10	Burke5	Burke10
Merger arbitrage	0.42	0.07	0.11	0.15	0.04	0.04
Equity market neutral	0.40	0.12	0.15	0.19	0.06	0.06
Short bias	−0.03	−0.01	−0.01	−0.01	−0.004	−0.004
Emerging markets	0.20	0.04	0.11	0.08	0.02	0.02
Equity hedge	0.32	0.09	0.11	0.14	0.05	0.04
Event driven	0.38	0.08	0.11	0.14	0.05	0.04
Macro	0.35	0.12	0.17	0.21	0.07	0.06
Relative value	0.50	0.08	0.12	0.19	0.05	0.05
Fixed income convertible arbitrage	0.27	0.03	0.06	0.10	0.02	0.02
Multistrategy	0.40	0.06	0.10	0.14	0.04	0.04

Source: Author's calculations

Table 5 Rankings of hedge fund strategies

Rank position	Sharpe	Calmar	Sterling5	Sterling10	Burke5	Burke10
1	Relative value	Equity market	Neutral	Macro	Macro	Macro
2	Merger arbitrage	Macro	Equity market neutral	Relative value	Equity market neutral	Equity market neutral
3	Multistrategy	Equity hedge	Relative value	Equity market neutral	Relative value	Relative value
4	Equity market neutral	Event driven	Equity hedge	Merger arbitrage	Equity hedge	Equity hedge
5	Event driven	Relative value	Event driven	Multistrategy	Event driven	Merger arbitrage
6	Macro	Merger arbitrage	Merger arbitrage	Equity hedge	Merger arbitrage	Event driven
7	Equity hedge	Multistrategy	Emerging markets	Event driven	Multistrategy	Multistrategy
8	Fixed income convertible arbitrage	Emerging markets	Multistrategy	Fixed income convertible arbitrage	Emerging markets	Fixed income convertible arbitrage
9	Emerging markets	Fixed income convertible arbitrage	Fixed income convertible arbitrage	Emerging markets	Fixed income convertible arbitrage	Emerging markets
10	Short bias	Short bias	Short bias	Short bias	Short bias	Short bias

Source: Author's study

Table 6 Spearman's rank correlation coefficients

Ratios	Sharpe	Calmar	Sterling5	Sterling10	Burke5	Burke10
Sharpe	1	0.44	0.47	0.76	0.54	0.59
Calmar	0.44	1	0.95	0.75	0.95	0.91
Sterling5	0.47	0.94	1	0.83	0.99	0.95
Sterling10	0.76	0.75	0.83	1	0.88	0.93
Burke5	0.54	0.95	0.99	0.88	1	0.98
Burke10	0.59	0.91	0.95	0.93	0.98	1

Underlined correlations are significant at $p < 0.05$
Source: Author's calculations

by the Sharpe ratio, the results achieved with it may not be appropriate. It will be the subject of author's further studies.

7 Conclusions

- The research shows that hedge fund strategies are characterized by negative skewness and high kurtosis in the examined period of 1990–2014.
- Because of that, the Sharpe ratio may not be a good measure of investment results, because it takes only the standard deviation into consideration while ignoring the third and the fourth central moment of the distribution.
- Rankings of hedge fund strategies made with the Sharpe ratio and with alternative measures are quite different, which shows that the Sharpe ratio may not be an adequate efficiency measure in the case of hedge fund strategies.
- Spearman's rank correlation coefficients between the Sharpe ratio and alternative efficiency measures are not statistically significant (with the exception of the correlation coefficient between the Sharpe ratio and the 10-day Sterling ratio).
- Results are surprising because the analysis done for the period of 1990–March 2011 or the one for January 2005–April 2011 have shown high correlations between the Sharpe ratio and alternative measures (e.g., [25]). It may be not only due to the different study period, but also to the substantial change of the risk-free interest rate.
- All in all, the need of further studies exists. The results suggest that alternative measures show different results in some circumstances than the Sharpe ratio, however it does not mean that they are more adequate.

Appendix

See Figs. 2 and 3.

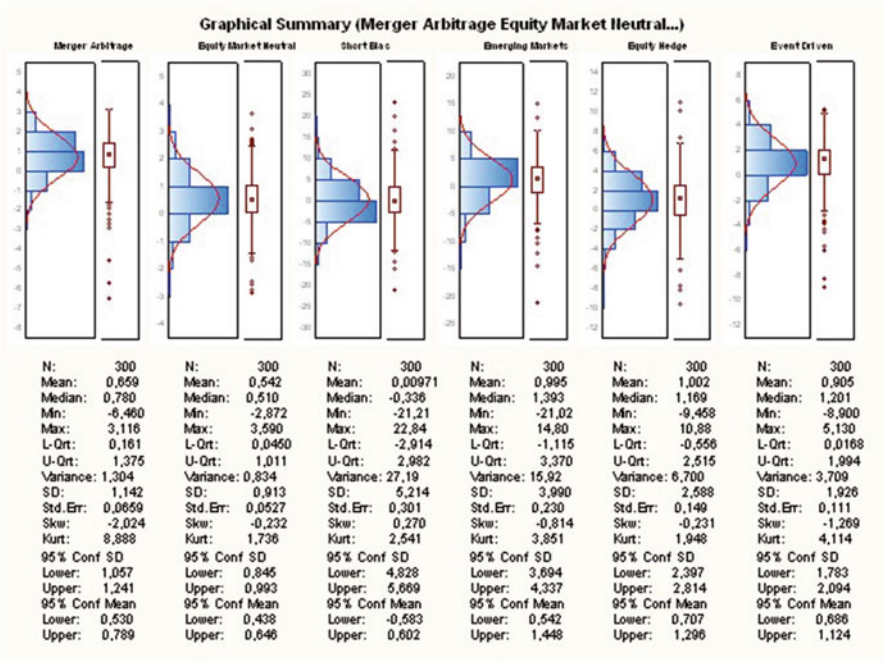


Fig. 2 Source: Author's study

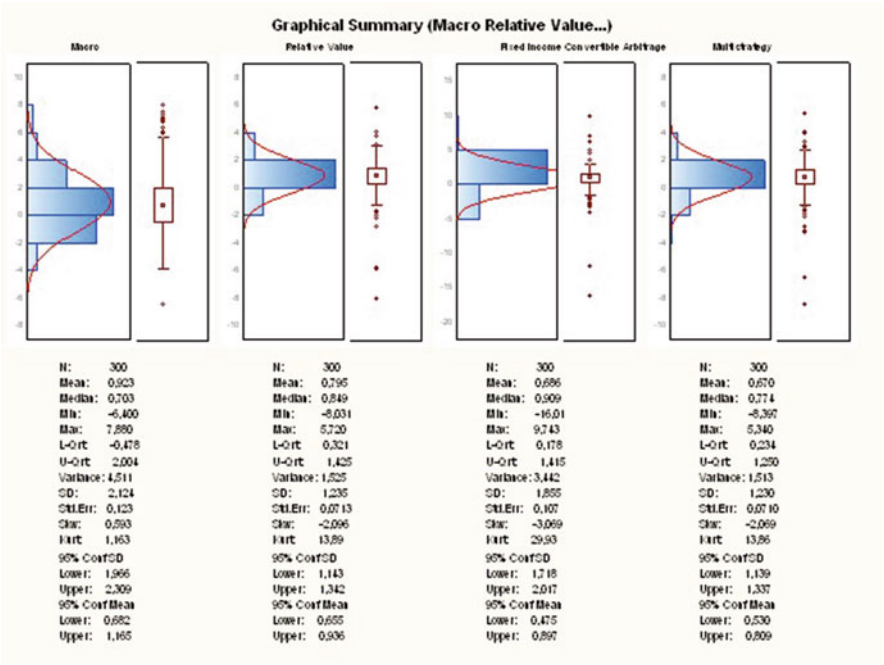


Fig. 3 Source: Author's study

References

1. Agarwal, V., Fos, V., Ijang, W.: Inferring reporting-related biases in hedge fund databases from hedge fund equity holdings. Research Collection BNP Paribas Hedge Fund Centre. Manag. Sci. **59**(6), 1271–1289 (2013). Available at: http://ink.library.smu.edu.sg/bnp_research/19
2. An Overview of Leverage, AIMA Canada Strategy Paper Series: Companion Document. no 4, p. 2 (2006)
3. Burke G.: A sharper Sharpe ratio. Futures **23**(3), 56 (1994)
4. Cao, Ch., Chen, Y., Liang, B., Lo, A.W.: Can hedge funds time market liquidity? J. Financ. Econ. **109**(2), 493–516 (2013)
5. Credit Suisse First Boston Group (January 2016). https://www.credit-suisse.com/us/en.html
6. Dowd, K.: Beyond value at risk. In: The New Science of Risk Management. Wiley, New York (1998)
7. Duffie, D., Wang, C., Wang, H.: Leverage Management. Stanford University, pp. 1–25 (2008)
8. Edwards F.R.: Hedge funds and the collapse of long–term capital management. J. Econ. Perspect. **13**(2), 189–210 (1999)
9. Eling, M., Schuhmacher, F.: Does the choice of performance measure influence the evaluation of hedge funds? J. Bank. Financ. **31**(9), 2632–2647 (2007)
10. Eling M.: Performance Measurement of Hedge Funds Using Data Envelopment Analysis. Working Papers on Risk Management and Insurance, No. 25. University of St. Gallen (2006)
11. Ernst and Young: The evolving dynamics of the hedge fund industry. Global Hedge Fund and Investor Survey, 1–40. (2015)
12. Financial Conduct Authority (FCA), Hedge Fund Survey, 14–22 (2015)
13. Gregoriou, G.N., Huebner, G., Papageorgiou, N., Rouah, F. (eds.): Hedge Funds. Insights in Performance Measurement, Risk Analysis, and Portfolio Management. Wiley, Hoboken (2005)
14. Gregoriou, G.N., Pascalau, R. (eds.): Derivatives Pricing, Hedge Funds and Term Structure Models. Palgrave, Macmillan, Hampshire (2011)
15. Guizot, A.: The Hedge Fund Compliance and Risk Management Guide. Wiley, Hoboken (2007)
16. Hespeler, F., Loiacono, G.: Monitoring systemic risk in the hedge fund sector. ESMA Working Paper No. 2, 1–39 (2015)
17. Hwang, S., Satchell, S.E.: The Asset Allocation Decision in a Loss Aversion World. Working Paper Series, WP01–14, Financial Econometrics Research Centre, 1–14 (2001)
18. Ijang, H., Kelly, B.: Tail Risk and Hedge Fund Returns. Fama–Miller Center for Research in Finance, Chicago Booth Paper No. 12–13, 1–43 (2012)
19. Jajuga, K. (ed.): Zarzadzanie ryzykiem, Wydawnictwo Naukowe PWN, Warszawa, p. 61 (2007)
20. Kaplan, P.D., Knowles, J.A.: Kappa: A Generalized Downside Risk–Adjusted Performance Measure. Morningstar Associates and York Hedge Fund Strategies, New York (2004)
21. Kestner: Getting a handle on true performance. Futures **25**(1), 44–46 (1996)
22. Markowitz, H.: Portfolio selection. J. Finance. **7**(1), 77–91 (1952)
23. McGuire, P., Remolona, E., Tsatsaronis, K.: Time-varying exposures and leverage in hedge funds. BIS Q. Rev. p. 68. (2005)
24. Pruchnicka-Grabias, I.: Hedge fund assets rates of return-theory and empirical tests. In: Barkovic, D., Runzheimer, B. (eds.) Interdisciplinary Management Research XI, Opatija. pp. 1188–1206 (2015)
25. Pruchnicka-Grabias I.: Lower partial moments and maximum drawdown measures in hedge fund risk-return profile analysis. Univ. J. Math. Math. Sci. **9**(1–2), 43–59. Pushpa Publishing House, Allahabad (2016)
26. Ross, S.: The arbitrage theory of capital asset pricing. J. Econ. Theory **13**(3), 341–360 (1976)
27. Sadka, R.: Liquidity risk and the cross-section of hedge-fund returns. J. Financ. Econ. **98**(1), 54–71 (2010)

28. Schneeweis T., Martin, G., Kazemi, H., Karavas, V.: The impact of leverage on hedge funds risk and return. J. Altern. Invest. **7**(4), 10–21 (2005)
29. Sharpe W.F.: Capital asset prices: a theory of market equilibrium under conditions of risk. J. Finance. **19**(3), 425–442 (1964)
30. Sharpe W.F.: The Sharpe ratio. J. Portf. Manag. **21**(1), 49–58 (1994)
31. Sortino, F.A., van der Meer, R., Plantiga, A.: The Dutch triangle. J. Portf. Manag. **26**(1), 50–59 (1999)
32. Stulz R.: Hedge funds: past, present and future. J. Econ. Perspect. **21**(2), 175–194 (2007)
33. Tran, V.Q.: Evaluating Hedge Fund Performance, pp. 54–64. Wiley, Hoboken (2016)
34. Young, T.W.: Calmar ratio: a smoother tool. Futures **20**(1), 40 (1991)
35. Pruchnicka-Grabias, I.: Zastosowanie miar maksymalnej straty na kapitale w badaniu efektywności funduszy hedgingowych. In: Kwartalnik Kolegium Społeczno – Ekonomicznego. Studia i prace, 3(23), pp. 133–245. Oficyna Ekonomiczna, Szkoła Główna Handlowa w Warszawie, Warszawa (2015)
36. Pruchnicka-Grabias, I.: Corporate Financial Risk Management. Szkoła Główna Handlowa w Warszawie, Warszawa (2015)
37. Pruchnicka-Grabias, I.: Maximum drawdown measures in hedge fund efficiency appraisal. Q. e-Finance **12**(4), 83–91 (2016). https://doi.org/10.1515/fiqf-2016-0010

Estimating the Extremal Coefficient: A Simulation Comparison of Methods

Marta Ferreira

Abstract Tail dependence is an important issue to evaluate risk. The multivariate extreme values theory is the most suitable to deal with the extremal dependence. The extremal coefficient measures the degree of dependence between the marginals of max-stable distributions, a natural class of models in this framework. The estimation of the extremal coefficient is addressed and a new estimator is compared through simulation with existing methods. An illustration with real data is presented.

1 Introduction

The Extreme Value Theory (EVT) is viewed with particular interest in several areas such as finance, insurance, engineering, environment, among others, where it is important to evaluate the risk of occurring extreme events.

The distinguishing feature of EVT is that it provides us a framework to compute the probability of events more extreme than any that have already been observed. See, for instance, [1] and [8].

Let $\{X_i\}_{i\geq 1}$ be an i.i.d. sequence of random variables (r.v.'s) with common distribution function (d.f.) F. If there exist real constants $\{b_n\}_{n\geq 1}$ and positive $\{a_n\}_{n\geq 1}$ such that the limit

$$\lim_{n\to\infty} P(\max(X_1,\ldots,X_n) \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x)\,. \qquad (1)$$

M. Ferreira (✉)
CMAT - University of Minho, Braga, Portugal

CEMAT, CEAUL - University of Lisbon, Lisbon, Portugal
e-mail: msferreira@math.uminho.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_4

exists for all continuity points of nondegenerate $G(x)$, then $G(x)$ is Generalized Extreme Value distribution (GEV), and has standard representation:

$$G(x) = \exp\left\{-(1+\xi x)^{-1/\xi}\right\}, \ 1+\xi x > 0.$$

Under the limit (1) we say that F belongs to the max-domain of attraction of $G(x)$, which resumes all the possible limit distributions: (reversed) Weibull ($\xi < 0$ - light tail and finite right-end-point), Gumbel ($\xi = 0$ - exponential tail) and Fréchet ($\xi > 0$ - heavy tail and infinite right-end-point). Observe that the result can be easily formulated for the minimum based on $\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n)$.

In a multivariate context we have to deal with dependence between marginals. The current Pearson's correlation may not describe well the dependence structure on the tails (see, e.g., [4] for examples). The multivariate EVT offers itself as a natural instrument to develop tail measures. The univariate result above can be expanded to a d-variate setting, where the maximum of a sequence of random vectors is the vector of componentwise maxima. Let $\{(X_1^{(n)}, \ldots, X_d^{(n)})\}_{n\geq 1}$ be i.i.d. copies of $(X_1, \ldots, X_d)$, with common d.f. F. If there exist real constants $\{b_j^{(n)}\}_{n\geq 1}$ and positive $\{a_j^{(n)}\}_{n\geq 1}$, $j = 1, \ldots, d$, and a d.f. G with non-degenerate margins, such that,

$$\begin{aligned} &P\left(\max_{1\leq i\leq n} X_1^{(i)} \leq a_1^{(n)} x_1 + b_1^{(n)}, \ldots, \max_{1\leq i\leq n} X_d^{(i)} \leq a_d^{(n)} x_d + b_d^{(n)}\right) \\ &= F^n(a_1^{(n)} x_1 + b_1^{(n)}, \ldots, a_d^{(n)} x_d + b_d^{(n)}) \underset{n\to\infty}{\longrightarrow} G(x_1, \ldots, x_d), \end{aligned} \tag{2}$$

exists for all continuity points of $G(x_1, \ldots, x_d)$, then it must be a multivariate extreme value (MEV) distribution (in the two-dimensional case it is denoted BEV), given by

$$G(x_1, \ldots, x_d) = \exp[-\ell\{-\log G_1(x_1), \ldots, -\log G_d(x_d)\}],$$

where $\ell : \mathbb{R}_+^d \to \mathbb{R}_+$ is the stable tail dependence function. This function characterizes the dependence of a multivariate extreme value distribution, which is no longer parametrically defined as in the univariate case. The function ℓ must be convex ($\ell(\upsilon\mathbf{v}+(1-\upsilon)\mathbf{w}) \leq \upsilon\ell(\mathbf{v})+(1-\upsilon)\ell(\mathbf{w})$, $\upsilon \in [0,1]$), homogeneous of order 1 ($\ell(s\cdot) = s\ell(\cdot)$, $0 < s < \infty$) and bounded by $\max(x_1, \ldots, x_d) \leq \ell(x_1, \ldots, x_d) \leq x_1 + \ldots + x_d$, $\forall (x_1, \ldots, x_d) \in [0, \infty)^d$ (upper bound corresponds to independence; lower bound means complete dependence). If the result (2) holds, we also say that F belongs to the max-domain of attraction of G. This function is max-stable, in the sense of $G^n(a_1^{(n)} x_1 + b_1^{(n)}, \ldots, a_d^{(n)} x_d + b_d^{(n)}) = G(x_1, \ldots, x_d)$, also valid for the univariate case. Indeed, the marginals G_j, $j = 1, \ldots, d$, of G are GEV functions.

We can formulate (2) based on a copulas' approach which allows us to look only on the dependence structure. A copula C is a d.f. whose margins are

uniformly distributed on [0, 1], namely, if C_F is the copula of $(X_1, \ldots, X_d)$, then $C_F(u_1, \ldots, u_d) = F\left(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\right)$, assuming that F_j^{-1}, $j = 1, \ldots, d$, are continuous [13]. Thus, formulation (2) implies for the respective copulas, C_F and C_G:

$$C_F^n(u_1^{1/n}, \ldots, u_d^{1/n}) \underset{n\to\infty}{\longrightarrow} C_G(u_1, \ldots, u_d),$$

where

$$C_G(u_1, \ldots, u_d) = \exp\{-\ell(-\log u_1, \ldots, -\log u_d)\} \tag{3}$$

is called a multivariate extreme value (MEV) copula (or BEV copula in case $d = 2$). However, the reciprocal is not true since each marginal must also belong to some max-domain of attraction.

Observe that extreme value models/copulas are fully identified by the tail dependence function. Formulation in (3) is not unique and other tail dependence functions can be used, for instance, the Pickands dependence function A, such that, for $(u_1, \ldots, u_d) \in (0, 1]^d \setminus \{(1, \ldots, 1)\}$,

$$C_G(u_1, \ldots, u_d) = \exp\left(\left(\sum_{j=1}^{d} \log u_j\right) A\left(\frac{\log u_1}{\sum_{j=1}^{d} \log u_j}, \ldots, \frac{\log u_{d-1}}{\sum_{j=1}^{d} \log u_j}\right)\right).$$

Function $A : \mathscr{S}_{d-1} \to [1/d, 1]$ is a restriction of ℓ to the unit simplex $\mathscr{S}_{d-1} = \{(w_1, \ldots, w_d) \in [0, 1]^d : \sum_{j=1}^{d} w_j = 1\}$. It also satisfies convexity and is such that $A(\mathbf{e}_j) = 1$ for $j = 1, \ldots, d$, and $\max(w_1, \ldots, w_d) \le A(\mathbf{w}) \le 1$, $\forall \mathbf{w} \in \mathscr{S}_{d-1}$. The Pickands dependence function relates with ℓ through

$$A(w_1, \ldots, w_{d-1}) \equiv A\left(\frac{x_1}{\sum_{j=1}^{d} x_j}, \ldots, \frac{x_{d-1}}{\sum_{j=1}^{d} x_j}\right) = \frac{\ell(x_1, \ldots, x_d)}{\sum_{j=1}^{d} x_j}.$$

Other formulations of the dependence function within MEV models were also considered in literature. A complete survey on this topic can be found in [2].

Estimation of the Pickands dependence function and the stable tail dependence function has been largely addressed in literature. Non-parametric methods can be seen in [5, 9] and the references therein. Reference [2] includes parametric methods.

The extremal coefficient [14, 15] measures the degree of dependence between the marginals of a MEV G, being given by

$$C_G(u, \ldots, u) = u^{\varepsilon}.$$

Observe that $1 \le \varepsilon \le d$, with the bounds $\varepsilon = 1$ and $\varepsilon = d$, respectively meaning complete dependence and independence between the marginals. If we apply logarithms to both members, by the homogeneity property of ℓ, we arrive at

$$\varepsilon = \ell(1, \ldots, 1),$$

and thus

$$\varepsilon = d\,A(1/d, \ldots, 1/d).$$

The extremal coefficient is a possible measure to evaluate tail dependence. It relates with other measures, namely, the bivariate tail dependence coefficient (TDC), usually denoted λ, often used in financial contexts to assess markets risk (see, e.g., [12] and the references therein). The TDC measures the probability of occurring extreme values for one random variable (r.v.) given that another assumes an extreme value too:

$$\lambda = \lim_{u \uparrow 1} P(F_1(X_1) > u | F_2(X_2) > u) = \lim_{u \uparrow 1} P(\cap_{i \in \{1,2\}} \{F_i(X_i) > u\})/(1-u).$$

We have that $\lambda = 2 - l(1,1) = 2(1 - A(0.5)) = 2 - \varepsilon$.

Here we address the estimation of the extremal coefficient and compare a new estimator with existing ones, based on a simulation study. We apply the procedures to real stock market indexes, in order to evaluate the contagion risk of large losses and gains.

2 Examples and Estimators

In the sequel we list some MEV models and respective coefficients (see, e.g., [2] and [7]):

- *Logistic model*

$$\ell(x, y) = (x^{1/\alpha} + y^{1/\alpha})^{\alpha},$$

  where $0 < \alpha \leq 1$; $\varepsilon = 2^{\alpha}$ and we have independence if $\alpha = 1$ and complete dependence if $\alpha \to 0$;
- *Asymmetric Logistic model*

$$\ell(x, y) = (1 - \psi_1)x + (1 - \psi_2)y + ((\psi_1 x)^{1/\alpha} + (\psi_2 y)^{1/\alpha})^{\alpha},$$

  where $0 < \alpha \leq 1$, $0 \leq \psi_1, \psi_2 \leq 1$; $\varepsilon = 2 - \psi_1 - \psi_2 + ((\psi_1)^{1/\alpha} + (\psi_2)^{1/\alpha})^{\alpha}$ and we have independence if either $\alpha = 1$, $\psi_1 = 0$ or $\psi_2 = 0$ and complete dependence if $\alpha \to 0$ and $\psi_1 = \psi_2 = 1$; (we obtain the logistic model whenever $\psi_1 = \psi_2 = 1$);
- *Hüsler-Reiss model*

$$\ell(x, y) = x\Phi(\beta^{-1} + 0.5\beta \log(x/y)) + y\Phi(\beta^{-1} + 0.5\beta \log(y/x)),$$

where $\beta > 0$; $\varepsilon = 2\Phi(\beta^{-1})$ and we have independence if $\beta \to 0$ and complete dependence if $\beta \to \infty$;

- *Negative Logistic model*

$$\ell(x, y) = x + y - (x^{-\beta} + y^{-\beta})^{-1/\beta},$$

where $\beta > 0$; $\varepsilon = 2 - 2^{-1/\beta}$ and we have independence if $\beta \to 0$ and complete dependence if $\beta \to \infty$;

- *Asymmetric Negative Logistic model*

$$\ell(x, y) = x + y - ((\psi_1 x)^{-\beta} + (\psi_2 y)^{-\beta})^{-1/\beta},$$

where $\beta > 0, 0 < \psi_1, \psi_2 \leq 1$; $\varepsilon = 2 - ((\psi_1)^{-\beta} + (\psi_2)^{-\beta})^{-1/\beta}$ and we have independence if either $\beta \to 0$, $\psi_1 \to 0$ or $\psi_2 \to 0$ and complete dependence if $\psi_1 = \psi_2 = 1$ and $\beta \to \infty$; (we obtain the negative logistic model whenever $\psi_1 = \psi_2 = 1$);

- *Bilogistic model*

$$\ell(x, y) = x r^{1-\gamma} + y(1 - r)^{1-\delta},$$

where r is the root of $(1 - \gamma)x(1 - r)^{\delta} - (1 - \delta)y r^{\gamma} = 0$, $0 < \gamma, \delta < 1$; $\varepsilon = r^{1-\gamma} + (1 - r)^{1-\delta}$ and we have independence if $\gamma = \delta \to 1$ and complete dependence if $\gamma = \delta \to 0$;

- *Negative bilogistic model*

$$\ell(x, y) = x + y - x r^{1+\gamma} - y(1 - r)^{1+\delta},$$

where r is the root of $(1 + \gamma)x r^{\gamma} - (1 + \delta)y(1 - r)^{\delta} = 0$, $\gamma, \delta > 0$; $\varepsilon = 2 - r^{1+\gamma} - (1 - r)^{1+\delta}$ and we have independence if $\gamma = \delta \to \infty$ and complete dependence if $\gamma = \delta \to 0$;

- *Dirichelet model*

$$\ell(x, y) = x(1 - B(q; \alpha + 1, \beta)) + y B(q; \alpha, \beta + 1),$$

where $q = \alpha y/(\alpha y + \beta x)$ and $B(q; \alpha, \beta)$ is the beta d.f. evaluated at q and having shape parameters α and β, $\alpha, \beta > 0$; $\varepsilon = 1 - B(q; \alpha + 1, \beta) + B(q; \alpha, \beta + 1)$ and we have independence if $\alpha = \beta \to 0$ and complete dependence if $\alpha = \beta \to \infty$;

- *Asymmetric mixed model*

$$\ell(1 - t, t) = 1 - (\alpha + \beta)t + \alpha t^2 + \beta t^3,$$

where $\alpha, \alpha + 3\beta \geq 0$ and $\alpha + \beta, \alpha + 2\beta \leq 1$; $\varepsilon = 2 - \alpha/2 + 3\beta/4$ and we have independence if $\alpha = \beta = 0$ and complete dependence cannot be achieved (the dependence increases for increasing α and fixed β).

- *d-dimensional logistic model and asymmetric logistic, respectively,*

$$\ell(x_1,\ldots,x_d) = \left(\sum_{j=1}^{d} x_j^{1/\alpha}\right)^{\alpha}$$

$$\ell(x_1,\ldots,x_d) = \sum_{J\in\mathscr{D}} \left(\sum_{j\in J} (\psi_{J,j}x_j)^{1/\alpha_J}\right)^{\alpha_J},$$

where $\mathscr{D}$ is the set of non-empty subsets of $D = \{1,\ldots,d\}$, $0 < \alpha, \alpha_J \leq 1$ are the dependence parameters and $\psi_{J,j} \geq 0$ are the asymmetry parameters such that $\sum_{J\ni j} \psi_{J,j} = 1$, for $j \in D$ and $J \in \mathscr{D}$. We have for the non and asymmetric cases, respectively, $\varepsilon = d^{\alpha}$ and $\varepsilon = \sum_{J\in\mathscr{D}} \left(\sum_{j\in J} (\psi_{J,j})^{1/\alpha_J}\right)^{\alpha_J}$, with a similar discussion for dependence/independence as in the bivariate form.

Non-parametric estimators avoid the parametric specification of the tail dependence function. The most known estimators in literature are the Pickands [11] and Capéraà-Fougéres-Genest [3], respectively given by

$$\widehat{\varepsilon}^{\,P} = d\left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\eta}_i\right)^{-1},$$

and

$$\widehat{\varepsilon}^{\,CFG} = d\exp\left\{-\gamma - \frac{1}{n}\sum_{i=1}^{n} \log(\widehat{\eta}_i)\right\},$$

with $\gamma = 0.5772\ldots$ representing the Euler-Mascheroni constant,

$$\widehat{\eta}_i = d \bigwedge_{j=1}^{d} -\log \widehat{U}_{i,j}$$

and where $\widehat{U}_{i,j} = F_{n,j}(X_{i,j})$, with

$$F_{n,j}(x) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}(X_{i,j} \leq x),\ j = 1,\ldots,d,$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. The empirical d.f. estimates the unknown marginals where denominator $n+1$ instead of n relates accuracy (see [2]).

Corrected versions of these estimators correspond to, respectively,

$$\widehat{\varepsilon}_c^{\,P} = d\left(\frac{1}{\widehat{\varepsilon}^{\,P}} - \frac{1}{d}\sum_{j=1}^{d}\left(\frac{1}{n}\sum_{i=1}^{n}\left(-\log\widehat{U}_{i,j}\right) - 1\right)\right)^{-1}$$

and

$$\widehat{\varepsilon}_c^{\,CFG} = d\widehat{\varepsilon}^{\,CFG}\exp\left\{-\frac{1}{d}\sum_{j=1}^{d}\left(-\gamma - \frac{1}{n}\sum_{i=1}^{n}\log\left(-\log\widehat{U}_{i,j}\right)\right)\right\}.$$

These corrections concern tail dependence function endpoint constraints, $A(0) = A(1) = 1$. For large sample sizes, the end-point corrections are negligible. See [2] and [9].

The Hall and Tajdivi estimator [10] consists in another correction of the Pickands formula and is defined by

$$\widehat{\varepsilon}^{\,HT} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\log(n+1)/i\right)}{\widehat{\varepsilon}^{\,P}}.$$

More details can also be found in [9].

More recently, the work in [6] allows the derivation of a new estimator for ε, given by

$$\widehat{\varepsilon}^{\,FF} = d\left(\frac{1}{1 - n^{-1}\sum_{i=1}^{n}\bigvee_{j=1}^{d}\widehat{U}_{i,j}^{d}} - 1\right).$$

In the next section, we analyze the performance of this new method. More precisely, we conduct a simulation study and compare $\widehat{\varepsilon}^{\,FF}$ with estimators $\widehat{\varepsilon}_c^{\,CFG}$, $\widehat{\varepsilon}_c^{\,P}$ and $\widehat{\varepsilon}^{\,HT}$, based on the MEV models listed above.

3 Simulation Study

The simulation study was based on the generation of 1000 random samples of size $n = 100, 1000$ of each of the models listed in Sect. 2. In the multivariate models, we considered dimension $d = 3$. The root mean squared error (rmse) and absolute bias were obtained and are reported in Tables 1, 2, 3 (the numbers in brackets correspond to the rmse). The three-dimensional Logistic and Asymmetric Logistic were denoted, respectively, 3-Log and 3-Alog. The values of the parameters

were chosen in order to cover the cases of complete dependence ($\varepsilon \approx 1$), tail independence ($\varepsilon \approx d$) and $\varepsilon \approx (d+1)/2$, whether $d = 2$ or $d = 3$. In the unit bound case in Table 3, the asymmetric models coincide with the respective symmetric versions and thus are omitted. The asymmetric mixed model was also excluded in this case since it cannot reach complete dependence (the largest value is 1.5 already reported in Table 1). Under the unit bound case, the estimators also present a global better performance. Among the existing methods, the CFG has an overall upper performance. We can see that the new estimator is competitive with the CFG. Particularly, it presents the smallest bias in most cases.

4 Application to Real Data

The data correspond to the daily closing prices of the stock indices, Germany DAX, Switzerland SMI, France CAC and UK FTSE, collected in the period 1991–1998 (weekends and holidays are omitted). The negative/positive log-returns are considered so as to evaluate the contagion risk that large losses/gains occurring in one market may cause in another. More precisely, in order to obtain a sample which could be modeled by a MEV law, it is considered the monthly maximum. The scatterplots in Figs. 1 and 2 exhibit some dependence, specially in the negative log-returns. This can also be corroborated with the estimates in Tables 4 and 5. The largest contagion risk for gains (positive log-returns) appears between SMI and CAC, whilst, for losses (negative log-returns), we found it between DAX and SMI. When considering the groups of three, the largest influence occurs in the triple DAX-SMI-CAC concerning gains and, for losses, in the triple DAX-SMI-FTSE.

Appendix

See Figs. 1 and 2 and Tables 1, 2, 3, 4, and 5.

Table 1 Absolute bias and root mean squared error (in brackets) of estimators $\widehat{\varepsilon}_c^{CFG}$, $\widehat{\varepsilon}_c^{P}$, $\widehat{\varepsilon}^{HT}$, and $\widehat{\varepsilon}^{FF}$

	CFG	P	HT	FF
$n = 1000$				
Log	0.00217 (0.01553)	0.00038 (0.01966)	0.00102 (0.01971)	0.00276 (0.01688)
Alog	0.00260 (0.02569)	0.00208 (0.02751)	0.00277 (0.02764)	0.00266 (0.02546)
HR	0.00659 (0.02462)	0.00533 (0.03243)	0.00550 (0.03256)	0.00053 (0.02691)
Neglog	0.00627 (0.02263)	0.00468 (0.02787)	0.00543 (0.02807)	0.00129 (0.02238)
Aneglog	0.00265 (0.02319)	0.00052 (0.03424)	0.00085 (0.03435)	0.00300 (0.02637)
Bilog	0.00164 (0.01576)	0.00145 (0.01896)	0.00288 (0.01915)	0.00002 (0.01586)
Negbilog	0.00205 (0.01041)	0.00136 (0.01296)	0.00294 (0.01323)	0.00081 (0.01065)
Dir	0.00455 (0.02425)	0.00439 (0.02857)	0.00532 (0.02879)	0.00075 (0.02406)
Amix	0.00724 (0.02093)	0.00714 (0.02377)	0.00833 (0.02420)	0.00456 (0.02163)
3-Log	0.00562 (0.03052)	0.00438 (0.02888)	0.00686 (0.02937)	0.00164 (0.02776)
3-Alog	0.00909 (0.04706)	0.00512 (0.04803)	0.00674 (0.04833)	0.00092 (0.04412)
$n = 100$				
Log	0.01240 (0.04573)	0.00778 (0.05174)	0.01720 (0.05441)	0.00354 (0.04824)
Alog	0.02865 (0.08808)	0.01443 (0.09597)	0.01918 (0.09842)	0.00384 (0.09157)
HR	0.04421 (0.08385)	0.02900 (0.10808)	0.03051 (0.11063)	0.00436 (0.08688)
Neglog	0.02656 (0.08330)	0.02380 (0.09701)	0.02906 (0.09990)	0.00295 (0.08359)
Aneglog	0.05350 (0.10369)	0.03904 (0.12079)	0.04189 (0.12393)	0.01619 (0.10488)
Bilog	0.00778 (0.04537)	0.00131 (0.05773)	0.01081 (0.05924)	0.00710 (0.04876)
Negbilog	0.00234 (0.03473)	0.00388 (0.03962)	0.00659 (0.04021)	0.00688 (0.03577)
Dir	0.02390 (0.07276)	0.01013 (0.08243)	0.01637 (0.08467)	0.00200 (0.07490)
Amix	0.01564 (0.05846)	0.00303 (0.06632)	0.01092 (0.06793)	0.00271 (0.06259)
3-Log	0.02483 (0.09143)	0.00688 (0.08689)	0.02336 (0.08999)	0.00068 (0.08671)
3-Alog	0.05250 (0.14449)	0.01240 (0.14227)	0.02324 (0.14551)	0.00815 (0.14088)

Table 2 Absolute bias and root mean squared error (in brackets) of estimators $\widehat{\varepsilon}_c^{\,CFG}$, $\widehat{\varepsilon}_c^{\,P}$, $\widehat{\varepsilon}^{\,HT}$, and $\widehat{\varepsilon}^{\,FF}$, for tail independence

	CFG	P	HT	FF
$n = 1000$				
Log	0.00386 (0.02591)	0.00023 (0.04239)	0.00023 (0.04253)	0.00325 (0.03326)
Alog	0.00747 (0.02386)	0.00768 (0.03595)	0.00771 (0.03607)	0.00174 (0.02878)
HR	0.00845 (0.02525)	0.00697 (0.03702)	0.00699 (0.03714)	0.00120 (0.02938)
Neglog	0.00867 (0.02597)	0.00910 (0.03678)	0.00913 (0.03691)	0.00153 (0.02970)
Aneglog	0.00952 (0.02775)	0.00917 (0.03617)	0.00920 (0.03629)	0.00357 (0.03202)
Bilog	0.00541 (0.02460)	0.00116 (0.03519)	0.00121 (0.03531)	0.00232 (0.02909)
Negbilog	0.00948 (0.02803)	0.00571 (0.03975)	0.00573 (0.03988)	0.00211 (0.03272)
Dir	0.00953 (0.02588)	0.00713 (0.03755)	0.00719 (0.03769)	0.00278 (0.02886)
Amix	0.01102 (0.02685)	0.01545 (0.03920)	0.01550 (0.03933)	0.00677 (0.03077)
3-Log	0.01268 (0.03734)	0.00390 (0.05210)	0.00392 (0.05228)	0.00949 (0.04686)
3-Alog	0.02744 (0.05073)	0.01868 (0.05886)	0.01879 (0.05907)	0.01283 (0.05309)
$n = 100$				
Log	0.04730 (0.08707)	0.02791 (0.11714)	0.02839 (0.11968)	0.00177 (0.09559)
Alog	0.04820 (0.08334)	0.03416 (0.10896)	0.03480 (0.11138)	0.00705 (0.08834)
HR	0.06330 (0.09536)	0.05111 (0.11442)	0.05212 (0.11686)	0.02667 (0.09329)
Neglog	0.04914 (0.09052)	0.03213 (0.11853)	0.03269 (0.12108)	0.00718 (0.09614)
Aneglog	0.04892 (0.08751)	0.03848 (0.12238)	0.03918 (0.12501)	0.00906 (0.09609)
Bilog	0.04349 (0.08276)	0.03238 (0.11475)	0.03327 (0.11733)	0.00372 (0.09310)
Negbilog	0.05521 (0.08761)	0.03953 (0.10945)	0.04029 (0.11181)	0.01684 (0.08927)
Dir	0.04606 (0.08456)	0.03251 (0.10908)	0.03341 (0.11152)	0.00696 (0.08659)
Amix	0.05057 (0.08892)	0.03649 (0.12328)	0.03715 (0.12591)	0.01117 (0.09550)
3-Log	0.16173 (0.20723)	0.09853 (0.22673)	0.10037 (0.23158)	0.05838 (0.19029)
3-Alog	0.14043 (0.18638)	0.08810 (0.19677)	0.09013 (0.20104)	0.03478 (0.16397)

Table 3 Absolute bias and root mean squared error (in brackets) of estimators $\widehat{\varepsilon}_c^{\,CFG}$, $\widehat{\varepsilon}_c^{\,P}$, $\widehat{\varepsilon}^{\,HT}$, and $\widehat{\varepsilon}^{\,FF}$, for $\varepsilon \approx 1$

	CFG	P	HT	FF
$n = 1000$				
Log	0.00068 (0.00076)	0.00227 (0.00233)	0.00058 (0.00079)	0.00053 (0.00061)
HR	0.00068 (0.00078)	0.00227 (0.00234)	0.00058 (0.00080)	0.00050 (0.00059)
Neglog	0.00070 (0.00079)	0.00238 (0.00244)	0.00069 (0.00086)	0.00056 (0.00064)
Bilog	0.00070 (0.00079)	0.00223 (0.00228)	0.00054 (0.00071)	0.00056 (0.00065)
Negbilog	0.00071 (0.00080)	0.00229 (0.00234)	0.00060 (0.00076)	0.00054 (0.00062)
Dir	0.00046 (0.00199)	0.00215 (0.00344)	0.00046 (0.00273)	0.00036 (0.00212)
3-Log	0.00107 (0.00118)	0.00319 (0.00327)	0.00092 (0.00117)	0.00083 (0.00095)
$n = 100$				
Log	0.00178 (0.00299)	0.01319 (0.01353)	0.00194 (0.00356)	0.00193 (0.00275)
HR	0.00205 (0.00319)	0.01278 (0.01318)	0.00152 (0.00358)	0.00205 (0.00293)
Neglog	0.00166 (0.00273)	0.01247 (0.01276)	0.00121 (0.00297)	0.00188 (0.00254)
Bilog	0.00215 (0.00313)	0.01355 (0.01398)	0.00229 (0.00413)	0.00239 (0.00311)
Negbilog	0.00193 (0.00316)	0.01306 (0.01348)	0.00180 (0.00380)	0.00205 (0.00297)
Dir	0.00258 (0.00764)	0.01155 (0.01457)	0.00033 (0.00890)	0.00344 (0.00782)
3-Log	0.00226 (0.00352)	0.01693 (0.01726)	0.00177 (0.00377)	0.00254 (0.00355)

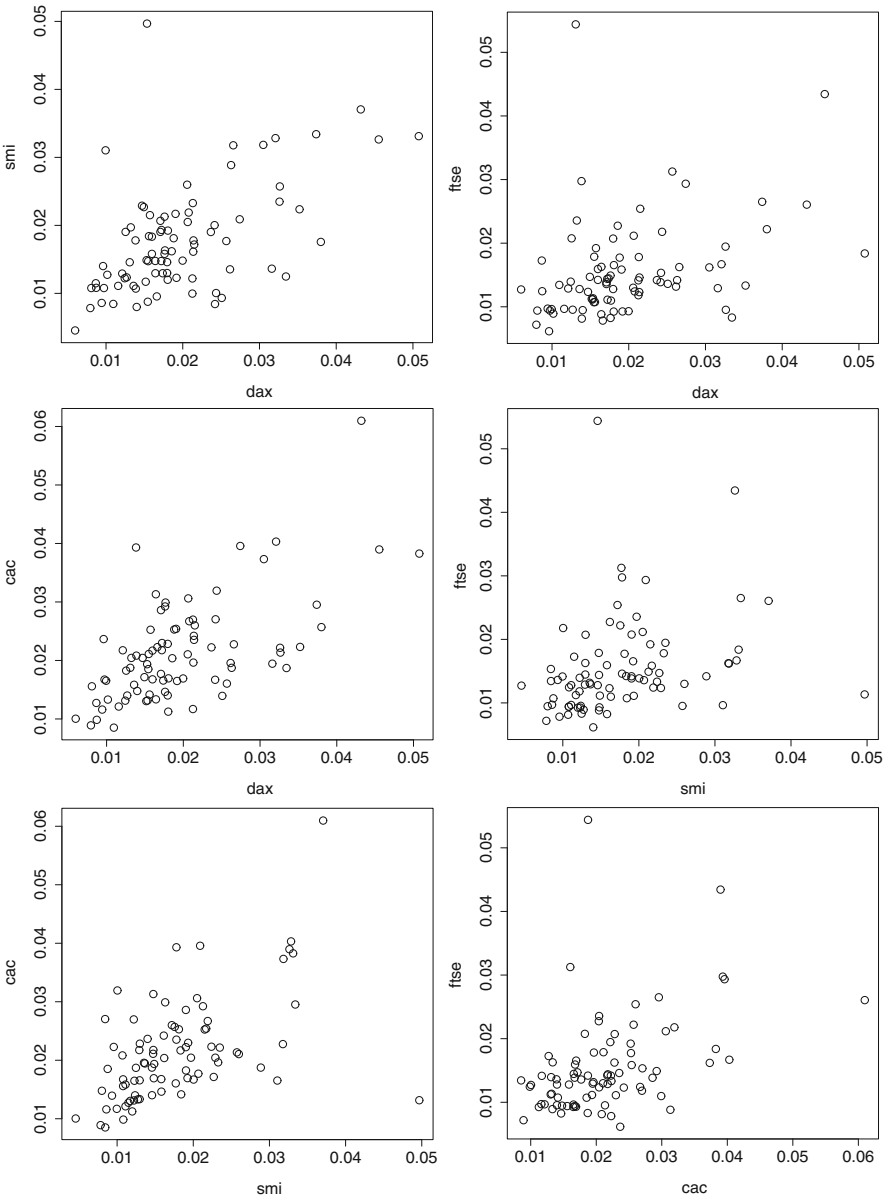


Fig. 1 Scatterplots of monthly maxima of the positive log-returns

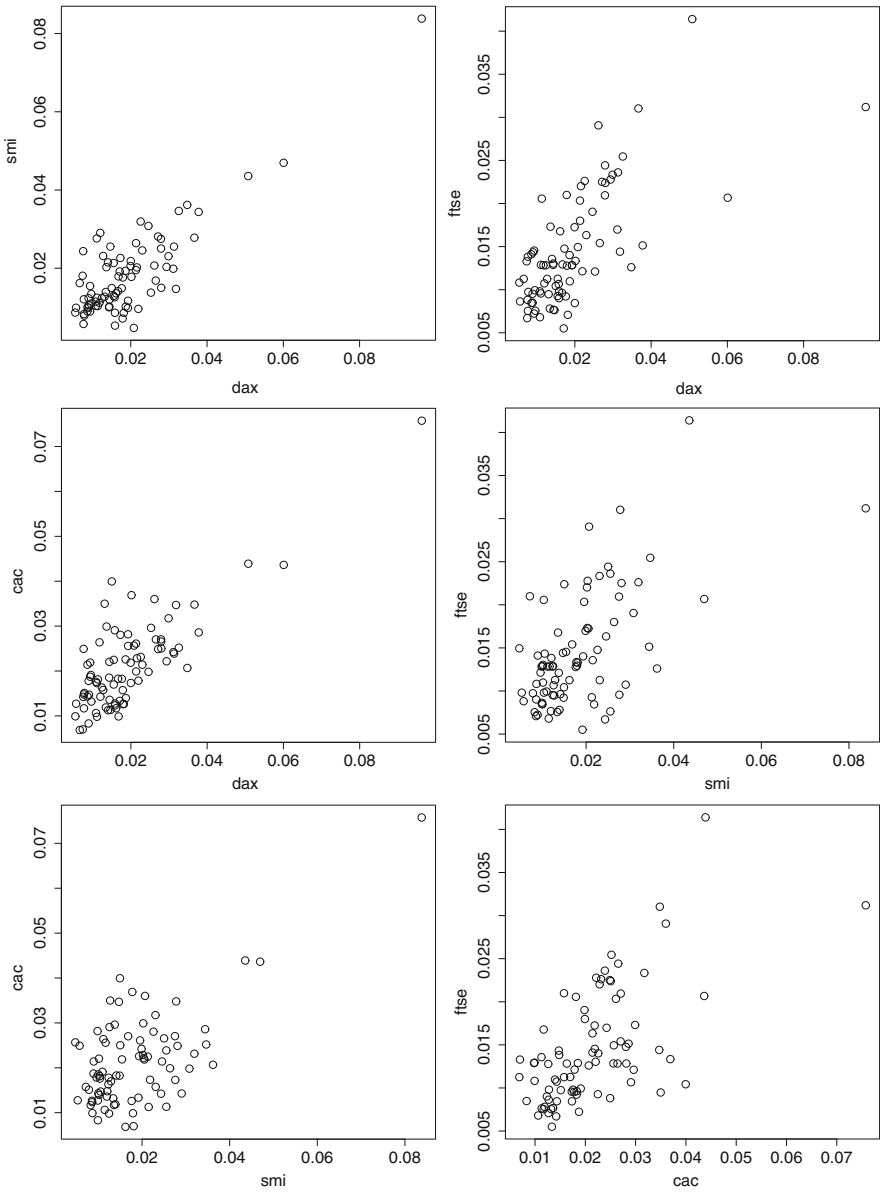


Fig. 2 Scatterplots of monthly maxima of the negative log-returns

Table 4 Estimates based on the positive log-returns

	CFG	P	HT	FF
DAX-SMI	1.51065	1.43912	1.42864	1.53281
DAX-CAC	1.53758	1.42372	1.41307	1.52922
DAX-FTSE	1.62908	1.57516	1.56645	1.63566
SMI-CAC	1.51419	1.42021	1.40952	1.51232
SMI-FTSE	1.66818	1.58037	1.57175	1.63858
CAC-FTSE	1.59196	1.59800	1.58964	1.60489
DAX-SMI-CAC	1.93313	1.73183	1.71288	1.91667
DAX-SMI-FTSE	2.10582	1.96864	1.95110	2.09493
DAX-CAC-FTSE	2.06665	1.96696	1.94940	2.07494
SMI-CAC-FTSE	2.06946	1.97731	1.95984	2.06701
DAX-SMI-CAC-FTSE	2.46149	2.28181	2.25646	1.91667

Table 5 Estimates based on the negative log-returns

	CFG	P	HT	FF
DAX-SMI	1.40796	1.46971	1.45958	1.45277
DAX-CAC	1.45738	1.47093	1.46082	1.48521
DAX-FTSE	1.42637	1.48738	1.47747	1.43936
SMI-CAC	1.62382	1.67586	1.66879	1.66517
SMI-FTSE	1.52124	1.56141	1.55251	1.53641
CAC-FTSE	1.51582	1.51732	1.50780	1.51055
DAX-SMI-CAC	1.86354	1.96386	1.94627	1.95163
DAX-SMI-FTSE	1.74872	1.84216	1.82375	1.81324
DAX-CAC-FTSE	1.80315	1.88928	1.87116	1.84349
SMI-CAC-FTSE	1.97593	2.00136	1.98408	2.02286
DAX-SMI-CAC-FTSE	2.13619	2.25712	2.23169	1.95163

Acknowledgements This research was financed by Portuguese Funds through FCT—Fundação para a Ciência e a Tecnologia, within the Project UID/MAT/00013/2013 and Project UID/MAT/00006/2013 and by the Research Centre CEMAT through the Project UID/Multi/04621/2013.

References

1. Alves, I.F., Rosário, P.: Parametric and semi-parametric approaches to extreme rainfall modelling. In: Kitsos, C.P., Oliveira, A.T., Rigas, A., Gulati, S. (eds.) Theory and Practice of Risk Assessment, pp. 279–291. Springer, Berlin (2015)
2. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes – Theory and Applications. Wiley, Chichester (2004)
3. Capéraà, P., Fougères, A.-L., Genest, C.: A nonparametric estimation procedure for bivariate extreme value copulas. Biometrika **84**, 567–577 (1997)
4. Embrechts, P., McNeil, A., Straumann, D.: Correlation and dependence in risk management: properties and pitfalls. In: Dempster, M.A.H. (eds.) Risk Management: Value at Risk and Beyond, pp. 176–223. Cambridge University Press, New York (2002)
5. Ferreira M.: Estimating multivariate extremal dependence: a new proposal. Theory Probab. Math. Stat. **93**, 156–162 (2015)
6. Ferreira, H., Ferreira, M.: On extremal dependence of block vectors. Kybernetika **48(5)**, 988–1006 (2012)
7. Genest, C., Segers, J.: Rank-based inference for bivariate extreme-value copulas. Ann. Stat. **37(5B)**, 2990–3022 (2009)
8. Gomes, D.P., Neves, M.M.: Adaptive choice and resampling techniques in extremal index estimation. In: Kitsos, C.P., Oliveira, A.T., Rigas, A., Gulati, S. (eds.) Theory and Practice of Risk Assessment, pp. 321–332. Springer, Berlin (2015)
9. Gudendorf, G., Segers, J.: Nonparametric estimation of multivariate extreme-value copulas. J. Stat. Plann. Inference **142**, 3073–3085 (2012)
10. Hall, P., Tajvidi, N.: Distribution and dependence-function estimation for bivariate extreme-value distributions. Bernoulli **6**, 835–844 (2000)
11. Pickands, J.: Multivariate extreme value distributions (with a discussion). In: Proceedings of the 43rd Session of the International Statistical Institute. Bulletin of the Institute of International Statistics, vol. 49(2), pp. 859–878, 894–902 (1981)
12. Schmidt, R., Stadtmüller, U.: Nonparametric estimation of tail dependence. Scand. J. Stat. **33**, 307–335 (2006)
13. Sklar, M.: Fonctions de répartition à n dimensions et leurs marges. Publ. Inst. Stat. Univ. Paris **8**, 229–231 (1959)
14. Smith, R.L.: Max-stable processes and spatial extremes. University of North Carolina, USA (1990, Preprint)
15. Tiago de Oliveira, J.: Structure theory of bivariate extremes, extensions. Est. Mat. Estat. Econ. **7**, 165–195 (1962/1963)

On a Business Confidence Index and Its Data Analytics: A Chilean Case

Víctor Leiva, Camilo Lillo, and Rodrigo Morrás

Abstract In this work, we present a methodology based on a Chilean business confidence index, which allows us to describe aspects of the market at a global level, as well as at industrial and sector levels of Chilean great brands. We introduce some issues related to business intelligence, customer surveys, market variables, and the confidence index mentioned. In addition, we carry out analytics of real-world data using this index, whose results show the competitiveness of some Chilean great brands.

1 Introduction

The concept of customer confidence is difficult to define, but it is highly linked to the service quality, and more currently, to business intelligence (BI). There is no consensus among different areas (such as economics, marketing, or psychology) about the definition of confidence [15]. However, recently some authors have defined confidence as a multidimensional construct that is often related to characteristics, such as benevolence, competence, honesty, and integrity [29, 36].

Consumer confidence also is defined as the degree of optimism that consumers are expressing for the state of the economy, through their saving and spending activity [31]. The first measures of consumer confidence were developed by [20], who mentioned that it is a broad measure of expected changes in income. It was

V. Leiva (✉)
School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
e-mail: victor.leiva@pucv.cl; www.victorleiva.cl

C. Lillo
Department of Computer Science, Universidad de Playa Ancha, Valparaíso, Chile

R. Morrás
Business School – Center of Experience and Services (CES), Universidad Adolfo Ibáñez, Santiago, Chile

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_5

not simply the expected size of a consumer future income, but the certainty or uncertainty that was associated with those expectations.

Consumer confidence is present and is of interest in diverse areas, from fast food brands to government institutions at local and international scales. This is because consumer confidence provides valuable information for business, financial, and political sectors, making it an important aspect to be kept in mind. For some industrial sectors, the consumer confidence is associated with her(his) satisfaction, but it is not possible to state a causal relationship between confidence and satisfaction. Furthermore, consumer satisfaction and confidence are key aspects to determine customer loyalty and commitment [17, 38]. Some studies show that consumer confidence has an impact on her(his) spending [2]. In service quality, these variables play an important role, where consumer confidence is affected directly by the experience with the brand (buyer–seller relationship). In addition, consumer confidence is the main determinant of relationship commitment between buyer and seller [8]. However, consumer confidence also can be affected by political factors or financial crises, among others [12]. USA, China, and some countries of the European Union, such as Germany and Portugal, have already advanced in consumer confidence and service quality topics [14, 18, 28, 34]. In Chile, services have been one of the fastest growing sectors in the economy. Furthermore, projections are showing a growing which will remain over time. For this reason, establishing methodologies about consumer confidence in the context of service quality for the Chile case is needed.

BI is a set of tools and methods related to statistics, informatics, mathematics, optimization, and business, which transform data into information. Then, this information is transformed into knowledge to optimize the process of decision-making in business. In this context, BI can be oriented to service, which is known as service oriented to business intelligence (SOBI). SOBI is based on customer requirements that were captured usually through periodic surveys, but also through unstructured data collected in real-time from the internet (for example, by means of e-mails, tweets, and comments in social networks). These unstructured data must be transformed into structured data and then analyzed with BI tools, for example using machine learning and/or statistical methods [3, 11].

The main objectives of this work are (i) to present the Chilean Business Confidence Index (CBCI) of the Center of Experience and Services (CES) of the Universidad Adolfo Ibáñez (UAI), CES-UAI in short, Chile, which we introduce in the next section; and (ii) to carry out analytics of real-word data related to the CBCI and collected by the CES-UAI. This analytics is useful for BI.

Section 2 mentions aspects about BI and its connection with services. Section 3 introduces the methodology based on the CBCI. Section 4 presents empirical results of the CBCI for Chilean industries by case studies. Section 5 conducts the mentioned analytics using BI tools. Section 6 discusses our conclusions and future research.

2 Business Intelligence

In this section we provide some details related to BI and its connection with the area of services.

BI can be summarized in a one word: data (*datum* in Latin) which means "what is given." Data are a symbolic representation of a qualitative or quantitative variable, whose symbol may be alphabetic or numeric, respectively. The data alone are useless. Only when they are transformed into information, we can extract reasonable conclusions from them. This information obtained from the data must be mixed with the experience to generate knowledge that allows us to make useful decisions [37].

In the modern times, the data transformation process into information is performed through computers [27]. In this way, alphabetic characters or symbols related to data of qualitative variables must be changed by numerical characters. This conversion is known technically as "coding." However, summarizing a complex concept as BI through one word as "data" can be confused, because "data" are strongly associated with "statistics," but BI is not restricted only to statistics. This concept considers a wider spectrum that involves several other sub-concepts related to informatics, statistics, and mathematics, such as detailed below.

First, the original data sources are often formed by several databases. These can have different formats and be stored into different servers. Even more, these databases may be: (i) structured (consisting of non-numeric or numeric characters stored in tables); (ii) non-structured (consisting of text, videos, pictures, tweets, among others); or (iii) semi-structured (consisting by structured and non-structured data). This structural aspect of the data is closely linked to "big data" (or massive data) concept, on which many people is calling big data revolution, but that we will not discuss in this article in detail. However, we can say that big data are information assets, characterized by their large volume, velocity and variety (3Vs), requiring innovative and efficient solutions to improve the knowledge process when decision making in organizations. The objective of big data is to provide high technology (hardware and software) to store, process, and analyze large amounts of data (mega, giga, tera, peta, exa, zetta, and yottabyte) and to create value in an organization. Thus, with today's technologies, such as digital equipment, analytical sensors and radio frequency identification, big data are collected efficiently, rapidly, and automatically [5]. In summary, the big data term is used frequently to describe large, diverse, and complex data sets, which are generated from different types of instruments, sensors, or computer-based transactions. This results in great opportunities for knowledge discovery. Nevertheless, facing the incoming big data era, many important concepts need to be updated, particularly, service quality. The reader interested in big data can find for more details in [25, 33]. Thus, BI begins (first stage) with the identification of the databases that will allow the structured data set to be analyzed and stored in the "data warehouse." The second stage of BI is the way of forming the data set (stored in the data warehouse), which will be used to carry out the analytics (data processing) and to generate knowledge discovering in databases (KDD). Then, once the databases are identified, the data must be extracted

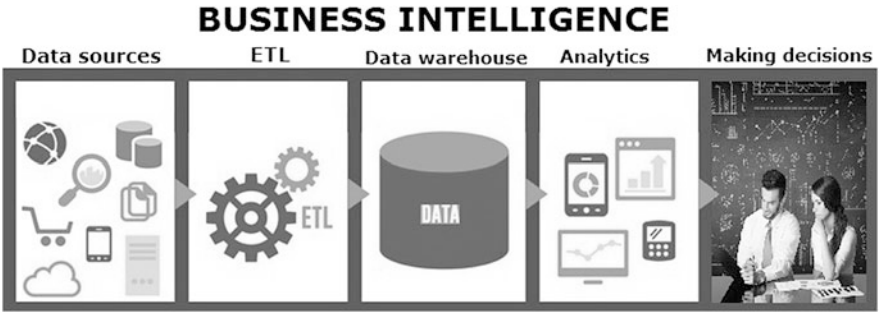


Fig. 1 Scheme of BI

to be then analyzed. This second stage of BI requires the use of a concept known as extract, transform, and load (ETL). It corresponds to a process which allows the organizations to move data from multiple sources, reformat, clean (or purge) and load them into the data warehouse to form the data set to be analyzed. The third stage of BI is to build the data warehouse. Thus, once we have it, the fourth stage of BI is to perform the data analytics. This concept corresponds to a generic term used in the context of BI and refers to a set of techniques that allow structured data to be processed and analyzed with the aim of transforming them into information. The fifth and final stage of BI is to make decisions that support the business process based on the information obtained from the data (quantitative aspect) and the experience of the decision makers (qualitative aspect). Figure 1 illustrates the steps of BI. For more information about these and other concepts, see [37].

The elements of BI are the following:

- **Data**: primary elements of the information, which do not have meaning for themselves, but they are useful to support the decision-making process.
- **Information**: processed data with meaning (relevance, purpose, and context).
- **Knowledge**: mixture of experience and information that is useful for decision making.

In BI, the focus of analysis is the customer (client), which it does not occur with other similar concept to BI known as "data science," where the customer is not necessarily the focus of analysis. Thus, in BI, the concept of customer relationship management (CRM) arises naturally [6]. CRM analytic combines businesses and technologies to analyze data of customers, discovering new patterns of behavior or market trends. The CRM analytic identifies problems in consumer service quality, for example confidence problems. When the levels of confidence explained by the CBCI are high, the attitude and predisposition of the customers towards the brand is positive. That is, the interactions and service experience are easier to achieve, responding to the expectations of customers. Then, the CBCI is a measure that allows the brands to anticipate when image and credibility in the company must be taken into account for improving its performance. The steps of BI are applied in the methodology defined in Sect. 3.

3 Methodology of the Chilean Business Confidence Index

In this section, we present the methodology used for the CBCI, from the collection of data to the construction of the index.

3.1 Target Population and Sample Design

When the CBCI is estimated, the target population under study is composed of all Chilean people older than 18 years, who live in cities with at least 130,000 inhabitants. The size of statistical sample is determined for each brand, which correspond a number of 128 brands in 2016. Each brand belongs to a sector, while each sector corresponds to a type of industry. Due to reasons of confidentiality, brand names are kept as anonymous. Chilean industries and sectors used to calculate the CBCI are shown in Fig. 2. A sample size of $n = 300$ customers is established considering statistical aspects. The survey is carried out annually to evaluate brands. However, the survey is also replicated semiannually at global, industrial, and sector levels. Note that the questions of the questionnaire used to calculate the CBCI are taken from the first semester of 2012.

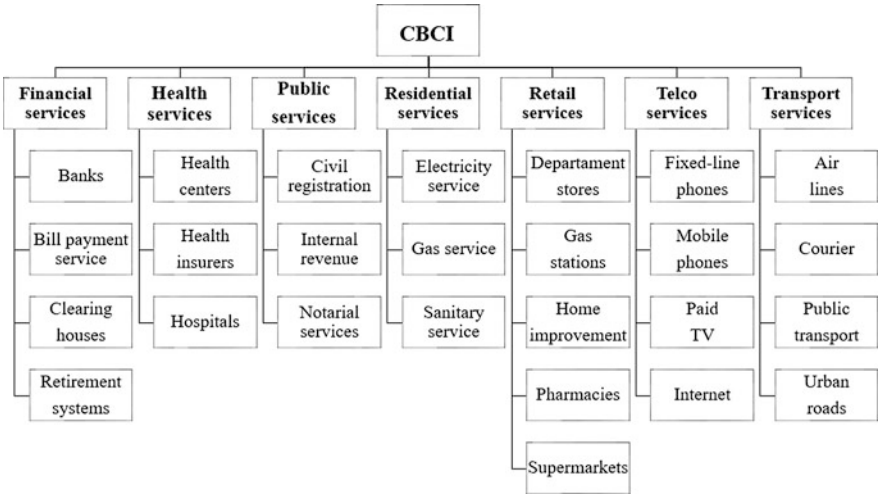


Fig. 2 Structure of Chilean industries and sectors used to calculate the CBCI in 2016

3.2 Data, Interview and Re-interview

The data are collected by using the computer-assisted telephone interviewing (CATI) method. This method has the advantage of immediate availability of results, which can be quickly analyzed and interpreted. By using the CATI method, it is possible to detect measurement errors easily, whereas coverage and non-response errors are more frequent. For more details about the CATI method, see [21].

Through phone calls by using the CATI method, customers respond to a structured questionnaire regarding a brand. However, a customer can be interviewed more than once in the sample and respond to the questionnaire repeatedly, which is known as re-interview. The re-interview concept corresponds to a procedure of replicating the questionnaire to the same consumer in another instant of time, which saves time and costs. In our case, the re-interview is aimed to another brand that belongs to another sector and industry than the previous brand for which the consumer was interviewed. The fact that a same customer responds the questionnaire more than once it can produce a statistical bias and the respondents may also repeat the type of response due to exhaustion. This usually happens in questionnaires sent by mail [7]. The errors associated with the re-interview may be quantified by means of statistical models; see [9, pp. 264–265]. In a survey with re-interviews, two assumptions must be considered: (i) the mean and variance of the response error associated with the re-interviews and the original interview are identical; and (ii) the covariance between the errors in the different occasions is zero. In our study, we have checked by means of statistical hypotheses testing that both assumptions are fulfilled. We will leave a more elaborated study of this topic for another work. For more information about the re-interviews, see [9, Chapters 11 and 15].

3.3 Customer Survey

The CBCI reported by the CES-UAI is composed by four questions. These questions and their corresponding random variables are detailed in Algorithm 2. Note that questions 1 and 2 are related to confidence and transparency, respectively. Then, they correspond to a rational perception and are based on the customer experience with the brand. However, questions 3 and 4 are related to concerns and compliance, respectively. Thus, they correspond to an emotional aspect and are based on the feelings of the consumer.

3.4 Building the CBCI

Due to the complexity of the confidence concept, it is usually measured through a consumer confidence index (CCI), which is built based on non-related and related questions from business and consumer surveys. A CCI is helpful because it captures consumer buying patterns [10]. There are currently two important and well-known CCIs: (i) the University of Michigan Consumer Sentiment Index (UMCSI) and (ii) the Conference Board Consumer Confidence Index (CBCCI). For definitions and comparison between these indices, the interested reader is referred to [26].

Algorithm 1 Thinking exclusively about {the name of the brand is indicated}, mention with a score from 1 to 7 how much you agree with the following statements:

1: I can trust on it (Y_1).
2: It is transparent, it does not deceive me and it does not hide anything (Y_2).
3: It cares about the well-being of its customers (Y_3).
4: When it promises anything, it complies (Y_4).

Such as in the case of UMCSI and CBCCI, responses to questions for the CBCI can be categorized as negative, neutral, or positive [19]. In order to build the confidence index on a brand, Algorithm 2 must be followed. Furthermore, we may calculate the sector, industrial, and global CBCI following Algorithms 3, 4, and 5, respectively. Note that these three algorithms are nested in Algorithm 2. In order to calculate the confidence index in the sector h, we need to calculate the index on each brand that belongs to the sector h. This also happens for the industries, that is, in order to calculate the confidence index in the industry t, we need to calculate the index on each sector that belongs to the industry t. For the global CBCI, it is necessary to calculate the index in each sector. This procedure can be replicated each year, obtaining a time series of the CBCI (global, sector or brand); see Sect. 4. Then, through the BI framework, each data set stored annually in the data warehouse should be treated by an ETL.

4 Case Study I: CBCI on 2016

In this section we present some current results of the CBCI.

We analyze the CBCI for the second semester of 2016 to global, industrial, and sector levels, which is shown in Fig. 3. From this figure, note that the values above the black points correspond to the CBCI in the sectors. In addition, we can change expression given in (2) of Algorithm 2 by the proportion of positive (or negative) responses to obtain an indicator of positivism (or negativity) using the CBCI. These values are displayed in the extremes of the bars of Fig. 3. In order to compare the sectors with the global CBCI, a dashed horizontal line is sketched. Observe that

Algorithm 2 Constructing the CBCI for the brand k

1: For the brand k and the question Y_i, with $i = 1, \ldots, 4$, encode as follows:

$$U_i = \begin{cases} -1, \text{ if } Y_i = 1, 2, 3 \text{ or } 4, \text{ indicating a negative response;} \\ 0, \text{ if } Y_i = 5, \text{ indicating a neutral response;} \\ 1, \text{ if } Y_i = 6 \text{ or } 7, \text{ indicating a positive response.} \end{cases} \quad (1)$$

2: Calculate the percentage of positive responses and subtract the percentage of negative responses. This is analogous to calculating

$$\overline{U_i} = \frac{1}{n_k} \sum_{j=1}^{n_k} U_{ij} \times 100, \quad (2)$$

where U_{ij} is defined in (1) and corresponds to the encoded variable Y_i for the customer j, with n_k being the sample size of the brand k.

3: Repeat steps 1 and 2 for $i = 1, 2, 3, 4$.

4: Compute the CBCI composed by the four variables through the average, that is, by

$$B_k = \frac{1}{4} \sum_{i=1}^{4} \overline{U_i}, \quad (3)$$

where U_{ij} is defined in (2).

Algorithm 3 Constructing the CBCI for the sector h

1: Repeat Algorithm 2 for each brand that belongs to the sector h.

2: Calculate the CBCI as the average of the brand index, that is, as

$$S_h = \frac{1}{b_h} \sum_{k=1}^{b_h} B_k, \quad (4)$$

where B_k is defined in (3) and b_h is the number of brands that belong to the sector h.

sectors may be grouped into two classes: (i) the membership sectors (dark gray) and (ii) transactional sectors (light gray). The membership class corresponds to sectors where the relationship between customer and company is contractual. However, in the transactional class, a contract is not needed. In general, for transactional sectors, there is a lower relationship between customer and brand, and then a lower level of satisfaction and confidence from customers. However, this does not happen in the Chilean reality; for more details on the membership and transactional sectors and

Algorithm 4 Constructing the CBCI for the industry t

1: Repeat Algorithm 3 for each sector that belongs to the industry t.
2: Calculate the CBCI as the average of the sector index, that is, as

$$I_t = \frac{1}{s_t} \sum_{h=1}^{s_t} S_h,$$

where S_h is defined in (4) and s_t is the number of sectors that belong to the industry t.

Algorithm 5 Constructing the global CBCI

1: Use Algorithm 3 for all sectors.
2: Compute the global CBCI as the average of all sector indices, that is, as

$$\text{CBCI} = \frac{1}{s} \sum_{h=1}^{s} S_h,$$

where S_h is defined in (4), using now all sectors which are denoted by a number s.

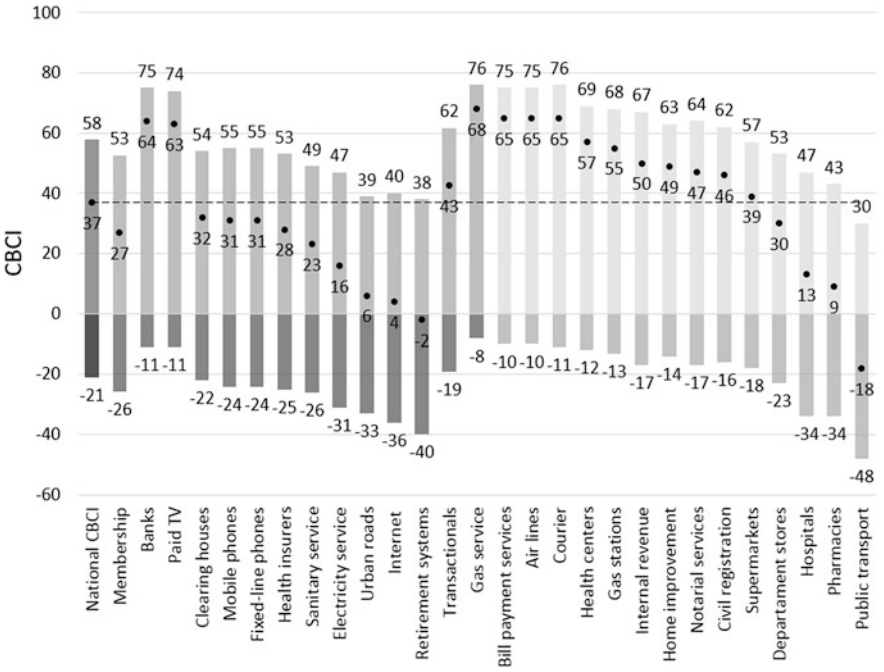


Fig. 3 Values of the CBCI during the second semester of 2016 to global and sector levels

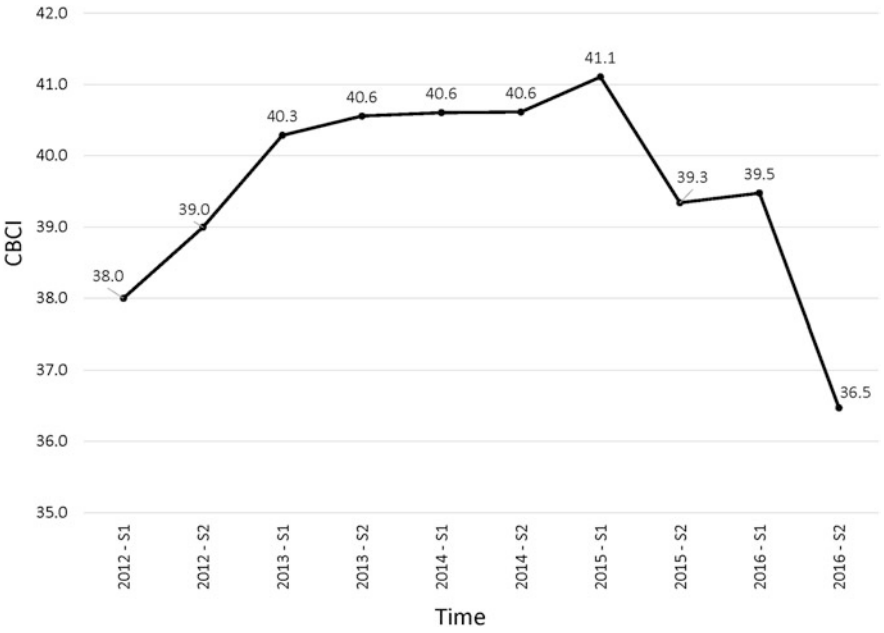
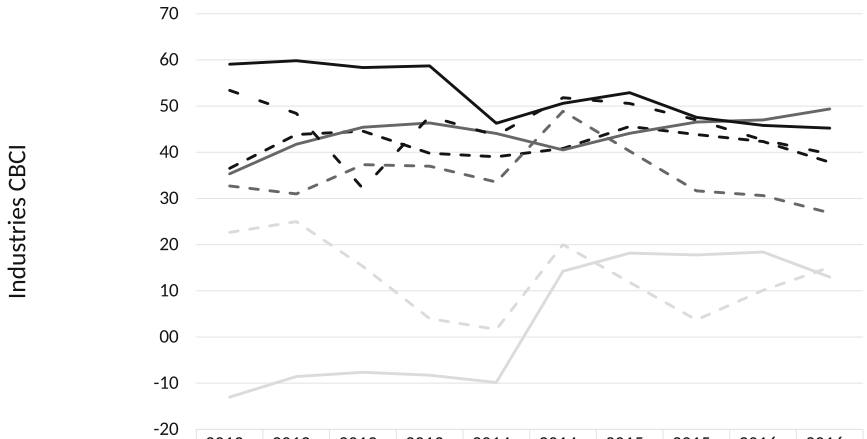


Fig. 4 Semiannual time series of the global CBCI for the indicated date

brands, see [16]. Observe that the worst evaluated sector is the public transport, with an value of −18 for the CBCI, which corresponds to Transantiago's problem [30].

In addition to a static result, it is necessary to evaluate the evolution of the behavior detected previously. Figures 4 and 5 show semiannual results over time of the CBCI at global and industrial levels; respectively. From Fig. 4, note that, since the first half of 2015, there has been a downward trend in the CBCI, behavior which is identical for industries related to financial, logistic, public, residential, retail, telecommunication and transport services. Also, we present the behavior over time of the financial industry in Fig. 6. Observe that, in the financial industry, there is also a general declination since 2015, where credit cards, clearing houses, retail banking and retirement pension system follow this pattern. Observe that the most notorious changes are found in the retirement pension system, with a jump of 27 points (from 24 to −3) in the CBCI between the first semester of 2015 and the second semester of 2016. Due to a confidentiality issue, we do not present the results at a brand level. Despite using only descriptive statistics in this section, note that some results can be already obtained as market trends. Through the same procedure shown in this section, we compare brands of a sector in order to establish new service policies in the businesses of these brands.



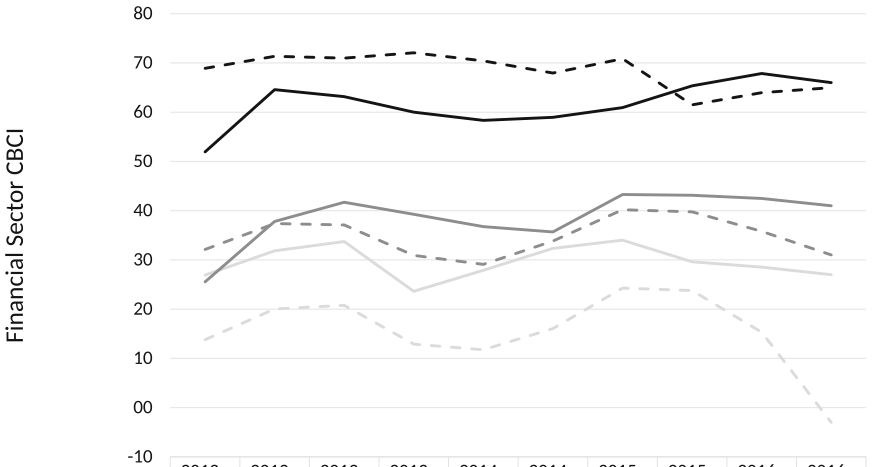
	2012 - S1	2012 - S2	2013 - S1	2013 - S2	2014 - S1	2014 - S2	2015 - S1	2015 - S2	2016 - S1	2016 - S2
Financial	37	44	45	40	39	41	46	44	42	38
Residentials	33	31	37	37	34	49	40	32	31	27
Health	35	42	45	46	44	41	44	47	47	49
Telecommunications	23	25	15	04	02	20	12	04	10	15
Transport and logistic	-13	-09	-08	-08	-10	14	18	18	18	13
Retail	53	48	32	48	44	52	51	47	43	40
Public services	59	60	58	59	46	51	53	48	46	45

Fig. 5 Semiannual time series of the industrial CBCI for the indicated date

5 Case Study II: Regression Modeling of the CBCI

Market behavior can be described by partial least squares regression models and/or structural equation models [1, 13, 32]. However, these models commonly have several random variables and it is difficult to verify all their assumptions (for example: constant variance and/or normality of the response variable). In this section, we show an illustration about the CBCI at brand level (Y), which is modeled by a service covariate (X). This covariate is also measured in the customer survey of the CES-UAI as follows: "Regarding to {the name of the brand is indicated}, how well do you agree with the sentence: was it a pleasant experience?". Use a score from 1 to 7. Note that this covariate is directly related to the experience that the customer had with the brand. This type of service covariate is an excellent predictor of confidence. To determine X at the brand level, we use Algorithm 2.

In this case study, we consider a data set corresponding to 2016, with $n = 128$ Chilean brands and 30,000 clients. We obtain the CBCI for each brand based on the 30,000 clients, which provides us $n = 128$. Table 1 presents a descriptive summary on the data of the CBCI for the brands, whereas Fig. 7 displays the histogram (left), box-plot (center), and scatter-plot (right) of the CBCI for brands. From Table 1 and



	2012 - S1	2012 - S2	2013 - S1	2013 - S2	2014 - S1	2014 - S2	2015 - S1	2015 - S2	2016 - S1	2016 - S2
Banks	52	65	63	60	58	59	61	65	68	66
Commercial cards	27	32	34	24	28	32	34	30	29	27
Clearing houses	32	37	37	31	29	34	40	40	36	31
Retail banking	26	38	42	39	37	36	43	43	42	41
Pension system	14	20	21	13	12	16	24	24	15	-03
Account payment	69	71	71	72	70	68	71	61	64	65

Fig. 6 Semiannual time series of the CBCI in financial sector for the indicated date

Table 1 Descriptive summary of the CBCI for 128 brands with data of 2016 from the CES-UAI

n	Min.	Median	Mean	Max.	Standard deviation	Coefficient of skewness	Excess of kurtosis
128	−17	37.5	37.81	83	22.23	−0.02	−0.70

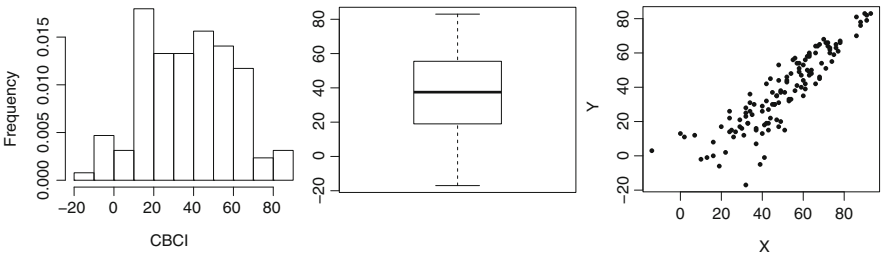


Fig. 7 Histogram (left), box-plot (center) and scatter-plot (right) of the CBCI for 128 brands with data of 2016 from the CES-UAI

Fig. 7, note that the empirical distribution of the CBCI is symmetrical and similar to the normal distribution, although more platykurtic (the normal distribution is mesokurtic with excess of kurtosis equal to 0, while brands have a CBCI with an excess of kurtosis equal to -0.7). Observe also that there is a linear relationship between X and Y. In addition, note that as the CBCI increases, its variability decreases. Thus, we are dealing with a heteroscedastic model. The simple linear regression model based on the normal distribution is one of the most used. However, it is well known that this model does not allow for heteroscedasticity (non-constant variance). Then, we describe both mean and variance with generalized additive models of location, scale and shape (GAMLSS) as detailed below [35].

The model to be considered is given by

$$g_1(\mu_i) = \eta_{1i} = \beta_{10} + \beta_{11}x_i, \quad g_2(\sigma_i) = \eta_{2i} = \beta_{20} + \beta_{21}x_i, \quad i = 1, \ldots, n, \tag{5}$$

where μ_i, σ_i are the mean and standard deviation of the CBCI and x_i the value of the service covariate, all of them for the customer i, whereas β_{jl}s are the regression coefficients. We compare the performance of different distributions of GAMLSS. Specifically, we consider the normal, generalized-t and skew exponential power type 1 (SEP1) distributions. We use for g_1 the identity link function, whereas the logarithmic and identity link functions are used for g_2. Once the models are fit, we compare them through model selection criteria based on loss of information such as Akaike (AIC) and Bayesian (BIC) information criteria. AIC and BIC allow us to compare models for the same data set and they are given by

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p, \quad \text{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + p\log(n),$$

where $\ell(\widehat{\boldsymbol{\theta}})$ is the logarithm of the likelihood function (log-likelihood) of the model with vector of parameters $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, n is the sample size, and p is the number of model parameters. AIC and BIC correspond to the log-likelihood function plus a component that penalizes such a function as the model has more parameters, making it more complex. A model with a smaller AIC or BIC is better; for more information about AIC and BIC, see [22]. Values for AIC and BIC considering the data are presented in Table 2. Note that a model with a lower information criterion, for both AIC and BIC, corresponds to the SEP1 with identity link function for g_2. The probability density function (PDF) of the SEP1 distribution, denoted by SEP1(μ, σ, ν, τ), is given by

$$f_Y(y; \mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_1}(\nu z); \quad y \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0, \nu \in \mathbb{R}, \tag{6}$$

Table 2 Values of AIC and BIC for the indicated model fitted to the CBCI data for 128 brands in 2016 from the CES-UAI

	Distribution					
	Normal		Generalized-t		SEP1	
Link function ($g_2(\sigma_i)$)	Identity	Log	Identity	Log	Identity	Log
AIC	927.86	932.78	931.67	936.79	927.20	933.88
BIC	944.31	944.19	948.78	953.90	939.27	951.00

where $z = (y-\mu)/\sigma$, f_{Z_1} and F_{Z_1} are the PDF and cumulative distribution function (CDF) of $Z_1 \sim \text{PE2}(0, \tau^{1/\tau}, \tau)$, a power exponential type 2 distribution with PDF given by

$$f_{Z_1}(z) = 2\tau^{(1-\tau)/\tau}\Gamma(1/\tau)\exp\left(-|z|^\tau/\tau\right); \quad z \in \mathbb{R}, \tau > 0.$$

The mean and variance of the PE2 distribution is given by $\text{E}[Y] = \mu + \sigma\text{E}[Z]$ and $\text{Var}[Y] = \sigma^2\text{Var}[Z] = \sigma^2(\text{E}[Z^2] - (\text{E}[Z])^2)$, respectively, where $Z = (Y-\mu)/\sigma$,

$$\text{E}[Z] = \text{sign}(\nu)\tau^{1/\tau}\frac{\Gamma(2/\tau)}{\Gamma(1/\tau)}\text{pBEo}(\nu^\tau/(1+\nu^\tau), 1/\tau, 2/\tau),$$

and

$$\text{E}[Z^2] = \frac{\tau^{2/\tau}\Gamma(3/\tau)}{\Gamma(1/\tau)},$$

with pBEo being the CDF of a beta distribution; for more information of SEP1 and EP2 distributions, see [4]. The SEP1 distribution parameters are estimated with the maximum likelihood (ML) method. Then, for the model formulated in (5), using $g_1 = g_2$ based on identity link functions and the PDF given in (6), we obtain the following ML estimates (with estimated asymptotic standard errors in parenthesis) for the corresponding parameters: $\widehat{\beta}_{10} = -24.68(6.16)$, $\widehat{\beta}_{11} = 1.15(0.07)$, $\widehat{\beta}_{20} = 24.37(6.71)$, $\widehat{\beta}_{21} = -0.24(0.06)$, $\widehat{\nu} = 0.53(0.18)$, and $\widehat{\tau} = 1.31(0.67)$. Note that all coefficients are significant at 5%. The assumptions of the model given in (5) are verified with the randomized quantile (RQ) residual, often used in GAMLSS models [24], defined as

$$r_i^{\text{RQ}} = \Phi^{-1}\left(F_Y(y_i; \widehat{\beta}_{10} + \widehat{\beta}_{11}x_i, \widehat{\beta}_{20} + \widehat{\beta}_{21}x_i, \widehat{\nu}, \widehat{\tau})\right), \quad i = 1, \ldots, 128, \tag{7}$$

where $F_Y(y; \mu, \sigma, \nu, \tau) = \int_0^y f_Y(u; \mu, \sigma, \nu, \tau)\text{d}u$ and Φ^{-1} is the inverse function of the standard normal CDF (quantile function), with $f_Y(y; \mu, \sigma, \nu, \tau)$ being defined in (6). Note that the RQ residual expressed in (7) must be compared to the normal distribution to evaluate its fitting to the data of 2016 from the CES-UAI for 128 brands. We use a theoretical probability versus empirical probability (PP) plot to do this evaluation. In addition, observe that the PP plot can be linked to the

Kolmogorov-Smirnov (KS) test, by means of which acceptance bands may be constructed inside of this plot. Algorithm 6 summarizes this construction [23]. Figure 8 (left) sketches a PP plot with 95% acceptance bands to verify the distributional assumption of the model given in (5). Note that KS p-value $= 0.7318$, which supports the normality assumption of the RQ residuals obtained from the heteroscedastic SEP1 regression model. This figure does not show unusual features and assumption that the response follows an SEP1 distribution seems to be suitable. From Fig. 8 (right), observe that no outlying observations are detected.

Algorithm 6 Goodness of fit to any distribution

1: Consider data $y_1, \ldots, y_n$ and order them as $y_{1:n}, \ldots, y_{n:n}$.
2: Estimate parameters $\boldsymbol{\theta}$ of the distribution by $\widehat{\boldsymbol{\theta}}$ with $y_1, \ldots, y_n$ and the ML method.
3: Compute $\widehat{v}_{j:n} = F(y_{j:n}; \widehat{\boldsymbol{\theta}})$, for $j = 1, \ldots, n$, with F being the corresponding CDF.
4: Calculate $\widehat{s}_j = \Phi^{-1}(\widehat{v}_{j:n})$.
5: Obtain $\widehat{u}_{j:n} = \Phi(\widehat{z}_j)$, with $\widehat{z}_j = (\widehat{s}_j - \bar{s})/d_s$, $\bar{s} = \sum_{j=1}^{n} \widehat{s}_j/n$ and $d_s = (\sum_{j=1}^{n} (\widehat{s}_j - \bar{s})^2/(n-1))^{1/2}$.
6: Draw the PP plot with points $w_{j:n} = (j - 0.5)/n$ versus $\widehat{u}_{j:n}$, for $j = 1, \ldots, n$.
7: Specify a significance level α.
8: Construct acceptance bands according to $(\max\{w - k_{1-\alpha} + 0.5/n, 0\}, \min\{w + k_{1-\alpha} - 0.5/n, 1\})$, where $k_{1-\alpha}$ is the $(1-\alpha) \times 100$th percentile of the KS distribution (adapted) and w is a continuous version of $w_{j:n}$.
9: Determine the p-value of the KS statistic and reject the null hypothesis of the corresponding distribution for the specified significance level α based on this p-value.
10: Corroborate coherence between steps 8 and 9.

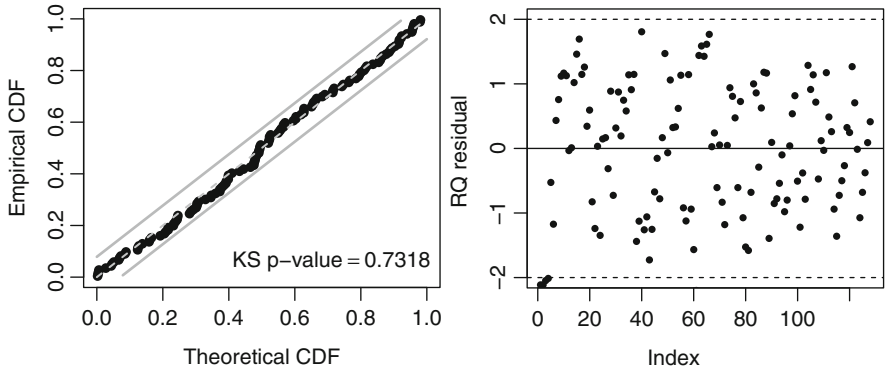


Fig. 8 PP plot with 95% acceptance bands for RQ residual (left), plot of index values against RQ residual (right) based on CBCI data for 128 brands in 2016 from the CES-UAI and the heteroscedastic SEP1 regression model

Figure 9 (left) shows a scatter-plot of the CBCI against pleasant experience indicator for Chilean brands, with the estimated model given in (5). Note that the fitted line (proposed model) has a good agreement with the data. In addition, from Fig. 9 (right), observe how the model captures heteroscedasticity. In summary, the proposed model allows us to categorize the competitiveness of a sector. Thus, brands with high competitiveness struggle to be the best, delivering good services, and consequently, the sector has a higher level of confidence. Then, a brand with a high competitiveness has a small variance in confidence, while a brand with low competitiveness has a large variance. Figure 10 shows a scatter-plot with the fitted model using fifth and 95th percentiles. This figure also presents brands in grey dots. In particular, as an example, black triangles correspond to brands in the internet sector, while black dots are related to brands in the gas sector. Due to reasons of confidentiality, as mentioned, we do not mention the name of these brands. Note that the gas sector is modeled with a low variability, because it is a sector with a

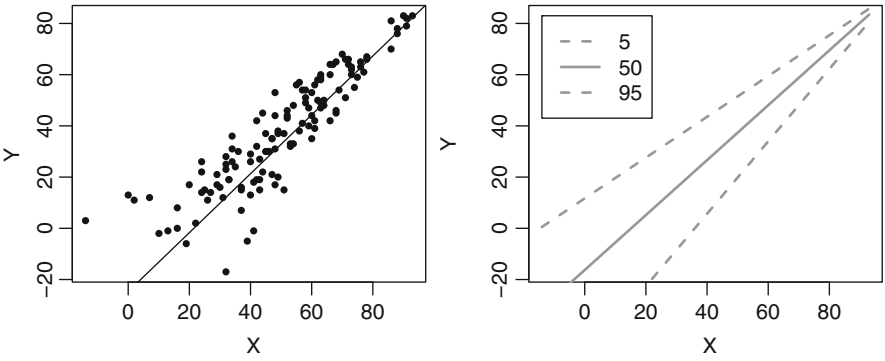


Fig. 9 Scatter-plot of CBCI data for 128 brands in 2016 from the CES-UAI with fitted SEP1 model (left) and predicted median CBCI brands and 5th and 95th percentiles (right)

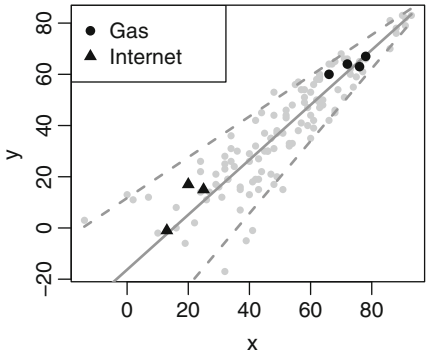


Fig. 10 Scatter-plot of CBCI data for 128 brands in 2016 from the CES-UAI with fitted SEP1 model and competitiveness of the sectors

high competitiveness. Figure 10 displays brands associated with the internet sector, from which is detected that this sector has a low competitiveness and therefore a large variance; see also Fig. 3. By a matter of confidentiality, the rest of the sectors are not shown and neither the name of the brand. From the point of view of BI, note that this result is totally related to the management and investment of new market plans to increase the competitiveness of a brand using SOBI.

6 Conclusions and Future Research

In this work we have presented a methodology related to a Chilean business confidence index, which is used to describe aspects of the market at global, industrial, and sector levels for Chilean brands. We have commented some issues related to business intelligence, customer and business surveys, market variables and of the mentioned confidence index. We have illustrated the methodology by business intelligence analytics related to a Chilean business confidence index, from the collection of data, characteristics of the business survey, calculation of the index, data analytics and decision making in service oriented to business intelligence. Two case studies have been considered in this illustration, which have provided a good idea about how the Chilean business confidence index works. The survey measures rational and emotional characteristics, which are related to brand empathy and customer experience, characteristics currently used in business intelligence for the analysis of sentiments. These characteristics are used to generate an index that can be used as an economic descriptor in the Chilean market. We have considered a heteroscedastic regression model to predict CBCI data, which is an evident aspect frequently did not considered in other analyses. This heteroscedasticity is described by generalized additive models of location, scale, and shape, which is employed to predict the Chilean business confidence index and its variability, which allowed us to determine whether a sector is highly competitive or not. This is important in service oriented to business intelligence and service quality, since it allows a direct reaction by the board of a brand to develop business management increasing its competitiveness and being the best in the local or international markets.

For future research, we propose to develop more sophisticated models for describing confidence, through partial least squares regression models, structural equation models, and/or mixtures of them. In addition, because the Chilean business confidence index was first generated during 2012, we do not have the appropriate sample size to perform time series. However, generalized additive models of location, scale, and shape allows us to describe this temporal structure of confidence, which is under study by the authors. Also, multivariate aspects and artificial intelligence models, as well as diagnostic methods, are being considered by the authors in future research.

Acknowledgements The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript. This research work was partially supported by FONDECYT 1160868 grant from the Chilean government.

References

1. Aktepe, A., Ersöz, S., Toklu, B.: Customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling. Comput. Ind. Eng. **86**, 95–106 (2015)
2. Allenby, G., Jen, L., Leone, R.: Economic trends and being trendy: the influence of consumer confidence on retail fashion sales. J. Bus. Econ. Stat. **14**, 103–111 (1996)
3. Aufaure, M., Zimányi, E.: Business Intelligence. Springer, Berlin (2012)
4. Azzalini, A.: Further results on a class of distributions which includes the normal ones. Statistica **46**, 199–208 (1986)
5. Baesens, B.: Analytics in a Big Data World: The Essential Guide to Data Science and its Applications. Wiley, New York (2014)
6. Brijs, B.: Business Analysis for Business Intelligence. Auerbach Publications, Boca Raton-FL (2012)
7. Bushery, J.M., Royce, M., Kasprzyk, D.: The schools and staffing survey: how reinterview measures data quality. Proc. Surv. Res. Methods Sect. Am. Stat. Assoc. **23**, 458–463 (1992)
8. Chumpitaz, R. Paparoidamis, N. Service quality, relationship satisfaction, trust, commitment and business to business loyalty. Eur. J. Mark. **41**, 836–867 (2007)
9. Cox, B., Binder, D., Nanjamma, B., Christianson, A., Colledge, M., Kott, P.: Business Survey Methods. Wiley, London (1995)
10. Curtin, R.: Consumer sentiment surveys: worldwide review and assessment. J. Bus. Cycle Meas. Anal. **1**, 7–42 (2007)
11. Daniel, M., Ferreira, R., Horta, N.: Company event popularity for financial markets using twitter and sentiment analysis. Expert Syst. Appl. **71**, 111–124 (2017)
12. Dees, S., Brinca, P.B.: Consumer confidence as a predictor of consumption spending: Evidence for the united states and the euro area. Int. Econ. **134**, 1–14 (2013)
13. Fornell, C., Johnson, M.D., Anderson, E.W., Cha, J., Bryant, B.E.: The American customer satisfaction index: nature, purpose, and findings. J. Mark. **60**, 7–18 (1996)
14. Hartmann, M., Klink, J., Simons, J.: Cause related marketing in the German retail sector: exploring the role of consumers trust. Food Pol. **52**, 108–114 (2015)
15. Hobbs, J., Goddard, E.: Consumers and trust. Food Pol. **52**, 71–74 (2015)
16. Hultén, B.: Customer segmentation: the concepts of trust, commitment and relationships. J. Target. Meas. Anal. Mark. **15**, 256–269 (2007)
17. Hunneman, A., Verhoef, P.C., Sloot, L.M.: The impact of consumer confidence on store satisfaction and share of wallet formation. J. Retail. **91**, 516–532 (2015)
18. Jiang, H., Zhang, Y.: An investigation of service quality, customer satisfaction and loyalty in China's airline market. J. Air Transp. Manag. **57**, 80–88 (2016)
19. Katona, G.: Psychological Analysis of Economic Behavior. McGraw Hill, New York (1951)
20. Katona, G.: The Powerful Consumer: Psychological Studies of the American Economy. McGraw-Hill, New York (1960)
21. Kenett, R., Salini, S.: Modern Analysis of Customer Surveys with Applications using R. Wiley, London-UK (2012)
22. Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2008)
23. Leiva, V., Saunders, S.C.: Cumulative damage models. Wiley StatsRef: Statistics Reference Online (2015)
24. Leiva, V., Ferreira, M., Gomes, M.I., Lillo, C.: Extreme value Birnbaum-Saunders regression models applied to environmental data. Stoch. Environ. Res. Risk. Assess. **30**, 1045–1058 (2016)
25. Liebowitz, J.: Big Data and Business Analytics. Auerbach Publications, New York (2013)
26. Ludvigson, S.: Consumer confidence and consumer spending. J. Econ. Perspect. **18**, 29–50 (2004)
27. Maheshwari, A.: Business Intelligence and Data Mining. Business Expert Press, New York (2015)

28. McGuckin, R.H.: Business Cycle Indicators Handbook. The Conference Board, New York (2001)
29. Moriuchi, E., Takahashi, I.: Satisfaction, trust and loyalty of repeat online consumer within the Japanese online supermarket trade. Aust. Mark. J. **24**, 146–156 (2016)
30. Muñoz, J.C., Batarce, M., Hidalgo, D.: Transantiago, five years after its launch. Res. Transp. Econ. **48**, 184–193 (2014)
31. Nekrasova, D.: Emotion and reason in making financial decisions. Int. J. Interdisc. Soc. Sci. **5**, 10 (2011)
32. Picón-Berjoyo, A., Ruiz-Moreno, C., Castro, I.: A mediating and multigroup analysis of customer loyalty. Eur. Manag. J. **34**, 701–713 (2016)
33. Prajapati, V.: Big Data Analytics with R and Hadoop. Packt Publishing, Birmingham-UK (2013)
34. Ramalho, E., Caleiro, A., Dionfsio, A.: Explaining consumer confidence in Portugal. J. Econ. Psychol. **32**:25–32 (2011)
35. Rigby, R., Stasinopoulos, D.: Generalized additive models for location, scale and shape. J. R. Stat. Soc. C **54**, 507–554 (2005)
36. Rubio, N., Villaseñor, N., Yagüe, M.: Creation of consumer loyalty and trust in the retailer through store brands: the moderating effect of choice of store brand name. J. Retail. Consum. Serv. **34**, 358–368 (2017)
37. Sherman, R.: Business Intelligence Guidebook: From Data Integration to Analytics. Morgan Kaufmann, New York (2014)
38. Stathopoulou, A., Balabanis, G.: The effects of loyalty programs on customer satisfaction, trust, and loyalty toward high- and low-end fashion retailers. J. Bus. Res. **69**, 5801–5808 (2016)

On the Application of Sample Coefficient of Variation for Managing Loan Portfolio Risks

Rahim Mahmoudvand and Teresa A. Oliveira

Abstract Banks and financial institutions are exposed with credit risk, liquidity risk, market risk, operational risk, and others. Credit risk often comes from undue concentration of loan portfolios. Among the diversity of tools available in literature for risk measurement, in our study the Coefficient of Variation (CV) was chosen taking into account that it reveals a very useful characteristic when loan portfolios comparison is desired: CV is unitless—it is independent of the unit of measure associated with the data. We obtain the lower and upper bounds for sample CV and the possibility of using it for measuring the risk concentration in a loan portfolio is investigated. The capital adequacy and the single borrower limit are considered and some theoretical results are obtained. Finally, we implement and illustrate this approach using a real data set.

1 Introduction

Providing loans, banks are exposed with many risks: credit risk, liquidity risk, market risk, operational risk, and others. Usually, the most important risk is credit risk. Credit risk is a critical area in banking and is of concern to a variety of stakeholders: institutions, consumers, and regulators. It has been the subject of considerable research interest in banking and finance communities, and has recently drawn the attention of statistical researchers, see [26]. Often credit risk comes from undue concentration of loan portfolios. Concentration risk in loan portfolios arises from uneven distribution of credit across sectors or providing large loans to individual borrowers, see [20]. The effects of crisis are still persistently reflected in the accompanying slow and weak Gross Domestic Production (GDP) growth

R. Mahmoudvand
Bu-Ali Sina University, Hamedan, Iran

T. A. Oliveira (✉)
CEAUL and Universidade Aberta, Lisbon, Portugal
e-mail: teresa.oliveira@uab.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_6

performance of most economies around the world, see [12]. The economic weakness in a region impacts the job market, the availability of credit, the price of consumer goods and services, wages, and a host of other items. This also might produce a large increase in the demands for loans. Therefore, modeling concentration risk in order to avoid default event is necessary for banks and financial systems.

Among the diversity of tools available in literature for risk measurement, in our study the Coefficient of Variation (CV) was chosen taking into account that it reveals a very useful characteristic when loan portfolios comparison is desired: CV is unitless—it is independent of the unit of measure associated with the data. CV has been often used in diverse areas as a measure of precision of data dispersion, once it allows comparing numerical distributions measured on different scales. In the literature it is easy to find studies in a big range of areas, such as Medicine, Agriculture, Industry, Insurance and Business, in which computation and or comparisons of CVs play a key rule for data behavior interpretation and modeling. We will refer some of these studies in which CV was used in the context of Risk Analysis and Risk Assessment, as well as some previous theoretical achievements under this topic, see [21].

Confidence intervals for CV in normal and log-normal populations were presented by [22] and [23]. Later on, [24] reports a meta-analysis of data for human and animal decision making under risk that uses the coefficient of variation (CV) as a measure of risk sensitivity. Forkman [10] presented results on the statistical inference for the CV in normally distributed data and [5] presented a paper on the coefficient of variation asymptotic distribution in case of non-iid random variables. In the context of Medicine [6] explored some closed-form confidence intervals for functions of the normal mean and standard deviation, and CV was considered. Banik and Kibrie [3] also used confidence intervals for estimating the population coefficient of variation.

More recently, [17] explored confidence intervals for the Coefficients of Variation with Bounded Parameters, and proposed an evaluation for such intervals in terms of coverage probability and of expected length via Monte Carlo simulation. These authors point out that if there is a presence of outliers in the data set, the width of the confidence interval obtained by including the bounds of the parameter space will be less effect from observations with extreme values.

Albatineh et al. [1] presented a simulated study considering the confidence interval estimation for the population coefficient of variation using ranked set sampling. An evaluation of the performance of several confidence interval estimators of the population coefficient of variation was presented, using ranked set sampling compared to simple random sampling. Simulation studies were based on normal, log-normal, skew normal, Gamma, and Weibull distributions with specified population parameters and for several sample sizes.

Hayter [13] presents the construction of two different kinds of confidence intervals for the coefficient of variation of a normal distribution, which were developed using inferences on the non-centrality parameter of a non-central t-distribution. The author considered:

1. the construction of a confidence interval that bounds the reciprocal of the coefficient of variation away from zero;
2. the construction of a confidence interval that provides an upper bound on the absolute value of the reciprocal of the coefficient of variation.

Several applications were considered by Hayter [13], such as financial analyses with Sharpe ratios and the measurement of signal-to-noise ratios, with specific attention being directed towards assessing win-probabilities for comparing two normal treatments.

In this paper we explore CV properties and application in the context of Credit Risk and we organized this paper as follows. In Sect. 2 some bounds for sample CV are obtained. In Sect. 3 one simple theoretical model for credit risk in portfolio of loans is presented. Also measures and inequalities for concentration risk, capital adequacy, and single obligor limit, based on sample CV are proposed. The empirical results are presented in Sect. 4. Section 5 presents a summary and some concluding comments.

2 Properties of Sample Coefficient of Variation

Recall that the coefficient of variation of a distribution with mean μ and variance σ^2 is defined as the ratio σ/μ. In practice, the coefficient of variation and the dispersion will be replaced by its estimator. If $\bar{X}$ and S^2 be mean and variance of sample, we use cv for sample CV and define by $cv = S/\bar{X}$.

2.1 Bounds for cv

We consider four different cases and find bounds for each one.

Case 1
Let $X_1, \ldots, X_n$ be nonnegative random variables. It is obvious that:

$$\left(\sum_{i=1}^{n} X_i\right)^2 \geq \sum_{i=1}^{n} X_i^2. \tag{1}$$

Using the Cauchy- Schwarz inequality, we have

$$\left(\sum_{i=1}^{n} X_i\right)^2 \leq n \sum_{i=1}^{n} X_i^2. \tag{2}$$

Equations (1) and (2) give

$$n\bar{X}^2 \leq \sum_{i=1}^{n} X_i^2 \leq n^2\bar{X}^2. \tag{3}$$

Considering the definition of sample variance and (3), we get

$$0 \leq cv \leq \sqrt{n} \tag{4}$$

The case cv=0 in (4) occurs when all the observations are equal and the case $cv = \sqrt{n}$ occurs when all the observations, except one of them, are zero.

Case 2
Let $X_1, \ldots, X_n$ be nonpositive random variables. In this case, we can similarly get the following bounds for cv:

$$-\sqrt{n} \leq cv \leq 0. \tag{5}$$

Case 3
Let $X_1, \ldots, X_n$ be random variables, for which we have $\sum_{i\neq j} X_i X_j \geq 0$. Then, we get:

$$\left(\sum_{i=1}^{n} X_i\right)^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j. \tag{6}$$

Using the condition $\sum_{i\neq j} X_i X_j \geq 0$ the equality (6) implies the inequality (1) and we obtain the following bound for cv:

$$|cv| \leq \sqrt{n}. \tag{7}$$

Case 4
Let $X_1, \ldots, X_n$ be random variables, in which we have $\sum_{i\neq j} X_i X_j \leq 0$. Then, a similar conclusion provides us the following bound for cv:

$$\sqrt{n} \leq |cv|\,. \tag{8}$$

Table 1 Probability of out of range estimates by cv when samples come from distribution $b(5, p)$

	n				
		10	20	50	100
p	0.01	0.601	0.363	0.093	0.004
	0.05	0.091	0.005	0.000	0.000
	0.10	0.191	0.090	0.009	0.001
	0.15	0.436	0.371	0.307	0.238

It is worth mentioning that under the conditions for cases 1–3, the bounds for sample cv of the mean can be written as below:

$$0 \leq \left| cv(\bar{X}) \right| \leq 1, \tag{9}$$

where we mean $cv(\bar{X}) = cv/\sqrt{n}$.

2.2 *Efficiency of cv*

Considering the coefficient of variation and some main concepts on probability theory the authors recommend [9], where some applications are also explored. Albercher et al. [2] showed that cv is an unbiased and consistence estimator for CV. However, [15] showed that in some cases cv provides a poor estimates for the population CV. As an example, assume that $X_1, \cdots, X_5$ is a random sample from an $N(0, 1)$. Then, they showed that $P(-3 < cv < 3) = 0.50$ whereas the population CV is infinite. Let us consider some examples from the Binomial distribution. Assume that $X_1, \ldots, X_n$ are i.i.d random variables from a $b(m, p)$. It is easy to see that the population CV is equal to $\sqrt{(1-p)/mp}$. Therefore, population CV is larger than 1 when $p < 1/(n+1)$. But, using case 1, we can conclude that $0 \leq cv \leq \sqrt{n}$. This means that there is a positive probability that sample cv produce poor estimates for population CV. Table 1 shows the probability $P(cv < 1)$ when we sample from $b(5, p)$. As it indicated, this probability decreases by sample size, but surprisingly, it decreases by p when p increases from 0.01 to 0.05 and then it increases again.

3 Application of cv for Measuring Concentration Risk in Loan Portfolio

One of the operational works in banks is paying loans to applicants. In approving each loan application, the Bank considers the purpose of the loan, assesses the repayment ability from the applicant's operating cash flows, business feasibility, capability of management, and collateral. However, usually there are some loans which are not paid back to bank. Credit risk is a risk where the borrower may not be able or willing to repay the debt he or she owes to the Bank, see [16].

Although these capitals may be a small percentage of the financial resources of banking institutions, it plays a crucial role in their long-term financing and solvency position and therefore in public credibility. This fact would justify the existence of a capital adequacy regulation in order to avoid bankruptcies and their negative externalities on the financial system although banks may respond to this regulation by increasing their risk exposure. On the other hand, too tight a regulation may lead banks to reduce their credit offer and, as a result, give rise to a fall in productive investment.

Risk measurement and management methods are still at an early stage and quite far from providing exact pictures of a bank's actual risk exposure. This is particularly true for credit risk models, which have been developed and applied only recently, see [26]. Regulators and academics alike have pointed out that the existing methods have to be improved before they can be used to determine a bank's regulatory capital, see [8, 11].

As it was mentioned, many credit risk models have been introduced for loan portfolios and one of the main problems in these models is determination of minimum capital adequacy. Bank management can apply the value at risk (VaR) concept to set capital requirements, see, e.g., [7, 14, 18]. VaR is a risk management tool which allows controlling for the probability of bankruptcy [4]. VaR is an estimate of total exposure to the various market risks interest rate, inflation, exchange rates, share prices, etc., see [19]. There are some methods for measuring VaR. But in this paper VaR is calculated by means of cv.

It should be mentioned that [25] used the method of coefficient of variation to assess the portfolio of loan in one of Iranian banks.

3.1 Capital Adequacy and Concentration Risk in the Loan Portfolios

Denote by n the number of demands for loan and denote by l_i the amount of loan for the applicant i. In a simple model of default risk, each obligor has only two possible end-of-period states, default and non-default. In the event of default, the lender suffers a loss of fixed size l_i for the applicant i. If L_i is a random variable which shows the total of the loss for applicant i, then we have $L_i = l_i B_i$, where B_i is a Bernoulli random variable with parameter p_i (in fact, p_i is the probability of default for loan i). For simplicity, suppose that $p_i = p$ for $i = 1, \ldots, n$ and $L_1, \ldots, L_n$ are independent. Denote by $L = \sum_{i=1}^{n} L_i$ the total loss in the loan portfolios, we can show that the distribution of L is asymptotically normal (see Appendix). Thus, we can get $\text{VaR}_\alpha = E[L] + z_\alpha \sqrt{var(L)}$. For a quantity CL to be the minimum capital adequacy, it must satisfy the following condition:

$$CL \geq \text{VaR}_\alpha \Rightarrow CL \geq E[L] + z_\alpha \sqrt{var(L)}. \tag{10}$$

Using the definition of population cv for L, it is easy to show

$$CL \geq p(1 + z_\alpha CV) \sum_{i=1}^{n} l_i. \tag{11}$$

Note that the right side of (11) is VaR. Therefore, VaR depends on three components: probability of default, level of confidence, and Coefficient of Variation. It means that we can control VaR by controlling these components. We can get the following equivalent formula for checking capital adequacy of bank by definition of sample CV.

$$cv^2 \leq \left(\left(\frac{CL - p \sum_{i=1}^{n} l_i}{z_\alpha \sqrt{p(1-p)} \sum_{i=1}^{n} l_i} \right)^2 - \frac{1}{n} \right) \frac{n^2}{n-1} \tag{12}$$

Let us see another application of cv for concentration risk in the loan of portfolios. Assume bank make loans by total volume $l = \sum_{i=1}^{n} l_i$ to n applicants. Consider three scenarios:

1. Bank decide to pay total volume to one applicant,
2. Bank decide to pay the same loan to all of applicants and
3. Bank decide to pay to all of the applicants, but not equally (maybe even some of them equal zero).

Let us see the behavior of cv for scenarios 1–3. For the first case, $cv = \sqrt{n}$. This corresponds to the maximum possible value of cv, see Eq. (4). It is clear that the first scenario is the riskiest decision for bank which coincides with the maximum cv. Scenario 2 produces $cv = 0$, i.e. the minimum possible value. It is also evident that the second decision has the minimum risk of default for bank. The last one depends on the value of cv and accordingly we can conclude the scenario has a high/low risk. We provide some interesting application of cv in the next two theorems.

Theorem 1 *Let $X_1, \ldots, X_n$ be nonnegative random variables and $\sum_{i=1}^{n} X_i = T$. In addition, suppose m ($< n$) is a natural number and $X_i \leq T/m$. Then, we have:*

$$cv \leq \sqrt{\frac{n(n-m)}{m(n-1)}}$$

and the upper bound is obtained if m observations are equal to T/m and the remaining are zero.

Proof We have

$$cv^2 = \frac{n^2}{n-1} \frac{\sum_{i=1}^{n} X_i^2}{\left(\sum_{i=1}^{n} X_i\right)^2} - \frac{n}{n-1} \leq \frac{n^2}{n-1} \frac{\sum_{i=1}^{n} X_i T/m}{T^2} - \frac{n}{n-1} = \frac{n(n-m)}{m(n-1)}$$

Getting upper bound is straightforward by substitution.

Theorem 2 *Assume* $cv \leq \gamma < \sqrt{n}$, *which* γ *is an arbitrary positive real value. Then*

$$X_i < \bar{X}\sqrt{(n-1)\gamma^2 + n} \quad , \quad i = 1, 2, \ldots, n.$$

Proof It is sufficient to note that

$$X_i^2 < \sum_{i=1}^{n} X_i^2 = \bar{X}^2 \left((n-1)cv^2 + n\right) \leq \bar{X}^2 \left((n-1)\gamma^2 + n\right).$$

Theorem 3 *Assume* $cv \leq \gamma < \sqrt{n}$ *and* $\sum_{i=1}^{n} X_i = T$. *Then we have*

$$X_i \leq \bar{X}\left(1 + \frac{(n-1)\gamma}{\sqrt{n}}\right) \quad , \quad i = 1, 2, \ldots, n.$$

Proof Let $X_{\max}$ be the maximum value satisfying conditions of this theorem. It is easy to see that cv is an increasing function of $\sum_{i=1}^{n} X_i^2$. So, we get $cv = \gamma$ when $\sum_{i=1}^{n} X_i^2$ attains its maximum possible value. Thus, set $cv = \gamma$ to obtain $X_{\max}$. We have:

$$\sum_{i=1}^{n} X_i^2 = \frac{T^2}{n^2}\left[(n-1)\gamma^2 + n\right]$$

$$\Rightarrow X_{\max}^2 + \sum_{i \neq \max} X_i^2 = \frac{T^2}{n^2}\left[(n-1)\gamma^2 + n\right]$$

Using the results of Case 1 in Sect. 2, $\sum_{i \neq \max} X_i^2$ is minimized (and then $X_{\max}$ is maximized) when $X_i = (T - X_{\max})/(n-1)$, $i \neq \max$. So, we have

$$X_{\max}^2 + \frac{(T - X_{\max})^2}{n-1} = \frac{T^2}{n^2}\left[(n-1)\gamma^2 + n\right]$$

$$\Rightarrow nX_{\max}^2 - 2TX_{\max} + \frac{T^2}{n^2}\left[n - (n-1)^2\gamma^2\right] = 0$$

$$\Rightarrow X_{\max} = \frac{T}{n}\left[1 + \frac{(n-1)\gamma}{\sqrt{n}}\right]$$

which completes the proof.

Let us justify the importance of these theorems. Beside we know that the upper bound for cv is $\sqrt{n}$, in many situations it may occur the need to consider a more strict restriction for the risk assumption. For those cases, consideration of γ instead of $\sqrt{n}$ is useful and to attain this aim we need to consider the conditions in Theorems 2 and 3.

4 An Example

One of Iranian Banks provided some information about loan portfolios during the year 2006. This information was the amount and the number of paid loans in during this period for four types of loans: Mozarebeh, Joaleh, Forosh Aghsati, and Gharzolhasaneh. The capitals of Bank for these loans are 10 Billion IRR, 1.5 Billion IRR, 0.50 Billion IRR, and 0.30 Billion IRR, respectively. According to the history of loan portfolio in this bank, we estimated the probability of defaults in four types. The results are as below for the above-mentioned type of loans:

$$p_{\text{moz}} = 0.0618 \qquad p_{\text{joa}} = 0.0153$$
$$p_{\text{foro}} = 0.0011 \qquad p_{\text{ghar}} = 0.0143$$

Calculated cv for four types are 0.97, 1.03, 0.32, and 0.72 respectively. Hence we can say type of Joaleh has a big concentration risk (comparing with others) and Forosh Aghsati has a small risk or concentration risk.

Using Eq. (12) the maximum cv obtained are 0.62, 1.35, 0.86, and 2.52 for these loan types, respectively. Then, according to Theorem 3 all loans must be less than 426,245,884, 229,508,199, 457,119,506, and 70,141,729 in four types of loans, respectively. There were not any loans in portfolios with greater than these amounts.

5 Conclusion

In this paper the boundary feature of sample is discussed and some bounds for different cases are obtained. As an interesting application of this discussion, some problems in credit risk models in loan portfolios were studied. Some inequalities for checking the adequacy of capital in banking systems were obtained. Numerical study showed that this approach can easily evaluate the concentration risk in portfolio of loans.

Appendix

Let us first recall the Lyapanov Theorem.

Theorem 4 *Let $X_1, \ldots, X_n$ be independent random variables with different distribution. If*

(a) $E\left[|X_i - E(X_i)|^3\right] < \infty$

(b) $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left[|X_i - E(X_i)|^3\right]}{\left(\sum_{i=1}^{n} var(X_i)\right)^{3/2}} = 0,$

then, we have

$$\frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E(X_i)}{\sqrt{\sum_{i=1}^{n} var(X_i)}} \to N(0, 1).$$

Proof See, e.g., Feller [9].

In order to apply this Theorem for L, we must check conditions (a) and (b). For the first condition, we have:

$$E\left[|L_i - E(L_i)|^3\right] = E\left[|l_i B_i - l_i p|^3\right] = l_i^3 p(1-p)\left(p^2 + (1-p)^2\right) < \infty.$$

For condition (b), we have:

$$\begin{aligned}
&\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E\left[|L_i - E(L_i)|^3\right]}{\left(\sum_{i=1}^{n} var(L_i)\right)^{3/2}} \\
&= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} l_i^3 p(1-p)\left(p^2 + (1-p)^2\right)}{\left(\sum_{i=1}^{n} l_i^2 p(1-p)\right)^{3/2}} \\
&= \frac{\left(p^2 + (1-p)^2\right)}{\sqrt{p(1-p)}} \lim_{n\to\infty} \frac{\sum_{i=1}^{n} l_i^3}{\left(\sum_{i=1}^{n} l_i^2\right)^{3/2}} \\
&\leq \frac{\left(p^2 + (1-p)^2\right)}{\sqrt{p(1-p)}} \lim_{n\to\infty} \frac{n \max l_i}{(n \min l_i)^{3/2}} = 0.
\end{aligned}$$

Therefore, using Lyapanov theorem, we can conclude that

$$\frac{L - E(L)}{\sqrt{var(L)}} \to N(0, 1).$$

Acknowledgements Funded by FCT-Fundação para a Ciência e a Tecnologia, Portugal, through the project UID/MAT/00006/2013.

References

1. Albatineh, N.A., Golam Kibria, B.M., Wilcox, M.L., Zogheib, B.: Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study. J. Appl. Stat. **41**(4), 733–751 (2014)
2. Albercher, H., Ladoucette, S.A., Teogels, J.L.: Asymptotic of the sample coefficient of variation and the sample dispersion, KU. Leuven USC report 2006–04 (Submitted, 2006)
3. Banik, S., Kibrie, B.M.G.: Estimating the population coefficient of variation by confidence intervals. Commun. Stat. Simul. Comput. **40**(8), 1236–1261 (2011)

4. Broll, U., Wahl, J.E.: Optimum bank equity capital and value at risk. In: Scholz, C., Zentes, J. (eds.) Strategic Management, a European Approach, pp. 69–82. Springer, Wiesbaden (2002)
5. Curto, J.D., Pinto, J.C.: The coefficient of variation asymptotic distribution in case of non-iid random variables. J. Appl. Stat. **36**(1), 21–32 (2009)
6. Donner, A., Zou, G.Y.: Closed-form confidence intervals for functions of the normal mean and standard deviation. Stat. Methods Med. Res. **21**(4) 347–359 (2010)
7. Duffie, D., Pan, J.: An overview of value at risk. J. Deriv. **4**, 7–49 (1997)
8. Erlenmaier, U.: Risk Management in banking, credit risk management and bank closure policies. Ph.D. Thesis, Heidelberg, Germany (2001)
9. Feller, W.: An Introduction to Probability Theory and It's Application. Wiley, New York (1971)
10. Forkman, J.: Statistical inference for the coefficient of variation in normally distributed data, Research Report, Centre of Biostochastics, Swedish University of Agricultural Sciences, 29 (2006)
11. Gordy, M.B.: A comparative anatomy of credit risk models. J. Bank. Financ. **24**, 119–149 (1998)
12. Hacioglu, U., Dincer, H.: Global Financial Crisis and Its Ramifications on Capital Markets. Springer, Berlin (2017)
13. Hayter, A.J.: Confidence bounds on the coefficient of variation of a normal distribution with applications to win-probabilities. J. Stat. Comput. Simul. **85**(18), 3778–3791 (2015)
14. Jorion, P.: Value at Risk: The New Benchmark for Controlling Market Risk. McGraw-Hill, New York (1997)
15. Mahmoudvand, R., Hassani, H., Wilson, R.: Is the sample coefficient of variation a good estimator for the population coefficient of variation? World Appl. Sci. J. **2**(5), 519–522 (2007)
16. Pyle, D.H.: Bank risk management: theory. In: Conference of Risk Management and Regulation in Banking, Jerusalem (1997)
17. Sappakitkamjorn, J., Niwitpong, S.: Confidence intervals for the coefficients of variation with bounded parameters. World Acad. Sci. Eng. Technol. Int. J. Math. Comput. Phys. Electr. Comput. Eng. **7**(9), 1416–1421 (2013)
18. Saunders, A.: Credit Risk Management. Wiley, New York (1999)
19. Schreiber, B.Z., Wiener, Z., Zaken, D.: The implementation of value at risk in placecountry-region Israel's banking system. Bank Isr. Bank. Syst. Rev. **7**, 61–87 (1999)
20. Skridulyte, R., Freitakas, E.: The measurement of concentration risk in loan portfolios. J. Sci. Papers Econ. Sociol. **5**(1), 51–61 (2012)
21. Tian, L.: Inferences on the common coefficient of variation. Stat. Med. **24**, 2213–2220 (2005)
22. Vangel, M.G.: Confidence intervals for a normal coefficient of variation. Am. Stat. **50**(1), 21–26 (1996)
23. Verrill, S.: Confidence bounds for normal and log-normal distribution coefficient of variation. Research Paper, EPL-RP-609, Madison, Wisconsin (2003)
24. Weber, E.U., Shafir, S., Blais, A.: Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. Psychol. Rev. **111**, 430–445 (2004)
25. Zarei, A.: Measuring of credit risk portfolio of bank by method of coefficient variation bound. Soc. Sci. **11**(16) 3908–3913 (2016)
26. Zhang, A.: Statistical methods in credit risk modelling. Ph.D. Thesis in Statistics, The University of Michigan (2009)

Acceptance-Sampling Plans for Reducing the Risk Associated with Chemical Compounds

Fernanda Figueiredo, Adelaide Figueiredo, and M. Ivette Gomes

Abstract In various manufacturing industries it is important to investigate the presence of some chemical or harmful substances in lots of raw material or final products, in order to evaluate if they are in conformity to requirements. In this work we highlight the adequacy of the inflated Pareto distribution to model measurements obtained by chromatography, and we define and evaluate acceptance-sampling plans under this distributional setup for lots of large dimension. Some technical results associated with the construction and evaluation of such sampling plans are provided as well as an algorithm for an easy implementation of the sampling plan that exhibits the best performance.

1 Introduction and Motivation

The presence of some chemical or harmful substances in many consumer or manufacturing products, such as haloanisoles that can cause musty or moldy off-flavors and, consequently, deteriorate the quality of the products, but also fumes, gases, vapors, or even dust, usually present in all workplaces, can endanger the health or safety of persons. To ensure the proper management of the risks associated with these substances, some of them have been prohibited, or at least, subject to a tight inspection, apart from being intensively controlled. See, for instance, [5, 10] and [14]. As it is said in Montgomery [9, p. 629] and Ryan [12, p. 101], acceptance

F. Figueiredo (✉)
Faculty of Economics, University of Porto, and CEAUL, Porto, Portugal
e-mail: otilia@fep.up.pt

A. Figueiredo
Faculty of Economics and LIAAD-INESC TEC, University of Porto, Porto, Portugal
e-mail: adelaide@fep.up.pt

M. I. Gomes
FCUL and CEAUL, University of Lisbon, Lisbon, Portugal
e-mail: ivette.gomes@fc.ul.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_7

sampling (AS) is not a substitute for adequate process monitoring and control, and its purpose is not quality improvement but only decide about the acceptance or rejection of products (lots) based on conformity to requirements. In particular, in several manufacturing industries, sensory analysis together with analytical methods based on chromatography measurements is usually done for identification and quantification of these chemical compounds in lots of raw material, and in some cases along the different phases of the process production.

In this study we consider three data sets with chromatography measurements associated with the concentrations of a chemical substance, performed on samples of items taken from large batches from a process production. Usually the chromatography instruments have small precision to measure with accuracy very low or very high concentrations of such chemical substances, and a common practice is to truncate the results below or above a certain threshold, which happens in the case of our data sets. Apart from being truncated, the type of data under analysis suggests an underlying inflated continuous distribution with a heavy right-tail, which leads us to the consideration of inflated Pareto models (which provide a reasonable nice fit to all data sets under analysis, discussed in Sect. 2.3). Inflated distributional models, in particular containing many zero values, have been commonly used in many areas of application, including agriculture, biology, ecology, environment, fishery and medicine, among others. For details on applications of this type of models, see, for instance, the pioneer works of [1] and [10], and other more recent works, such as [2, 5, 7, 8, 11, 13] and [15], among others. Then, we develop AS plans for lots of large dimension under this distributional setup. Acceptance control charts cannot be applied to our case, and it was not at all our purpose to use total control techniques, which would increase drastically the size of the article.

The most common AS plans are classified as plans for variables or attributes, being the quality characteristics measured on a numerical scale or the inspected items expressed as defective or non-defective, respectively. Both types of AS plans can be single, double, multiple and sequential, and designed so that they produce equivalent results. Given that the different types of AS plans have advantages and disadvantages to each other, when selecting the type of sampling procedure, one must consider the particular problem to solve and the desired efficiency taking into consideration the restrictions associated with its implementation. Here we point out that a variables sampling plan usually requires a sample of smaller size than an attributes sampling plan for the same level of protection, although the sampling/observation unit costs can be higher in the variables sampling plan. It is also important to refer that the main disadvantage of variables AS plans is that the distribution of the quality characteristic under study has to be known (or estimated). Other details about AS plans can be found in [3, 6, 9, 14] and [12].

The previous considerations lead us to consider and evaluate single AS plans for inflated Pareto models. Due to the long time needed to perform chromatography analysis, double or sequential sampling plans do not seem adequate. Moreover, the simple sampling plans are easier to implement than the others, and as we will see later in Sect. 3.3, lead to quite satisfactory results. In a previous work (see [5]), using a real data set with measurements obtained from chromatography analysis, similar

to the ones here considered, we compared the performance of specific and complex variables AS plans, using the bootstrap methodology (see [4] for details) to construct replicates of the lots combined with Monte Carlo simulations. Although the results obtained in the aforementioned study were quite satisfactory, being possible to design and evaluate the performance of AS plans under a specific distributional setup, we still think sensible to define other AS plans.

The paper is organized as follows. After presenting a small introduction and the motivation for the study, Sect. 2 presents the problem under study and the available data sets. It is also provided some information about the inflated Pareto distribution, including the maximum likelihood estimation of its parameters, and the p-values obtained with the chi-square goodness of fit test, which lead us to conclude that this distributional model provides a good fitting for the data. In Sect. 3, two variables AS plans are developed for lots of items, assuming that the quality characteristic is a *random variable* (rv) with inflated Pareto distribution, and their performance when applied to a real data set is analyzed. Some distributional results about the Pareto and the inflated Pareto distributions used in the construction and evaluation of the previous sampling plans are also presented, as well as an algorithm for an easy implementation of the AS plan that exhibits the best performance. The paper ends with some conclusions in Sect. 4.

2 Modeling Chromatography Measurements with an Inflated Pareto Distribution

Next we will present the real problem under study, the available data of chromatography measurements associated with the concentrations of a chemical substance, and some results of goodness of fit tests that lead us to the consideration of an inflated Pareto distribution to model such type of data.

2.1 A Brief Description of the Problem and the Data Sets

Consider a company that wants to design and evaluate AS plans for lots of raw material. Let X be a continuous rv with unknown distribution, associated with the measurements of the quality characteristic, in our case the concentration of a chemical substance in an item of raw material. Suppose that the company establishes that all the items must have values of $X \leq 4$, and that, due to the sensitivity and precision of the measurement instrument, a chromatograph, all the observed values will be greater than or equal to 0.5. Assume that the lots for inspection are of very large size, N, say thousands of items.

To implement and compare different AS plans, in order to choose the one that best fits the objectives of the company, suppose we only have access to a historical data set of measurements of the concentration of the chemical substance. Here, for

Table 1 Measurements of the concentration of the chemical substance in three different types of raw material, given in data sets A, B, and C of size 1600, 752, and 848 items respectively

Classes of measurements	Set A number of items (%)	Set B number of items (%)	Set C number of items (%)
0.5	908 (56.7%)	403 (53.6%)	505 (59.6%)
]0.5,1.0]	357 (22.3%)	157 (20.9%)	200 (23.6%)
]1.0,2.0]	187 (11.7%)	91 (12.1%)	96 (11.3%)
]2.0,3.0]	53 (3.3%)	31 (4.1%)	22 (2.6%)
]3.0,4.0]	16 (1.0%)	12 (1.6%)	4 (0.5%)
]4.0,5.0]	20 (1.3%)	13 (1.7%)	7 (0.8%)
]5.0,7.5]	17 (1.1%)	12 (1.6%)	5 (0.6%)
]7.5,10.0]	16 (1.0%)	10 (1.3%)	6 (0.7%)
]10.0,20.0]	10 (0.6%)	8 (1.1%)	2 (0.2%)
>20.0	16 (1.0%)	15 (2.0%)	1 (0.1%)

illustration of the type of data under study, we consider three data sets, A, B, and C, presented in Table 1, corresponding to the measurements of the concentration of the chemical substance in three different types of raw material. Then, when analyzing the performance of the proposed sampling plans, developed in Sect. 3, we only consider set A, the larger sample.

From Table 1 we observe that 95%, 92.3%, and 97.6% of the items type A, B, and C, respectively, have a concentration level of this chemical substance smaller than or equal to 4.0, and therefore, approximately 5%, 7.7%, and 2.4% of these items do not satisfy the requirements of the company. It is also important to refer that 56.7%, 53.6%, and 59.6% of the measurements in set A, B, and C respectively, are equal to 0.5, which means either the absence of the chemical substance or wrong quantification due to the low sensitivity of the equipment. Finally all sets of data reveal an underlying distribution with a heavy right tail. This preliminary analysis of the data samples leads us to the consideration of inflated Pareto models to describe such type of data.

2.2 Inflated Pareto Model and ML Estimation

Let X be a mixed rv from an inflated Pareto distribution, with *cumulative distribution function* (cdf) given by

$$F(x;\, p, \xi, \delta) = p + (1-p)(1-(x/\delta)^{-1/\xi}), \quad x \geq \delta, \tag{1}$$

and *probability density function* (pdf) given by

$$f(x;\, p, \xi, \delta) = \begin{cases} (1-p)(\xi\delta)^{-1}(x/\delta)^{-1/\xi-1}, & x > \delta, \\ p, & x = \delta, \end{cases} \tag{2}$$

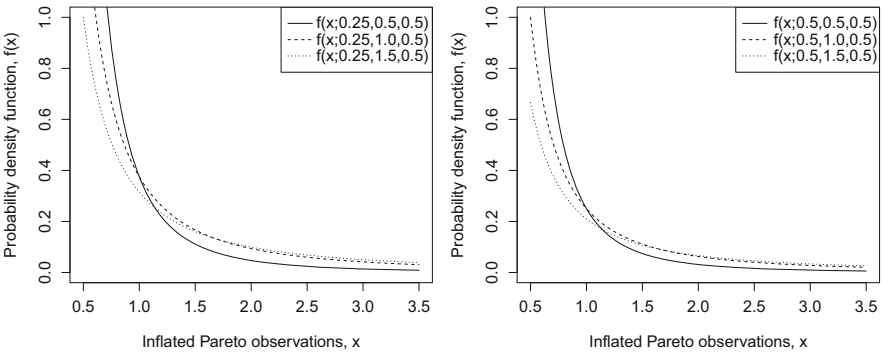


Fig. 1 Probability density function of inflated Pareto distributions with parameters $p = 0.25$ (left), 0.5 (right), $\delta = 0.5$ and $\xi = 0.5, 1, 1.5$

where δ and ξ are, respectively, the scale and the shape parameter, both positive, and the parameter p is the probability associated with the singular distribution at $x = \delta$. Note that if p and δ are fixed, as larger ξ is, larger is the weight of the right-tail, and higher the probability of getting high values. Look, for instance, for the shape of the inflated Pareto pdf represented in Fig. 1.

To estimate the parameters of an inflated Pareto distribution, with $f(x; p, \xi, \delta)$ given in (2), consider a random sample $(X_1, X_2, \ldots, X_n)$ of size n. The *maximum likelihood* (ML) estimates of the parameters p, ξ, and δ maximize the logarithm of the likelihood function, defined by

$$\ln L(p, \xi, \delta) = n_1 \ln p + n_2 \ln(1-p) - n_2 \ln(\xi\delta) - (1/\xi + 1) \sum_{i=1}^{n_2} \ln(x_i/\delta), \qquad (3)$$

where n_1 and n_2 denote the number of observations, respectively, equal and greater to δ in the overall sample of size n. Thus, these estimates are the solution of the system of likelihood equations

$$\begin{cases} \dfrac{\partial \ln L(p,\xi,\delta)}{\partial p} = \dfrac{n_1}{p} - \dfrac{n_2}{1-p} = \dfrac{n_1 - np}{p(1-p)} = 0 \\ \dfrac{\partial \ln L(p,\xi,\delta)}{\partial \xi} = \xi n_2 - \displaystyle\sum_{i=1}^{n_2} \ln(x_i/\delta) = 0 \\ \dfrac{\partial \ln L(p,\xi,\delta)}{\partial \delta} = \dfrac{n_2}{\xi\delta} > 0 \end{cases} \iff \begin{cases} \widehat{p} = \dfrac{n_1}{n} \\ \widehat{\xi} = \dfrac{1}{n_2} \displaystyle\sum_{i=1}^{n_2} \ln(x_i/\delta) \\ \widehat{\delta} = \min x_i, \end{cases}$$

given that these values maximize the logarithm of the likelihood function, and consequently, the likelihood function.

2.3 Inflated Pareto Models Fitted to the Historical Data Sets

According to the historical data, we have fitted inflated Pareto models to the data, with cdf given in (1). We have considered $\delta = 0.5$, because the equipment has no precision to measure with accuracy values below this threshold, and for this reason all the observations smaller or equal to 0.5 were registered as 0.5. We have further estimated the other parameters, ξ and p, by ML. Thus, we proceeded as follows: we split the sample of size n into observations equal to $\delta = 0.5$ (subsample of size n_1), and greater than δ (subsample of size n_2);

- To estimate p, we considered the proportion of observations equal to $\delta = 0.5$ in the overall sample, i.e., $\hat{p} = n_1/n$; we have got $\hat{p} = 908/1600 = 0.5675$ for set A, $\hat{p} = 403/752 = 0.5359$ for set B, and $\hat{p} = 505/848 = 0.5955$ for set C.
- To estimate ξ, we considered $\hat{\xi} = \sum_{i=1}^{n_2} \ln(x_i/\delta)/n_2 = \overline{y}$, with $y_i = \ln(x_i/\delta)$; we have got $\hat{\xi} = 0.9288$ for set A, $\hat{\xi} = 1.1035$ for set B, and $\hat{\xi} = 0.7507$ for set C.

In Fig. 2, we present the histograms associated with the measurements of data sets A, B, and C, and the estimated pdf curves corresponding to the inflated Pareto distributions fitted to the data. As we can observe, these models with cdf given in (1) seem to be adequate to fit such type of measurements. The application of the chi-square goodness of fit test led us to a similar conclusion, taking into consideration the obtained p-values, equal to $P(\chi^2_7 > 13.18) = 0.0678$ for set A, $P(\chi^2_7 > 5.26) = 0.6271$ for set B, and $P(\chi^2_5 > 7.03) = 0.2184$ for set C, where χ^2_k denotes an rv with a chi-square distribution with k degrees of freedom. According to the fitted model, the estimated probability of occurrence of a value higher than 4.0 is 4.61%, 7.05%, and 2.53% for items of raw material of type A, B, and C, respectively.

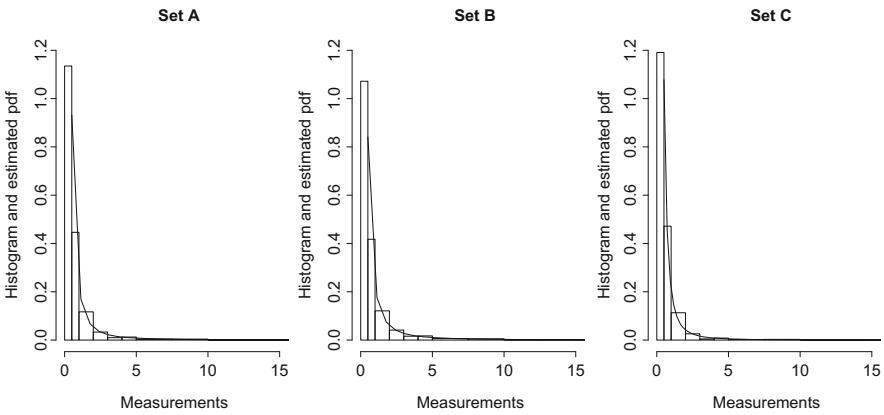


Fig. 2 Histograms and estimated pdf of the inflated Pareto distributions fitted to the A, B, and C data sets

3 Acceptance Sampling Plans for Inflated Pareto Data

Suppose large lots of items, of size N, coming from an inflated Pareto process X, with cdf given in (1), and assume that we have only one *upper specification limit* USL, for the quality characteristic X. After a prior estimation of p and δ, on the basis of a historical data set, assume these parameters fixed and known. In case of having only one *lower specification limit*, LSL, or instead, two specification limits, USL and LSL, the development of AS plans is similar but with more computations.

3.1 Some Preliminaries

The most common AS plans are outlined for controlling the fraction of defective items, in our case θ, given by

$$\theta = P(X > \text{USL}) = (1 - p)(\delta/\text{USL})^{1/\xi}, \tag{4}$$

or a process parameter associated with the production of defectives, such as the parameter ξ, that can be written as function of θ and USL, through the expression

$$\xi = \ln(\delta/\text{USL})/\ln(\theta/(1 - p)). \tag{5}$$

Note that for δ and p fixed, θ will be small if ξ is small.

To develop AS plans, we must consider consistent estimators for θ or ξ, with known distribution. Noting that the ML estimate of ξ is obtained with the observations of the sample greater than δ, the Pareto distribution has a crucial role in the development and implementation of such sampling plans.

First, consider a random sample of size n taken from the lot, $(X_1, \ldots, X_n)$, and then, consider the subsample of the observations greater than δ, say $(X_1, \ldots, X_{n_2})$, of size n_2. The rv's X_i, $1 \leq i \leq n_2$ of this subsample follow a Pareto distribution with cdf given by $F(x; \xi, \delta) = 1 - (x/\delta)^{-1/\xi}$, $x > \delta$. Note that the size n_2 of the subsample is an rv with a binomial distribution, $B(n, 1 - p)$, being the mean size of this subsample, $E(n_2)$, given by $n \times (1 - p)$.

Based on the subsample $(Y_i = \ln(X_i/\delta)\,,\ 1 \leq i \leq n_2)$, where the rv's Y_i, $1 \leq i \leq n_2$ are distributed as an exponential distribution with cdf $F_Y(y) = 1 - e^{-y/\xi}$, $y > 0$, we will consider the following consistent estimators of ξ to develop variables AS plans: the sample mean statistic (the ML estimator of ξ),

$$\overline{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i, \tag{6}$$

with $2n_2\overline{Y}/\xi$ following a $\chi^2_{2n_2}$ distribution, and the statistic T, defined by

$$T = \frac{Y_{n_2:n_2}}{\log n_2 + \gamma}, \tag{7}$$

where $\gamma = 0.5772$ is the Euler constant and $Y_{n_2:n_2}$ denotes the maximum of the sample. This statistic has a cdf given by

$$P(T \leq t) = P\left(Y_{n_2:n_2} \leq t(\log n_2 + \gamma)\right) = (1 - \exp(-t(\log n_2 + \gamma)/\xi))^{n_2}, \; t > 0. \tag{8}$$

The motivation for considering this statistic is the following: being the lots of very large size (thousand of items), the size of the samples taken for inspection is also large (for a standard level of control, the sampling rate is 1 per 10,000 items), and consequently, the value of the sample mean, $\overline{y}$, can be very small even when we have several items above the upper specification limit; the statistic T is also a consistent estimator, related to the maximum of the sample, and therefore, we have decided to investigate if it is a better option to develop AS plans.

3.2 Design of Variables AS Plans

Designing a single sampling plan usually consists of determining the values of the sample size and the acceptance constant which allow us to obtain a sampling plan with a specified performance, in general, predetermined producer's and consumer's risks, or a specific *operating characteristic* (OC) curve. Sometimes, imposed by operational and cost constraints, as happens in our case study, it is necessary to design sampling plans for a fixed sample size and in order to obtain a fixed producer's risk. In this case we only have to determine the acceptance constant.

Let $P(A|\theta)$ denote the probability of acceptance of a lot with a fraction defective θ. The *acceptable quality level* (AQL) is the poorest level of quality for the supplier's process that the producer would consider to be acceptable as a process average. The producer's risk, α, is defined by

$$\alpha = P(\overline{A}|\theta = \text{AQL}). \tag{9}$$

Thus, AS plans designed for a fixed sample size n and to obtain a fixed α-risk for a given quality level AQL are designed such that

$$P(A|\theta = \text{AQL}) = P\left(A|\xi = \frac{\ln(\delta/\text{USL})}{\ln(\text{AQL}/(1-p))}\right) = 1 - \alpha. \tag{10}$$

In our case, as referred before, for a fixed sample size n, the number n_2 of observations in the sample that will be greater than δ is an rv. Thus, the acceptance value of the plan, k, is not constant, but depends on the observed value of n_2. To analyze the performance of the following sampling plans in Sect. 3.3, we will consider, for a fixed n and p (remember that after a prior estimation of p and δ, on the basis of an available historical data set, these parameters are assumed fixed and known), some possible values of n_2 around its mean value, and we determine the associated acceptance constant value $k \equiv k(n_2)$.

The AS plans we are going to consider and evaluate are the following ones:

Plan I Accept the lot if $\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \leq k$. From (6), the parameter $k \equiv k(n_2)$ of this plan must satisfy the condition

$$P\left(\overline{Y} \leq k|\xi = \frac{\ln(\delta/\text{USL})}{\ln(\text{AQL}/(1-p))}\right) = 1-\alpha,$$

and is given by

$$k = \frac{1}{2n_2}\frac{\ln(\delta/\text{USL})}{\ln(\text{AQL}/(1-p))}F^{-1}_{\chi^2_{2n_2}}(1-\alpha), \tag{11}$$

where $F^{-1}_{\chi^2_{2n_2}}$ denotes the inverse of the cdf of a $\chi^2_{2n_2}$ distribution.

Plan II Accept the lot if $T = Y_{n_2:n_2}/(\log n_2 + \gamma) \leq k$, $\gamma \simeq 0.5772$, with $Y_{n_2:n_2}$ denoting the maximum of the sample. From (7) and (8), the parameter $k \equiv k(n_2)$ of this plan must satisfy the condition

$$P(T \leq k|\xi) = \left(1 - \exp\left(-\frac{k(\ln n_2 + \gamma)\ln(\text{AQL}/(1-p))}{\ln(\delta/\text{USL})}\right)\right)^{n_2} = 1-\alpha,$$

and is given by

$$k = -\frac{\ln(\delta/\text{USL})\ln\left(1-(1-\alpha)^{1/n_2}\right)}{(\ln n_2 + \gamma)\ln\left(\text{AQL}/(1-p)\right)}. \tag{12}$$

3.3 *Performance of the Previous Sampling Plans*

In the context of the problem in Sect. 2.1, we assume lots of very large size N, and an upper specification limit USL $= 4$ for the items. Thus, items associated with measurements of concentration of the chemical substance above 4 are considered defective (or nonconforming). To illustrate the performance of the previous sampling plans I and II, we only consider the larger historical data set A of the measurements of the chemical substance in items of one type of raw material. To

Table 2 Acceptance constant $k \equiv k(n_2)$ for some possible values of n_2 according to the sample size n of the sampling plans implemented to obtain an α-risk = 5% for a given AQL (2.5% or 1%)

	AQL = 1%	AQL = 1%	AQL = 2.5%	AQL = 2.5%
(n, n_2)	Plan I	Plan II	Plan I	Plan II
(50, 10)	0.86695	1.14383	1.14561	1.51149
(50, 15)	0.80545	1.07072	1.06434	1.41488
(50, 20)	0.76949	1.02892	1.01682	1.35964
(50, 25)	0.74527	1.00087	0.98482	1.32257
(50, 30)	0.72757	0.98029	0.96144	1.29538
(100, 35)	0.71392	0.96430	0.94340	1.27426
(100, 40)	0.70299	0.95139	0.92895	1.25720
(100, 45)	0.69398	0.94066	0.91704	1.24301
(100, 50)	0.68639	0.93153	0.90701	1.23095
(100, 55)	0.67988	0.92364	0.89842	1.22052

determine AS plans for the fraction of defective items in the batches, we assume δ known, equal to 0.5, and p fixed, equal to the obtained ML estimate, $\widehat{p} = 0.5675$, computed with the available historical data set that mimics the quality of the process production. We consider that the deterioration of the quality of the lots of items is essentially due to changes in the parameter ξ of the distribution. As we referred before, for fixed p, the probability of occurring very high values of X increases with ξ.

The previous sampling plans I and II, based on the statistics $\overline{Y}$ and T defined in (6) and (7), respectively, were designed for some fixed possible values of n_2 around its mean value $E(n_2) = n \times (1 - p)$, taking into account a sample of size $n = 50, 100$ and the estimated value of p, equal to 0.5675. The acceptance constants $k \equiv k(n_2)$, presented in Table 2, were determined through the Eqs. (11) and (12), in order to obtain a predetermined α-risk for a given AQL level. We note that values of n_2 above or below the ones presented in Table 2 occur with a probability near zero, i.e., $P(10 \leq n_2 \leq 30 | n = 50) \simeq 100\%$ and $P(35 \leq n_2 \leq 55 | n = 100) \simeq 100\%$.

To compare the performance of these sampling plans we analyzed the OC curve, i.e., the curve fitted to the points $(\theta, P(A|\theta))$, for $\theta = 0, 1/N, \ldots, 1$. This curve shows the discriminatory power of the sampling plan. Comparing the OC curves, the most severe plan is the one associated with the OC curve that decreases faster.

For illustration, we represented in Figs. 3 and 4, the OC curves of the plans I and II designed to have an α-risk = 5% when AQL = 2.5% and AQL = 1%. When analyzing the OC curves in these figures we observe that plan I, for the desired levels of protection, is significantly better than plan II, and its performance increases with the value of n_2 and n, as expected. As n increases, the OC curves become closer for the different possible values of n_2. For smaller AQL values we get similar conclusions, with improvements in the performance of the sampling plans.

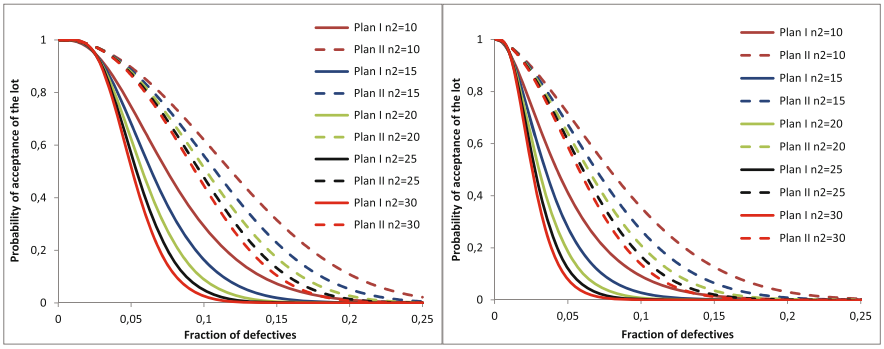


Fig. 3 OC curves of the sampling plans I (solid line) and II (dashed line) designed for $n = 50$ and α-risk = 5% for AQL = 2.5% (left) and AQL = 1% (right), being n_2 the number of observations greater than δ in the sample

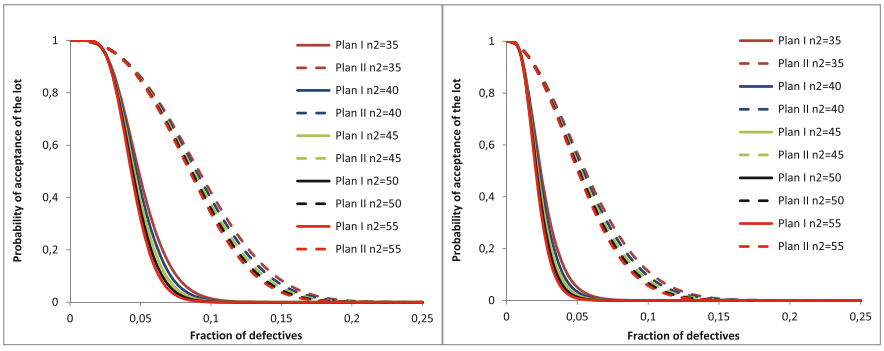


Fig. 4 OC curves of the sampling plans I (solid line) and II (dashed line) designed for $n = 100$ and α-risk = 5% for AQL = 2.5% (left) and AQL = 1% (right), being n_2 the number of observations greater than δ in the sample

3.4 Algorithm for the Implementation of Plan I, for Inflated Pareto Data

To promote the use of the acceptance-sampling plan I developed for inflated Pareto data, we provide the following algorithm for an easy implementation of the sampling plan, designed to obtain a fixed α-risk for a given quality level AQL and a fixed sample size n.

Algorithm:

1. Consider a prior sample that mimics the quality of the process, and estimate the parameters of the model with cdf given in (1);
2. Fix the acceptable quality level AQL, the α-risk, and the sample size n;
3. Sample from the process until obtaining n observations; then, determine the number of observations greater than δ in the sample, say n_2.

4. Determine the acceptance constant $k \equiv k(n_2)$, using Eq. (11);
5. Compute the values $Y_i = \ln(X_i/\delta),\ 1 \leq i \leq n_2$;
6. Compute the control statistic $\overline{Y}$ using Eq. (6);
7. Take the decision: if $\overline{Y} \leq k(n_2)$, accept the lot; otherwise, reject the lot.

4 Conclusions

In this paper we refer the importance of inflated models in applications, and in particular, we present some motivation for the use of the inflated Pareto distribution, that is a very manageable mixture-type model, with simple distributional properties. We derive variables AS plans for inflated Pareto data, assuming that the deterioration of the quality of the lots is essentially due to changes in the shape parameter of the distribution, and considering the other parameters of the model fixed and known. Simple analytical expressions are provided to determine the acceptance constant for a fixed sample size that allow to achieve the desired performance in terms of the producer's risk for a given AQL level. We illustrate the performance of such sampling plans in terms of the obtained OC curves. To promote and facilitate the use of the sampling plan with the best performance by practitioners, an algorithm for its implementation is provided. Finally, the implementation of control charts to monitor inflated Pareto data is also of great interest, as well as the design of AS plans for inflated Pareto models in the case of all the parameters unknown. These topics will be addressed in a future work.

Acknowledgements This research was partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology), through the projects UID/MAT/0006/2013 (CEA/UL) and UID/EEA/50014/2013. We also acknowledge the valuable suggestions from a referee.

References

1. Aitchison, J.: On the distribution of a positive random variable having a discrete probability mass at the origin. J. Am. Stat. Assoc. **50**, 901–908 (1955)
2. Baksh, M.F., Bohning, D., Lerdsuwansri, R.: An extension of an over-dispersion test for count data. Comput. Stat. Data Anal. **55**, 466–474 (2011)
3. Carolino, E., Barão, I.: Robust methods in acceptance sampling. REVSTAT **11**(1), 67–82 (2013)
4. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman and Hall, London (1993)
5. Figueiredo, F., Figueiredo, A., Gomes, M.I.: Comparison of sampling plans by variables using the bootstrap and Monte Carlo simulations. AIP Conf. Proc. **1618**, 535–538 (2014)
6. Gomes, M.I.: Acceptance sampling. In: Lovric, M. (ed.), International Encyclopedia of Statistical Science, Part 1, pp. 5–7. Springer, Berlin (2011)
7. Lachenbruch, P.A.: Comparison of two-part models with competitors. Stat. Med. **20**, 1215–1234 (2001)

8. Loganathan, A., Shalini, B.K.: Determination of single sampling plans by attributes under the conditions of zero-inflated Poisson distribution. Commun. Stat. Simul. Comput. **43**, 538–548 (2014)
9. Montgomery, D.C.: Introduction to Statistical Quality Control: A Modern Introduction, 6th edn. Wiley, London (2009)
10. Owen, W.J., DeRouen, T.A.: Estimation of the mean for lognormal data containing zeros and left-censored values, with applications to the measurement of worker exposure to air contaminants. Biometrics **36**(4), 707–719 (1980)
11. Rakitzis, A., Maravelakis, P., Castagliola, P.: CUSUM control charts for the monitoring of zero-inflated binomial processes. Qual. Reliab. Eng. Int. (2015). https://doi.org/10.1002/qre.1764
12. Ryan, T.P.: Statistical Methods for Quality Improvement, 2nd edn. Wiley, London (2000)
13. Sileshi, G.: The excess-zero problem in soil animal count data and choice of appropriate models for statistical inference. Pedobiologia **52**, 1–17 (2008)
14. Whitaker, T.B., Doko, M.B., Maestroni, B.M., Ogunbanwo, B.F.: Evaluating the performance of sampling plans to detect Fumonisin B_1 in maize lots marketed in Nigeria. J. AOAC Int. **90**(4), 1050–1059 (2007)
15. Zidan, M., Wang, J.-C., Niewiadomska-Bugaj, M.: Comparison of k independent, zero-heavy lognormal distributions. Can. J. Stat. **39**(4), 690–702 (2011)

Risk of Return Levels for Spatial Extreme Events

Luísa Pereira and Cecília Fonseca

Abstract The impact of environmental extreme events, ranging from disturbances in ecosystems to economic impacts on society and losses of life, motivated the study of extremes of random fields.

In this paper the main question of interest is about risk: if occurs one exceedance of a high level in a given location, $\mathbf{x} \in \mathbb{R}^2$, and the maximum over a neighborhood of $\mathbf{x}$ does not exceed the level then, what will be the probability that an exceedance occurs in another location? We define a coefficient as a measure of this probability which allows us to evaluate the risk of return levels. This coefficient is independent of the univariate marginal distribution of the random field and can be related to well-known dependence coefficients, which will provide immediate estimators. The performance of the proposed estimator is analyzed with a max-stable maxima of moving maxima random field. We illustrate the results with an application to annual maxima temperatures over Texas.

1 Introduction

Extreme Value Theory is the branch of probability and statistics aimed at characterizing the behavior of extremes in series of observations. It has its beginnings in the early to middle part of the last century.

Although there are well-developed approaches to model univariate and multivariate extremal processes, in recent years, there have been significant advances in

L. Pereira (✉)
Universidade da Beira Interior, Centro de Matemática e Aplicações (CMA-UBI), Covilhã, Portugal
e-mail: lpereira@ubi.pt

C. Fonseca
Instituto Politécnico da Guarda, Centro de Matemática e Aplicações (CMA-UBI), Guarda, Portugal
e-mail: cfonseca@ipg.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_8

the modeling of extreme events in the spatial and space-time domains. Perhaps, one reason for this is the realization among stakeholders (climate scientists, environmental engineers, insurance companies, etc.) that in an evolving climate there may be changes in the sizes and frequencies of rare events, rather than in the averages, which can lead to the most devastating losses of life, damage to infrastructure, and so forth.

There are many geostatistical tools and methods for modeling and interpreting spatial attributes. However, their basis in Gaussian distributions makes them unsuitable for extremal modeling, because the Gaussian density function has a light tail and therefore can badly underestimate probabilities associated with extreme events. So, it is natural to ask what distributions can arise as limits for maxima of independent variables?

Fisher and Tippett [8] show that the suitably rescaled maxima of independent random variables, and a wide variety of random processes, follow the generalized extreme value distribution (GEV) defined as

$$H(y) = \begin{cases} \exp\left[-(1+\xi(y-\mu)/\tau_+)^{-1/\xi}\right], & \xi \neq 0 \\ \exp\left[-\exp\left(-(y-\mu)/\tau\right)\right] & , \xi = 0 \end{cases}, \tag{1}$$

where $a_+ = \max(a, 0)$, $\mu \in \mathbb{R}$ is the location parameter, $\tau > 0$ is the scale parameter, and $\xi \in \mathbb{R}$ is the shape parameter which determines the weight of the upper tail of the density, with increasing ξ corresponding to higher probabilities of large events.

The Eq. (1) satisfies the max-stability property, that is, for any $m \in \mathbb{N}$ there exist real numbers $a_m > 0$ and b_m such that

$$H^m(a_m x + b_m) = H(x), \quad x \in \mathbb{R}.$$

This necessary condition for a limiting distribution for maxima is satisfied only by GEV, giving it strong mathematical support as a suitable distribution for fitting to maxima of scalar random variables.

The subfamilies of the GEV distribution defined by $\xi = 0$, $\xi > 0$ and $\xi < 0$ correspond, respectively, to the Gumbel, Fréchet, and Weibull distributions, also known as type I, II, and III extreme value distributions.

Next, we will introduce the mathematical tools to modeling spatial extreme events.

In what follows $Z(\mathbf{x})$ is a random field defined over a discrete subset of $\mathbb{R}^2$ (Fig. 1, on right) or over a regular grid identified with Z^2 (Fig. 1, on left).

Figure 2 illustrates a realization of a random field of maxima precipitation in different locations of Bourgogne. The data are from Naveau et al. [13].

The natural statistical models for spatial extremes are the max-stable random fields. They are the natural analogues of the GEV distribution for modeling extreme events in the spatial and space-time domains. It relies on extensions of GEV distribution that satisfy an appropriate generalization of the max-stability property.

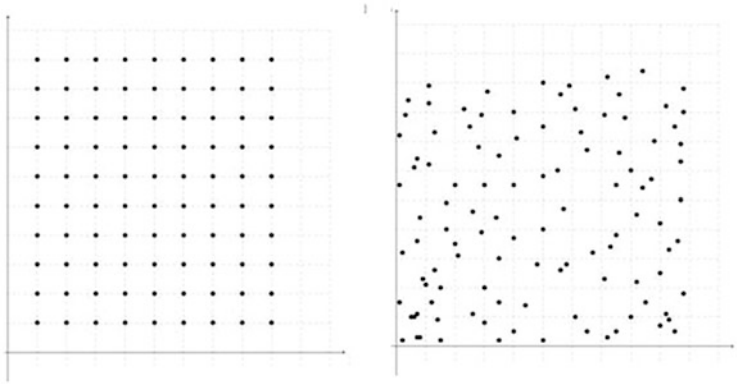


Fig. 1 Regular grid (on left) and random pattern (on right)

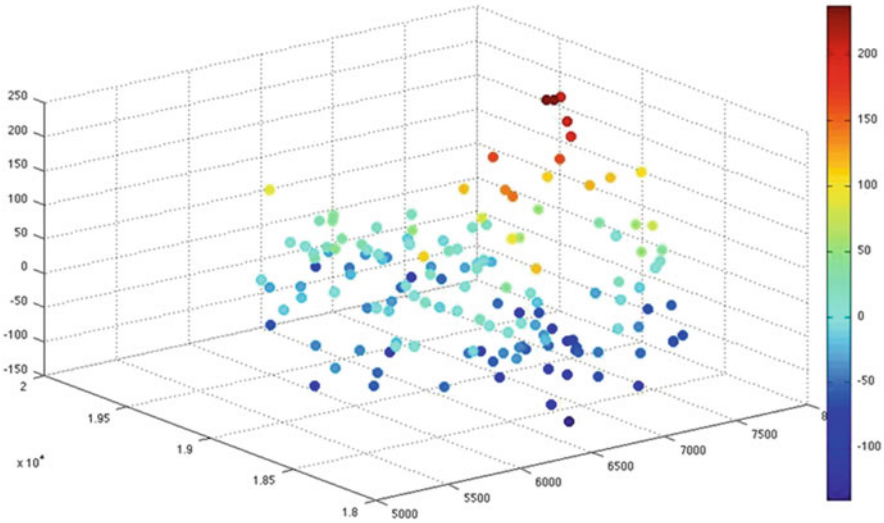


Fig. 2 51-year maxima of daily precipitation in Bourgogne of France

Briefly, a max-stable random field $\{Z(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^d\}$, $d \in \mathbb{N}$, is the limit process of maxima of independent and identically distributed random fields $Y^{(j)}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$, $j = 1, 2, \ldots, n$. Namely, for suitable $\{a_n(\mathbf{x}) > 0\}_{n\geq 1}$ and $\{b_n(\mathbf{x})\}_{n\geq 1}$ sequences of real constants,

$$Z(\mathbf{x}) = \lim_{n\to+\infty} \frac{\bigvee_{j=1}^{n} Y^{(j)}(\mathbf{x}) - b_n(\mathbf{x})}{a_n(\mathbf{x})}, \qquad \mathbf{x} \in \mathbb{R}^d,$$

provided the limit exists, where $\bigvee_{i=1}^{k} a_i$ denotes the maximum of $\{a_1, a_2, \ldots, a_k\}$.

For each choice of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, the distribution of $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_k))$ is a multivariate extreme value distribution $G_{\mathbf{x}_1,\ldots,\mathbf{x}_k}$, where its margins are univariate extreme value distribution functions themselves.

Max-stable random fields have been widely applied to real data in environmental, atmospheric, and geological sciences [2–4, 18, 19].

Here $\mathbf{Z} = \{Z(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2\}$ denotes a max-stable random field over $\mathbb{R}^2$. Since one can transform one max-stable distribution into another one by a monotone transformation we assume, without loss of generality, that the margins of $\{Z(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^2\}$ have a unit Fréchet distribution, $F(x) = \exp(-x^{-1})$, $x > 0$ (Resnick [14]).

The tail dependence function of $G_{\mathbf{x}_1,\ldots,\mathbf{x}_k}$,

$$l_{\mathbf{x}_1,\ldots,\mathbf{x}_k}(w_1, \ldots, w_k) = \lim_{u \downarrow 0} \frac{1 - G_{\mathbf{x}_1,\ldots,\mathbf{x}_k}\left(G_{\mathbf{x}_1}^{-1}(1 - uw_1), \ldots, G_{\mathbf{x}_k}^{-1}(1 - uw_k)\right)}{u},$$

$(w_1, \ldots, w_k) \in \mathbb{R}_+^k$, characterizes fully the dependence among its marginals distributions $G_{\mathbf{x}_j}$, $j = 1, \ldots, k$ (Resnick [14], Beirlant et al. [1]), but it cannot be easily inferred from data. So, several dependence coefficients have been considered in order to resume the dependence among the marginals of $G_{\mathbf{x}_1,\ldots,\mathbf{x}_k}$: extremal coefficients, tail dependence coefficients, and madogram (Li [12], Smith [18], Schlather and Tawn [15], Cooley et al. [5], Naveau et al. [13], Fonseca et al. [9, 10], Ferreira and Ferreira [6, 7], among others).

The scalar $l_{\mathbf{x}_1,\ldots,\mathbf{x}_k}(1, \ldots, 1)$, denoted by $\epsilon_{\{\mathbf{x}_1,\ldots,\mathbf{x}_k\}}$, is the extremal coefficient defined in Schlater and Tawn [15], which summarizes the extremal dependence between the variables of the max-stable random field $\mathbf{Z}$ indexed in the region $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. This coefficient is equal to k for the independent case, and to 1 for the full dependence case. Otherwise, its value varies between 1 and k depending on the degree of dependence. Its value can be thought as the number of effectively independent locations among the k under consideration.

Another way to assess the amount of extremal dependence is through the concept of tail dependence. Multivariate tail dependence coefficients have been used to describe the amount of dependence in the orthant tail of a multivariate distribution (Schmid and Schmidt [16], Li [12], Ferreira and Ferreira [6], among others). Recently, the most referred in literature is the upper tail dependence coefficient of Li [12], defined as

$$\lambda_{A,B} = \lim_{u \uparrow 1} P\left(\bigcap_{\mathbf{x} \in A} \{F(Z(\mathbf{x})) > u\} \middle| \bigcap_{\mathbf{y} \in B} \{F(Z(\mathbf{y})) > u\} \right),$$

provided the limit exists, where A and B are discrete subsets of $\mathbb{R}^2$. This coefficient is a generalization of the upper tail dependence coefficient of Sibuya [17] corresponding to $A = \{\mathbf{x}\}$ and $B = \{\mathbf{y}\}$,

$$\lambda_{\{\mathbf{x}\},\{\mathbf{y}\}} = \lim_{u \uparrow 1} P\left(F(Z(\mathbf{x})) > u \middle| F(Z(\mathbf{y})) > u \right).$$

It characterizes the dependence in the tail of the random pair $(Z(\mathbf{x}), Z(\mathbf{y}))$, i.e., $\lambda_{\{\mathbf{x}\},\{\mathbf{y}\}} > 0$ corresponds to tail dependence and $\lambda_{\{\mathbf{x}\},\{\mathbf{y}\}} = 0$ means tail independence.

In this paper the main question concerns the risk of return levels. Thereby, in Sect. 2 we introduce a coefficient as a measure of the probability of occurring an exceedance of a high level u, in a location $\mathbf{y} \in \mathbb{R}^2$, given that also occurs an exceedance of u in a location $\mathbf{x} \in \mathbb{R}^2$ but the maximum over a neighborhood of $\mathbf{x}$ does not exceed the level u. Its main properties are presented, namely its relation with the extremal and upper tail dependence coefficients mentioned above. In Sect. 3 we present an estimator for the risk coefficient of return of a high level and its performance is analyzed with a max-stable maxima of moving maxima (M4) random field. Finally, Section 4 illustrates our approach through an application to annual maxima temperatures over Texas.

2 Risk of Return Levels and Dependence of Spatial Extreme Events

The next definition introduces the risk coefficient of return of a high level and then we present its relations with the dependence coefficients mentioned above.

Definition 1 Let $\{Z(\mathbf{x}), \mathbf{x} \in \mathbb{R}^2\}$ be a max-stable random field with unit Fréchet margins, F, and $R_{\mathbf{x}}$ a surrounding region of $\mathbf{x}$ with a finite number of locations. The risk coefficient of return of a high level at the location $\mathbf{x}+\mathbf{h}$, $\mathbf{h} \in \mathbb{R}^2$, is defined by

$$\delta_{\vee}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \lim_{u \uparrow 1} P\left(F(Z(\mathbf{x}+\mathbf{h})) > u \,\middle|\, F(Z(\mathbf{x})) > u, \bigvee_{\mathbf{y} \in R_{\mathbf{x}}} F(Z(\mathbf{y})) \leq u \right),$$

provided the limit exists.

Remark 1 We can define the risk coefficient of return of a low level in a similar way, i.e.,

$$\delta_{\wedge}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \lim_{u \downarrow 0} P\left(F(Z(\mathbf{x}+\mathbf{h})) \leq u \,\middle|\, F(Z(\mathbf{x})) \leq u, \bigwedge_{\mathbf{y} \in R_{\mathbf{x}}} F(Z(\mathbf{y})) > u \right),$$

where $\bigwedge_{i=1}^{k} a_i$ denotes the minimum of $\{a_1, \ldots, a_k\}$.

Remark 2 When applied to stationary sequences of random variables $\{Z(i) : i \in \mathbb{N}\}$, the risk coefficient $\delta_{\vee}(i+h|i, R_i)$, with $R_i = \{i+1, i+2, \ldots, i+h-1\}$, $h \in \mathbb{N}$, is the probability of a return period.

Remark 3

1. If the random variables $Z(\mathbf{x})$ and $Z(\mathbf{y})$ with $\mathbf{y} \in R_{\mathbf{x}}$ are totally dependent, then the risk coefficient of a return level is not defined since for each $\mathbf{y} \in R_{\mathbf{x}}$ we have

$$\begin{aligned}\lim_{u\uparrow 1} P\left(F(Z(\mathbf{x})) > u, F(Z(\mathbf{y})) \leq u\right) &= \lim_{u\uparrow 1} P\left(F(Z(\mathbf{y})) \leq u\right) \\ &\quad - P^{\epsilon_{\{\mathbf{x},\mathbf{y}\}}}\left(F(Z(\mathbf{y})) \leq u\right) \\ &= 0.\end{aligned}$$

2. If the random variables $Z(\mathbf{x})$, $Z(\mathbf{x}+\mathbf{h})$ and $Z(\mathbf{y})$ with $\mathbf{y} \in R_{\mathbf{x}}$ are independent, we have $\delta_{\vee}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = 0$.

The next result highlights the connection between $\delta_{\vee}(\mathbf{h}+\mathbf{x}|\mathbf{x}, R_{\mathbf{x}})$ and the extremal and multivariate upper tail dependence coefficients.

Proposition 1 *Let* $\{Z(\mathbf{x}), \mathbf{x} \in \mathbb{R}^2\}$ *be a max-stable random field with unit Fréchet margins,* F*, and* $R_{\mathbf{x}}$ *a surrounding region of* $\mathbf{x}$ *with a finite number of locations. Then,*

1.

$$\delta_{\vee}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \frac{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}\}} + \epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}+\mathbf{h}\}} - \epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x},\mathbf{x}+\mathbf{h}\}} - \epsilon_{R_{\mathbf{x}}}}{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}\}} - \epsilon_{R_{\mathbf{x}}}};$$

2.

$$\delta_{\vee}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \left(2 - \epsilon_{\{\mathbf{x},\mathbf{x}+\mathbf{h}\}}\right) \times \frac{1 - \sum_{\emptyset\neq J\subseteq R_{\mathbf{x}}} (-1)^{|J|+1} \lambda_{J,\{\mathbf{x},\mathbf{x}+\mathbf{h}\}}}{1 - \sum_{\emptyset\neq J\subseteq R_{\mathbf{x}}} (-1)^{|J|+1} \lambda_{J,\{\mathbf{x}\}}}.$$

Proof

1. We have

$$\begin{aligned}&P\left(F(Z(\mathbf{x}+\mathbf{h})) > u, F(Z(\mathbf{x})) > u, \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y})) \leq u\right) \\ &= P\left(\bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y})) \leq u\right) - P\left(\bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y})) \leq u, \bigwedge_{\mathbf{y}\in\{\mathbf{x},\mathbf{x}+\mathbf{h}\}} F(Z(\mathbf{y}) \leq u\right) \\ &= u^{\epsilon_{R_{\mathbf{x}}}}\left(1 - u^{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}\}} - \epsilon_{R_{\mathbf{x}}}} - u^{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}+\mathbf{h}\}} - \epsilon_{R_{\mathbf{x}}}} + u^{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x},\mathbf{x}+\mathbf{h}\}} - \epsilon_{R_{\mathbf{x}}}}\right). \quad (2)\end{aligned}$$

On the other hand, it holds

$$
\begin{aligned}
& P\left(F(Z(\mathbf{x}))>u, \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y}))\leq u\right) \\
&= P\left(\bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y}))\leq u\right) - P\left(\bigvee_{\mathbf{y}\in R_{\mathbf{x}\cup\{\mathbf{x}\}}} F(Z(\mathbf{y}))\leq u\right) \\
&= u^{\epsilon_{R_{\mathbf{x}}}}\left(1-u^{\epsilon_{R_{\mathbf{x}}\cup\{\mathbf{x}\}}-\epsilon_{R_{\mathbf{x}}}}\right). \qquad (3)
\end{aligned}
$$

Therefore, dividing (2) by (3), the result follows from the L'Hôpital's rule.

2. Since

$$
\begin{aligned}
& P\left(F(Z(\mathbf{x}+\mathbf{h}))>u, F(Z(\mathbf{x}))>u, \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y}))\leq u\right) \\
= \quad & P\left(F(Z(\mathbf{x}+\mathbf{h}))>u, F(Z(\mathbf{x}))>u\right) \\
& \times\left(1-P\left(\bigcup_{\mathbf{y}\in R_{\mathbf{x}}}\{F(Z(\mathbf{y}))>u\}\,|F(Z(\mathbf{x}+\mathbf{h}))>u, F(Z(\mathbf{x}))>u\right)\right) \\
= \quad & \left(1-2u+u^{\epsilon_{\{\mathbf{x},\mathbf{x}+\mathbf{h}\}}}\right)\left(1-\sum_{\emptyset\neq J\subseteq R_{\mathbf{x}}}(-1)^{|J|+1}\lambda_{J,\{\mathbf{x}+\mathbf{h},\mathbf{x}\}}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
& P\left(F(Z(\mathbf{x}))>u, \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} F(Z(\mathbf{y}))\leq u\right) \\
&= P\left(F(Z(\mathbf{x}))>u\right)\left(1-P\left(\bigcup_{\mathbf{y}\in R_{\mathbf{x}}}\{F(Z(\mathbf{y}))>u\}\,|F(Z(\mathbf{x}))>u\right)\right) \\
&= (1-u)\left(\sum_{\emptyset\neq J\subseteq R_{\mathbf{x}}}(-1)^{|J|+1}\lambda_{J,\{\mathbf{x}\}}\right),
\end{aligned}
$$

the result follows. □

In the following, we present the expression of the risk coefficient of return levels for random fields with known distribution functions for its margins.

Example 1 Consider the $M4$ random field defined in Fonseca et al. [10], as

$$Z(\mathbf{x}) = \bigvee_{l=1}^{+\infty} \bigvee_{m=-\infty}^{+\infty} a_{lm\mathbf{x}} Y_{l,1-m}, \quad \mathbf{x} \in Z^2,$$

where $\{Y_{l,n}\}_{l\geq 1, n\in Z}$ is a family of independent unit Fréchet random variables and, for each $\mathbf{x} \in Z^2$, $\{a_{lm\mathbf{x}}\}_{l\geq 1, m\in Z}$ are non-negative constants such that $\sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} a_{lm\mathbf{x}} = 1$.

The distribution function of $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_p))$ is characterized by the copula

$$C(u_{\mathbf{x}_1}, \ldots, u_{\mathbf{x}_p}) = \prod_{l=1}^{+\infty} \prod_{m=-\infty}^{+\infty} \bigwedge_{\mathbf{x}\in\{\mathbf{x}_1,\ldots,\mathbf{x}_p\}} u_{\mathbf{x}}^{a_{lm\mathbf{x}}}, \quad u_{\mathbf{x}_i} \in [0,1],\ i = 1, \ldots, p.$$

For each pair of regions $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ and $B = \{\mathbf{y}_{k+1}, \ldots, \mathbf{y}_{k+s}\}$ we have

$$l_{A,B}(w_1, \ldots, w_k, w_{k+1}, \ldots, w_{k+s}) = \sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} \bigvee_{i=1}^{k} w_i^{-1} a_{lm\mathbf{x}_i} \vee \bigvee_{i=k+1}^{k+s} w_i^{-1} a_{lm\mathbf{y}_i},$$

$w_i \in \mathbb{R},\ i = 1, \ldots, k+s$.

Therefore,

$$\delta_\vee(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \frac{\sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}} \vee a_{lm\mathbf{x}} + \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}} \vee a_{lm(\mathbf{x}+\mathbf{h})}}{\sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}} \vee a_{lm\mathbf{x}} - \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}}}$$
$$- \frac{\sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}} \vee \bigvee_{\mathbf{z}\in\{\mathbf{x},\mathbf{x}+\mathbf{h}\}} a_{lm\mathbf{y}} + \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}}}{\sum_{l=1}^{+\infty} \sum_{m=-\infty}^{+\infty} \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}} \vee a_{lm\mathbf{x}} - \bigvee_{\mathbf{y}\in R_{\mathbf{x}}} a_{lm\mathbf{y}}}$$

Example 2 Let us consider the symmetric logistic model, introduced in Gumbel [11], one of the oldest parametric models of tail dependence. Its distribution function with unit Fréchet margins is defined by

$$F(w_1, \ldots, w_k; \theta) = \exp\left\{ - \left(\sum_{j=1}^{k} w_j^{-\frac{1}{\theta}} \right)^{\theta} \right\}, \quad for\ w_1, \ldots, w_k > 0\ \ and\ \theta \in [0,1],$$

and the corresponding tail dependence function is given by

$$l_A(w_1, \ldots, w_k; \theta) = \left(w_1^{-\frac{1}{\theta}} + \ldots + w_k^{-\frac{1}{\theta}} \right)^{\theta}, \quad A = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}.$$

Therefore,

$$\delta_{\vee}(\mathbf{x}+\mathbf{h}\,|\mathbf{x}, R_{\mathbf{x}}) = \frac{2\,(1+|R_{\mathbf{x}}|)^{\theta} - (2+|R_{\mathbf{x}}|)^{\theta} - |R_{\mathbf{x}}|^{\theta}}{(1+|R_{\mathbf{x}}|)^{\theta} - |R_{\mathbf{x}}|^{\theta}},$$

where $|A|$ denotes the cardinality of the event A.

3 Estimation

Let $(Z^{(i)}(\mathbf{x}_1), \ldots, Z^{(i)}(\mathbf{x}_k))$, $i = 1, \ldots, n$, be independent copies of $(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_k))$. Proposition 1 gives rise to the following estimator for the risk coefficient of a return level:

$$\widehat{\delta}_{\vee}(\mathbf{x}+\mathbf{h}|\mathbf{x}, R_{\mathbf{x}}) = \frac{\widehat{\epsilon}_{R_{\mathbf{x}}\cup\{\mathbf{x}\}} + \widehat{\epsilon}_{R_{\mathbf{x}}\cup\{\mathbf{x}+\mathbf{h}\}} - \widehat{\epsilon}_{R_{\mathbf{x}}\cup\{\mathbf{x},\mathbf{x}+\mathbf{h}\}} - \widehat{\epsilon}_{R_{\mathbf{x}}}}{\widehat{\epsilon}_{R_{\mathbf{x}}\cup\{\mathbf{x}\}} - \widehat{\epsilon}_{R_{\mathbf{x}}}},$$

where

$$\widehat{\epsilon}_A = \frac{\overline{M(A)}}{1-\overline{M(A)}},$$

$\overline{M(A)}$ is the sample mean,

$$\overline{M(A)} = \frac{1}{n}\sum_{i=1}^{n} \bigvee_{\mathbf{x}\in A} \widehat{F}_{\mathbf{x}}(Z^{(i)}(\mathbf{x}))$$

and $\widehat{F}_{\mathbf{x}}$, $\mathbf{x} \in \mathbb{R}^2$ is the (modified) empirical distribution of $F_{\mathbf{x}}$,

$$\widehat{F}_{\mathbf{x}}(u) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbb{1}_{\{Z^{(i)}(\mathbf{x})\le u\}}.$$

The estimator is strongly consistent given the consistency of the estimator $\widehat{\epsilon}_A$ already stated in Ferreira and Ferreira [6].

To assess the performance of the estimator of the risk coefficient of a return level, we shall consider the Example 1 with a finite number of signature patterns and a finite range of sequential dependencies, as is presented in Fonseca et al. [9].

Example 3 Let us consider that for each location $\mathbf{x} = (x_1, x_2) \in Z^2$ we have

(1) $x_1 > x_2 \wedge x_1$ even $\wedge\, x_2$ odd ,

$$a_{11\mathbf{x}} = a_{12\mathbf{x}} = a_{21\mathbf{x}} = a_{22\mathbf{x}} = \tfrac{1}{8}$$

$$a_{31\mathbf{x}} = a_{32\mathbf{x}} = a_{41\mathbf{x}} = a_{42\mathbf{x}} = \tfrac{1}{8}$$

(2) $x_1 \leq x_2 \wedge x_1$ even $\wedge\, x_2$ odd ,

$$a_{11\mathbf{x}} = a_{12\mathbf{x}} = \tfrac{2}{17},\ \ a_{21\mathbf{x}} = \tfrac{5}{17},\ \ a_{22\mathbf{x}} = \tfrac{4}{17}$$

$$a_{31\mathbf{x}} = a_{32\mathbf{x}} = a_{41\mathbf{x}} = a_{42\mathbf{x}} = \tfrac{1}{17}$$

(3) $x_1 > x_2 \wedge\ x_1$ odd $\wedge\, x_2$ even ,

$$a_{11\mathbf{x}} = \tfrac{1}{20},\ \ a_{12\mathbf{x}} = \tfrac{2}{20},\ \ a_{21\mathbf{x}} = \tfrac{3}{20},\ \ a_{22\mathbf{x}} = \tfrac{4}{20}$$

$$a_{31\mathbf{x}} = \tfrac{5}{20},\ \ a_{32\mathbf{x}} = \tfrac{3}{20},\ \ a_{41\mathbf{x}} = a_{42\mathbf{x}} = \tfrac{1}{20}$$

(4) $x_1 \leq x_2 \wedge\ x_1$ odd $\wedge\, x_2$ even ,

$$a_{11\mathbf{x}} = \tfrac{1}{36},\ \ a_{12\mathbf{x}} = \tfrac{2}{36},\ \ a_{21\mathbf{x}} = \tfrac{3}{36},\ \ a_{22\mathbf{x}} = \tfrac{4}{36}$$

$$a_{31\mathbf{x}} = \tfrac{5}{36},\ \ a_{32\mathbf{x}} = \tfrac{6}{36},\ \ a_{41\mathbf{x}} = \tfrac{7}{36},\ \ a_{42\mathbf{x}} = \tfrac{8}{36}$$

(5) $x_1 > x_2 \wedge x_1$ even $\wedge\, x_2$ even ,

$$a_{11\mathbf{x}} = \tfrac{1}{40},\ \ a_{12\mathbf{x}} = \tfrac{2}{40},\ \ a_{21\mathbf{x}} = \tfrac{3}{40},\ \ a_{22\mathbf{x}} = \tfrac{4}{40}$$

$$a_{31\mathbf{x}} = \tfrac{5}{40},\ \ a_{32\mathbf{x}} = \tfrac{6}{40},\ \ a_{41\mathbf{x}} = \tfrac{7}{40},\ \ a_{42\mathbf{x}} = \tfrac{12}{40}$$

(6) $x_1 \leq x_2 \wedge x_1$ even $\wedge\, x_2$ even ,

$$a_{11\mathbf{x}} = \tfrac{1}{45},\ \ a_{12\mathbf{x}} = \tfrac{2}{45},\ \ a_{21\mathbf{x}} = \tfrac{3}{45},\ \ a_{22\mathbf{x}} = \tfrac{4}{45}$$

$$a_{31\mathbf{x}} = \tfrac{6}{45},\ \ a_{32\mathbf{x}} = \tfrac{8}{45},\ \ a_{41\mathbf{x}} = \tfrac{9}{45},\ \ a_{42\mathbf{x}} = \tfrac{12}{45}$$

(7) $x_1 > x_2 \wedge\ x_1$ odd $\wedge\, x_2$ odd ,

$$a_{11\mathbf{x}} = \tfrac{1}{50},\ \ a_{12\mathbf{x}} = \tfrac{7}{50},\ \ a_{21\mathbf{x}} = \tfrac{3}{50},\ \ a_{22\mathbf{x}} = \tfrac{4}{50}$$

$$a_{31\mathbf{x}} = \tfrac{6}{50},\ \ a_{32\mathbf{x}} = \tfrac{8}{50},\ \ a_{41\mathbf{x}} = \tfrac{9}{50},\ \ a_{42\mathbf{x}} = \tfrac{12}{50}$$

(8) $x_1 \leq x_2 \wedge\ x_1$ odd $\wedge\, x_2$ odd ,

$$a_{11\mathbf{x}} = \tfrac{1}{60},\ \ a_{12\mathbf{x}} = \tfrac{7}{60},\ \ a_{21\mathbf{x}} = \tfrac{3}{60},\ \ a_{22\mathbf{x}} = \tfrac{14}{60}$$

$$a_{31\mathbf{x}} = \tfrac{6}{60},\ \ a_{32\mathbf{x}} = \tfrac{8}{60},\ \ a_{41\mathbf{x}} = \tfrac{9}{60},\ \ a_{42\mathbf{x}} = \tfrac{12}{60}.$$

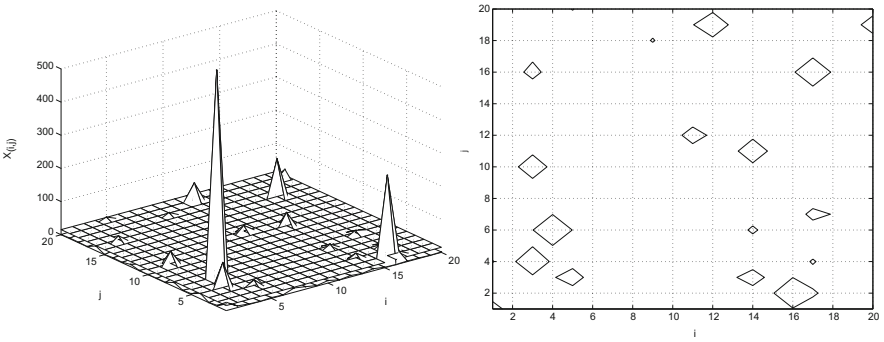


Fig. 3 Simulation of the M4 as defined in Example 3 (left) and the contour at $x_{(i,j)} = 15.3417$, the 95% quantile (right)

Table 1 Results of the risk coefficient, $\delta_\vee(\mathbf{x}+\mathbf{h}\,|\mathbf{x}, R_\mathbf{x})$, where $\mathbf{x} = (2, 2)$ and $R_\mathbf{x} = \{(2, 3), (3, 3), (3, 2)\}$

	$x_2 + h_2$			
$x_1 + h_1$	2	3	4	5
2	–	–	1	−1.5e−15
3	–	–	0.576923	−1.5e−15
4	0.634615	0	1	−1.5e−15
5	−1.5e−15	0.553846	−1.5e−15	−1.5e−15

The values of $(a_{l1\mathbf{x}}, a_{l2\mathbf{x}})$, $l = 1, \ldots, 4$, define the four signature patterns of the random field (Fig. 3).

For $\mathbf{x} = (2, 2)$ and $R_\mathbf{x} = \{(2, 3), (3, 2), (3, 3)\}$ we obtain the risk coefficients given in Table 1.

The results of the application of the estimator $\widehat{\delta}_\vee(\mathbf{x}+\mathbf{h}\,|\mathbf{x}, R_\mathbf{x})$ are presented in Table 2.

As we can see from the values of the mean square error (MSE) the estimates obtained from our estimator are quite close to the true values of $\delta_\vee(\mathbf{x}+\mathbf{h}\,|\mathbf{x}, R_\mathbf{x})$ which highlights the good performance of our estimator.

4 Application

We compute the estimates for the risk coefficient of a return level, for maxima temperature $Z(\mathbf{x})$ recorded over Texas. We focus on a subset of 12 locations (Fig. 4) covering a common period from 1948 up to 2015, without substantial interruptions.

For each location the maximum for the temperatures in warm season (April to September) and cold season (October to March) are extracted for the whole period, from the National Climatic Data Center of NOAA (National Oceanic and Atmospheric Administration).

Table 2 Results with 100 replications of 1000 i.d.d. max-stable M4 random fields of the Example 1, where $\mathbf{x} = (2, 2)$ and $R_{\mathbf{x}} = \{(2, 3), (3, 3), (3, 2)\}$

y=x+h	(4,2)	(4,3)	(4,4)	(3,4)	(2,4)	(5,2)	(5,3)	(5,4)	(5,5)	(4,5)	(3,5)	(2,5)
$\widehat{\delta_{\vee}}(\mathbf{y}\,\vert\mathbf{x}, R_{\mathbf{x}})$	0.652546	2.89e−4	1	0.575092	1	−1.7e−16	0.560383	−1.4e−16	−1.1e−16	−1.6e−16	−9.4e−17	−1.8e−16
MSE	0.000717	1.4e−5	0	0.000341	0	3.1e−30	0.001270	3.2e−30	3.1e−30	3.3e−30	3.5e−30	3.1e−30

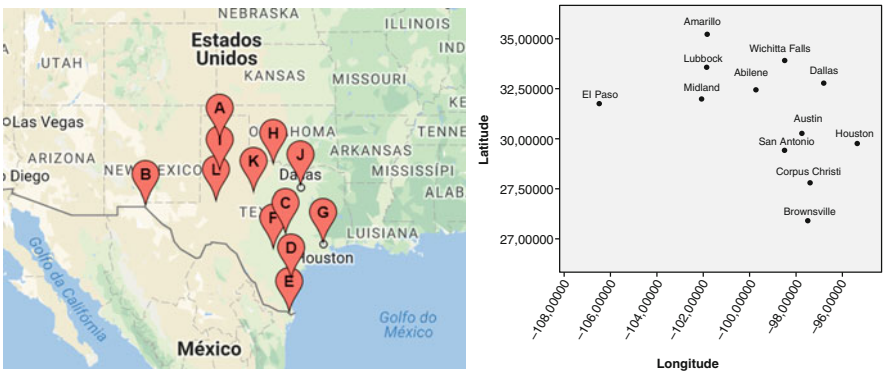


Fig. 4 Locations of the stations where temperature data were collected, obtained from the NOAA's National Climatic Data Center (left) and their representation in Lambert coordinates (right)

Table 3 Estimates of the risk coefficient, where $\mathbf{x}$ denotes the location Wichita Falls and $R_{\mathbf{x}} = \{\text{Lubbock, Dallas, Abilene, Midland}\}$

h	**x+h**	$d(\mathbf{x}, \mathbf{x}+\mathbf{h})$	$\widehat{\delta}_{\vee}(\mathbf{x}+\mathbf{h}\,\vert\mathbf{x}, R_{\mathbf{x}})$(WS)	$\widehat{\delta}_{\vee}(\mathbf{x}+\mathbf{h}\,\vert\mathbf{x}, R_{\mathbf{x}})$ (CS)
(1.308, −3.338)	Amarillo	339	0.1826	0.3107
(−2.155, −7.994)	El Paso	785	0	0.1148
(−3.647, 0.75)	Austin	412	0	0.1574
(−6.113, 1.097)	C. Christi	689	0	0
(−8.012, 0.996)	Brownsville	897	0.01059	0
(−4.49; 0.000)	S. Antonio	500	0	0.1481
(−4.15; 3.13)	Houston	549	0	0

In what follows $\mathbf{x}$ denotes the location Wichita Falls and $R_{\mathbf{x}} = \{\text{Lubbock, Dallas, Abilene, Midland}\}$. Our analysis indicates that the risk of a return level varies according to the following factors: the distance between the locations $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$ and the season (cold or warm).

From Table 3 we conclude that the risk of a return level is almost always higher in cold season than in warm season and for greater distances between the locations $\mathbf{x}$ and $\mathbf{x}+\mathbf{h}$, there is lower risk of a return of a high temperature.

References

1. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
2. Buishand, T.A., de Haan, L., Zhou, C.: On spatial extremes: With application to a rainfall problem. Ann. Appl. Stat. **2**, 624–642 (2008)
3. Coles, S.G.: Regional modelling of extreme storms via max-stable processes. J. R. Stat. Soc. **55**, 797–816 (1993)

4. Coles, S.G., Tawn, J.A.: Modelling extremes of the areal rainfall process. J. R. Stat. Soc. Ser. B **58**, 329–347 (1996)
5. Cooley, D., Naveau, P., Poncet, P.: Variograms for spatial max-stable random fields. REVSTAT **10**, 135–165 (2006)
6. Ferreira, H., Ferreira, M.: On extremal dependence of block vectors. Kybernetika **48**, 988–1006 (2012)
7. Ferreira, M., Ferreira, H.: On extremal dependence: some contributions. TEST **21**, 566–583 (2012)
8. Fisher, R., Tippett, L.: Limiting forms of th frequency distribution of the largest or smallest member of a sample. Proc. Camb. Philos. Soc. **24**, 180–190 (1928)
9. Fonseca, C., Ferreira, H., Pereira, L., Martins, A.P.: Stability and contagion measures for spatial extreme value analyses. Kybernetika **50**, 914–928 (2014)
10. Fonseca, C., Pereira, L., Ferreira, H., Martins, A.P.: Generalized madogram and pairwise dependence of maxima over two regions of a random field. Kybernetika **51**, 193–211 (2015)
11. Gumbel, E.J.: Bivariate exponential distributions. J. Am. Stat. Assoc. **55**, 698–707 (1960)
12. Li, H.: Orthant tail dependence of multivariate extreme value distributions. J. Multivar. Anal. **46**, 262–282 (2009)
13. Naveau, P., Guillou, A., Cooley, D., Diebolt, J.: Modelling pairwise dependence of maxima in space. Biometrika **96**, 1–17 (2009)
14. Resnick, S.I.: Extreme Values, Regular Variation and Point Processes. Springer, Berlin (1987)
15. Schlather, M., Tawn, J.: A dependence measure for multivariate and spatial extreme values: properties and inference. Biometrika, **90**, 139–156 (2003)
16. Schmid, F., Schmidt, R.: Multivariate conditional versions of Spearman's rhoand related measures of tail dependence. J. Multivar. Anal. **98**, 1123–1140 (2007)
17. Sibuya, M.: Bivariate extreme statistics. Ann. Inst. Stat. Math. **11**, 195–210 (1960)
18. Smith, R.L.: Max-stable processes and spatial extremes. Univ. North Carolina, USA (Preprint, 1990)
19. Zhang, Z., Smith, R.L.: On the estimation and application of max-stable processes. J. Stat. Plan. Inference **140**, 1135–1153 (2010)

Nonparametric Individual Control Charts for Silica in Water

Luís M. Grilo, Mário A. Santos, and Helena L. Grilo

Abstract The soluble silica content in the demineralized water is a continuous variable measured and controlled in the Chemical Laboratory of a Portuguese thermoelectric central, in order to keep the equipment operating under the best conditions, allowing, in particular, to extend its useful life. In this case study, this variable could be considered approximately normal distributed and because we just have one measure, for each group of the sample, an individual control chart to monitor the silica content is obtained based on average moving range. Once the available sample size is small and it is hard to fit a model, robust control limits using a nonparametric method based on empirical quantiles (which according to some simulations studies perform also well under the normality of the observations) are also estimated with the bootstrap procedure. The comparison of the control limits obtained with different approaches and with(out) outliers is very important for technicians since the value of silica should be as small as possible. The process capability study, also developed, shows that the process does not stay within the engineering specification limits, although it seems stable.

1 Introduction

Demineralized water is indispensable for the energy production process in a Portuguese thermoelectric central, since the formation of high pressure water steam is obtained by heat transfer between the boiler and demineralized water, which

L. M. Grilo (✉)
Unidade Departamental de Matemática e Física, Instituto Politécnico de Tomar, Tomar, Portugal

FCT, UNL, Centro de Matemática e Aplicações (CMA), Caparica, Portugal
e-mail: lgrilo@ipt.pt

M. A. Santos
Unidade Departamental de Engenharia, Instituto Politécnico de Tomar, Tomar, Portugal

H. L. Grilo
Centro de Sondagens e Estudos Estatísticos, Instituto Politécnico de Tomar, Tomar, Portugal

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_9

circulates in adjacent pipes. This water steam provides the movement of the turbine blades producing mechanical energy, which is then converted into electricity in a generator. The variable silica (in $\mu g\ SiO_2/L$), soluble in water, has to be removed because this chemical compound has a tendency to form deposits on the walls of the equipment and piping. Furthermore, when water boils in the boiler, this silica has a high abrasive effect on its metal walls and being entrained in the steam duct will cause wear on turbine blades. The water is taken directly from a river close to the thermoelectric central, which is subjected to a set of unit operations in water treatment facility. The treatment occurs in the passage through anion exchange resins. These same resins are specifically selected to adsorb the required species. The treatment process works by passing the water through a column, where it comes into contact with the active material (resins). The unwanted ionic species which are dissolved in the solution are adsorbed selectively to the surface. The concentration of soluble silica in demineralized water depends on the concentration of soluble silica in water collected in the river and on the saturation level of the anionic exchange resins. To monitor eventual changes in this industrial process we just have, namely for economic reasons (considering time and money), a sample of one measurement (one data point is collected at each point in time). Thus, in order to determine whether the process is operating normally or needs to be adjusted, Shewhart individual control charts (X) for the variable "silica" are obtained with individual observations stem from a process which is statistically in-control (as in [3–7, 9, 10]), computing the control limits based on the average moving range (AMR), since the empirical distribution of this variable could be considered approximately normal. Because the sample size is small, we also estimate robust control limits using a nonparametric method based on empirical quantiles (EQ), to turn the X control charts into more sensitive ones to persistent assignable causes. These alternative control charts are a special case of the bootstrapping control charts and are not only quite robust against deviations from normality but also perform reasonably well under normality of the observations [9, 11]. After removing the outliers, control limits are also estimated and compared. A capability study is also developed, with(out) two outliers, in order to analyze the process ability to produce outputs within specification limits.

2 Data Analysis

In Table 1 we have the available dataset, which is a small sample size of $n = 10$ measurements of soluble silica content in the demineralized water ($\mu g\ SiO_2/L$).

Table 1 The measurements of soluble silica content in the demineralized water ($\mu g\ SiO_2/L$)

Subject	1	2	3	4	5	6	7	8	9	10
Dataset	4.0	4.9	5.4	8.3	7.8	5.7	3.0	4.6	3.7	4.4

Table 2 Some descriptive statistics of silica (µg SiO_2/L)

Silica	Statistic	Std. error
Mean	5.180	0.540
5% trimmed mean	5.128	
Median	4.750	
Std. deviation	1.709	
Variation coef. (%)	32.992	
Minimum	3.000	
Maximum	8.300	
Skewness	0.907	0.687
Kurtosis	0.059	1.334

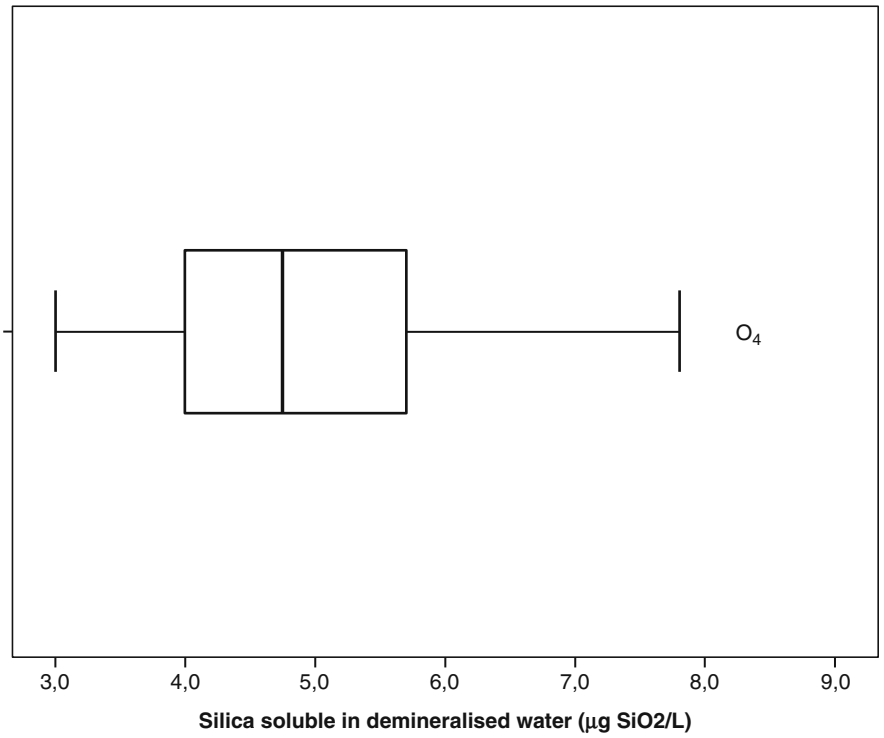


Fig. 1 Box-plot of silica in water (µg SiO_2/L)

Some descriptive statistics obtained with the sample of 10 individual measurements of the silica in water are shown in Table 2. The mean, the trimmed mean, and the median (location measures) are close, although the median stays below. There is a considerable dispersion given by the variation coefficient (i.e., the ratio between the standard deviation and the mean), which is approximately 33%. The shape of the empirical distribution is approximately mesokurtic, since the coefficient value is very close to zero, and it is positive skewed (the skewness coefficient value is positive and higher than 0.5), where a moderate outlier is also identified (Fig. 1).

Table 3 The Shapiro-Wilk normality test of silica (µg SiO_2/L)

Silica	Shapiro-Wilk		
	Statistic	df	p-value
Dataset	0.910	10	0.278

Here we consider that the moderate outliers are between 1.5 and 3 interquartile ranges down from the first quartile or up from the third quartile.

The p-value obtained with the Shapiro-Wilk normality test (Table 3) led us to not reject the null hypothesis of normality, for the usual significance levels considered.

3 Control Limits Methods

The X control chart is obtained to monitor changes that modify the silica mean. In this type of control charts we have the solid central line as the average value and the two dashed lines representing the lower and upper control limits, respectively, denoted by LCL and UCL. The control limits reflect the expected amount of variation in the sample means when only common causes of variation are presented. If the process is in-control, nearly all of the sample points will fall between those limits.

When the cumulative distribution function (c.d.f.) F is associated with normal model, usually represented by Φ, with mean μ and standard deviation σ, the control limits of the Shewhart X chart are

$$LCL = \mu - \Phi^{-1}_{\left(\frac{\alpha}{2}\right)}\sigma, \qquad UCL = \mu + \Phi^{-1}_{\left(1-\frac{\alpha}{2}\right)}\sigma, \tag{1}$$

where Φ^{-1} is the standard normal quantile function and the α level represents the false alarm rate. The parameters μ and σ in (1) are usually unknown. However, if an independent and identically distributed (i.i.d.) random sample $(X_1, X_2, \ldots, X_n)$ is available, we can estimate these parameters using the classical estimators, which are, respectively, the sample mean and the sample standard deviation,

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma} = S' = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}.$$

The sample standard deviation, S', is asymptotically efficient for an i.i.d. normal random sample, but it is also sensitive to trends and oscillations, which is a disadvantage. An estimator less sensitive is the AMR (average moving range), where the difference between each data point, X_i, and its predecessor, X_{i-1}, is calculated as $|X_i - X_{i-1}|$ and for m individual values, there are $m - 1$ ranges. The arithmetic mean of these values is computed as

$$\overline{MR} = \frac{1}{n-1}\sum_{i=2}^{n} |X_i - X_{i-1}|, \tag{2}$$

which can be scaled by $d_2(2) = 2/\sqrt{\pi}$, in order to obtain an unbiased estimator for σ under normality, i.e.

$$\hat{\sigma} = \frac{\overline{MR}}{d_2(2)} = \frac{1}{d_2(2)} \frac{\sum_{i=2}^{n} |X_i - X_{i-1}|}{n-1} = \frac{\sqrt{\pi}}{2} \overline{MR}. \tag{3}$$

3.1 Control Limits Based on Average Moving Range

The control limits based on the AMR, in (2) and (3), for the traditional X control charts are [2]

$$LCL_{\text{AMR}} = \overline{X} - \Phi^{-1}_{\left(\frac{\alpha}{2}\right)} \frac{\sqrt{\pi}}{2} \overline{MR}, \quad UCL_{\text{AMR}} = \overline{X} + \Phi^{-1}_{\left(1-\frac{\alpha}{2}\right)} \frac{\sqrt{\pi}}{2} \overline{MR}. \tag{4}$$

To avoid an excessive number of false alarms, that occurs in most industrial processes, Shewhart proposed the value of 0.0027 for the α level which corresponds to $\Phi^{-1}_{\left(1-\frac{\alpha}{2}\right)} = 3$ and considering $d_2 = 2/\sqrt{\pi}$ we can rewrite (4) as

$$LCL_{\text{AMR}} = \overline{X} - 3\frac{\overline{MR}}{d_2(2)} \approx \overline{X} - 2.66\,\overline{MR}, \tag{5}$$

and

$$UCL_{\text{AMR}} = \overline{X} + 3\frac{\overline{MR}}{d_2(2)} \approx \overline{X} + 2.66\,\overline{MR}. \tag{6}$$

According to [10], independently of the observations probability distribution, the AMR control charts tend to perform reasonably well for moderate sample sizes, which is not a situation of this case study, where we just have a small sample size.

3.2 Control Limits Based on Empirical Quantiles

The control limits of the EQ (empirical quantiles) for X chart will be defined according to [9], where a natural estimator of the q-quantile of the unimodal unknown c.d.f. F is the empirical quantile $\hat{F}_n^{-1}(q)$, defined as

$$\hat{F}_n^{-1}(q) = \inf\left\{x \mid \hat{F}_n(x) \geq q\right\}, \qquad 0 < q < 1$$

where $\hat{F}_n$ is the empirical c.d.f. that puts mass $1/n$ at each X_i, $1 \leq i \leq n$, i.e.,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}, \qquad -\infty < x < +\infty,$$

where I represents the indicator function, i.e. $I_{\{x \leq y\}}$ equals 1 if $x \leq y$ and 0 otherwise. Thus, the obvious estimators of the lower and upper control limits based on the EQ are, respectively,

$$LCL_{\text{EQ}} = \hat{F}_n^{-1}\left(\tfrac{\alpha}{2}\right) = X_{\lfloor \frac{\alpha}{2}n+1 \rfloor} \quad \text{and} \quad UCL_{\text{EQ}} = \hat{F}_n^{-1}\left(1-\tfrac{\alpha}{2}\right) = X_{\lceil (1-\frac{\alpha}{2})n \rceil}. \tag{7}$$

Here $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denotes the order statistics of $X_1, X_2, \ldots, X_n$ which is the initial sample, $\lfloor \cdot \rfloor$ denotes the floor of the argument (that is the largest integer that does not exceed the argument) and $\lceil \cdot \rceil$ denotes the ceiling of the argument (that is the smallest integer not less than the argument). The nonparametric control charts become attractive if large datasets are available, i.e. we need at least 1000 observations in order to attain reasonably performance. Nevertheless, this may be surpassed with the bootstrap approach which is a computational intensive technique based on the philosophy that the unknown c.d.f. F of a random variable will be replaced by an empirical c.d.f. $\hat{F}_n$. Thus, we apply the bootstrap procedure to obtain, for example, $UCL_{\text{EQ}} = \hat{F}_n^{-1}\left(1-\frac{\alpha}{2}\right)$ as an estimate of the $UCL_{\text{EQ}} = F_n^{-1}\left(1-\frac{\alpha}{2}\right)$ in control charts for individual observations (as in [4, 7]). This nonparametric method has the advantage of being easy to compute and distribution-free when there is an in-control situation as we have here.

There are some new interesting and elaborate methods about control charts for individual observations/measurements [1, 8], but they are not so easy to implement computationally and to interpret by non-statisticians. Nevertheless, these methods might be used for comparison in future work.

4 Comparison of Control Limits for X Charts

An X control chart, as Fig. 2, is used to detect trends and shifts in the data, and thus in the process. The data is time-ordered; that is, the data appear in the sequence in which they were generated. The bilateral control chart for silica displays the individual measurements, the estimates control limits obtained with the AMR method, LCL_{AMR} and UCL_{AMR}, as well as the lower and upper specification limits (respectively, LSL and USL).

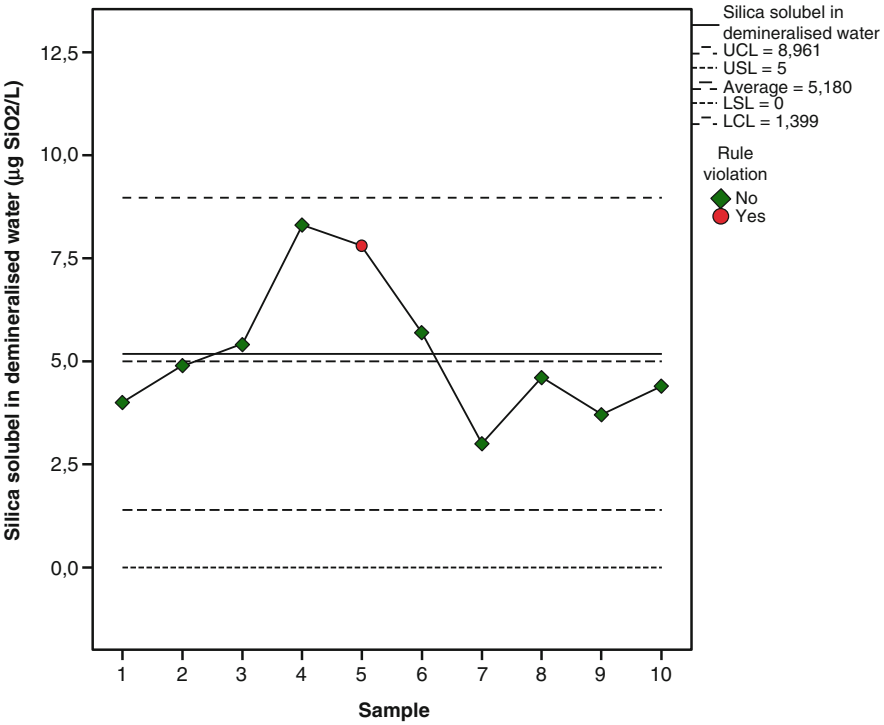


Fig. 2 Individual control chart of silica in water (µg SiO_2/L)

4.1 Control Limits Computed with Outliers

To calculate the control limits with AMR method we use (5) and (6), the mean value in Table 2 and the average moving range value, $\overline{MR} \approx 1.422$, applying (2) to the data in Table 1. Merely one point violates the control rules (Table 4 and Fig. 2). In Table 5 we have the estimates of the control limits for X chart of silica, considering both methods (AMR and EQ), the equations in (5), (6), and (7), respectively, and the usual α level of 0.0027. To obtain the control limits based on the EQ method, we simply draw 5000 bootstrapped samples, with the same size as the dataset ($n = 10$), with replacement from a population made up of the observed dataset. Then, we determine the mean of each sample, which creates the sampling distribution of the mean, and using (7) we obtain the control limits. These control limits, more sensitive to small shifts in the mean, could be superimposed in Fig. 2 to see how narrower these limits are and to visualize the number of individual values outside the limits. To compare how relevant are the different results we can take into account the range of limits in Table 5, for both methods (AMR and EQ), where the one that corresponds to the EQ control limits is considerably smaller.

Table 4 Rule violations for run of silica(µg SiO_2/L)

Rule violations for run	
Case number	Violations for points
5	2 points out of the last 3 above +2 sigma

1 point violates control rules

Table 5 Control limits for individual charts of silica (µg SiO_2/L)

Control limits			
Method	*LCL*	*UCL*	Range of limits
AMR	1.399	8.961	7.562
EQ	3.877	6.860	2.983

In Fig. 2, a considerable part of the control limits obtained with the AMR method is outside of the specification limits (where the ideal lower specification limit should be $LSL = 0$ µg SiO_2/L and the upper specification limit should never surpass the $USL = 5$ µg SiO_2/L, according to the established company's specifications limits for this variable). A similar situation occurs with control limits estimated with the EQ method (Table 5). Note that we have six points within the specification limits (i.e., 60%).

In the moving range charts, not presented here, all points are within the control limits and no pattern is identified.

4.2 Control Limits Computed Without Outliers

The company's upper specification limit for the soluble silica in demineralized water is 5 µg SiO_2/L, but some higher values were obtained. However, it is important to note that when this situation occurs the responsible department proceed to the regeneration of anion exchange resins. In fact, if we remove the moderate outlier (with value 8.3), visible in Fig. 1, then another observation appears as a moderate outlier (with value 7.8). According to the laboratory technicians both atypical observations should be removed, because although the outliers could not be a measurement error, the experts consider that they should not appear, because the process must be monitored rigorously. Thus, we decide to remove them and with the smaller sample of 8 individual measurements of silica in water we obtain the descriptive statistics in Table 6. The central tendency measures (mean, trimmed mean and median) are almost equal to 4.5 and the dispersion is now smaller, as expected, given the value of variation coefficient (approximately 20%). The empirical distribution is approximately symmetric (slight skewed, as we can see in the box-plot of Fig. 3) and mesokurtic, since the coefficients values of skewness and kurtosis are, respectively, within the interval -0.5 to 0.5). In Table 7 we confirm with the Shapiro-Wilk test that we can consider the distribution approximately normal (with p-value $= 0.984$ and for the significance levels usually considered we do not reject the null hypothesis of normality).

Table 6 Some descriptive statistics of silica, without outliers (μg SiO_2/L)

Silica	Statistic	Std. error
Mean	4.463	0.315
5% trimmed mean	4.475	
Median	4.500	
Std. deviation	0.891	
Variation coef. (%)	19.964	
Minimum	3.000	
Maximum	5.700	
Skewness	−0.233	0.752
Kurtosis	−0.489	1.481

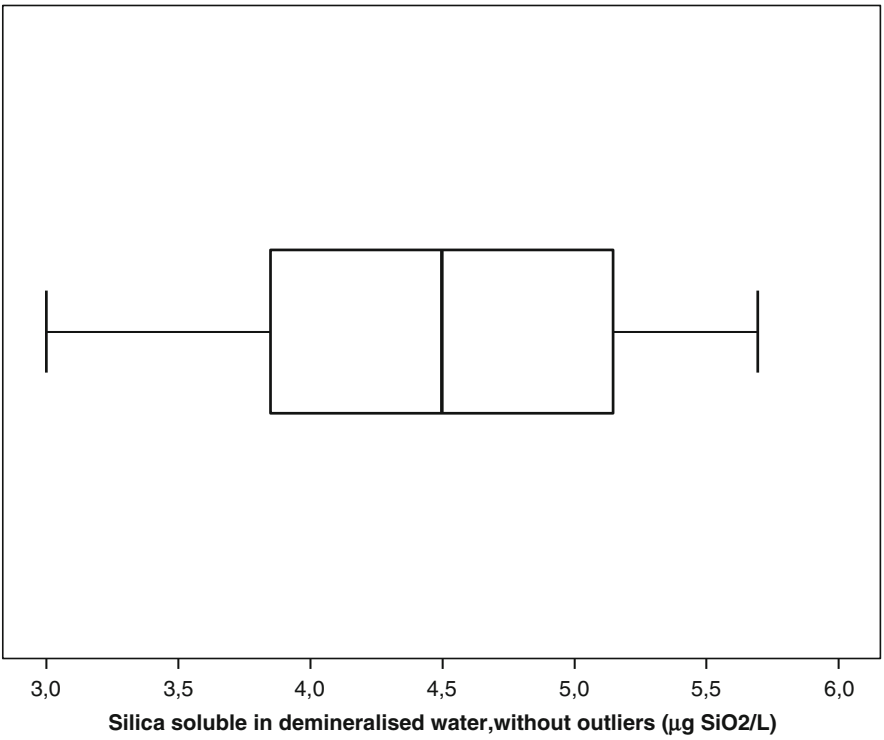


Fig. 3 Box-plot of silica in water (μg SiO_2/L), without outliers

Table 7 The Shapiro-Wilk normality test of silica, without outliers (μg SiO_2/L)

Silica	Shapiro-Wilk		
	Statistic	df	*p*-value
Dataset	0.985	8	0.984

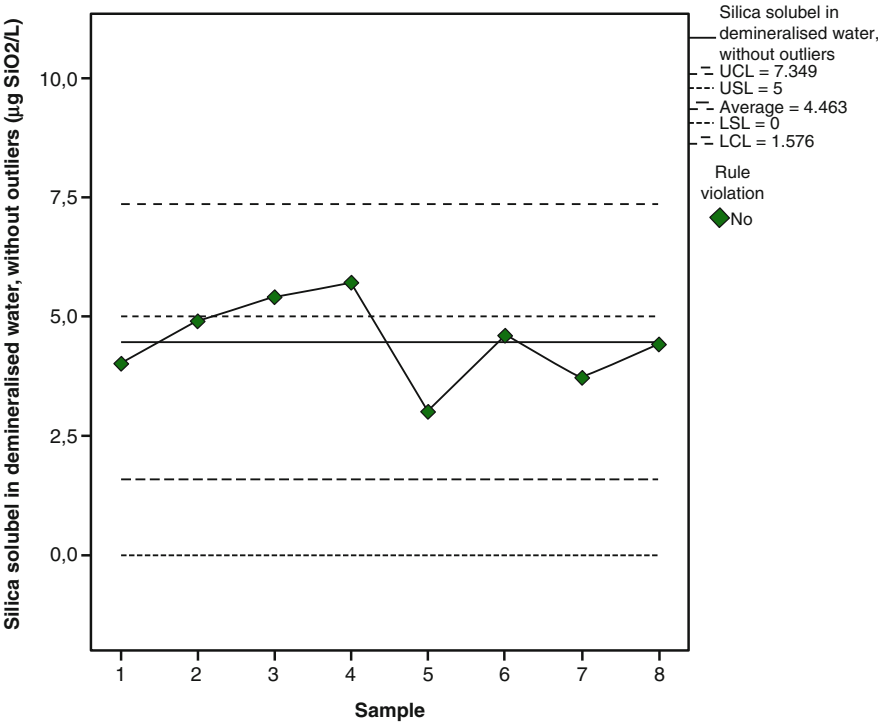


Fig. 4 Individual control chart of silica in water (µg SiO_2/L), without outliers

Table 8 Control limits for individual charts of silica (µg SiO_2/L), without outliers

Control limits			
Method	*LCL*	*UCL*	Range of limits
AMR	1.576	7.349	5.773
EQ	3.588	5.331	1.743

In Fig. 4 we have the X control chart for silica in water after removing two outliers, where the estimates of control limits LCL_{AMR} and UCL_{AMR} are obtained using the mean value in Table 6 and the average moving range value $\overline{MR} \approx 1.086$ (applying (2) to the data in Table 1, after removing the two outliers mentioned before). Now, the average line is inside of specification limits and we do not have points violating the control rules. Moreover, two points are outside the specification limits (i.e., 25%). In Table 8 we also have the estimates of the control limits using the method EQ, where we draw 5000 bootstrapped samples, with the same size of 8. As expected, when we compare the results obtained with outliers (Table 5), the range of the estimated control limits decreases (Table 8).

5 Process Capability

To analyze the ability of the process to produce outputs within specification limits we consider some statistics measures of process capability. In Table 9 we have the values computed with the following well-known capability indices, which are ratios of the process spread and specification spread:
C_p—capability of the process is the ratio of the difference between the specification limits and the observed process variation whose estimate is obtained by

$$\hat{C}_p = \frac{USL - LSL}{6 \times \frac{\overline{MR}}{d_2(2)}};$$

C_{pk}—capability of the process related to both dispersion and centeredness whose estimate is given by

$$\hat{C}_{pk} = \min\left\{\frac{USL - \overline{x}}{3 \times \frac{\overline{MR}}{d_2(2)}}, \frac{\overline{x} - LSL}{3 \times \frac{\overline{MR}}{d_2(2)}}\right\};$$

k—measure the deviation of the process mean from the midpoint of the specification limits which is given by

$$\hat{k} = \frac{|m - \overline{x}|}{(USL - LSL)/2} \text{ where } m = \frac{USL + LSL}{2}.$$

Since these ratios are unitless values we use them to compare the capability of the process with or without outliers (Table 9). Given that $C_p < 1$ in both cases, the process is too variable (in particular for the original dataset). Note that, according to many practitioners a value of C_p less than 1 is unacceptable (usually they consider 1.33 as a minimum acceptable).

When $k = 0$ the process is perfectly centered, once the mean is the same as the midpoint. Thus, we can consider that "k" quantifies the amount of which a distribution is centered (the minimum value of "k" is 0). The large value of k,

Table 9 The capability indices estimated for individual charts of silica (μg SiO_2/L)

Capability indices	Original dataset	Dataset without outliers
C_p^a	0.661	0.866
k	1.072	0.785
C_{pk}^a	−0.048	0.186

The normal distribution is assumed. $LSL = 0$ and $USL = 5$
[a]The estimated capability sigma is based on the mean of the sample moving ranges

especially for the original dataset, combined with a small value of C_p (estimated), indicates that the process does not stay within the specification.

Since the estimates of C_{pk} are very small for both cases (the value obtained with the original dataset is negative), then the process mean falls outside of the specification limits.

Processes that are in control should have a process capability that is near the process performance and although we know that it is possible to have data that falls outside the specification limits (LSL, USL) and still have a capable process, here the process is barely capable of consistently meeting the requirements (even for the case of the dataset without outliers).

6 Final Remarks

As in all processes, in the measurement of silica in water there is also some variability. This way, the control charts allow for quick identification of irregularities and possible intervention and correction, reducing costs and extending the useful life of equipment. The individual control charts and the approaches considered here to obtain different control limits (with average moving range and empirical quantiles methods) are very important for the members of the Chemical Laboratory of the thermoelectric central, once they allow to evaluate the silica variable, despite the small sample size available.

The process capability also gives considerable information on how much the process should be improved, since here it is necessary to continually try to minimize the variation of the process.

The comparison of results, with(out) outliers in dataset, shows their impact in the values of the control limits and also in the capability indices.

Given the results achieved and their importance to reduce the expenses on the maintenance of the equipment and piping, it was possible to sensitize the thermoelectric technicians to collect measurements of silica with higher frequency in the future (despite the involved costs), making possible to check the results obtained here in order to have a better control of the entire process.

Acknowledgements This work was partially supported by the Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/ /2013 (Centro de Matemática e Aplicações).

References

1. Chen, Y.: A new type of Bayesian nonparametric control charts for individual measurements. J. Stat. Theory Pract. **10**, 226–238 (2016)
2. Duncan, A.J.: Quality Control and Industrial Statistics, 5th edn. Irwin Homewood, IL (1986)

3. Grilo, L.M., Grilo, H.L.: Individual and moving range control charts in the production of olive oil. AIP Conf. Proc. **1648**, 1–4 (2015)
4. Grilo, L.M., Grilo, H.L.: Comparison of individual charts to monitor peroxide index of olive oil. Pro. Adv. Math. Comput. Sci. Appl. (**57**), 272–275 (2016)
5. Grilo, L.M., Mateus, D.M.R., Alves, A.C., Grilo, H.L.: Robust control charts in industrial production of olive oil. AIP Conf. Proc. **1618**, 539–542 (2014)
6. Grilo, L.M., Grilo, H.L., Marques, C.J.: Industrial production of gypsum: quality control charts. In: Theory and Practice of Risk Assessment: ICRA 5, Tomar, Portugal, 2013. Springer Proceedings in Mathematics & Statistics, vol. 136, pp. 225–234. Springer, Berlin (2015)
7. Grilo, L.M., Silva, D.S., Nogueira, I.M., Grilo, H.L., Oliveira, T.A.: Individual control charts in paperboard industry. AIP Conf. Proc. **1790**, 140009 (2016)
8. Ning, W., Yeh, A.B., Wu, X., Wang, B.: A nonparametric phase I control chart for individual observations based on empirical likelihood ratio. Qual. Reliab. Eng. Int. **31**, 37–55 (2015)
9. Vermaat, M.B., Does, R.J.M.M., Klaassen, C.A.J.: A comparison of Shewhart individuals control charts based on normal, non-parametric, and extreme-value theory. Qual. Reliab. Eng. Int. **19**(4) 337–353 (2003)
10. Wheeler, D.J.: Advanced Topics in Statistical Process Control. SPC Press, Knoxville (1995)
11. Willemain, T.R., Runger, G.C.: Designing control charts using an empirical reference distribution. J. Qual. Technol. **28**(1), 31–38 (1996)

Revisiting Resampling Methods in the Extremal Index Estimation: Improving Risk Assessment

D. Prata Gomes and M. M. Neves

Abstract Extreme value theory is an area of primordial importance for modelling extreme risks, allowing to estimate and predict beyond the range of data available. Among several parameters of interest, the *extremal index* is a crucial parameter in a dependent set-up, characterizing the degree of local dependence in the extremes of a stationary sequence. Its estimation has been addressed by several authors but some difficulties still remain. Resampling computer intensive methodologies have been recently considered in a reliable estimation of parameters of rare events. However classical bootstrap cannot be applied and block bootstrap procedures need to be considered. The block size for resampling strongly affects the estimates and needs to be properly chosen. Here, procedures for the choice of the block size for resampling are revisited and an improvement of the methods used in previous works for that choice is also considered. A simulation study will illustrate the performance of the aforementioned procedures. A real application is also presented.

1 Introduction and Basic Notions

Extreme risks are associated to very bad outcomes, have disastrous impact and occur with a low probability. They appear in areas such as natural disasters, security, ecology, market risks, as a few examples. These events are by definition unusual, so they are a challenge for statisticians that look for adequate models for quantifying the intensity, the extent of extreme data, eventually beyond the range of available

D. Prata Gomes (✉)
Centro de Matemática e Aplicações (CMA), FCT, UNL, Lisboa, Portugal

Departamento de Matemática, FCT, UNL, Lisboa, Portugal
e-mail: dsrp@fct.unl.pt

M. M. Neves
Instituto Superior de Agronomia (ISA), and CEAUL, Universidade de Lisboa, Lisboa, Portugal
e-mail: manela@isa.ulisboa.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_10

data. *Extreme Value Theory* (EVT) is an area of primordial importance providing models that allow to estimate and predict beyond the range of data available.

In its beginnings EVT found applications in areas such as structural engineering, environment, climate, hydrology, ocean engineering, material strength. Nowadays many other areas such as finance, internet traffic, biology, ecology have revealed the importance of using EVT for modeling extreme occurrences and for assessing the associated risks.

The central result in classical EVT states that given the sample $(X_1, \ldots, X_n)$ of independent and identical (i.i.d.) random variables, with a distribution function (d.f.) F, if there are sequences $\{a_n > 0\}$ e $\{b_n\}$ such that

$$\lim_{n\to\infty} P\left(\left(X_{n:n} - b_n\right)/a_n \le x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \tag{1}$$

$\forall x \in R$, where $X_{n:n} \equiv \max(X_1, \ldots, X_n)$ and G is a nondegenerate distribution function. Then $G \equiv \mathrm{EV}_\xi$ is the so-called *Extreme Value* (EV) d.f., given by

$$\mathrm{EV}_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), & 1+\xi x > 0 \text{ if } \xi \neq 0 \\ \exp(-\exp(-x)), & x \in R \qquad\quad \text{if } \xi = 0. \end{cases} \tag{2}$$

The EV d.f., in (2), incorporates the three Fisher–Tippett types: the Gumbel for $\xi = 0$, the limit for exponentially-tailed distributions; the Fréchet family, for $\xi > 0$, the limit for negative polynomial heavy-tailed distributions and the Weibull family for $\xi < 0$, the limit for short-tailed distributions. The shape parameter ξ is called *the extreme value index* and it measures the heaviness of the right-tail, $\overline{F} := 1 - F$. As ξ increases the right tail becomes heavier.

Other parameters are also of great interest such as: the *probability of exceedance* of a high level, the *return period* of a high level, the *right endpoint* of an underlying model F or a *high quantile* of probability $1 - p$ (p small).

The limit result in (1) was derived for i.i.d random variables. However, in many practical applications a more realistic situation is the one where dependent observations can appear, e.g. extremes conditions often persist over several consecutive observations. The most natural generalization of an independent case is the dependent setup—where the variables may be mutually dependent, but whose stochastic properties are homogeneous through time.

Suppose that $\{X_n\}_{n\ge1}$ is a strictly stationary sequence of random variables with marginal d.f. F and $\{Y_n\}_{n\ge1}$ is an i.i.d. sequence with the same parent d.f. F. From [27] a strictly stationary sequence, $\{X_n\}_{n\ge1}$, is said to have an *extremal index* (EI), $\theta \in (0, 1]$, if for all $\tau > 0$, there exists a sequence of thresholds $u_n \equiv u_n(\tau)_{n\ge1}$ such that

$$P\left(Y_{n:n} \le u_n\right) = F^n(u_n) \underset{n\to\infty}{\longrightarrow} e^{-\tau} \quad \text{and} \quad P\left(X_{n:n} \le u_n\right) \underset{n\to\infty}{\longrightarrow} e^{-\theta\tau},$$

where $Y_{n:n} \equiv \max(Y_1, \ldots, Y_n)$.

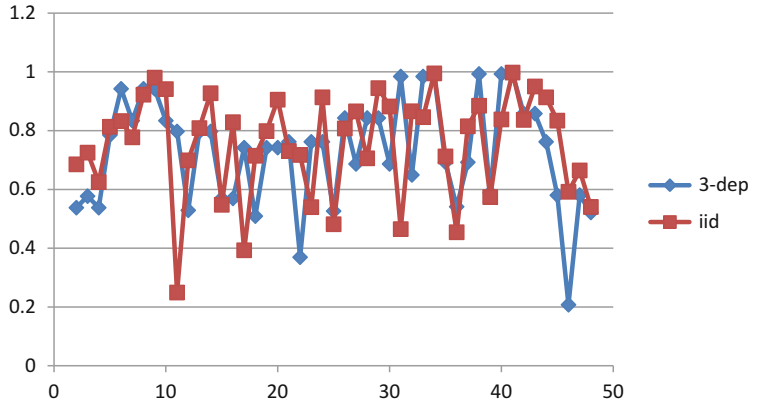


Fig. 1 One realization of an i.i.d. process and a 3-dependent process

For illustration of the behaviour of a stationary process for some values of θ let us consider the following examples:

Example 1 A *Moving Maximum Process* I, [10].

Let $\{Y_n\}_{n\geq -2}$ be a sequence of i.i.d. uniform variables on $[0, 1[$ with F the common d.f.. Let $\{X_n\}_{n\geq 1}$ be the *3-dependent moving maxima* sequence, defined as

$$\text{[M1]} \qquad X_n = \max(Y_{n-3}, Y_{n-1}, Y_n), \quad n \geq 1. \tag{3}$$

The marginal underlying distribution for X_n is F^3 and for X_n we have $\theta = 1/3$. Consider also $\{Z_n\}_{n\geq 1}$ an i.i.d. sequence with the same d.f. F^3.

Figure 1 shows one realization of X_n and Z_n. Three-sized clusters of exceedances of high levels can be seen.

See a *shrinkage* of the largest observations for the 3-dependent sequence.

Example 2 : A *Max-Autoregressive Process* I, [2].

Let $\{Y_n\}_{n\geq 1}$ be a sequence of i.i.d., with d.f. standard Fréchet. For $0 < \theta \leq 1$, let

$$\text{[M2]} \qquad X_1 = Y_1, \quad X_n = \max\{(1-\theta)X_{n-1}, \theta Y_n\} \quad n \geq 2. \tag{4}$$

For $u_n = nx,\ x > 0,\ P\left(X_{n:n} \leq u_n\right) \to \exp\left(-\theta/x\right)$, as $n \to \infty$, so the EI of the sequence is θ.

Figure 2 shows partial realizations of the process M2 with $\theta = 0.9;\ 0.5$ and 0.1, respectively. The maxima show increasing clustering as $\theta \to 0$.

A result relating the limiting distribution of the maxima of a stationary sequence $\{X_n\}_{n\geq 1}$, provided that it has limited long-range dependence at extreme levels, and the associated independent sequence, $\{Y_n\}_{n\geq 1}$, was established in Theorem 2.5, [27].

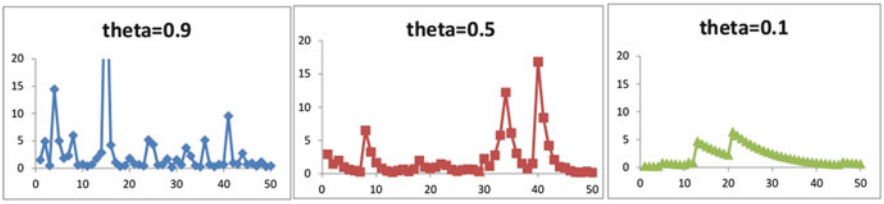


Fig. 2 One realization of the M2 process with $\theta = 0.9;\ 0.5$ and 0.1

It states that under the $D(u_n)$ condition (see [26, 29]), with $u_n = a_n x + b_n$, as $n \to \infty$,

$$P\left(Y_{n:n} \le u_n\right) \longrightarrow G_1(x) \quad \text{if and only if} \quad P\left(X_{n:n} \le u_n\right) \longrightarrow G_2(x),$$

where $G_2 = G_1^{\theta}$, for a constant θ such that $0 < \theta \le 1$.

So, given that $G_1 = \mathrm{EV}_{\xi}$, an EV d.f., with location, scale and shape parameters (μ, σ, ξ), the limit law $G_2 = \mathrm{EV}_{\xi}^{\theta}$, is also an EV d.f. with location, scale and shape parameters $(\mu_\theta, \sigma_\theta, \xi_\theta)$ given by

$$\mu_\theta = \mu - \sigma \frac{1 - \theta^{\xi}}{\xi}, \qquad \sigma_\theta = \sigma \theta^{\xi}, \qquad \xi_\theta = \xi.$$

This parameter, θ, affects then other parameters of extreme events, but it has its own importance because it measures the relationship between the dependence structure of the data and the behaviour of the exceedances over a high threshold u_n, being directly related to the clustering of exceedances. Actually, $\theta = 1$ for i.i.d. sequences and $\theta \to 0$ whenever dependence increases. Notice a "shrinkage of maximum values" as dependence increases in Fig. 2.

The EI estimation, such as the estimation of other parameters of extreme values, is usually done in a semi-parametric approach based on the probabilistic asymptotic results in the tail of the unknown distribution. However several difficulties appear. Those semi-parametric estimates are usually performed on the basis of the largest k order statistics in the sample, and the estimators show strong dependence on that value k, with a high variance for small values of k and a high bias for large values of k. This brings a real need for the adequate choice of k, one of the problems here addressed.

Jackknife and Bootstrap procedures have revealed to give good results in the reduction of the bias of an estimator allowing to obtain more stable paths of the estimates for improving the estimation of the parameters of extreme events, see [5, 13, 15, 16, 36–38], to mention a few works.

After a brief reference to some EI estimators and their asymptotic properties, the goal of this work is to improve the performance of those estimators through computational procedures based on resampling blocks of observations, that need to be adequately chosen. Actually, the i.i.d. nonparametric bootstrap needs to be

adapted to the dependent context. One of the procedures proposed for the bootstrap resampling in this situation, see, for example, [24], consists of defining blocks for resampling, instead of resampling the individual observations. But the performance of the bootstrap estimator crucially depends on the block size that must be supplied by the user. Several authors such as [3, 19] and [25] proposed ways of estimating the optimal block size. Here we follow [25], who proposed a nonparametric plug-in (NPPI) method for the empirical choice of the optimal block size for the block bootstrap estimation of some characteristics of an estimator. Some simulation results and applications have been shown in [36, 37] and [38], for the bias and [39], for the variance of an estimator. Here, the bias and the variance are dealt together considering the *Mean Squared Error*. Some improvements on the previous works have been obtained. Some results from a complete simulation study are shown as well a real application in an important area of risk analysis—the financial area.

2 Semiparametric Estimation of the EI

Several interpretations of θ have appeared, providing several suggestions for its estimation. The most common interpretation of θ is that as being *the reciprocal of the "mean time of duration of extreme events"*, i.e., $\theta = 1/$*limiting mean cluster size*, which is directly related to the exceedances of high levels, see [22, 28].

Identifying clusters by the occurrence of downcrossings or upcrossings, we can write

$$\theta = \lim_{n\to\infty} P\big(X_2 \le u_n | X_1 > u_n\big) = \lim_{n\to\infty} P\big(X_2 > u_n | X_1 \le u_n\big). \tag{5}$$

Given a sample, $(X_1, \ldots, X_n)$, the empirical counterpart of the above interpretation led to the classical up-crossing (down-crossing) estimator, UC-estimator, $\widehat{\Theta}^{UC}$, (DC-estimator, $\widehat{\Theta}^{DC}$), see [11, 12, 31],

$$\begin{aligned}\widehat{\Theta}^{UC}(u_n) &:= \frac{\sum_{i=1}^{n-1} I\,(X_i \le u_n < X_{i+1})}{\sum_{i=1}^{n} I(X_i > u_n)} \\ &\equiv \frac{\sum_{i=1}^{n-1} I\,(X_i > u_n, X_{i+1} \le u_n)}{\sum_{i=1}^{n} I(X_i > u_n)} := \widehat{\Theta}^{DC}(u_n),\end{aligned} \tag{6}$$

for a suitable threshold u_n, where $I(A)$ denotes, as usual, the indicator function of A. Consistency of this estimator is obtained provided that the high level u_n is a normalized level, i.e. if with $\tau \equiv \tau_n > 0$, the underlying d.f. F verifies

$$F(u_n) = 1 - \tau/n + o(1/n), \quad n \to \infty \text{ and } \tau/n \to 0.$$

Other forms of identifying clusters have motivated other estimators. Let us mention two very popular estimators: *the blocks estimator* and *the runs estimator*. For the definition, properties and detailed study of these estimators, see [20, 21, 41, 42].

2.1 Improving the Estimation of the EI Through the Generalized Jackknife Methodology

Let us focus our attention in the $\widehat{\Theta}^{UC}$ in (6). This estimator shows a very strong bias and a very sharp *Mean Squared Error*

$$\text{MSE}\left(\widehat{\Theta}^{UC}\right) = E\left[(\widehat{\Theta}^{UC} - \theta)^2\right] = \text{Bias}^2\left(\widehat{\Theta}^{UC}\right) + \text{Var}\left(\widehat{\Theta}^{UC}\right),$$

which reveals a need for a very accurate way of choosing k in order to obtain a reliable estimate of θ.

Gomes et al. [14] considered the Generalized Jackknife methodology (see [18]) that uses properties of the bias and the variance of any estimator for developing estimators with bias and mean squared error often smaller than those of an initial set of estimators and proposed reduced biased estimators of the EI. Given the sample $(X_1, \ldots, X_n)$ and the associated ascending order statistics, $X_{1:n} \leq \cdots \leq X_{n:n}$, these authors considered the deterministic level $u \equiv u_n$ substituted by the stochastic one, $X_{n-k:n}$, and wrote the UC-estimator, in (6), as a function of k,

$$\widehat{\Theta}^{UC} \equiv \widehat{\Theta}^{UC}(k) := \frac{1}{k}\sum_{i=1}^{n-1} I\,(X_i \leq X_{n-k:n} < X_{i+1}). \tag{7}$$

For some dependent structures, [14] showed that the bias of $\widehat{\Theta}^{UC}$ has two dominant components of orders k/n and $1/k$,

$$\text{Bias}[\widehat{\Theta}^{UC}(k)] = \varphi_1(\theta)\left(\frac{k}{n}\right) + \varphi_2(\theta)\left(\frac{1}{k}\right) + o\left(\frac{k}{n}\right) + o\left(\frac{1}{k}\right), \tag{8}$$

whenever $n \to \infty$ and $k \equiv k(n) \to \infty$, $k = o(n)$.

Using that information on the bias of $\widehat{\Theta}^{UC}$ in (7), [14] considered first a generalized jackknife EI estimator of order 2, based on $\widehat{\Theta}^{UC}$ computed at the three levels, k, $\lfloor k/2 \rfloor + 1$ and $\lfloor k/4 \rfloor + 1$, where $\lfloor x \rfloor$ denotes, as usual, the integer part of x. They got the estimator

$$\widehat{\Theta}^{GJ} \equiv \widehat{\Theta}^{GJ}(k) := 5\widehat{\Theta}^{UC}([k/2]+1) - 2\big(\widehat{\Theta}^{UC}([k/4]+1) + \widehat{\Theta}^{UC}(k)\big). \tag{9}$$

This is an asymptotically unbiased estimator of θ, in the sense that it can remove the two dominant components of bias referred to in (8).

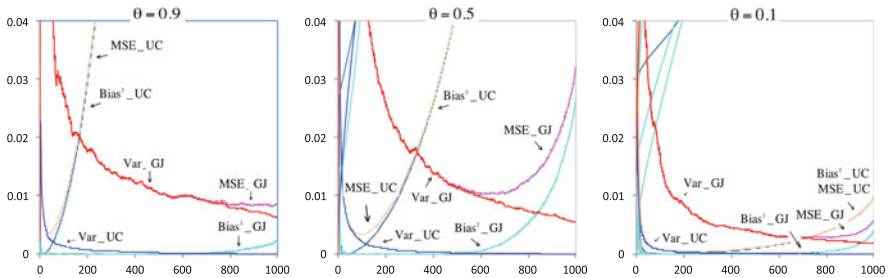


Fig. 3 MSE, Var and Bias^2 of $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ for $\theta = 0.9, 0.5, 0.1$ in the *Max-Autoregressive Process*, Example 2

More generally, [14] considered the levels k, $\lfloor \delta k \rfloor + 1$ and $\lfloor \delta^2 k \rfloor + 1$, depending on a *tuning parameter* δ, $0 < \delta < 1$, and the class of estimators,

$$\widehat{\Theta}^{GJ(\delta)}(k) := \frac{(\delta^2+1)\,\widehat{\Theta}^{UC}\left([\delta k]+1\right) - \delta\left(\widehat{\Theta}^{UC}\left([\delta^2 k]+1\right) + \widehat{\Theta}^{UC}(k)\right)}{(1-\delta)^2}. \tag{10}$$

Actually $\widehat{\Theta}^{GJ(1/2)}(k) \equiv \widehat{\Theta}^{GJ}$, in (9). Among the members of the class in (10), those authors have been heuristically led to the choice $\delta = 1/4$. Distributional properties of $\widehat{\Theta}^{GJ(1/4)}(k)$ have been obtained by [14] through simulation techniques, see also [17, 32] for a review.

In Fig. 3 the simulated bias, variance and MSE obtained through a Monte Carlo simulation with 1000 replicas, of $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ estimators, for a sample size $n = 1000$ from a *Max-Autoregressive Process*, Example 2, with extremal indexes $\theta = 0.9$; $\theta = 0.5$ and $\theta = 0.1$ are plotted.

Remark 1 A few remarks, also presented in [14, 33, 37–39] regarding these estimators can be pointed out:

- The $\widehat{\Theta}^{UC}$ estimator shows a very strong bias, that is the dominant component of the MSE.
- MSE($\widehat{\Theta}^{UC}$) is very sharp, which reveals a need for a very accurate way of choosing k in order to obtain a reliable estimate of θ.
- MSE($\widehat{\Theta}^{GJ}$) is not so sharp as MSE($\widehat{\Theta}^{UC}$), suggesting less dependence on the value k for obtaining the estimate of θ.
- $\widehat{\Theta}^{GJ}$ shows a more stable simulated mean value (not shown here, see, for example, [14, 32]), near the target value of the parameter but at expenses of a very high variance.

Recently, bootstrap procedures and the choice of the sample fraction for the semi-parametric estimation of parameters of extreme events have been considered. Let us refer to some recent works such as [5, 7, 13, 15, 17], to cite only a few.

3 Resampling Techniques for Stationary Sequences

In its classical form, the bootstrap methodology [8] has proven to be a powerful nonparametric tool when based on i.i.d. observations. But [40] showed that it could be inadequate under dependence. So the two different situations need to be taken into account: resampling from an i.i.d. sequence or resampling from a dependent sequence.

Several attempts have been made to extend the bootstrap procedure to the dependent case. The main result was achieved when resampling of single observations was replaced by block resampling. The motivation for this scheme is to preserve the dependence structure of the underlying model within each block. Several ways of blocking have been proposed: *Nonoverlapping Block Bootstrap* (NBB), [4]; *Moving Block Bootstrap* (MBB), [23, 30]; *Circular Block Bootstrap* (CBB), [34] and *Stationary Bootstrap* (SB), [35]. The first three methods consider to resample blocks of observations with nonrandom block length. The last one considers a random block length and hence, has a slightly more complicated structure.

For each way of blocking it is necessary to consider a length $b \equiv b(n)$ to resample blocks of observations, but the accuracy of block bootstrap estimators depends critically on the block size for resampling that must be supplied by the user (see [25]).

The nonparametric plug-in (NPPI) method for the empirical choice of the optimal block size for the block bootstrap estimation, proposed by [25], is used. The method employs nonparametric resampling procedures to estimate the relevant constants in the leading term of the "optimal" block length, so it does not require the knowledge and/or derivation of explicit analytical expressions for those constants. The method applied here is based on the Jackknife-After-Bootstrap (JAB) of [9] and [24].

In previous works the method was applied to control the bias and the variance through the block bootstrap estimator. Here it was extended to the MSE of the estimators $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ in (7) and (9), respectively.

3.1 The NPPI Procedure: A Brief Overview

Given the random sample $(X_1, \ldots, X_n)$, let $\widehat{\Theta}_n$ be an estimator of θ. Let us consider now the parameter $\phi_n \equiv \mathrm{MSE}(\widehat{\Theta}_n)$ and the corresponding block bootstrap estimator based on blocks of size b (we shall consider here the MBB),

$$\widehat{\phi}_n^*(b) = \mathrm{MSE}_* \left(\widehat{\Theta}_n^*(b)\right),$$

where MSE_* denotes the conditional mean squared error, given the data.

Lahiri et al. [25], Section 2, remember that for many population parameters (denoted here by ϕ_n), the variance of the corresponding block bootstrap estimator is an increasing function of the block size, b, while its bias is a decreasing function

of b, and that under suitable regularity conditions, the variance and the bias of the block bootstrap estimator admit expansions of the form

$$n^{2a}\ \mathrm{Var}\left(\widehat{\phi}_n^*(b)\right) = C_1 n^{-1} b + o\left(n^{-1} b\right), \tag{11}$$

$$n^{a}\ \mathrm{Bias}\left(\widehat{\phi}_n^*(b)\right) = C_2 b^{-1} + o\left(b^{-1}\right), \tag{12}$$

as $n \to \infty$ and over a suitable set $\kappa_n \subset \{2, \ldots, n\}$ of possible block length b. C_1, C_2 and a are constants depending on the characteristics of EI estimator under study. As for the bias and the variance of an estimator, [25] suggested $a = 1$, so C_1 and C_2, from (11) and (12), can be approximately given by

$$C_1 \sim n b^{-1} \mathrm{Var}\left(\widehat{\phi}_n^*(b)\right) \quad \text{and} \quad C_2 \sim b\ \mathrm{Bias}\left(\widehat{\phi}_n^*(b)\right).$$

A consistent estimation of $\mathrm{Var}\left(\widehat{\phi}_n^*(b)\right)$ and $\mathrm{Bias}\left(\widehat{\phi}_n^*(b)\right)$ was proposed, see [19, 25], as

$$\widehat{\mathrm{Var}}_n \equiv \widehat{\mathrm{VAR}}_{\mathrm{JAB}}(\widehat{\phi}_n^*(b)) \quad \text{and} \quad \widehat{\mathrm{Bias}}_n = 2\left(\widehat{\phi}_n^*(b) - \widehat{\phi}_n^*(2b)\right),$$

where $\widehat{\mathrm{VAR}}_{\mathrm{JAB}}$ is the Jackknife-After-Bootstrap variance estimator. Parameters C_1 and C_2 can be consistently estimated by

$$\widehat{C}_1 = n b^{-1} \widehat{\mathrm{VAR}}_{\mathrm{JAB}}(\widehat{\phi}_n^*(b)) \quad \text{and} \quad \widehat{C}_2 = 2b\left(\widehat{\phi}_n^*(b) - \widehat{\phi}_n^*(2b)\right).$$

References [19] and [25] showed that the optimal block size b_n^0 has the form

$$b_n^0 = \left(\frac{2C_2^2}{C_1}\right)^{1/3} n^{1/3}(1 + o(1)),$$

so the NPPI estimator of the optimal block size is then given by

$$\widehat{b}_n^0 = \left(\frac{2\widehat{C}_2^2}{\widehat{C}_1}\right)^{1/3} n^{1/3}. \tag{13}$$

The JAB methodology, allowing to assess the accuracy of bootstrap estimators for dependent data, was derived by [24]. The key step is to delete resampled blocks instead of blocks of original data values.

If $\widehat{\phi}_n^*(b)$ is the MBB estimator of ϕ_n and $\ell = n - b + 1$ the number of "observable" blocks of length b:

- Let m be an integer such that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$, denoting the number of bootstrap blocks to be deleted.

- Write $M = \ell - m + 1$ and for $i = 1, \ldots, M$ let us define the set $I_i = \{1, \ldots, \ell\} \backslash \{i, \ldots, i + m - 1\}$, denoting the index set of all blocks obtained by deleting the m blocks.
- Resample $[n/b]$ from the reduced collection $\{B_j : j \in I_i\}$ and compute the ith jackknife block-deleted estimate, $\widehat{\phi}_n^{*i} \equiv \widehat{\phi}_n^{*i}(b), i = 1, \ldots, M$.

The JAB variance estimator for $\widehat{\phi}_n^*(b)$ is then defined as

$$\widehat{\mathrm{VAR}}_{\mathrm{JAB}}(\widehat{\phi}_n^*(b)) = \frac{m}{(\ell - m)M} \sum_{i=1}^{M} \left[\widetilde{\phi}_n^{*i}(b) - \widehat{\phi}_n^*(b)\right]^2, \tag{14}$$

where $\widetilde{\phi}_n^{*i}(b) = m^{-1}\left[\ell\widehat{\phi}_n^*(b) - (\ell - m)\widehat{\phi}_n^{*i}(b)\right]$ is the ith block-deleted jackknife pseudo-value of $\widehat{\phi}_n(b), i = 1, \ldots, M$.

According to suggestions given in [25] to obtain $\widehat{C}_1$ and $\widehat{C}_2$, as a first approach we used $b = \lceil n^{1/5} \rceil$ and $m = \lceil n^{1/3} b^{2/3} \rceil$.

4 Finite Sample Behaviour for Simulated Data and a Real Data Set

From the models described below, samples of sizes of $n = 500$ and $n = 1000$ were generated for some values of the parameters. We have performed a simulation study on the bases of 1000 replicates. Table 1 shows, for each model and values of the parameters considered in the simulation, the results obtained from the application of the NPPI procedure and the bootstrap estimates $\widehat{\theta}^{\star UC}$ and $\widehat{\theta}^{\star GJ}$ calculated through an adaptive procedure that chooses the sample fraction for obtaining the estimated EI based on a stability criterion, for the choice of a k_{opt}, see the description and some illustrative examples in [33].

Table 1 $\widehat{\theta}^{\star UC}$ and $\widehat{\theta}^{\star GJ}$ estimates of θ for samples simulated from models M1, M2, M3 and M4, b_{opt} block size and the associated block bootstrap estimates (with b_{opt}), calculated through the stability criterion, for $n = 500$, and $n = 1000$

Models	$n = 500$ ($b_{\mathrm{ini}} = 3$)				$n = 1000$ ($b_{\mathrm{ini}} = 4$)			
	b_{opt} for $\widehat{\theta}^{\star UC}$		b_{opt}	$\widehat{\theta}^{\star GJ}$	b_{opt} for $\widehat{\theta}^{\star UC}$		b_{opt}	$\widehat{\theta}^{\star GJ}$
M1 ($\theta = 1/3$)	32	–	19	0.8345	74	–	48	0.5804
M2 ($\theta = 0.5$)	49	–	28	0.5718	73	–	30	0.4981
M2 ($\theta = 0.9$)	20	–	5	0.9018	16	–	15	0.8177
M3 ($\theta = 0.5$)	58	–	40	0.5260	91	–	70	0.5047
M3 ($\theta = 0.909$)	21	–	4	0.8975	16	–	13	0.9570
M4 ($\theta = 0.5$)	46	–	23	0.5846	77	–	26	0.4987
M4 ($\theta = 0.9$)	18	–	8	0.9880	29	–	34	0.8941

When applying the NPPI method, the estimated "optimal" block length depends on the value of k. So, the mode of the estimated block size values was adopted as the "optimal" block length, b_{opt}, for resampling and the block bootstrap estimates were obtained using the sample generated and that value, b_{opt}. The first values for $b \equiv b_{\text{ini}}$ to initialize the NPPI procedure were $b_{\text{ini}} = 3 (n = 500)$ and $b_{\text{ini}} = 4 (n = 1000)$, as previously referred to.

- **Models M1** and **M2** defined in (3) and (4), respectively.
- **Model M3**, *Moving Maximum Process II*, [6].
  Let $\{Y_n\}_{n\geq 0}$ be a sequence of i.i.d., with d.f. standard Fréchet. For $a \geq 0$, let

$$\text{[M3]} \qquad X_0 = Y_0, \quad X_n = (a+1)^{-1} \max\left\{aY_{n-1}, Y_n\right\}, \quad n = 1, 2, \ldots. \tag{15}$$

  If $a \leq 1$ we have $\theta = 1/(a+1)$, otherwise $\theta = a/(a+1)$. Then $\theta = \max\{1, a\}/(a+1)$ and $1/2 \leq \theta \leq 1$.
- **Model M4**, *Max-Autoregressive Process II*, [1].
  Let $\{Y_n\}_{n\geq 1}$ be a sequence of i.i.d., with d.f. standard Fréchet and X_0 a random variable with d.f. $H_0(z) = \exp\left(-z^{-1}(\beta^{-1} - 1)\right)$. For $0 < \beta < 1$, let

$$\text{[M4]} \qquad X_n = \beta \max\left\{X_{n-1}, Y_n\right\}, \quad n = 1, 2, \ldots. \tag{16}$$

  The EI of this process is $\theta = 1 - \beta$.

$\widehat{\theta}^{UC}$ estimates show a very strong bias, so the block bootstrap estimates, although revealing a more smooth pattern, cannot remove such a heavy bias (see Fig. 4 as an illustration). For this reason estimates will not be given. On the other side, the block bootstrap estimator $\widehat{\Theta}^{\star GJ}$ reveals itself as a very promising one, with more stable paths around the true value of the parameter, in the most situations (see Fig. 4 (left, center)).

A real data set, in the field of Finance was also used for illustrating the procedures. The data refers to the *Daily Adjusted Closing Price for S&P 500 index*,

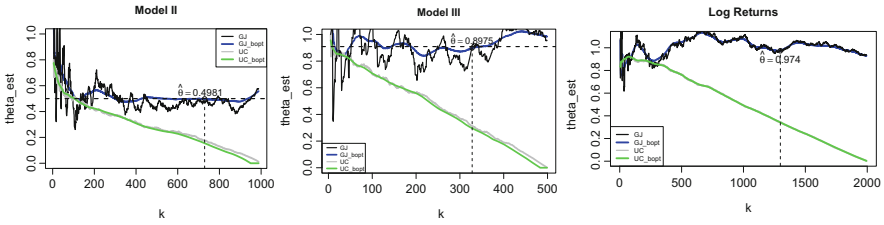


Fig. 4 Sample paths (estimates and bootstrap estimates with b_{opt}) for θ in Model 2 ($\theta = 0.5$), Model 3 ($\theta = 0.9090$) and for the real data set of *log-returns*, from left to right. $\widehat{\theta}^{\star GJ}$ represented were calculated through the stability criterion. For the real data set of *log-returns*, $\widehat{\theta}^{\star GJ}(k_{\text{opt}}) = 0.974(1294)$

x_t from October 21, 1982, until October 2, 1990 ($n = 2000$). As usual, in these studies, we will deal with the *Log-Returns*, $r_t = \ln x_t - \ln x_{t-1}$.

Figure 4 (right) shows the estimates $\widehat{\theta}^{GJ}$ and $\widehat{\theta}^{UC}$ sample paths, the associated $\widehat{\theta}^{\star GJ}$ and $\widehat{\theta}^{\star UC}$ calculated with $b_{\text{opt}} = 15$ and $b_{\text{opt}} = 16$, respectively, and is represented the estimate calculated with the procedure mentioned above.

Acknowledgements This research is partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2013 (CEA/UL) and PEst-OE/MAT/UI0297/2013 (CMA/UNL).

References

1. Alpuim, M.: An extremal Markovian sequence. J. Appl. Prob. **26**, 219–232 (1989)
2. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.L.: Statistics of Extremes. Theory and Applications. Wiley, Chichester (2004)
3. Bühlmann, P., Künsch, H.: Block length selection in the bootstrap for time series. J. Comput. Stat. Data Anal. **31**, 295–310 (1999)
4. Carlstein, E.: The use of subseries methods for estimating the variance of a general statistics from a stationary times series. Ann. Stat. **14**, 1171–1179 (1986)
5. Danielsson, J., de Haan, L., Peng, L., de Vries, C.G.: Using a bootstrap method to choose the sample fraction in the tail index estimation. J. Multivar. Anal. **76**, 226–248 (2001)
6. Davison, A.: Statistics of Extremes. Courses 2011–2012. École Polytechnique Fédérale de Lausanne, Lausanne (2011)
7. Draisma, G., de Haan, L., Peng, L., Pereira, T.: A bootstrap-based method to achieve optimality in estimating the extreme value index. Extremes **2**(4), 367–404 (1999)
8. Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)
9. Efron, B.: Jackknife-after-bootstrap standard errors and influence functions (with discussions). J. R. Stat. Ser. B **54**, 83–111 (1992)
10. Ferreira, H.: The upcrossing index and the extremal index. J. Appl. Probab. **43**, 927–937 (2006)
11. Gomes, M.I.: Statistical inference in an extremal Markovian Model. In: COMPSTAT 1990, pp. 257–262. Physica-Verlag, Heidelberg (1990)
12. Gomes, M.I.: On the estimation of parameters of rare events in environmental time series. In: Barnett, V., Feridun Turkman, K. (eds.) Em Statistics for Environment, pp. 225–241. Wiley, New York (1993)
13. Gomes, M.I., Oliveira, O.: The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction. Extremes **4**(4), 331–358 (2001)
14. Gomes, M.I., Hall, A., Miranda, C.: Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. J. Comput. Statist. Data Anal. **52**, 2022–2041 (2008)
15. Gomes, M.I., Mendonça, S., Pestana, D.: Adaptive reduced-bias tail index and VaR estimation via the bootstrap methodology. Commun. Stat. Theory Methods **40**(16), 2946–2968 (2011)
16. Gomes, M.I., Martins, M.J., Neves, M.M.: Generalised Jackknife-based estimators for univariate extreme-value modeling. Commun. Stat. Theory Methods **42**(7), 1227–1245 (2013)
17. Gomes, M.I., Figueiredo, F., Martins, M.J., Neves, M.M.: Resampling methodologies and reliable tail estimation. S. Afr. Statist. J. **49**, 1–20 (2015)
18. Gray, H.L., Schucany, W.R.: The Generalized Jackknife Statistic. Marcel Dekker, New York (1972)
19. Hall, P., Horowitz, J.L., Jing, B.-Y.: On blocking rules for the bootstrap with dependent data. Biometrika **50**, 561–574 (1995)
20. Hsing, T.: Estimating the parameters of rare events. Stoch. Process. Appl. **37**, 117–139 (1991)

21. Hsing, T.: Extremal index estimation for a weakly dependent stationary sequence. Ann. Stat. **21**, 2043–2071 (1993)
22. Hsing, J.T., Hüsler, J., Leadbetter, M.R.: On the exceedance point process for a stationary sequence. Probab. Theory Relat. Fields **78**(1), 97–112 (1988)
23. Künsch, H.: The jackknife and the bootstrap for general stationary observations. Ann. Math. **17**, 1217–1241 (1989)
24. Lahiri, S.: On the jackknife after bootstrap method for dependent data and its consistency properties. Economet. Theor. **18**, 79–98 (2002)
25. Lahiri, S., Furukawa, K., Lee, Y.-D.: Nonparametric plug-in method for selecting the optimal block lengths. Stat. Methodol. **4**, 292–321 (2007)
26. Leadbetter, M.R.: On extreme values in stationary sequences. Z. Wahrsch. Verw. Gebiete **28**, 289–303 (1974)
27. Leadbetter, M.R.: Extremes and local dependence in stationary sequences. Z. Wahrsch. Verw. Gebiete **65**(2), 291–306 (1983)
28. Leadbetter, M.R., Nandagopalan, L.: On exceedance point process for stationary sequences under mild oscillation restrictions. In: Hüsler, J., Reiss, R.D. (eds.) Extreme Value Theory. Proceedings, Oberwolfach 1987. Lecture Notes in Statistics, vol. 52, pp. 69–80. Springer, Berlin (1989)
29. Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and related properties of random sequences and series. Springer, New York (1983)
30. Liu, R., Singh, K.: Moving blocks jackknife and bootstrap capture weak dependence. In: Lepage, R., Billard, L. (eds.) Exploring the Limits of Bootstrap, pp. 225–248. Wiley, New York (1992)
31. Nandagopalan, S.: Multivariate extremes and estimation of the extremal index. PhD Thesis, University of North Carolina, Chapel Hill (1990)
32. Neves, M.M.: Bootstrap and Jackknife methods in extremal index estimation: a review. In: Gonçalves, E., Oliveira, P.E., Tenreiro, C. (eds.) Contributions in Statistics and Inference: Celebrating Nazaré Mendes Lopes' Birthday. Textos de Matemática, vol. 47, pp. 49–66. DMUC, Coimbra (2015)
33. Neves, M.M., Gomes, M.I., Figueiredo, F., Prata Gomes, D.: Modeling extreme events: sample fraction adaptive choice in parameter estimation. J. Stat. Theory Pract. **9**(1), 184–199 (2015)
34. Politis, D.R., Romano, J.P.: A circular block-resampling procedure for stationary data. In: Lepage, R., Billard, L. (eds.) Exploring the Limits of Bootstrap, pp. 263–270. Wiley, New York (1992)
35. Politis, D.R., Romano, J.P.: The stationary bootstrap. J. Am. Stat. Assoc. **89**(428), 1303–1313 (1994)
36. Prata Gomes, D., Neves, M.M.: Resampling methodologies and the estimation of parameters of rare events. In: Numerical Analysis and Applied Mathematics 2011 (ICNAAM 2011), AIP Conference Proceedings, vol. 1389, pp.1475–1478 (2011)
37. Prata Gomes, D., Neves, M.M.: Bootstrap and other resampling methodologies in statistics of extremes. Commun. Stat. Simul. Comput. **44**(10), 2592–2607 (2015)
38. Prata Gomes, D., Neves, M.M.: Adaptive choice and resampling techniques in extremal index estimation. In: Kitsos, C., Oliveira, T., Rigas, A., Gulati, S. (eds.) Theory and Practice of Risk Assessment. Proceedings in Mathematics and Statistics, pp. 321–332. Springer, Berlin (2015)
39. Prata Gomes, D., Neves, M.M.: Computer intensive methods for improving the extremal index estimation. In: Numerical Analysis and Applied Mathematics 2014 (ICNAAM 2014), AIP Conference Proceedings, vol. 1648, pp. 540005-1–540005-4 (2015)
40. Singh, K.: On the asymptotic accuracy of the Efron's bootstrap. Ann. Stat. **9**, 345–362 (1981)
41. Smith, R., Weissman, I.: Estimating the extremal index. J. R. Stat. Soc. B **56**, 515–528 (1994)
42. Weissman, I., Novak, S.: On blocks and runs estimators of the extremal index. J. Stat. Plann. Inf. **66**, 281–288 (1998)

Improving Asymptotically Unbiased Extreme Value Index Estimation

Frederico Caeiro, Ivanilda Cabral, and M. Ivette Gomes

Abstract The extreme value index characterizes the tail behaviour of a distribution, and indicates the size and frequency of certain extreme events under a given probability model. In this work, we are interested in improvements attained through the reduction of bias of the extreme value index estimators related to Lehmer's mean of the log-excesses. A comparison with other reduced bias estimators, namely the corrected-Hill estimator, in Caeiro et al. (Revstat 3(2):111–136, 2005), is also performed.

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a sample of independent and identically distributed (iid) random variables (rv's) with a common distribution function (df) F with a Pareto type right tail. Then the survival function $\overline{F} := 1 - F$ is a *regularly varying function* with an index of regular variation equal to $-1/\xi, \xi > 0$, i.e., $\lim_{t\to\infty} \overline{F}(tx)/\overline{F}(t) = x^{-1/\xi}$, and we write $\overline{F} \in \mathscr{R}_{-1/\xi}$. Consequently F is in the max domain of attraction

F. Caeiro (✉)
Faculdade de Ciências e Tecnologia and CMA, Universidade Nova de Lisboa, Caparica, Portugal
e-mail: fac@fct.unl.pt

I. Cabral
Departamento de Ciência e Tecnologia, Universidade de Cabo Verde, Praia, Cabo Verde
e-mail: ivanilda.cabral@docente.unicv.edu.cv

M. Ivette Gomes
CEAUL, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal
e-mail: ivette.gomes@fc.ul.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_11

of the general extreme value model, with df given by

$$\mathrm{EV}_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), \ 1+\xi x > 0, & \text{if } \xi \neq 0 \\ \exp(-\exp(-x)), \ x \in R, & \text{if } \xi = 0, \end{cases} \tag{1}$$

with $\xi > 0$, and we use the notation $F \in \mathscr{D}_{\mathscr{M}}\left(\mathrm{EV}_\xi\right)$. In this paper we deal with the estimation of the shape parameter ξ, the so-called *extreme value index* (EVI) which characterizes the weight of the right tail. Relevant literature on the estimation of the EVI can be found in [1, 2, 10].

Let $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ denote the associated non-decreasing order statistics from the sample of size n. For these Pareto-type models, we refer the classic maximum likelihood Hill EVI-estimator [15], the average of the log-excesses, $V_{ik} := \ln X_{n-i+1:n} - \ln X_{n-k:n}$, $1 \leq i \leq k < n$, above a high random threshold $X_{n-k:n}$,

$$\hat{\xi}^{\mathrm{H}}(k) = \mathrm{H}(k) := \frac{1}{k}\sum_{i=1}^{k} V_{ik}, \quad k = 1, 2, \ldots, n-1. \tag{2}$$

Hill's estimator is consistent for intermediate high thresholds, or equivalently, for intermediate k, i.e. for a non-decreasing sequence of integers $k \equiv k_n$, $1 \leq k < n$, such that

$$k \to \infty \quad \text{and} \quad k/n \to 0, \quad \text{as} \quad n \to \infty. \tag{3}$$

Due to the high variance for high thresholds, high bias for low thresholds, the mean square error (MSE) has usually a very peaked pattern. As a consequence it is difficult to determine the optimal k, under a minimum mean square error criterion. To mitigate the problem of the choice of k, crucial in applications, other estimators have been introduced by several authors. Here we shall consider a simple class of semi-parametric estimators of ξ related to the Lehmer's mean [14] of the log-excesses. Such class of estimators is given by

$$\hat{\xi}^{\mathrm{L}_\alpha}(k) \equiv \mathrm{L}_\alpha(k) := \frac{M_{k,n}^{(\alpha)}}{\alpha M_{k,n}^{(\alpha-1)}}, \qquad \left[M_{k,n}^{(\alpha)} := \sum_{i=1}^{k} V_{ik}^{\alpha}, \quad M_n^{(0)}(k) \equiv 1 \right] \tag{4}$$

parameterized in the tuning parameter $\alpha > 0.5$ and is consistent for all $\xi > 0$. Notice that since Lehmer's mean is a generalization of the arithmetic mean, $\mathrm{L}_\alpha(k)$ in (4) is a generalization of the classic Hill estimator in (2) ($\mathrm{L}_1(k) \equiv \mathrm{H}(k)$). The class of estimators in (4) was already studied in [13, 16, 17] and, if $\alpha \geq 1$, belongs to the class of EVI-estimators introduced in [4] (see also [6, 11]) with functional

expression

$$\hat{\xi}^{(\delta,\alpha)}(k) := \frac{\Gamma(\alpha)}{M_{k,n}^{(\alpha-1)}} \left(\frac{M_{k,n}^{(\delta\alpha)}}{\Gamma(\delta\alpha+1)} \right)^{1/\delta}, \quad \delta > 0, \ \alpha \geq 1, \quad k = 1, 2, \ldots, n-1, \tag{5}$$

where Γ denotes the complete Gamma function. Indeed, if we choose $\delta = 1$, in (5), we obtain $\mathrm{L}_\alpha(k)$ in (4). As noticed in [4], if $\delta > 1$ there is a value $\alpha \in [1, \infty[$ such that the dominant component of the asymptotic bias of $\hat{\xi}^{(\delta,\alpha)}(k)$, in (5), is null. If $\delta \leq 1$, the dominant component of asymptotic bias of $\hat{\xi}^{(\delta,\alpha)}(k)$ is always non-null. For $\delta = 2$ in (5), we obtain a class studied in [3]. Since $\mathrm{L}_\alpha(k)$ in (4) is biased for every α we also play with the direct reduction of the dominant bias component, working with the *reduced bias* (RB) Lehmer's EVI-estimators introduced in [13],

$$\hat{\xi}^{\mathrm{L}_\alpha^{\mathrm{RB}}}(k) = \mathrm{L}_\alpha^{\mathrm{RB}}(k) := \mathrm{L}_\alpha(k) \left(1 - \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{(1-\hat{\rho})^\alpha} \right), \quad \alpha > 0.5 \tag{6}$$

with $(\hat{\beta}, \hat{\rho})$ an adequate estimator of the vector of second-order parameters (β, ρ), to be defined in Sect. 2. Details on the estimation of second-order parameters (β, ρ) can be found in [9] and the references therein. When $\alpha = 1$, $\mathrm{L}_1^{\mathrm{RB}}(k)$ is the Corrected Hill (CH) EVI-estimator in [7], defined by

$$\hat{\xi}^{\mathrm{CH}}(k) = \mathrm{CH}(k) \equiv \mathrm{L}_1^{\mathrm{RB}}(k) := \mathrm{H}(k) \left(1 - \frac{\hat{\beta}(n/k)^{\hat{\rho}}}{1-\hat{\rho}} \right). \tag{7}$$

For adequate levels k and an adequate external estimation of the vector of second-order parameters, (β, ρ), the use of $\mathrm{CH}(k)$ enables us to eliminate the dominant component of asymptotic bias of the EVI-estimator $\mathrm{H}(k)$, keeping its asymptotic variance (see [5, 12] for further details).

The remainder of this paper is organized as follows: In Sect. 2, after the introduction of a few technical details in the field of extreme value theory, we deal with the asymptotic non-degenerate behaviour of the EVI-estimators. We further proceed with the study of the RB EVI-estimators, in (6), providing information on the dominant non-null asymptotic bias, under a third-order framework. Section 2 ends with the asymptotic comparison at optimal levels of the RB EVI-estimators. Some overall conclusions are drawn in Sect. 3.

2 Asymptotic Properties of the EVI-Estimators

2.1 A Brief Review of Several Conditions for Heavy Tailed Models

The model F underlying the data is heavy-tailed whenever $\overline{F} \in \mathscr{R}_{-1/\xi}$, $\xi > 0$, or equivalently, $U \in \mathscr{R}_{\xi}$ where U is the tail quantile function defined by $U(t) = F^{\leftarrow}(1 - 1/t)$, $t \geq 1$, with $F^{\leftarrow}(x) = \inf\{y : F(y) \geq x\}$. We thus assume the validity of any of the common first-order conditions for any $\xi > 0$:

$$F \in \mathscr{D}_{\mathscr{M}}\left(\mathrm{EV}_{\xi}\right) \iff \overline{F} \in \mathscr{R}_{-1/\xi} \iff U \in \mathscr{R}_{\xi}. \tag{8}$$

The *second-order parameter* ρ rules the rate of convergence in (8) and can be defined as the non-positive parameter appearing in the limiting relation

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \psi_{\rho}(x) := \begin{cases} \left(x^{\rho} - 1\right)/\rho, & \text{if } \rho < 0, \\ \ln x, & \text{if } \rho = 0, \end{cases} \tag{9}$$

often assumed to hold for every $x > 0$, with $A \in \mathscr{R}_{\rho}$ ultimately decreasing. This second-order condition is required for the derivation of the non-degenerate asymptotic bias of the EVI-estimators, under a semi-parametric framework.

To obtain information on the normal asymptotic behaviour of estimators of second-order parameters and on the asymptotic bias of RB EVI-estimators, it is sensible to further assume a third-order condition, ruling the rate of convergence in (9), and which guarantees that

$$\lim_{t\to\infty} \frac{\frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} - \psi_{\rho}(x)}{B(t)} = \psi_{\rho+\rho'}(x), \tag{10}$$

where $B \in \mathscr{R}_{\rho'}$, and $\rho' \leq 0$ is a third order parameter. Notice that the previous conditions hold for most heavy-tailed models used in applications, such as the $\mathrm{EV}_{\xi}(x)$, in (1), the generalized Pareto, $\mathrm{GP}_{\xi}(x) = 1 + \ln \mathrm{EV}_{\xi}(x)$, $x \geq 0$, the Fréchet, $F(x) = \exp(-x^{-1/\xi})$, $x > 0$, $(\xi > 0)$ the Burr, $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x \geq 0$ $(\xi > 0,\ \rho < 0)$ and the Student's t, among others. In this paper we shall assume that (10) holds with $\rho = \rho' < 0$ and that we can choose

$$A(t) = \xi \beta t^{\rho}, \qquad B(t) = \beta' t^{\rho'} = \frac{\zeta A(t)}{\xi}, \quad \zeta = \frac{\beta'}{\beta}, \tag{11}$$

with β and β', the scale second and third order parameters, respectively.

2.2 Asymptotic Behaviour of the Estimators at a Level k

To study the asymptotic behaviour of the EVI estimators, let us first introduce the auxiliary rv

$$\hat{\xi}^{\mathrm{L}_\alpha^{\mathrm{RB}^*}}(k) = \mathrm{L}_\alpha^{\mathrm{RB}^*}(k) := \mathrm{L}_\alpha(k)\left(1 - \frac{\beta (n/k)^\rho}{(1-\rho)^\alpha}\right), \tag{12}$$

related to the EVI estimator in (6) and depending on the vector of second-order parameters (β, ρ). If $F \in \mathscr{D}_{\mathscr{M}}\left(\mathrm{EV}_\xi\right)$ and k is intermediate, the EVI estimators in (2), (4), (6), (7) and the rv in (12) are consistent for $\xi > 0$. Under the validity of the second-order condition in (9), and with intermediate k such that $\lambda_A := \lim_{n\to\infty} \sqrt{k} A(n/k)$ is finite, trivial adaptations of the results in [7] and [16] enable us to guarantee the asymptotic normality of all the aforementioned rv's. More precisely, $\hat{\xi}^\bullet(k)$ with $\bullet$ denoting H, L_α, $\mathrm{L}_\alpha^{\mathrm{RB}}$, CH and $\mathrm{L}_\alpha^{\mathrm{RB}^*}$, respectively given in (2), (4), (6), (7) and (12), are asymptotically normal, i.e.

$$\sqrt{k}\left(\hat{\xi}^\bullet(k) - \xi\right) \xrightarrow[n\to\infty]{d} \mathscr{N}\left(\lambda_A b_\bullet, \sigma_\bullet^2\right),$$

where $\mathscr{N}\left(\mu, \sigma^2\right)$ stands for a normal rv with mean value μ and variance σ^2. For the non-RB EVI-estimators, $b_\bullet$ is given by

$$\begin{aligned} &b_{\mathrm{H}} = \frac{1}{1-\rho}, \qquad b_{\mathrm{L}_\alpha} = \frac{1}{(1-\rho)^\alpha}, \qquad b_{\mathrm{L}_\alpha^{\mathrm{RB}^*}} = 0, \\ &\sigma_{\mathrm{H}}^2 = \xi^2, \qquad \sigma_{\mathrm{L}_\alpha}^2 = \frac{\xi^2 \Gamma(2\alpha - 1)}{\Gamma^2(\alpha)}, \qquad \sigma_{\mathrm{L}_\alpha^{\mathrm{RB}^*}}^2 = \sigma_{\mathrm{L}_\alpha}^2. \end{aligned} \tag{13}$$

Moreover, if β and ρ are consistently estimated through $\hat{\beta}$ and $\hat{\rho}$, with $\hat{\rho} - \rho = o_p(1/\ln n)$, we get a null dominant component of bias for the RB EVI-estimators $\mathrm{L}_\alpha^{\mathrm{RB}}$ and CH, that is $b_{\mathrm{L}_\alpha^{\mathrm{RB}}} = b_{\mathrm{CH}} = 0$. The variance is kept at the same level of the associated non-RB EVI-estimators, that is $\sigma^2_{\mathrm{L}_\alpha^{\mathrm{RB}}} = \sigma^2_{\mathrm{L}_\alpha}$ and $\sigma^2_{\mathrm{CH}} = \sigma^2_{\mathrm{H}}$.

We shall next proceed with the study of the estimators under the third order condition.

Theorem 1 *If we assume that the third order condition, in* (10)*, holds for intermediate levels k, then we can write,*

$$\sqrt{k}\left(\hat{\xi}^\bullet(k) - \xi\right) \xrightarrow[n\to\infty]{d} \mathscr{N}(0, \sigma_\bullet^2) + b_\bullet \sqrt{k} A(n/k) + c_\bullet \sqrt{k} A^2(n/k)(1 + o_p(1)),$$

where $\bullet$ *can denote H,* L_α *or* $L_\alpha^{RB^*}$, $b_\bullet$ *and* $\sigma_\bullet^2$ *are given in* (13) *and* $c_\bullet$ *is given by*

$$c_H = c_{L_1} = \frac{1}{\xi(1-2\rho)}, \qquad c_{L_\alpha} = \frac{1}{\xi\rho}\left(\frac{\zeta\rho+1-2\rho}{(1-2\rho)^\alpha} - \frac{1-\rho(1-\rho)^{\alpha-1}}{(1-\rho)^{2\alpha-1}}\right),$$
$$c_{L_\alpha^{RB^*}} = \frac{1}{\xi\rho}\left(\frac{\zeta\rho+1-2\rho}{(1-2\rho)^\alpha} - \frac{1-\rho(1-\rho)^{\alpha}}{(1-\rho)^{2\alpha}}\right). \tag{14}$$

Theorem 2 *Under the conditions of Theorem 1 and if we consider consistent estimators* $(\hat\beta, \hat\rho)$ *of* (β, ρ) *both computed at a level* k_1 *such that* $k = o(k_1)$, *assume that* $(\hat\rho - \rho)\ln n = o_p(1)$ *and* $(\hat\beta - \beta)/\beta \overset{p}{\sim} -(\hat\rho - \rho)\ln(n/k_1)$, *a condition that holds for several estimators of* β, *then*

$$\hat\xi^{L_\alpha^{RB}}(k) - \hat\xi^{L_\alpha^{RB^*}}(k) \overset{p}{\sim} -\frac{A(n/k)}{(1-\rho)^\alpha}\left(\hat\rho - \rho\right)\ln(k/k_1).$$

Consequently, $\hat\xi^{L_\alpha^{RB}}(k)$ *is consistent if* $\left(\hat\rho - \rho\right)\ln(k/k_1) = o_p(1/A(n/k))$ *and has asymptotic normal distribution if* $\left(\hat\rho - \rho\right)\ln(k/k_1) = o_p(1/\sqrt{k}A(n/k))$.

The proof of the two previous theorems is identical to the proof of Theorems 2, 3 and 4 in [9].

2.3 Asymptotic Comparison of the RB EVI-Estimators at Optimal Levels

Since RB estimators have a null dominant component of asymptotic bias and keep the same variance as the associated non-RB EVI-estimator, we shall next proceed to the comparison of RB estimators, at their optimal levels. An asymptotic comparison, at their optimal levels, between $\hat\xi^{\mathrm{L}_\alpha}(k)$ and $\hat\xi^{\mathrm{H}}$ can be found in [16].

The comparison is again done in a way similar to the one used in [8, 16, 17], among others, for the classical EVI-estimators and in [9] for specific sets of RB EVI-estimators. Let $\hat\xi_n^\bullet(k)$ denote any arbitrary RB semi-parametric estimator of the EVI, ξ, for which we have

$$\hat\xi^\bullet(k) \overset{d}{=} \xi + \frac{\sigma_\bullet}{\sqrt{k}}Z_k^\bullet + c_\bullet\, A^2(n/k) + o_p(A^2(n/k)), \tag{15}$$

for any intermediate sequence of integers k, and where $Z_k^\bullet$ is asymptotically standard normal. Then, $\sqrt{k}(\hat\xi^\bullet(k) - \xi) \to N(\lambda c_\bullet, \sigma_\bullet^2)$ provided that k is such that $\sqrt{k}\,A^2(n/k) \to \lambda$, with λ finite, as $n \to \infty$. We then write $\mathrm{Bias}_\infty(\hat\xi_n^\bullet(k)) := c_\bullet\, A^2(n/k)$, and $\mathrm{Var}_\infty(\hat\xi_n^\bullet(k)) := \sigma_\bullet^2/k$. The so-called *asymptotic mean square*

error (AMSE) is then given by $\text{AMSE}\left(\hat{\xi}_n^{\bullet}(k)\right) := \sigma_{\bullet}^2/k + c_{\bullet}^2\, A^4(n/k)$. Regular variation theory enables us to show that, whenever $c_{\bullet} \neq 0$, there exists a function $\varphi(n) = \varphi(n, \xi, \rho)$, such that

$$\lim_{n\to\infty} \varphi(n)\, \text{AMSE}\left(\hat{\xi}_{n0}^{\bullet}\right) = \left(\sigma_{\bullet}^2\right)^{-\frac{4\rho}{1-4\rho}} \left(c_{\bullet}^2\right)^{\frac{1}{1-4\rho}} =: \text{LMSE}\left(\hat{\xi}_{n0}^{\bullet}\right),$$

with LMSE standing for limiting mean square error, $\hat{\xi}_{n0}^{\bullet} := \hat{\xi}_{n,k_0^{\bullet}(n)}^{\bullet}$ and

$$k_0^{\bullet}(n) := \arg\inf_k \text{AMSE}\left(\hat{\xi}_n^{\bullet}(k)\right).$$

It is then sensible to consider the following: Given two biased estimators $\hat{\xi}_n^{(1)}(k)$ and $\hat{\xi}_n^{(2)}(k)$, for which a distributional representation of the type of the one in (15) holds, with constants (σ_1, c_1) and (σ_2, c_2), $c_1, c_2 \neq 0$, respectively, both computed at their optimal levels, the *Asymptotic Root Efficiency* (AREFF) of $\hat{\xi}_{n0}^{(1)}$ relatively to $\hat{\xi}_{n0}^{(2)}$ is

$$\text{AREFF}_{1|2} \equiv \text{AREFF}_{\hat{\xi}_{n0}^{(1)}|\hat{\xi}_{n0}^{(2)}} := \sqrt{\frac{\text{LMSE}\left[\hat{\xi}_{n0}^{(2)}\right]}{\text{LMSE}\left[\hat{\xi}_{n0}^{(1)}\right]}} = \left(\left(\frac{\sigma_2}{\sigma_1}\right)^{-4\rho} \left|\frac{b_2}{b_1}\right|\right)^{\frac{1}{1-4\rho}}. \tag{16}$$

The AREFF indicator, in (16), has been conceived so that the highest the AREFF indicator is, the better is the first estimator. We have

$$\text{AREFF}_{\text{L}_\alpha^{\text{RB}}|\text{CH}} = \left\{ \left(\frac{\Gamma^2(\alpha)}{\Gamma(2\alpha-1)}\right)^{-2\rho} \left| \frac{\zeta\rho(1-\rho)^{2\alpha}(1-2\rho)^{\alpha-1} - \rho(1-\rho)^{2\alpha-2}(1-2\rho)^{\alpha}}{(\zeta\rho+1-2\rho)(1-\rho)^{2\alpha} - (1-\rho(1-\rho)^{\alpha})(1-2\rho)^{\alpha}} \right| \right\}^{\frac{1}{1-4\rho}}$$

Since $\text{AREFF}_{\text{L}_\alpha^{\text{RB}}|\text{CH}}$ depends on the value of the parameters α, ρ and ζ, for technical simplicity we shall consider $\zeta = 1$, the value of ζ for the generalized Pareto and Burr models. In Fig. 1 we picture the values of $\text{AREFF}_{\text{L}_\alpha^{\text{RB}}|\text{CH}}$ in the (α, ρ)-plane.

As can be seen in Fig. 1, for models with $\zeta = 1$ the gain in efficiency is not very high. For every ρ in $(-2, 0)$, and independently of ξ, there exists always a region of α-values where $\hat{\xi}^{\text{L}_\alpha^{\text{RB}}}$ is asymptotically more efficient than $\hat{\xi}^{\text{CH}}$ estimator, both computed at their optimal levels. To have $\text{AREFF}_{\text{L}_\alpha^{\text{RB}}|\text{CH}} > 1$, the figure suggests that we should choose $0.5 < \alpha \leq 1$ if $\rho > -1.9$ and $\alpha \geq 1$ otherwise. The optimal choice of α, for a given ρ, can only be computed with a numerical approach, and is outside the scope of this paper.

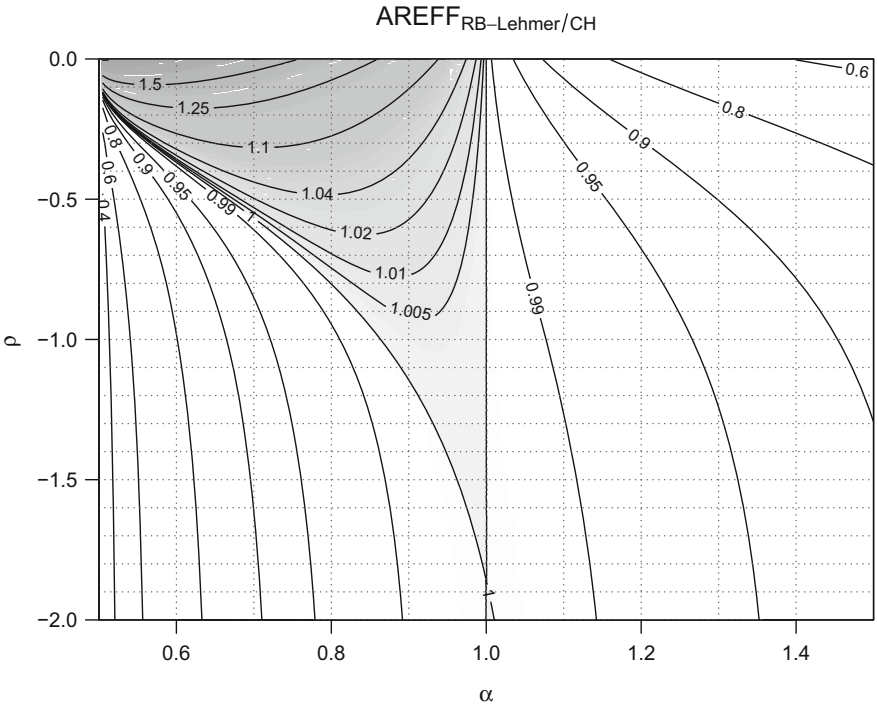


Fig. 1 Contour plot of $\text{AREFF}_{\text{L}_\alpha^{\text{RB}}|\text{CH}}$ in the (α, ρ)-plane

3 Conclusions

As concluded in [16], the optimal $\hat{\xi}^{\text{L}_\alpha}$ can beat the optimal Hill estimator, $\hat{\xi}^{\text{H}}$, in the whole plane (ξ, ρ). The same conclusion holds when we compare $\hat{\xi}^{\text{L}_\alpha^{\text{RB}}}$ to the $\hat{\xi}^{\text{CH}}$ EVI-estimator. In the simulation study presented in [16], for $\hat{\xi}^{\text{L}_\alpha}$, the authors noticed that the highest efficiency was obtained for large values of α, away from the asymptotically optimal α. Therefore, further research on the choice of the parameters α for the class of estimators $\hat{\xi}^{\text{L}_\alpha^{\text{RB}}}$ is needed. It is important to compare the simulated and asymptotically optimal α and to study heuristics to select the value of the tuning parameter α.

Acknowledgements Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through projects UID/MAT/00006/2013 (CEA/UL) and PEst-OE/MAT/UI0297/2013 (CMA/UNL).

References

1. Beirlant J., Goegebeur Y., Segers J., Teugels J.: Statistics of Extremes. Theory and Applications. Wiley, Chichester (2004)
2. Beirlant, J., Caeiro, F., Gomes, M.I.: An overview and open research topics in statistics of univariate extremes. Revstat **10**(1), 1–31 (2012)
3. Caeiro, F., Gomes, M.I.: A class of asymptotically unbiased semi-parametric estimators of the tail index. Test **11**(2), 345–364 (2002)
4. Caeiro, F., Gomes, M.I.: Bias reduction in the estimation of parameters of rare events. Theory Stoch. Process. **8**(24), 67–76 (2002)
5. Caeiro, F., Gomes, M.I., Henriques Rodrigues, L.: Reduced-bias tail index estimators under a third order framework. Commun. Stat. Theory Methods **38**(7), 1019–1040 (2009)
6. Caeiro, F., Gomes, M.I.: Comparison of asymptotically unbiased extreme value index estimators: a Monte Carlo simulation study. AIP Conf. Proc. **1618**, 551–554 (2014)
7. Caeiro, F., Gomes, M.I., Pestana, D.D.: Direct reduction of bias of the classical Hill estimator. Revstat **3**(2), 111–136 (2005)
8. Caeiro, F., Prata Gomes, D.: Adaptive estimation of a tail shape second order parameter: a computational comparative study. In: Simos, T.E., Kalogiratou, Z., Monovasilis, T. (eds.) AIP Conference Proceedings, vol. 1702, p. 030005 (2015)
9. Caeiro, F., Gomes, M.I., Beirlant, J., de Wet, T.: Mean-of-order p reduced-bias extreme value index estimation under a third-order framework. Extremes **19**, 561–589 (2016). https://doi.org/10.1007/s10687-016-0261-5
10. Gomes, M.I., Guillou, A.: Extreme value theory and statistics of univariate extremes: a review. Int. Stat. Rev. **83**(2), 263–292 (2015)
11. Gomes, M.I., Caeiro, F., Figueiredo, F.: Bias reduction of a tail index estimator through an external estimation of the second-order parameter. Statistics **38**(6), 497–510 (2004)
12. Gomes MI, Pestana D, Caeiro F.: A note on the asymptotic variance at optimal levels of a bias-corrected Hill estimator. Stat. Probab. Lett. **79**, 295–303 (2009)
13. Gomes M.I., Penalva, H., Caeiro, F., Neves M.M.: Non-reduced versus reduced-bias estimators of the extreme value index – efficiency and robustness. In: Colubi, A., Blanco, A., Gatu, C. (eds.) Proceedings of COMPSTAT 2016: 22th International Conference on Computational Statistics, Oviedo, pp. 279–290 (2016)
14. Havil, J.: Gamma: Exploring Euler's Constant. Princeton University Press, Princeton, New Jersey (2003)
15. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Statist. **3**, 1163–1174 (1975)
16. Penalva, H., Caeiro, F., Gomes, M.I., Neves, M.M.: An efficient naive generalisation of the Hill estimator—discrepancy between asymptotic and finite sample behaviour. Notas e Comunicações CEAUL 02/2016 (2016)
17. Penalva, H., Gomes, M.I., Caeiro, F., Neves, M.: A couple of non reduced bias generalized means in extreme value theory: an asymptotic comparison. REVSTAT (2018, accepted)

Hazard Rate and Future Lifetime for the Generalized Normal Distribution

Thomas L. Toulias and Christos P. Kitsos

Abstract The target of this paper is to discuss a generalized form of the well-known Law of Frequency Error. This particular Law of Frequency of Errors is what is known as "Gaussian" or "Normal" distribution and appeared to have an aesthetic appeal to all the branches of Science. The Generalized Normal Distribution is presented as a basis to our study. We derive also the corresponding hazard function as well as the future lifetime of the Generalized Normal Distribution (GND), while new results are also presented. Moreover, due to some of the important distribution the GND family includes, specific results can also be extracted for some other distributions.

1 Introduction

The evolution of the well-known Normal Distribution (or Gaussian) from the well-known Law of Frequency Error is essential. Galton in 1889 pointed out that: *"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error'. The law would have been personified by the Greeks and deified, if they had known of it."*, [5, pg.66]. This particular Law appeared to have an aesthetic appeal to all the branches of Science, since the time was first discussed in [2].

Gauss, not only in astronomical measurements, but also in geodesy (studying the systematic errors of angle measurements) worked with what was latter named "Gaussian" or "Normal" distribution [6]. The first attempt on the so important Central Limit Theorem was by Abraham de Moivre in 1733 in [2] but, as usual, this work did not receive special interest. Later in 1920, George Polya named the method, [15]. Edgeworth, although more philosopher than a statistician, published some work on Ethics and then his first paper on Statistics was in 1833 [3]. That was

T. L. Toulias (✉) · C. P. Kitsos
Technological Educational Institute of Athens, Egaleo, Athens, Greece
e-mail: xkitsos@teiath.gr

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_12

the basis of his work *"The philosophy of chance"* in 1884 in the 11th edition of Encyclopedia Britannica, [4]. He introduced the terms "modulus" and "fluctuation" (the $\sqrt{2\sigma}$ and $2\sigma^2$, respectively).

This short discussion is to prove the importance on the Normal distribution, even in the early days. There were certainly attempts to provide a more general form for the Normal, see [13], but these attempts were rather "technical" generalizing of what a normal distribution includes, i.e. generalizing the coefficient or generalizing the exponent. Only recently a generalized form emerged through the theoretical background of the Logarithm Sobolev Inequalities.

Indeed, this generalized form of the Law of Frequency Error can be expressed through an exponential power generalization of the usual multivariate Normal distribution introduced by Kitsos and Tavoularis in [8], which has an information theoretic background as it is derived through the study of a generalized Fisher's entropy type information measure.

This three-parameter distribution, called the γ-order Normal distribution (or GND), is discussed in Sect. 2, while in Sect. 3, the hazard rate and the future lifetime of the GND distribution family are derived and discussed.

2 The Generalized Normal Distribution

New entropy type information measures were introduced in [8], generalizing the known Fisher's entropy type information measure; see also [9–11, 16]. In principle, the information measures are divided into three main categories: parametric (a typical example Fisher's information), non-parametric (with Shannon information measure to be the most well known) and entropy type, see [11]. The information-theoretic extraction of the p-variate γ-order Generalized Normal distribution (or GND) is based on the generalized form of the well-known Fisher's entropy type information measure, see [8, 10] for details. Recall now the definition of the probability distribution function (p.d.f.) of the GND family of distributions, [8]:

Definition 1 The p-variate random variable X follows the γ-order Generalized Normal distribution, i.e. $X \sim \mathcal{N}_\gamma^p(\mu, \Sigma)$, with location parameter vector $\mu \in \mathbb{R}^p$, shape parameter $\gamma \in \mathbb{R} \setminus [0, 1]$ and positive definite scale parameter matrix $\Sigma \in \mathbb{R}^{p\times p}$, when the density function f_X of X is of the form

$$f_X(x) = f_X(x;\ \mu, \Sigma, \gamma, p) := C_X \exp\left\{-\tfrac{\gamma-1}{\gamma} Q_X(x)^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad x \in \mathbb{R}^p, \tag{1}$$

where Q_X is the p-quadratic form $Q_X(x) := (x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}$, $x \in \mathbb{R}^p$, while the normalizing factor C_X is defined as

$$C_X = C_X(\Sigma, \gamma, p) := \frac{\Gamma\left(\frac{p}{2}+1\right)}{\pi^{p/2}\,\Gamma\left(p\frac{\gamma-1}{\gamma}+1\right)\sqrt{|\Sigma|}} \left(\tfrac{\gamma-1}{\gamma}\right)^{p\frac{\gamma-1}{\gamma}}, \tag{2}$$

where $|\mathrm{A}| := \det \mathrm{A}$ denotes the determinant of any $\mathrm{A} \in \mathbb{R}^{p\times p}$.

From the p.d.f. f_X as above, notice that the location vector of X is essentially the mean vector of X, i.e. $\mu = \mu_X := \mathrm{E}(X)$. Moreover, for the shape parameter value $\gamma = 2$, $\mathscr{N}_2^p(\mu, \Sigma)$ is reduced to the well-known multivariate normal distribution, where Σ is now the covariance of X, i.e. $\mathrm{Cov}\, X = \Sigma$. Recall that

$$\mathrm{Cov}\, X = \frac{\Gamma\Big((p+2)\frac{\gamma-1}{\gamma}\Big)}{p\,\Gamma^3\Big(p\frac{\gamma-1}{\gamma}\Big)}\Big(\frac{\gamma}{\gamma-1}\Big)^{2\frac{\gamma-1}{\gamma}}\Sigma, \tag{3}$$

for the positive definite scale matrix Σ; see [11]. In order to parameterize the GND family such that the scale parameter matrix Σ to be the always the covariance of each γ-order Normal distribution, i.e. $\mathscr{N}_\gamma^p(\mu_X, \Sigma_X)$ with $\mu_X = \mathrm{E}(X)$ and $\Sigma_X = \mathrm{Cov}\, X$, like the usual multivariate Normal distribution is expressed, the p.d.f. should be of the form

$$f_X(x) = f_X(x;\ \mu_X, \Sigma_X, \gamma, p) := C_X \exp\left\{-k_\gamma^p\, Q_X(x)^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad x \in \mathbb{R}^p, \tag{4}$$

with $Q_X(x) := (x-\mu_X)\Sigma_X^{-1}(x-\mu_X)^{\mathrm{T}}$, $x \in \mathbb{R}^p$, where

$$k_\gamma^p := \left[\frac{\Gamma\Big((p+2)\frac{\gamma-1}{\gamma}\Big)}{p\,\Gamma\Big(p\frac{\gamma-1}{\gamma}\Big)}\right]^{\frac{\gamma}{2(\gamma-1)}}, \tag{5}$$

while the normalizing factor C_X should be then written in the form

$$C_X = C_X(\Sigma_X, \gamma, p) := \tfrac{1}{2}\,\Gamma(p/2)\sqrt{\frac{\Gamma\Big((p+2)\frac{\gamma-1}{\gamma}\Big)}{p\pi^p\,\Gamma\Big(p\frac{\gamma-1}{\gamma}\Big)|\Sigma_X|}\Big(\frac{\gamma-1}{\gamma}\Big)^{(p-1)\frac{\gamma-1}{\gamma}-1}}. \tag{6}$$

Various attempts to generalize the usual Normal distribution are known. The introduced univariate γ-order Normal $\mathscr{N}_\gamma\big(\mu, \sigma^2\big)$ coincides with the existent generalized normal distribution in [13, 18]. Recall also the univariate power exponential distribution $\mathscr{PE}(\mu, \sigma, \beta)$, [1], with p.d.f.

$$f(x) = f(x;\ \mu, \sigma, \beta) := \frac{\exp\left\{-\frac{1}{2}\big|\frac{x-\mu}{\sigma}\big|^{\frac{2}{1+\beta}}\right\}}{2^{\frac{\beta+3}{2}}\,\sigma\,\Gamma\big(\frac{\beta+3}{2}\big)}, \quad x \in \mathbb{R}. \tag{7}$$

The multivariate case of the γ-order Normal $\mathscr{N}_\gamma^p(\mu, \Sigma)$ coincides with the existent multivariate power exponential distribution $\mathscr{PE}^p(\mu, \Sigma', b)$, as introduced in [7], where $\Sigma' := 2^{2(\gamma-1)/\gamma}\Sigma$ and $b := \frac{1}{2}\gamma/(\gamma-1)$. See also [17].

Note that the family of the $\mathcal{N}_\gamma^p$ distributions acts to the generalized form of the Information Inequality just like the usual Normal distribution acts on the usual Information Inequality, i.e. by providing the corresponding equality; see [8, Cor. 3.2], [10, 16].

The family of $\mathcal{N}_\gamma^p(\mu, \Sigma)$ distributions, i.e. the family of the elliptically contoured γ-order Generalized Normals, provides a smooth bridging between some important multivariate (and elliptically countered) distributions. Indeed, [11]:

Theorem 1 *For the elliptically contoured p-variate γ-order Normal distribution $\mathcal{N}_\gamma^p(\mu, \Sigma)$ with $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$, we obtain the following special cases:*

Case $\gamma := 0$. For the limiting case of the shape parameter $\gamma \to 0^-$, the degenerate Dirac distribution $\mathcal{D}(\mu)$ with pole at μ is derived in dimensions $p := 1, 2$, while for $p \geq 3$ the p.d.f. of $\mathcal{N}_0(\mu, \Sigma)$ is flattened (p.d.f. is everywhere zero).

Case $\gamma := 1$. For the limiting case of $\gamma \to 1^+$, the elliptically contoured Uniform distribution $\mathcal{U}^p(\mu, \Sigma)$ is obtained, which is defined over the p-ellipsoid $A : (x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}} \leq 1$, $x \in \mathbb{R}^p$.

Case $\gamma := 2$. For the "normality" case of $\gamma := 2$ the usual p-variate Normal distribution $\mathcal{N}^p(\mu, \Sigma)$ is obtained.

Case $\gamma := \pm\infty$. For the limiting case of $\gamma \to \pm\infty$, the elliptically contoured Laplace distribution $\mathcal{L}^p(\mu, \Sigma)$ is derived.

We clarify that, according to Probability Theory, the term "elliptically contoured" distributions refers to multivariate distributions where their corresponding density functions are formed by quadratic functions, like Q_X as in the definition (1). This means that the "sliced" intersection of the p-dimensional hyper-surface of an elliptically contoured p.d.f., embedded in a $(p+1)$-dimensional Euclidean space, with a p-dimensional hyper-plane corresponding each time to the probability level $p \in [0, 1]$, forms a $(p-1)$-ellipsoid. For the special case of $\Sigma := \sigma^2\mathbb{I}_p$, the p-ellipsoids above are reduced to p-spheres and the corresponding multivariate distribution is said to be a "spherically contoured" distribution.

One of the merits of the $\mathcal{N}_\gamma$ family is that it can provide "heavy-tailed" distributions as the change of shape parameter γ influences the "probability mass" at the tails. This makes the introduced shape parameter important to the Risk Analysis, as the "risk" is known relevant to the shape parameter. Practically, it is easy to understand that only in case $\gamma = 2$ the confidence intervals are "confidence" at the chosen significant level. Otherwise we have to adjust the confidence interval, with the appropriate new standard deviation, for the chosen shape parameter; see [11].

Throughout this paper we shall use the notation $g = g(\gamma) := (\gamma - 1)/\gamma$, $\gamma \in \mathbb{R} \setminus [0, 1]$.

The cumulative distribution function (c.d.f.) of the GND, as in (1), can be calculated as, [12],

$$F_X(x) = 1 - \frac{\Gamma\big(g, gz^{1/g}\big)}{2\,\Gamma(g)}, \quad z = z(x;\ \mu, \sigma) := \frac{x-\mu}{\sigma}, \quad x \in \mathbb{R}. \tag{8}$$

Alternatively, using positive arguments for the upper (complementary) incomplete gamma function $\Gamma(a, x)$, $x \in \mathbb{R}$, $a \in \mathbb{R}_+$ (which is more computationally oriented approach), it holds that

$$F_X(x) = \tfrac{1+\operatorname{sgn} z}{2} - (\operatorname{sgn} z)\frac{\Gamma\left(g, g|z|^{1/g}\right)}{2\,\Gamma(g)}, \quad x \in \mathbb{R}. \tag{9}$$

In such a case, the quantile function is then given by

$$\begin{aligned} Q_X(P) &:= \inf\left\{x \in \mathbb{R} :\ F_X(x) \geq P\right\} \\ &= \operatorname{sgn}(2P-1)\sigma\left[\tfrac{1}{g}\,\Gamma^{-1}\left(g, |2P-1|\right)\right]^g, \quad P \in (0,\ 1). \end{aligned} \tag{10}$$

Table 1 provides the probability values $F_X(x) = \Pr\{X \leq x\}, x = -3, -2, \ldots, 3$, of an r.v. $X \sim \mathscr{N}_\gamma(0, 1)$ for various shape parameter γ values. The column for $x = 0$ is omitted as $F_X(0) = 1/2$ for every $X \sim \mathscr{N}_\gamma(0, 1)$ (recall that $\mathscr{N}_\gamma(0, 1)$ is a symmetric distribution around its mean 0). Moreover, the last two columns provide the 1st and 3rd quartile points $Q_X(1/4)$ and $Q_X(3/4)$ of X, i.e. $\Pr\left\{X \leq Q_X(k/4)\right\} = k/4$, $k = 1, 3$, for various γ values. Indeed, 'heavy-tailed" distributions are obtained, as γ value increases towards $+\infty$, with the probability mass increasing until it reaches the corresponding probability values of the Laplace distribution $\mathscr{L}(0, 1)$; see, for example, the $F_X(k)$, $k = -3, -2, -1$, values in Table 1 which correspond to the left tail probability mass values of $\mathscr{N}_{\gamma>0}(0, 1)$. Such comments might be proved helpful to a researcher who tries to see the Risk

Table 1 Probability mass values $F_X(x)$ for various $x \in \mathbb{R}$ as well as the 1st and 3rd quartiles $Q_X(1/4)$, $Q_X(3/4)$, for certain r.v. $X \sim \mathscr{N}_\gamma(0, 1)$

γ	$F_X(-3)$	$F_X(-2)$	$F_X(-1)$	$F_X(1)$	$F_X(2)$	$F_X(3)$	$Q_X(\frac{1}{4})$	$Q_X(\frac{3}{4})$
-50	0.0260	0.0690	0.1846	0.8154	0.9310	0.9740	-0.6936	0.6936
-10	0.0304	0.0742	0.1869	0.8131	0.9258	0.9696	-0.6951	0.6951
-5	0.0357	0.0802	0.1895	0.8105	0.9198	0.9643	-0.6967	0.6967
-2	0.0502	0.0950	0.1958	0.8042	0.9050	0.9498	-0.7004	0.7004
-1	0.0699	0.1131	0.2030	0.7970	0.8869	0.9301	-0.7042	0.7042
$-1/2$	0.0970	0.1361	0.2116	0.7884	0.8639	0.9030	-0.7082	0.7082
$-1/10$	0.1656	0.1889	0.2299	0.7701	0.8111	0.8344	-0.7142	0.7142
1	0	0	0	1	1	1	-0.5	0.5
2	0.0013	0.0228	0.1587	0.8413	0.9772	0.9987	-0.6745	0.6745
3	0.0071	0.0402	0.1699	0.8301	0.9598	0.9929	-0.6833	0.6833
4	0.0112	0.0480	0.1742	0.8258	0.9520	0.9888	-0.6865	0.6865
5	0.0138	0.0523	0.1765	0.8235	0.9477	0.9862	-0.6881	0.6881
10	0.0193	0.0604	0.1805	0.8195	0.9396	0.9807	-0.6909	0.6909
50	0.0238	0.0663	0.1833	0.8167	0.9337	0.9762	-0.6927	0.6927
$\pm\infty$	0.0249	0.0677	0.1839	0.8161	0.9323	0.9751	-0.6931	0.6931

for some "heavy tailed" data set. He has, firstly, to evaluate the appropriate shape parameter, and then form confidence intervals and risk levels. Moreover, as γ value decreases towards 0, the probability mass is further increasing towards the degenerate Dirac distribution $\mathscr{D}(0)$; see, for example, the corresponding $F_X(k)$, $k = -3, -2, -1$, values for the left tail probability mass values of $\mathscr{N}_{\gamma<0}(0, 1)$.

We shall try now to clarify to practitioners, and not only, how the GND "includes" a number of other distributions. This is not very common in practice, where "one" distribution is assumed. That is, the benefit here—depending on the data set—is that the shape parameter "forms" the appropriate distribution which fits the data. Figure 1 illustrates Theorem 1 with $X_\gamma \sim \mathscr{N}_\gamma(0, 1)$ in a compact form, by depicting a surface formed by all the c.d.f. curves $F_{X_\gamma}(x)$, $x \in [-3, 3]$, for every $\gamma \in [-10, 0) \cup [1, 10]$. The known c.d.f. of the Uniform ($\gamma = 1$) and Normal ($\gamma = 2$) distributions are also depicted. The c.d.f. of $\mathscr{N}_{\gamma=\pm 10}(0, 1)$, which approximates the Laplace distribution $\mathscr{L}(0, 1) = \mathscr{N}_{\pm\infty}(0, 1)$, as well as the c.d.f. of $\mathscr{N}_{-0.005}(0, 1)$, which approximates the degenerate Dirac distribution $\mathscr{D}(0)$, are clearly presented. Notice the smooth-bringing of F_{X_γ} between these significant distributions which are included into the $\mathscr{N}_\gamma$ family of distributions for $\gamma \in \mathbb{R} \cup \{\pm\infty\} \setminus (0, 1)$. Moreover, upon the formed surface, the quantile functions $Q_{X_\gamma}(P)$ are depicted as curves parameterized by $\gamma \in [-10, 10]$, for $P := 0.05, 0.1, \ldots 0.95$, while the 1st and 3rd quartile curves $Q_{X_\gamma}(1/4)$ and $Q_{X_\gamma}(3/4)$ are distinguished.

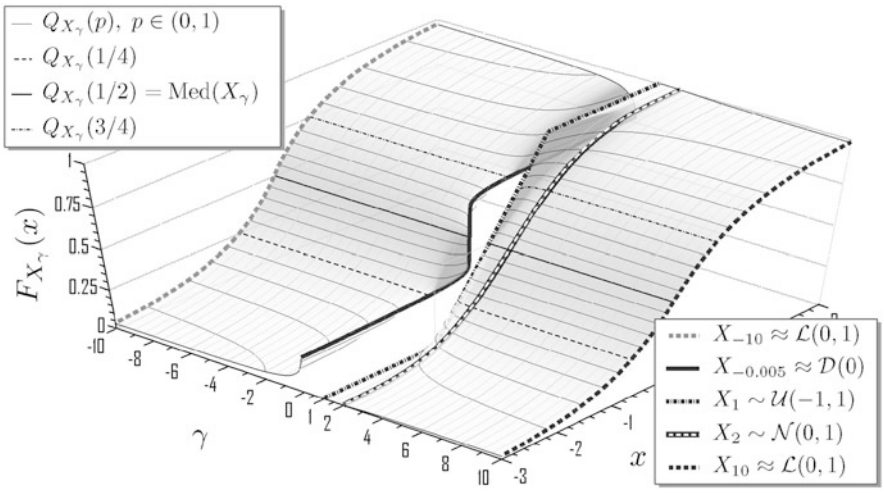


Fig. 1 Surface graph of all the c.d.f.-s $F_{X_\gamma}(x)$ along $x \in [-3, 3]$ and $\gamma \in [-10, 10]$, where $X_\gamma \sim \mathscr{N}_\gamma(0, 1)$, as well as the quantile curves $Q_{X_\gamma}(P)$, $P \in [0, 1]$

3 Hazard Rate and Future Lifetime for the GND

Consider a continuous random variable Y having c.d.f. $G(y)$ and p.d.f. $g(y)$, such that $G(0) := 0$. The conditional p.d.f. of Y given that $Y > y$, say $h(y)$, is

$$\Pr(Y|Y > y) = h(y) = \frac{g(y)}{1 - G(y)}, \quad y > 0, \tag{11}$$

is known as the hazard rate or Failure rate, so that to give emphasis that this conditional p.d.f. describes survival time distributions, given that an individual survey occurs at time y. Two are the classified distributions:

1. ICR = Increasing Failure Rate, and
2. DFR = Decreasing Failure Rate.

Typical example of IFR is the half-Normal distribution with failure rate equal to

$$\frac{1}{\sigma}\phi(x/\sigma)\big[1 - \Phi(x/\sigma)\big], \quad x > 0. \tag{12}$$

When $\sigma := 1$ the reciprocal failure is known as Mill's ratio. For a list of p.d.f.-s and the corresponding risk in Benchmark Dose Analysis, applied in Cancer problems, see [14]. Based on this introductory comments we try to extend the failure rate to the introduced GND. It is clear that the evaluation of the hazard rate of the GND will include, due to Theorem 1, the hazard rates of all the related distributions with it. Indeed, for the hazard rate and the cumulative hazard rate of the GND we have the following:

Proposition 1 *The hazard rate h_X of a univariate γ-order normally distributed r.v. $X \sim \mathcal{N}_\gamma\big(\mu, \sigma^2\big)$ is given by*

$$h_X(x) = \frac{g^{g-1}\,\mathrm{e}^{-g|z|^{1/g}}}{\sigma\,\Gamma\big(g, gz^{1/g}\big)}, \quad x \in \mathbb{R}. \tag{13}$$

The corresponding inverse hazard function h_X^{inv}, of r.v. X is given by

$$h_X^{\mathrm{inv}}(x) = \frac{g^{g-1}\,\mathrm{e}^{-g|z|^{1/g}}}{\sigma\,\Gamma(g) + \sigma\,\gamma\big(g, gz^{1/g}\big)}, \quad x \in \mathbb{R}, \tag{14}$$

where $\gamma(\cdot,\cdot)$ denotes the lower incomplete gamma function. Alternatively, using positive arguments for the upper/lower incomplete gamma function we obtain

$$h_X(x), h_X^{\mathrm{inv}}(x) = \frac{g^{g-1}\,\mathrm{e}^{-g|z|^{1/g}}}{\sigma\,\Gamma(g) \pm (\operatorname{sgn} z)\sigma\,\gamma\big(g, g|z|^{1/g}\big)}, \quad x \in \mathbb{R}, \tag{15}$$

where the minus sign corresponds to the hazard rate function while the plus sign to the inverse hazard rate function.

Proof From the definition of the hazard rate $h_X := f_X/(1 - F_X)$ and the inverse hazard rate $h_X^{\text{inv}} := f_X/F_X$ of an r.v. X, we derive through (1), (8), and the fact that $\gamma(a, x) := \Gamma(a) - \Gamma(a, x)$, $x \in \mathbb{R}$, $a \in \mathbb{R}_+$, the requested forms (13) and (14), respectively. Moreover, using positive arguments for the upper/lower incomplete gamma function through (9), we finally obtain (15).

Based on Proposition 1, the cumulative hazard function can be evaluated. Recall that the reliability function is the ratio f/h. Therefore a generalized form of reliability $R(t) = f(t)/h(t)$, $t \in \mathbb{R}_+$, can be produced. This needs further investigation, and we shall produce it elsewhere. Our target here is the study of the hazard function of the GND.

Proposition 2 *The cumulative hazard function H_X of an r.v. $X \sim \mathscr{N}_\gamma(\mu, \sigma^2)$ is given by*

$$H_X(x) = -\log \frac{\Gamma\big(g, gz^{1/g}\big)}{2\,\Gamma(g)}, \quad x \in \mathbb{R}. \tag{16}$$

Alternatively, we derive that

$$H_X(x) = \begin{cases} -\log\left\{1 - \dfrac{\Gamma\big(g, g|z|^{1/g}\big)}{2\,\Gamma(g)}\right\}, & \text{for } x \le \mu, \\ -\log \dfrac{\Gamma\big(g, gz^{1/g}\big)}{2\,\Gamma(g)}, & \text{for } x > \mu. \end{cases} \tag{17}$$

See Appendix for the proof.

As a result the average hazard rate of $X \sim \mathscr{N}_\gamma\big(\mu, \sigma^2\big)$ between the points $x_1 < x_2$ can then be given, through (16). Indeed:

$$h_X^{\text{avg}}(x_1, x_2) := \text{avg}(h)_{x_1}^{x_2} = \frac{1}{x_2 - x_1}\int_{x_1}^{x_2} h_X(x)\,\mathrm{d}\,x = -\frac{1}{(x_2 - x_1)}\log\frac{\Gamma\big(g, gz_2^{1/g}\big)}{\Gamma\big(g, gz_1^{1/g}\big)},$$

where $z_i = z_i(x_i) := (x_i - \mu)/\sigma$, $i = 1, 2$.

The following two examples clarify our claim that the hazard rates of some well-known distributions have been evaluated, due to Proposition 1. In particular, the hazard rates of the Laplace and the Uniform distributions are provided in the following.

Example 1 For the Laplace distributed r.v. $X \sim \mathscr{N}_{\pm\infty}\big(\mu, \sigma^2\big) = \mathscr{L}(\mu, \sigma)$, using the fact that $\Gamma(1, x) = \mathrm{e}^{-x}$, $x \in \mathbb{R}$, the corresponding hazard and inverse hazard rates, as in (13) and (14), for $\gamma \to \pm\infty$, or $g \to 1^+$, can be written as

$$h_X(x) = \begin{cases} \frac{1}{\sigma}\,\mathrm{e}^{2(x-\mu)/\sigma}, & \text{for } x \le \mu, \\ 1/\sigma, & \text{for } x > \mu, \end{cases} \quad \text{and} \tag{18}$$

$$h_X^{\text{inv}}(x) = \begin{cases} \dfrac{\mathrm{e}^{2(x-\mu)/\sigma}}{2\sigma\,\mathrm{e}^{(x-\mu)/\sigma} - \sigma}, & \text{for } x \le \mu, \\ \dfrac{1}{2\sigma\,\mathrm{e}^{(x-\mu)/\sigma} - \sigma}, & \text{for } x > \mu, \end{cases} \tag{19}$$

respectively, which are the hazard and inverse hazard rates of the Laplace distribution, as expected, while the corresponding cumulative hazard function is then given, through (17), by

$$H_X(x) = \begin{cases} -\log\left\{1 - \frac{1}{2}\,\mathrm{e}^{(x-\mu)/\sigma}\right\}, & \text{for } x \le \mu, \\ \log 2 + \dfrac{x-\mu}{\sigma}, & \text{for } x > \mu. \end{cases} \tag{20}$$

Notice that the value of $z = z(x) := (x\mu)/\sigma$, $x \in \mathbb{R}$, is essential for the above formulas.

Example 2 The hazard function of the Uniform distribution can be derived as special case from (13). For the uniformly distributed r.v. $X \sim \mathscr{U}(a, b)$, recall from Theorem 1 that $X \sim \mathscr{N}_1(\mu, \sigma^2) := \lim_{\gamma\to 1^+} \mathscr{N}_\gamma(\mu, \sigma^2) = \mathscr{U}(a, b)$, where $\mu := (a + b)/2$ and $\sigma := (b - a)/2$. Recall also the p.d.f. of a p-variate Uniform distributed r.v. $X \sim \mathscr{U}^p(\mu, \Sigma)$, which is formulated by $f_X(x) := \pi^{-p}|\Sigma|^{-1/2}\,\Gamma\big(1 + p/2\big)$, $x \in A$, where $A : Q(x) \le 1$, which is the area inside a p-ellipsoid defined by the quadratic function Q as in (1); see [10]. The corresponding p.d.f. of univariate ($p := 1$) r.v. X is then $f_X(x) = 1/(2\sigma)$, $x \in [\mu - \sigma,\ \mu + \sigma]$. The above p.d.f. can be written alternatively as $f_X(x) = 1/(b - a)$, $x \in [a, b]$, which is the usual formulation of the uniform distribution $\mathscr{U}(a, b)$. Moreover, the c.d.f. of $X \in \mathscr{U}(a, b)$ can be found from (8) for $\gamma \to 1^+$, i.e. for $g \to 0^+$. In order to extract the c.d.f., consider the following limit:

$$\lim_{z\to 0} \frac{\gamma(a, z)}{z^a} = \frac{1}{a}, \quad z \in \mathbb{R}, \quad a \in \mathbb{R}_+. \tag{21}$$

Through the known relation between upper and lower incomplete gamma function, i.e.

$$\gamma(a, z) = \Gamma(a) - \Gamma(a, z), \quad z, a \in \mathbb{R}, \tag{22}$$

(21) implies that

$$\Gamma(a, z) \overset{z\to 0}{\approx} \Gamma(a) - \frac{1}{a} z^a, \quad z \in \mathbb{R}, \quad a \in \mathbb{R}_+, \tag{23}$$

and thus

$$g\,\Gamma\big(g, gz^{1/g}\big) \overset{gz^{1/g}\to 0}{\approx} \Gamma(g + 1) - g^g z, \quad z \in \mathbb{R}, \quad a \in \mathbb{R}_+. \tag{24}$$

Assuming that $|z| := |(x-\mu)|/\sigma \leq 1$, or equivalently $x \in [\mu-\sigma,\ \mu+\sigma]$, then $g \to 0^+$ implies that $gz^{1/g} \to 0^+$. Therefore, (24) yields

$$\lim_{g\to 0^+} g\,\Gamma\left(g, gz^{1/g}\right) = 1 - z, \quad z := \frac{x-\mu}{\sigma} \in \mathbb{R}. \tag{25}$$

Let $X_\gamma \sim \mathscr{N}_\gamma(\mu, \sigma^2)$, $\gamma \in \mathbb{R} \setminus [0,\ 1]$. From Theorem 1 we can then be reduced from r.v. X_γ to $X \sim \mathscr{U}(a, b)$ in limit, as $F_X = \lim_{\gamma\to 1^+} F_{X_\gamma}$. Hence from (8) we obtain, through (25), that

$$F_X(x) = 1 - \tfrac{1}{2}\lim_{g\to 0^+} g\,\Gamma\left(g, g\left(\tfrac{x-\mu}{\sigma}\right)^{1/g}\right) = 1 - \tfrac{1}{2}\left(1 - \frac{x-\mu}{\sigma}\right), \quad x \in [\mu-\sigma,\ \mu+\sigma],$$

and since $\mu := (a+b)/2$ and $\sigma := (b-a)/2$, the above expression gives the well-known formula of the c.d.f. of a uniformly distributed r.v. $X \sim \mathscr{U}(a, b)$, i.e.

$$F_X(x) = \frac{x-\mu+\sigma}{2\sigma} = \frac{x-a}{b-a}, \quad x \in [a,\ b]. \tag{26}$$

Therefore, the hazard rate of $X \sim \mathscr{U}(a, b)$ is finally of the form

$$h_X(x) = \frac{f_X(x)}{1-F_X(x)} = \frac{1}{\mu+\sigma-x} = \frac{1}{b-x}, \quad x \in [a,\ b], \tag{27}$$

while the inverse hazard rate is then given by

$$h_X^{\text{inv}}(x) = \frac{f_X(x)}{F_X(x)} = \frac{1}{x-\mu+\sigma} = \frac{1}{x-a}, \quad x \in [a,\ b]. \tag{28}$$

Notice the symmetry in (27) and (28), and the already well-known result in (26), as these results are reduced from the generalized form in (8), since the GND family includes also the Uniform distribution.

Figure 2 illustrates the surface formed by all the hazard rates curves $h_{X_\gamma}(x)$, $x \in [-3,\ 3]$, for every $\gamma \in [-10,\ 0) \cup [1,\ 10]$ where $X_\gamma \sim \mathscr{N}_\gamma(0, 1)$. The hazard rate of the Uniform ($\gamma = 1$) and Normal ($\gamma = 2$) distributions is clearly depicted. Moreover, the hazard rate of $\mathscr{N}_{\gamma=\pm 10}(0, 1)$, which approximates the Laplace distribution $\mathscr{L}(0, 1) = \mathscr{N}_{\pm\infty}(0, 1)$, as well as the one of $\mathscr{N}_{-0.005}(0, 1)$, which approximates the degenerate Dirac distribution $\mathscr{D}(0)$, is also distinguished.

Another essential measure in Risk Analysis is the future lifetime. The future lifetime is the time remaining until death, given survival to age x_0. In reliability theory the terminology holds for humans, animals and machines. Given age $x_0 = 0$ (i.e. the age of birth, or the time someone just bought a new machine) it is clear that the expected future lifetime equals to the expected lifetime (for the human or for the machine). Sometimes in Industry, the Reliability term of "expected future lifetime" is replaced by "mean residual time". We proceed the application for the GND.

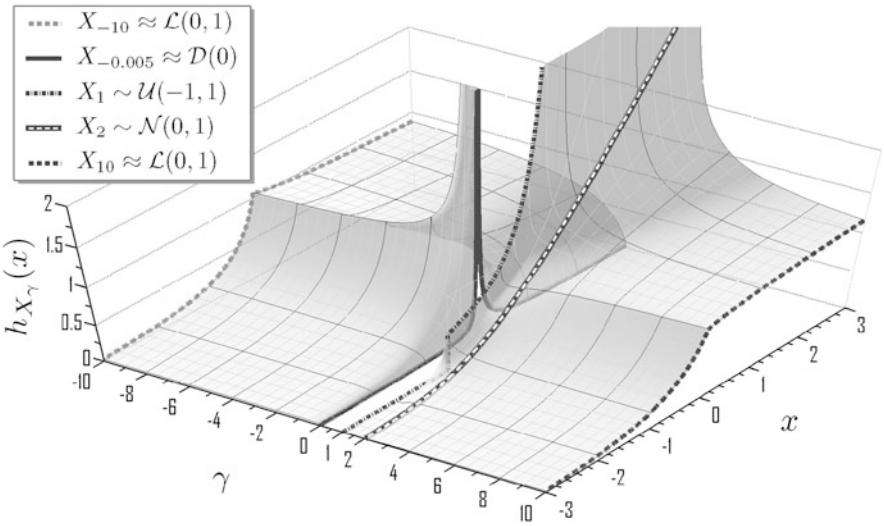


Fig. 2 Surface graph of all the hazard rates $h_{X_\gamma}(x)$ along $x \in [-3,\ 3]$ and $\gamma \in [-10,\ 10]$, where $X_\gamma \sim \mathscr{N}_\gamma(0,1)$

Proposition 3 *The future lifetime r.v.* X_0 *at point* $x_0 \in \mathbb{R}$ *of an r.v.* $X \sim \mathscr{N}_\gamma\left(\mu, \sigma^2\right)$ *has a density function of the form*

$$f_{X_0}(x) = \frac{g^{g-1} \exp\left\{-g \left|\frac{x}{\sigma} + z_0\right|^{1/g}\right\}}{\sigma\, \Gamma\left(g, g z_0^{1/g}\right)}, \quad z_0 := \frac{x_0 - \mu}{\sigma}, \quad x \in \mathbb{R}, \tag{29}$$

while the c.d.f. is given by

$$F_{X_0}(x) = 1 - \frac{\Gamma\left(g, g\left(\frac{x}{\sigma} + z_0\right)^{1/g}\right)}{\Gamma\left(g, g z_0^{1/g}\right)}, \quad x \in \mathbb{R}, \tag{30}$$

The corresponding expected future lifetime is then given by

$$\mathrm{E}(X_0) = \frac{2(\mu - x_0)\, \Gamma(g)}{\Gamma\left(g, g z_0^{1/g}\right)}. \tag{31}$$

See Appendix for the proof.

However, for a more physical meaning it is preferable to consider future lifetime which is related to an r.v. T with a (left) threshold at point 0, in order $T - t_0$ to represent the time remaining until failure given survival until time $t_0 \in \mathbb{R}_+$. The following proposition calculates the future lifetime of a generalized half-normal r.v. (having zero threshold) which is related to the GND. For this reason, we define the

γ-order half-normal generalized distribution (GHND) to be the distribution of an r.v. $|X|$ when $X \sim \mathcal{N}_\gamma\left(0, \sigma^2\right)$. We can then write that $|X| \sim \mathcal{HN}_\gamma\left(\sigma^2\right)$. Trivially, $\mathcal{HN}_2\left(\sigma^2\right)$ is the usual half-normal distribution denoted here with $\mathcal{HN}\left(\sigma^2\right)$. Note that, in case the mean of X is not zero, i.e. when $X \sim \mathcal{N}_\gamma\left(\mu, \sigma^2\right)$, then $|X| \sim \mathcal{FN}_\gamma\left(\mu, \sigma^2\right)$ meaning that r.v. $|X|$ follows now the Folded Normal distribution. For the future lifetime of a GHND we obtain the following.

Proposition 4 *The future lifetime r.v. T_0 at time $t_0 \in \mathbb{R}_+$ of a γ-order half normally distributed r.v. $T \sim \mathcal{HN}_\gamma\left(\sigma^2\right)$ has a density function of the form*

$$f_{T_0}(t) = \frac{2g^{g-1} \exp\left\{-g\left(\frac{t_0+t}{\sigma}\right)^{1/g}\right\}}{\sigma\, \Gamma\left(g, g(t_0/\sigma)^{1/g}\right)}, \quad t \in \mathbb{R}_+, \tag{32}$$

while the c.d.f. is given by

$$F_{T_0}(t) = 1 - \frac{\Gamma\left(g, g\left(\frac{t_0+t}{\sigma}\right)^{1/g}\right)}{\Gamma\left(g, g(t_0/\sigma)^{1/g}\right)}, \quad t \in \mathbb{R}_+, \tag{33}$$

The corresponding expected future lifetime is then given by

$$\mathrm{E}(T_0) = \frac{\sigma\, \Gamma\left(2g, g(t_0/\sigma)^{1/g}\right)}{g^g\, \Gamma\left(g, g(t_0/\sigma)^{1/g}\right)} - t_0. \tag{34}$$

See Appendix for the proof.

4 Discussion

In this paper, we study and calculated the hazard rate and the future life time, as well as the functions related to them, for the GND. The evaluated functions in (11), (12), (14), and (15) hold for all shape parameters γ. Certain values of the shape parameter lead to Laplace and Uniform distributions, and this is discussed in Examples 1 and 2. In Example 1 we came across well-known results, as it is a "simple" case of the GND family. The future lifetime, and the relevant material, is essential also in Risk Analysis, as well as in Reliability Theory. More general results were obtained, while the procedure is: choose the shape parameter, along with the appropriate mean and variance, and then calculate the corresponding produced results.

Appendix

Proof (of Proposition 2) Considering the definition of the cumulative hazard function $H_X := \int h_X$ for an r.v. X, and the fact that

$$\tfrac{\mathrm{d}}{\mathrm{d}\,x}\big(\log S_X\big) = \frac{S_X'}{S_X} = -\frac{f_X}{S_X} = -h_X,$$

where $S_X := 1 - F_X$ denotes the survival function for a r.v. X, we obtain that

$$H_X(x) := \int_{-\infty}^{x} h_X(u)\,\mathrm{d}\,u = \big[-\log S_X(u)\big]_{u=-\infty}^{x} = -\log S_X(x), \quad x \in \mathbb{R},$$

and applying (8) and (9) we finally derive (16) and (17), respectively.

Proof (of Proposition 3) Recall that the future lifetime at time $t_0 \in \mathbb{R}_+$ is defined to be the time remaining until failure (or death), given survival until time t_0. Let X_{t_0}, or X_0, be an r.v. describing the future lifetime of a system described by an r.v. X at time t_0, i.e. $X_0 := X - t_0$ provides the time to failure (of a system with r.v. X) at, or before, time $t + t_0$ given survival until time t_0. The c.d.f. of X_0, which is the probability of failure at, or before, time $t + t_0$ given survival until time t_0, is then written in the form

$$F_{X_0}(t) := \Pr\big(X \le t + t_0 \,\big|\, X > t_0\big) = \frac{\Pr\big(t < X \le t + t_0\big)}{\Pr\big(X > t_0\big)} = \frac{F_X(t_0 + t) - F_X(t_0)}{S_X(t_0)}, \tag{35}$$

for $t \in \mathbb{R}_+$, while the future lifetime probability density of X_0 is then

$$f_{X_0}(t) := \tfrac{\mathrm{d}}{\mathrm{d}\,t} F_{X_0}(t) = \frac{f_X(t + t_0)}{S_X(t_0)}, \quad t \in \mathbb{R}_+. \tag{36}$$

Assuming now that $X \sim \mathcal{N}_\gamma\big(\mu, \sigma^2\big)$ and $t_0 := x_0 \in \mathbb{R}$, the expressions (29) and (30) are derived from (36) and (35), through (1) and (8), respectively.

The corresponding expected future lifetime of X_0 at $x_0 \in \mathbb{R}$ is then given, according to (36), by

$$\mathrm{E}(X_0) := \int_{\mathbb{R}} x f_{X_0}(x)\,\mathrm{d}\,x = \frac{1}{S_X(x_0)} \int_{\mathbb{R}} x f_X(x + x_0)\,\mathrm{d}\,x. \tag{37}$$

Using the linear transformation $u = u(x) := x + x_0$, $x \in \mathbb{R}$, we obtain

$$\mathrm{E}(X_0) = \frac{1}{S_X(x_0)} \int_{\mathbb{R}} (u - x_0) f_X(u)\,\mathrm{d}\,u = \frac{\mu - x_0}{S_X(x_0)},$$

and thus (31) is derived by substituting (8) to the above.

Proof (of Proposition 4) The p.d.f. f_T of r.v. $T := |X|$ can be easily expressed as $f_T = 2f_X$, where f_X denotes the p.d.f. of the r.v. $X \sim \mathcal{N}_\gamma(0, \sigma^2)$. The corresponding c.d.f. F_T can also be expressed through the c.d.f. F_X, as

$$F_T(t) := \int_0^t f_T(u)\,\mathrm{d}\,u = 2\int_{-\infty}^t f_X(t)\,\mathrm{d}\,t - 2\int_{-\infty}^{+\infty} f_X(t)\,\mathrm{d}\,t = 2F_X(t) - 1, \quad t \in \mathbb{R}_+, \tag{38}$$

and through (8) we obtain

$$F_T(t) = 1 - \frac{\Gamma\big(g, g(t/\sigma)^{1/g}\big)}{\Gamma(g)}, \quad t \in \mathbb{R}_+, \tag{39}$$

while the survival function S_T of T is given by

$$S_T(t) = \frac{\Gamma\big(g, g(t/\sigma)^{1/g}\big)}{\Gamma(g)}, \quad t \in \mathbb{R}_+. \tag{40}$$

According now to (36) and (35) we easily derive respectively, through (1), (39), and (40), the requested expressions (32) and (33).

The corresponding expected future lifetime of T_0 at $t_0 \in \mathbb{R}$ is given, similarly to (37), by

$$\mathrm{E}(T_0) = \frac{1}{S_T(t_0)}\int_{\mathbb{R}_+} t f_T(t + t_0)\,\mathrm{d}\,t = \frac{1}{S_T(t_0)}\int_{t_0}^{+\infty}(u - t_0) f_T(u)\,\mathrm{d}\,u,$$

where $u = u(t) := t + t_0$. $t \in \mathbb{R}_+$. Integrating by parts, the above is written as

$$\begin{aligned}
\mathrm{E}(T_0) &= \frac{1}{S_T(t_0)}\big[(t - t_0)F_T(t)\big]_{t=t_0}^{+\infty} - \frac{1}{S_T(t_0)}\int_{t_0}^{+\infty} F_T(t)\,\mathrm{d}\,t \\
&= \frac{1}{S_T(t_0)}\left(\lim_{t\to+\infty} t\right) - t_0 - \frac{1}{S_T(t_0)}\int_{t_0}^{+\infty} F_T(t)\,\mathrm{d}\,t \\
&= \frac{1}{S_T(t_0)}\int_{t_0}^{+\infty} 1 - F_T(t)\,\mathrm{d}\,t = \frac{1}{S_T(t_0)}\int_{t_0}^{+\infty} S_T(t)\,\mathrm{d}\,t = \frac{I(t_0)}{\Gamma\big(g, g(t_0/\sigma)^{1/g}\big)},
\end{aligned} \tag{41}$$

where

$$I(t_0) := \int_{t_0}^{+\infty}\Gamma\big(g, g(t/\sigma)^{1/g}\big)\,\mathrm{d}\,t = g^{1-g}\sigma\int_{u_0:=u(t_0)}^{+\infty} u^{g-1}\,\Gamma(g, u)\,\mathrm{d}\,u,$$

and $u = u(t) := g(t/\sigma)^{1/g}$, $t \in \mathbb{R}_+$. Integration by parts yields

$$I(t_0) = g^{-g}\sigma\big[u^g\,\Gamma(g,u)\big]_{u=u_0}^{+\infty} - g^{-g}\sigma\int_{u_0}^{+\infty} u^g\,\mathrm{d}\,\Gamma(g,u). \tag{42}$$

Recall the known limit

$$\lim_{u\to+\infty}\frac{\Gamma(g,u)}{u^{g-1}\,\mathrm{e}^{-u}} = 1, \tag{43}$$

which implies that

$$\lim_{u\to+\infty} u^g\,\Gamma(g,u) = \lim_{u\to+\infty}\frac{u^{2g-1}}{e^u} = 0,$$

and through (22), the definite integral $I(t_0)$ in (42) can be written successively as

$$\begin{aligned} I(t_0) &= -(u_0/g)^g\sigma\,\Gamma(g,u_0) - g^{-g}\sigma\int_{u_0}^{+\infty} u^g\,\mathrm{d}\,\gamma(g,u) \\ &= -(u_0/g)^g\sigma\,\Gamma(g,u_0) - g^{-g}\sigma\int_{u_0}^{+\infty} u^{2g-1}\,\mathrm{e}^{-u}\,\mathrm{d}\,u. \end{aligned} \tag{44}$$

Splitting now the definite integral of (44) into $\int_0^{+\infty} - \int_0^{u_0}$ and apply the definitions of the gamma and the lower incomplete gamma functions respectively, we obtain

$$I(t_0) = -(u_0/g)^g\sigma\,\Gamma(g,u_0) + g^{-g}\sigma\big[\,\Gamma(2g) - \gamma(2g,u_0)\big],$$

and using (22) again,

$$I(t_0) = -(u_0/g)^g\sigma\,\Gamma(g,u_0) + g^{-g}\sigma\,\Gamma(2g,u_0). \tag{45}$$

Finally, substituting $u_0 := u(t_0) = g(t_0/\sigma)^{1/g}$ into (45), and then applying (45) into (41), the requested expected future lifetime of T_0 at time $t_0 \in \mathbb{R}_+$ is then given by (34).

References

1. Coin, D.: A method to estimate power parameter in exponential power distribution via polynomial regression. J. Stat. Comput. Simul. **83**(11), 1981–2001 (2013)
2. Daw, R.H., Pearson, E.S.: Studies in the history of probability and statistics, XXX: Abraham de Moivre's 1733 derivation of the normal curve: a bibliographical note. Biometrika **59**, 677–680 (1972)

3. Edgeworth, F.Y.: XLII. The law of error. Philos. Mag. Ser. 5 **16**(100), 300–309 (1883)
4. Edgeworth, F.Y.: IV. The philisophy of chance. Mind **os-IX**(34), 223–235 (1884)
5. Galton, F.: Natural Inheritance. McMilan, London (1889)
6. Gauss, C.F.: Bestimmung der Genauigkeit der Beobachtungen. Zeitschrift Astronomi **1**, 185–197 (1816)
7. Gómez, E., Gómez-Villegas, M.A., Martin, J.M.: A multivariate generalization of the power exponential family of distributions. Commun. Stat. Theory Methods **27**(3), 589–600 (1998)
8. Kitsos, C.P., Tavoularis, N.K.: Logarithmic Sobolev inequalities for information measures. IEEE Trans. Inf. Theory **55**(6), 2554–2561 (2009)
9. Kitsos, C.P., Toulias, T.L.: New information measures for the generalized normal distribution. Information **1**, 13–27 (2010)
10. Kitsos, C.P., Toulias, T.L.: Inequalities for the Fisher's information measures. In: Rassias, M.Th. (ed.) Handbook of Functional Equations: Functional Inequalities, pp. 281–313. Springer, Berlin (2014)
11. Kitsos, C.P., Toulias, C.P., Trandafir, P.C.: On the multivariate γ–ordered normal distribution. Far East J. Theor. Stat. **38**(1), 49–73 (2012)
12. Kitsos, C.P., Vassiliadis, V.G., Toulias, T.L.: MLE for the γ-order generalized normal distribution. Discuss. Math. Probab. Stat. **34**, 143–158 (2014)
13. Nadarajah, S.: A generalized normal distribution. J. Appl. Stat. **32**(7), 685–694 (2005)
14. Parham, F. Portier, C.: Benchmark Dose Approach. In: Edler, L., Kitsos, C. (eds.) Recent Advances Quantitative Methods in Cancer and Human Health Assessment. Wiley (2005)
15. Pólya, G.: Uber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem. Mathematische Zeitschrift. Springer (1920)
16. Toulias, T.L., Kitsos, C.P.: Generalizations of entropy and information measures. In: Darras, N.J., Rassias M.Th. (eds.) Computation, Cryptography and Network Security, pp. 495–526. Springer (2015)
17. Verdoolaege, G, Scheunders, P.: On the geometry of multivariate generalized Gaussian models. J. Math. Imaging Vision **43**(3), 180–193 (2011)
18. Yu, S., Zhang, A., Li, H.: A review of estimating the shape parameter of the generalized Gaussian distribution. J. Comput. Inf. Syst. **8**(21), 9055–9064 (2012)

Part II
Statistical Modeling and Risk Issues in Several Areas

Wavelet-Based Detection of Outliers in Poisson INAR(1) Time Series

Isabel Silva and Maria Eduarda Silva

Abstract The presence of outliers or discrepant observations has a negative impact in time series modelling. This paper considers the problem of detecting outliers, additive or innovational, single, multiple or in patches, in count time series modelled by first-order Poisson integer-valued autoregressive, PoINAR(1), models. To address this problem, two wavelet-based approaches that allow the identification of the time points of outlier occurrence are proposed. The effectiveness of the proposed methods is illustrated with synthetic as well as with an observed dataset.

1 Introduction

Time series, as any other data, may contain outliers which are observations that look discordant from most of the observations in the dataset. Neglecting the presence of outliers in a time series hinders statistical inference, leading to model misspecification and biased parameter estimation. Since the seminal work of Fox [7] two major approaches for dealing with outliers in time series may be distinguished. One approach advocates the use of robust estimators to reduce the effect of the outlying observations. However, this approach often leads to ignoring observations hence eventually masking the presence of important underlying phenomena, precluding risk analysis. Alternatively, several methodologies for detecting and estimating outliers and other intervention effects have been established for ARMA models. The emphasis has been on iterative procedures and likelihood based statistics, see, for instance, Chang et al. [5], Chen and Liu [6] and Tsay [17]. Also several tailored procedures have been proposed to some nonlinear time series models. However, the

I. Silva (✉)
Faculdade de Engenharia, Universidade do Porto and CIDMA, Porto, Portugal
e-mail: ims@fe.up.pt

M. E. Silva
Faculdade de Economia, Universidade do Porto and CIDMA, Porto, Portugal
e-mail: mesilva@fep.up.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_13

problem of detection and estimation of outliers in time series of counts has received less attention in the literature. Count time series occur in many areas such as telecommunications, actuarial science, epidemiology, hydrology and environmental studies where the detection of outliers may be invaluable in risk assessment.

One of the most popular classes of models for time series of counts is the class of INAR models proposed by Al-Osh and Alzaid [1] and McKenzie [11], extensively studied in the literature and applied to many real-world problems because of its easiness of interpretation. These models are apparently autoregressive models in which the usual multiplication has been replaced by a random operation, called thinning operation (for details, see Scotto et al. [13]) and the innovations are discrete-valued random variables. Barczy et al. [2, 3] proposed Conditional Least Squares estimation of the INAR(1) model parameters contaminated with outliers additive and innovational, assuming that the time points of the outliers occurrence are known, but their sizes are unknown. Recently, Silva and Pereira [15] suggested a Bayesian approach in order to detect additive outliers in PoINAR(1) models.

In this work, procedures to identify the times of outlier occurrence in PoINAR(1) time series using wavelets are proposed. Wavelets are basis functions that combine properties such as localization in time and scale, orthonormality, different degrees of smoothness, compact support and fast implementation, for details see Percival and Walden [12]. In particular, Discrete Wavelet Transform (DWT), which is a powerful tool for a time-scale multi-resolution analysis, is applied. DWT can be considered as filters of different cut-off frequencies used to analyse a signal at different scales. In a first approach, similar to that of Grané and Veiga [8], the so-called detail coefficients derived from DWT, using the Haar wavelet, are compared with a threshold. In a second approach, the parametric resampling method of Tsay [18] is used in order to obtain the empirical distribution of these detail coefficients.

The remainder of this work is organized as follows. Section 2 presents the first-order Poisson Integer-valued AutoRegressive model contaminated with additive and innovational outliers. A brief description of wavelets and DWT is given in Sect. 3. The proposed wavelet-based procedures to detect time of outlier occurrence are explained in Sect. 4. The proposed procedures are illustrated and compared with synthetic data in Sect. 5. Furthermore, the methods are also applied on an observed dataset. Finally, Sect. 6 concludes the paper.

2 Poisson INAR(1) Model Contaminated with Outliers

Motivated by the need of modelling correlated series of counts, several models for integer-valued time series were proposed in the literature. One of them is the INteger AutoRegressive model proposed by Al-Osh and Alzaid [1] and McKenzie [11]. This model is based on the binomial thinning operation, proposed by Steutel and Van Harn [16], which is defined on a non-negative integer-valued random variable X by $\alpha \circ X = \sum_{k=1}^{X} Y_k$, where $\alpha \in [0, 1]$ and $\{Y_k\}$, $k = 1, \ldots, X$,

is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables, independent of X, with $P(Y_k = 1) = 1 - P(Y_k = 0) = \alpha$. This sequence is called the counting series of $\alpha \circ X$. Note that, $\alpha \circ X | X \sim \text{Bi}(X, \alpha)$. For an account of the properties of the thinning operation, see Silva and Oliveira [14].

Let $\{X_t\}$ be a discrete time, positive integer-valued stochastic process. It is said to be a PoINAR(1) process if it satisfies the following equation,

$$X_t = \alpha \circ X_{t-1} + e_t, \tag{1}$$

where $e_t \sim \text{Poisson}(\lambda)$, is the so-called arrival process, $0 < \alpha < 1$, and for each t, all counting series of $\alpha \circ X_{t-1}$ are mutually independent and independent of $\{e_t\}$. Under these conditions, the process is strictly stationary and $X_t \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$ if $X_0 \sim \text{Poisson}(\frac{\lambda}{1-\alpha})$.

A time series is affected by an additive outlier (AO) if an external error or exogenous change occurs on a certain time point, affecting only this observation and not entering the dynamics of the process. Formally, a contaminated PoINAR(1) with $I \in \mathbb{N}$ additive outliers with magnitude $\omega_i \in \mathbb{N}$ at time points $s_i \in \mathbf{N},\ i = 1, \ldots, I$ can be defined as follows:

$$Y_t = X_t + \sum_{i=1}^{I} \delta_{i,s_i} \omega_i,$$

where X_t is a PoINAR(1) model satisfying (1) and $\delta_{k,m} = 1$, if $k = m$; $\delta_{k,m} = 0$, if $k \neq m$, is an indicator function.

On the other hand, an innovational outlier (IO) can be considered as an internal change or endogenous effect on the noise process, affecting all subsequent observations. Thus, the observed time series $Y_1, \ldots, Y_n$ is a PoINAR(1) process contaminated with $I \in \mathbb{N}$ innovational outliers with size ω_i at time points s_i, $i = 1, \ldots, I$ if it satisfies the following equation

$$Y_t = \alpha \circ Y_{t-1} + \eta_t,$$

with $\eta_t = e_t + \sum_{i=1}^{I} \delta_{i,s_i} \omega_i$, where $e_t \sim \text{Poisson}(\lambda)$ and I, s_i, ω_i and $\delta_{k,m}$ are defined as before.

Note that in both cases, the underlying outlier free process X_t is unobserved.

3 Brief Description of Discrete Wavelet Transform

A wavelet is a function that can be considered as a small wave which grows and decays in a limited time period, for details see Percival and Walden [12]. Similarly to Fourier analysis that uses sinusoidal functions to find the frequency components

contained in a signal, wavelet analysis uses shifted and scaled versions of a so-called wavelet mother to provide the time localization of each spectral component. Formally, a (mother) wavelet is any real-valued function $\psi(\cdot)$ defined on $\mathbb{R}$ satisfying $\int_{-\infty}^{\infty} \psi(u)\, du = 0$, $\int_{-\infty}^{\infty} \psi^2(u)\, du = 1$, and $0 < \int_0^{\infty} \frac{|\boldsymbol{\Psi}(f)|^2}{f}\, df < \infty$, where $\boldsymbol{\Psi}(f) = \int_{-\infty}^{\infty} \psi(u)\, \mathrm{e}^{-i\, 2\pi\, f\, u}\, du$ is the Fourier transform of $\psi(\cdot)$.

Following Percival and Walden [12], let $\mathbf{X} = \{X_t, t = 0, \ldots, N-1\}$ be a time series (or signal), with $N = 2^J$, $J \in \mathbb{N}$. The DWT coefficients $\mathbf{W} = \{W_n, n = 0, \ldots, N-1\}$ are defined by

$$\mathbf{W} = \mathscr{W}\mathbf{X} \quad \Leftrightarrow \quad [\mathbf{W}_1 \ \ldots \mathbf{W}_J \ \mathbf{V}_J]^T = [\mathscr{W}_1 \ \ldots \mathscr{W}_J \ \mathscr{V}_J]^T \mathbf{X},$$

where $\mathscr{W}$ is an $N \times N$ orthonormal matrix of dilations and translations of the mother wavelet $\psi(\cdot)$, defined as $\frac{1}{\sqrt{d}} \psi\left(\frac{u-t}{d}\right)$ with dilation d and translation t parameters taking dyadic values, i.e., $d = 2^j$ and $t = k2^j$, for $j, k \in \mathbb{Z}$. Note that, for $j = 1, \ldots, J$, $\mathbf{W}_j$ is a column vector with $N/2^j$ elements that contains all the DWT coefficients for scale $\tau_j = 2^{j-1}$, $\mathbf{V}_J$ contains the scaling coefficients $\mathbf{W}_{N-1}$, associated with average on scale $d_J = 2^J$, $\mathscr{W}_j$ has dimension $N/2^j \times N$ and $\mathscr{V}_J$ is $1 \times N$.

The wavelet coefficients of white noise or Gaussian data are themselves white noise or Gaussian random variables, respectively, see Percival and Walden [12]. Furthermore, as referred by Bilen and Huzurbazar [4] and Percival and Walden [12], wavelet coefficients in $\mathbf{W}_j$ are approximately uncorrelated even when the data is highly correlated and they allow the reconstruction of the time series. The synthesis of $\mathbf{X}$ (inverse DWT) is given by $\mathbf{X} = \mathscr{W}^T\mathbf{W} = \sum_{j=1}^{J} \mathscr{W}_j^T \mathbf{W}_j + \mathscr{V}_J^T \mathbf{V}_J = \sum_{j=1}^{J} \mathscr{D}_j + \mathscr{A}_J$, where $\mathscr{D}_j$ is called the jth *level wavelet detail* and $\mathscr{A}_J$ has all its elements equal to the sample mean of the time series. For $1 \le j \le J-1$, the jth *level wavelet smooth* is $\mathscr{A}_j = \sum_{k=j+1}^{J} \mathscr{D}_k + \mathscr{A}_J$, and can be considered as an approximation (smoother version) of $\mathbf{X}$.

In practice, the discrete wavelet transform (DWT) matrix $\mathbf{W}$ is computed through a so-called pyramid algorithm introduced by Mallat [9] that uses linear filtering and downsampling operations. More specifically, for a even width L, consider a wavelet filter $\{h_l : l = 0, \ldots, L-1\}$, which is a high-pass filter, and a scaling filter $g_l = (-1)^{l+1} h_{L-1-l}$, that is a low-pass filter. In the first step of the pyramidal algorithm, two sets of coefficients are produced by the convolution of $\mathbf{X}$ with the low-pass filter $\{g_l\}$ (producing the first level approximation coefficients $c\mathbf{A}_1$) and with the high-pass filter $\{h_l\}$ (deriving the first level detail coefficients $c\mathbf{D}_1$), and then a downsample is performed (retain every other filtered value). The next step divides the first level approximation coefficients in two sequences using the same procedure, replacing $\mathbf{X}$ by $c\mathbf{A}_1$ and computing $c\mathbf{A}_2$ and $c\mathbf{D}_2$. Therefore, at level j, the decomposition of $\mathbf{X}$ has the following structure $[c\mathbf{A}_j, c\mathbf{D}_j, c\mathbf{D}_{j-1}, \ldots, c\mathbf{D}_1]$.

The detail coefficients capture certain features of the time series, such as sudden changes, peaks, or spikes, presenting large values in the presence of these singularities, and therefore they can be used to detect outliers. In general, the first

level of decomposition is enough to analyse time series contaminated with outliers Bilen and Huzurbazar [4] and Grané and Veiga [8].

There are many mother wavelets. In this work, the Haar wavelet (among the many mother wavelets) is used. Since it can be considered as a square wave defined by

$$\psi(t) = \begin{cases} -1/\sqrt{2}, & -1 \leq t \leq 0 \\ 1/\sqrt{2}, & 0 < t \leq 1 \\ 0, & \text{otherwise,} \end{cases}$$

the Haar wavelet is more suitable for count data. In this case, low-pass filters correspond to moving averages of the observations and high-pass filters correspond to moving differences of the observations.

4 Procedures to Detect the Time of Outliers Occurrence

In this section, two wavelet-based methods for detecting the time of occurrence of outliers in PoINAR(1) processes are described. The procedures can be summarized in the following steps:

Step 1 Given an observed time series of counts, $\mathbf{Y} = \{Y_t, t = 0, \ldots, N\}$, fit a PoINAR(1) model and compute the resulting Pearson residuals[1] $\mathbf{Z} = \{\hat{z}_t, t = 1, \ldots, N-1\}$, given by $\hat{z}_t = \dfrac{Y_t - (\hat{\alpha} Y_{t-1} + \hat{\lambda})}{\sqrt{\hat{\alpha}(1-\hat{\alpha})Y_{t-1} + \hat{\lambda}}}$.

Step 2 The DWT is applied to the Pearson residuals to obtain the first level detail coefficients, $c\mathbf{D}_1 = (d_1, d_2, \ldots, d_{N/2})$.

Step 3a **Threshold approach**:

(i) Set the threshold k_1^a (discussed in Sect. 4.1).

(ii) The set of (ordered) indices, $\mathbf{S} = \{s_1, \ldots, s_I\}$, containing the positions of the detail coefficients which are above the threshold k_1^a is obtained. As in Grané and Veiga [8], the problem of masking[2] is avoided by searching the outliers recursively. This means that for each outlier detected, $\mathbf{Z}$ is reconstructed applying the inverse discrete wavelet transform (IDWT) to modified detail coefficients where the largest (in absolute value) detail coefficient above the threshold is set to zero. The procedure ends when no more outliers are detected.

[1] $Z_t = \frac{Y_t - \mathrm{E}[Y_t|Y_{t-1}]}{\sqrt{\mathrm{Var}(Y_t|Y_{t-1})}}$.

[2] Masking occurs when one outlier prevents others from being detected.

Step 3b **Parametric resampling approach**:

(i) Compute the acceptance envelope (discussed in Sect. 4.2).

(ii) The set of (ordered) indices, $\mathbf{S} = \{s_1, \ldots, s_I\}$, containing the positions of the detail coefficients which are outside of the acceptance envelope is calculated.

Step 4 The exact position of the outlier in the residual series is obtained as in Grané and Veiga [8]: let s be a generic element of $\mathbf{S}$, compute the sample mean of $\mathbf{Z}$ without the observations $2s$ and $2s-1$, given by $\bar{z}_{N-2} = \dfrac{1}{N-2} \displaystyle\sum_{i \neq 2s, 2s-1} \hat{z}_i$; the time of the outlier occurrence in the residual series is $2s$ if $|\hat{z}_{2s} - \bar{z}_{N-2}| > |\hat{z}_{2s-1} - \bar{z}_{N-2}|$, or equal to $2s-1$ otherwise.

As noted by Bilen and Huzurbazar [4] and Grané and Veiga [8], the first level coefficients detect only the beginning of an outliers patch and therefore, when searching for patches of outliers it is necessary to use the second level detail coefficients, $c\mathbf{D}_2$. Thus, in **Step 3a** there are two thresholds $k_1^{a_1}$ and $k_2^{a_2}$, corresponding to the first and second levels of detail coefficients, respectively. Similarly, there are two acceptance envelopes, one for $c\mathbf{D}_1$ and one for $c\mathbf{D}_2$, in **Step 3b**.

4.1 Setting the Threshold

In the non-Gaussian context of this work, there are no results available for the distribution of the detail coefficients. Thus Monte Carlo simulations are used to obtain the empirical distribution of the maximum of the detail coefficients (in absolute value) for the Pearson residuals of PoINAR(1) models. Then a threshold is computed as follows. For each (α, λ) in the set $\{(\alpha, \lambda) : \alpha = (2k+1) \times 10^{-1}, k = 0, \ldots, 4; \lambda = 2k+1, k = 0, \ldots, 14\}$, 20000 replications of the corresponding PoINAR(1) process are generated for each sample size $N = 2^J + 1$, for $J = 7, \ldots, 10$. The model is fitted, the Pearson residuals, $\hat{z}_i$, for $i = 1, \ldots, N-1$, are computed and the maximum of the first and second level detail coefficients are obtained. The thresholds $k_1^{a_1}$ and $k_2^{a_2}$ are set as the $100(1-a)$th percentiles of the corresponding empirical distributions, for $a = a_1$ or $a = a_2$. The results[3] indicate that the thresholds vary not only with the sample size N but also with the specific combination of the parameters α and λ. Therefore, adopting a conservative strategy, for each sample size N the thresholds are set to the minimum obtained for all the combinations of parameters in each level of decomposition. The obtained thresholds are shown in Table 1.

[3] Available from the authors.

Table 1 Empirical threshold values corresponding to 90th and 95th percentiles of the maximum of the detail coefficients (first and second level), in absolute value, for PoINAR(1) Pearson residuals

N	128	256	512	1024
$k_1^{0.05}$	3.469	3.694	3.886	4.118
$k_1^{0.1}$	3.182	3.450	3.657	3.840
$k_2^{0.05}$	3.157	3.347	3.518	3.691
$k_2^{0.1}$	2.936	3.138	3.320	3.504

4.2 Computing the Acceptance Envelope

Tsay [18] proposed to obtain the empirical distribution of a chosen functional using bootstrap samples generated from a fitted model, and then compare the observed value for the series with this distribution. For this purpose, an acceptance envelope is obtained from the $100(1 - \mathtt{a}/2)$th and $100\mathtt{a}/2$th percentiles of this empirical distribution. If the fitted model is adequate, the functional of interest of the original data should be within the envelope. In this work, the functionals of interest are the first and second level detail coefficients of the Pearson residuals of PoINAR(1) model. Thus, for several sample sizes $N = 2^J+1$, $J = 7, 8, 9$, and parameter values $\{(\alpha, \lambda) : \alpha \in \{0.1, 0.5, 0.9\}; \lambda \in \{1, 5, 9, 13\}\}$, 20000 realizations of PoINAR(1) process are generated and the corresponding Pearson residuals are estimated. For each series of Pearson residuals, the DWT is applied to obtain the first and second level detail coefficients, $c\mathbf{D}_1$ and $c\mathbf{D}_2$, and the acceptance envelopes are constructed from the 0.01th and 99.99th percentiles[4] of the empirical distribution of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively. Once again, the results[5] show that the acceptance envelopes vary not only with the sample size N but also with the combination of the parameter values (α, λ). Therefore, assuming a conservative strategy, for each sample size, an acceptance envelope with the minimum amplitude is chosen. The acceptance envelopes are available from the authors upon request.

5 Simulation Study and Illustration

This section presents the results of a simulation study designed to evaluate and compare the performance of the procedures described above (implemented in Matlab [10]). For these purposes, the percentage of correct detections and the average number of false outliers detected in 1000 repetitions are computed. In each

[4] In the performed simulation study, the detail coefficients present a large variability. Therefore, as a compromise between correct and false detection of outliers, it is found that a reasonable acceptance envelope is constructed from the 0.01th and 99.99th extreme percentiles.

[5] Available from the authors.

repetition, a realization of a PoINAR(1) process with parameters in the set $\{(\alpha, \lambda) : \alpha \in \{0.1, 0.5, 0.8\}; \lambda \in \{1, 3, 5\}\}$ is contaminated with single (1) or multiple (3) outliers either additive or innovational, randomly placed, with integer-valued magnitude $\omega = \lceil 5\sigma_X \rceil, \lceil 10\sigma_X \rceil$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x. The Pearson residual series are obtained and the procedures described in Sect. 4 are applied. Several sample sizes are considered, $N = 128, 256, 512$. Some of the results are shown in Tables 2 and 3 for the threshold $k_1^{0.05}$ and the acceptance envelope constructed from the 0.01th and 99.99th percentiles of the empirical distribution of $c\mathbf{D}_1$.

For the case of contamination with 1 outlier (Table 2), the complete set of results shows that the procedures are sensitive to the increasing of the magnitude of the outlier (AO or IO) but none of the approaches presents better performance than the other. The percentage of correct detection is similar for both types of outliers. When the outlier magnitude is equal to $\lceil 10\sigma_X \rceil$, for the threshold approach the minimum percentage of correct detections is 98.2% and 99.1% for AO and IO cases, respectively; while for the parametric resampling approach, the minimum values are 97.8% for the AO case and 98.8% for the IO case. The average number of false outlier detection is slightly bigger for the AO cases, where the maximum average number of false outliers detected is 0.794 for the threshold approach and 0.985 for

Table 2 Percentage of correct detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, contaminated with 1 additive outlier or 1 innovational outlier, with magnitude $\lceil 5\sigma_X \rceil$ and $\lceil 10\sigma_X \rceil$

			1 additive outlier				1 innovational outlier			
			% correct		Average false		% correct		Average false	
(α, λ)	N	ω	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.1, 1)	128	$\lceil 5\sigma_X \rceil = 6$	81.8	72.5	0.088	0.05	69.9	63.4	0.092	0.069
		$\lceil 10\sigma_X \rceil = 11$	98.2	97.8	0.07	0.05	99.7	98.8	0.094	0.061
	256	$\lceil 5\sigma_X \rceil = 6$	64	81.8	0.114	0.128	67.4	63.6	0.168	0.147
		$\lceil 10\sigma_X \rceil = 11$	98.7	99.1	0.102	0.105	99.9	99	0.122	0.144
	512	$\lceil 5\sigma_X \rceil = 6$	78.1	91.8	0.185	0.268	60.3	66.7	0.163	0.293
		$\lceil 10\sigma_X \rceil = 11$	100	100	0.166	0.239	100	100	0.18	0.284
(0.5, 3)	128	$\lceil 5\sigma_X \rceil = 13$	73	99	0.047	0.03	73.6	63.4	0.096	0.046
		$\lceil 10\sigma_X \rceil = 25$	100	99.9	0.002	0.013	99.9	100	0.077	0.049
	256	$\lceil 5\sigma_X \rceil = 13$	64.7	99.6	0.064	0.059	67.4	84.2	0.098	0.086
		$\lceil 10\sigma_X \rceil = 25$	99.8	99.9	0.085	0.143	100	100	0.132	0.103
	512	$\lceil 5\sigma_X \rceil = 13$	98.5	99.3	0.095	0.152	64.9	86.1	0.123	0.26
		$\lceil 10\sigma_X \rceil = 25$	99.7	100	0.158	0.087	100	100	0.113	0.225
(0.8, 5)	128	$\lceil 5\sigma_X \rceil = 25$	97.9	97.7	0.023	0.026	98.1	95.7	0.049	0.04
		$\lceil 10\sigma_X \rceil = 50$	100	100	0.51	0	100	100	0.053	0.028
	256	$\lceil 5\sigma_X \rceil = 25$	91.5	94.4	0.391	0.404	97.2	96.1	0.071	0.064
		$\lceil 10\sigma_X \rceil = 50$	100	100	0	0	100	100	0.059	0.067
	512	$\lceil 5\sigma_X \rceil = 25$	92.5	96.5	0.524	0.087	98.7	98.9	0.068	0.156
		$\lceil 10\sigma_X \rceil = 50$	100	100	0.001	0.004	100	100	0.077	0.12

Table 3 Percentage of correct detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, contaminated with 3 additive outlier or 3 innovational outlier, with magnitude $\lceil 5\sigma_X \rceil$ and $\lceil 10\sigma_X \rceil$

			3 additive outliers				3 innovational outliers			
			% correct		Average false		% correct		Average false	
(α, λ)	N	ω	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.5, 1)	128	$\lceil 5\sigma_X \rceil = 8$	80.9	30.3	0.021	0.041	83.1	73.7	0.082	0.05
		$\lceil 10\sigma \rceil = 15$	100	99.9	0.009	0.002	99.8	99.9	0.039	0.02
	256	$\lceil 5\sigma_X \rceil = 8$	79.9	77.4	0.032	0.052	77.8	78.6	0.124	0.146
		$\lceil 10\sigma_X \rceil = 15$	66.1	100.0	0.078	0.011	99.9	99.9	0.09	0.096
	512	$\lceil 5\sigma_X \rceil = 8$	88.2	79.5	0.085	0.158	69.9	66.1	0.198	0.383
		$\lceil 10\sigma_X \rceil = 15$	99.9	99.9	0.031	0.055	99.6	100.0	0.171	0.282
(0.8, 3)	128	$\lceil 5\sigma_X \rceil = 20$	99.6	89.6	0.011	0.026	97.2	97.3	0.021	0.01
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.011	0.006	100	100	0.011	0.003
	256	$\lceil 5\sigma_X \rceil = 20$	90.0	90.5	0.165	0.199	99.3	98.6	0.05	0.056
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.087	0.077	100	100	0.028	0.022
	512	$\lceil 5\sigma_X \rceil = 20$	91.4	94.3	0.384	0.576	97.7	97.0	0.062	0.128
		$\lceil 10\sigma_X \rceil = 39$	100	100	0.833	0	100	100	0.063	0.083
(0.1, 5)	128	$\lceil 5\sigma_X \rceil = 12$	57.5	44.0	0.026	0.013	54.7	47.7	0.042	0.026
		$\lceil 10\sigma_X \rceil = 24$	99.6	99.5	0.034	0.021	99.8	99.8	0.027	0.012
	256	$\lceil 5\sigma_X \rceil = 12$	54.2	50.7	0.039	0.043	51.2	51.9	0.057	0.054
		$\lceil 10\sigma_X \rceil = 24$	99.9	99.9	0.027	0.025	99.9	99.8	0.058	0.062
	512	$\lceil 5\sigma_X \rceil = 12$	39.9	57.3	0.07	0.114	43.6	29.5	0.062	0.129
		$\lceil 10\sigma_X \rceil = 24$	99.8	99.9	0.028	0.098	99.9	99.8	0.057	0.112

the parametric resampling approach. In the IO cases, the values are 0.184 and 0.379 for the first and second approaches, respectively.

In the case of contamination with 3 outliers, the results presented in Table 3 show that the percentage of correct detections decreases marginally with respect to Table 2. The analysis of the complete set of results for multiple outliers shows that in general for IO case the threshold approach seems preferable since it leads to a higher percentage of correct detections while the mean number of false detections is comparable to the parametric approach. On the other hand, for AO case the parametric approach leads to a higher percentage of correct detections but also to an increase of 70% in the mean number of false detections.

Finally, to examine the performance of the procedures to detect patches of outliers, Table 4 presents the percentages of correct (complete) detections and partial detections and the average number of false patches detected, in 1000 repetitions. As before, in each repetition, the Pearson residuals series are obtained from a realization of a PoINAR(1) model, for several samples sizes and combinations of parameter values. In each realization, a patch with 3 additive outliers, with magnitude equal to $\lceil 10\sigma_X \rceil$, is placed randomly. The threshold approach has been applied with the 90th percentiles of the empirical distribution of the maximum of the absolute value of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively $k_1^{0.1}$ and $k_2^{0.1}$ (see Table 1). For each level of

Table 4 Percentage of correct and partial detections and average number of false outliers detected, in 1000 repetitions of PoINAR(1) models with sample sizes $N+1$ for some parameter values, with a patch of 3 additive outliers, with magnitude $\lceil 10\sigma_X \rceil$

			% correct		% partial		Average false	
(α, λ)	ω	N	Thresh.	Env.	Thresh.	Env.	Thresh.	Env.
(0.8, 1)	$\lceil 10\sigma_X \rceil = 23$	128	100	63.9	0	34.7	0.001	1
		256	100	71.3	0	0	0	0.779
		512	100	99.9	0	0.1	0.001	0.986
(0.1, 3)	$\lceil 10\sigma_X \rceil = 19$	128	69.1	60.8	0.1	38.5	0.077	1
		256	98.8	100	0	0	0.053	0.313
		512	99.5	99.9	0	0.1	0.02	0.616
(0.5, 5)	$\lceil 10\sigma_X \rceil = 32$	128	100.0	99.9	0	0	0.023	0.999
		256	100.0	99.7	0	0	0.011	0.012
		512	99.8	100	0	0	0.01	0.167

decomposition, in the parametric resampling approach, the acceptance envelopes are constructed from the 0.01th and 99.99th percentiles of the empirical distribution of $c\mathbf{D}_1$ and $c\mathbf{D}_2$, respectively. The results indicate that the threshold approach presents a better performance. However, the percentage of the partial detection obtained in the parametric resampling approach indicates that the results can be improved by tuning the acceptance envelope of the second level of decomposition of DWT.

Note that, since the outliers (single, multiple or patch) are placed randomly, if they appear in the first observation, both approaches have a poor performance. The same happens when two outliers are placed in subsequent observations, since it can be considered as a patch.

As a final illustration of the described procedures, consider the real dataset with 241[6] observations concerning the number of different IP addresses (in periods of 2 min length) at the server of the Department of Statistics of the University of Würzburg on November 29th, 2005, between 10 a.m. and 6 p.m., represented in Fig. 1 and studied by Silva and Pereira [15] and Weiß[19]. The values of sample mean ($\bar{x} = 1.32$) and sample variance ($\hat{\sigma}^2 = 1.39$) and the analysis of the sample autocorrelation and partial autocorrelation functions indicate that a PoINAR(1) model can be fitted to this dataset. By applying both approaches to outlier occurrence time detection to this dataset, an outlier is detected at $t = 224$ (corresponding to $S = \{112\}$). Figure 2 represents the threshold and the acceptance envelope for this illustration. The detection of the outlier at $t = 224$ agrees with the results in Weiß[19] and Silva and Pereira [15]. The former reference indicates as true value $X_{224} = 1$ while in the latter reference the authors use a Bayesian approach that detects an outliers at $t = 224$ with probability 0.99 and estimates $\hat{\alpha} = 0.27$, $\hat{\lambda} = 0.89$ and $\omega = 7$.

[6]Since 241 is not a power of two, by default Matlab extends the signal by using symmetric-padding (symmetric boundary value replication).

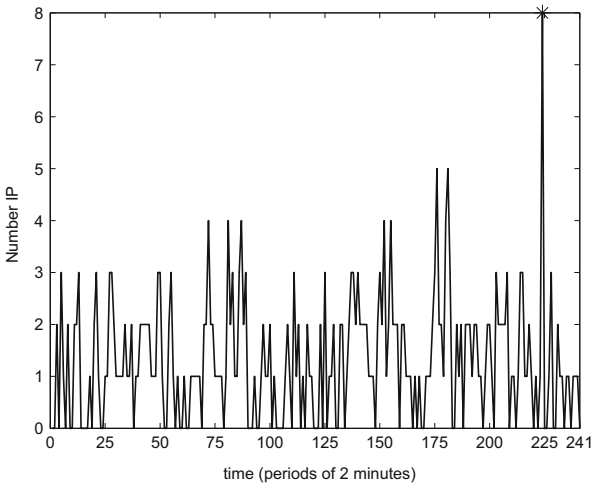


Fig. 1 Chronogram of the IP dataset

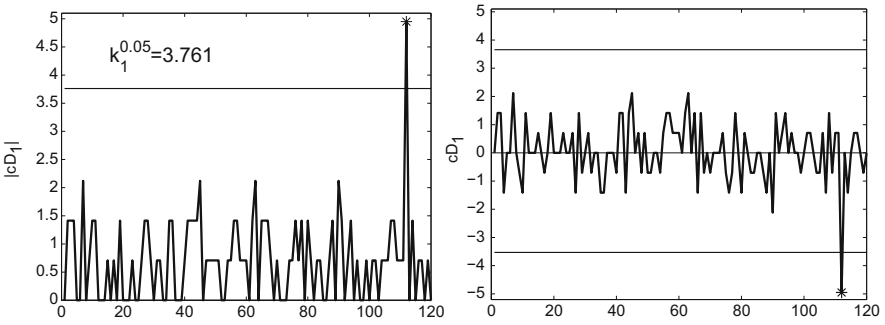


Fig. 2 Results of threshold approach (left panel) and parametric resampling approach (right panel) on the IP dataset

6 Final Remarks

Parametric wavelet-based methods for the detection of outlier occurrences are described. The procedures use the Haar DWT of the Pearson residuals of the PoINAR(1) model. In a first approach, a threshold based on the empirical distribution of the maximum of the (first and second levels) detail coefficients is used. In a second approach, an acceptance envelope constructed from the empirical distribution of these detail coefficients is obtained through parametric resampling methods. The procedures do not require previous knowledge on the number of outliers and are adequate to detect one or multiple outliers, of different types, additive or innovational and patches of additive outliers. However, an open issue is the discrimination of the two types of outliers.

DWT can only be applied when the sample size of the time series is a power of two. To overcome this limitation, the proposed approaches to outlier detection

can use the modified version of DWT, designated by Maximum Overlap DWT (MODWT), introduced by Percival and Walden [12], since MODWT can be applied for a time series of any length.

The performance of the proposed procedures is illustrated with synthetic and real count data. The results show that the methods are efficient and reliable. As far as it is known, this is the first work that treats patches of outliers in the counting time series context. Improvements are still possible by calibrating the percentiles of the empirical distributions used to detect the time of outlier occurrence, either in the threshold approach or in the parametric resampling approach. Different applications may need different significance levels.

The procedures proposed can be applied in other contexts and can also be extended to detect changes in the structure and dynamics of the processes.

Acknowledgements The authors would like to thank the referees for their comments which helped to improve the paper and to Aurea Grané for supplying the programs of the paper [8]. This work is partially supported by Portuguese funds through the CIDMA and the Portuguese Foundation for Science and Technology ("FCT—Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013.

References

1. Al-Osh, M.A., Alzaid, A.A.: First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. **8**, 261–275 (1987)
2. Barczy, M., Ispány, M., Pap, G., Scotto, M., Silva, M.E.: Innovational outliers in INAR(1) models. Commun. Stat. Theory Methods **39**, 3343–3362 (2010)
3. Barczy, M., Ispány, M., Pap, G., Scotto, M., Silva, M.E.: Additive outliers in INAR(1) models. Stat. Pap. **53**, 935–949 (2011)
4. Bilen, C., Huzurbazar, S.: Wavelet-based detection of outliers in time series. J. Comput. Graph. Stat. **11**, 311–327 (2002)
5. Chang, I., Tiao, G.C.,Chen, C.: Estimation of time series parameters in the presence of outliers. Technometrics **30**, 193–204 (1988)
6. Chen, C., Liu, L.M.: Joint estimation of model parameters and outlier effects in time series. J. Am. Stat. Assoc. **88**, 284–297 (1993)
7. Fox, A.J.: Outliers in time series. J. R. Stat. Soc. Ser. B **34**, 350–363 (1972)
8. Grané, A., Veiga, H.: Wavelet-based detection of outliers in financial time series. Comput. Stat. Data Anal. **54**, 2580–2593 (2010)
9. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 674–693 (1989)
10. MATLAB Release 2012a. The MathWorks, Inc., Natick, MA, USA
11. McKenzie, E.: Some simple models for discrete variate time series. Water Resour. Bull. **21**, 645–650 (1985)
12. Percival, D., Walden, A.: Wavelet Methods for Time Series Analysis. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York (2006)
13. Scotto, M.G., Weiß, C.H., Gouveia, S.: Thinning-based models in the analysis of integer-valued time series: a review. Stat. Model. **15**, 590–618 (2015)
14. Silva, M.E., Oliveira, V.L.: Difference equations for the higher-order moments and cumulants of the INAR(1) model. J. Time Ser. Anal. **25**, 317–333 (2004)

15. Silva, M.E., Pereira, I.: Detection of additive outliers in Poisson INAR(1) time series. In: Bourguignon, J.P. et al. (eds.) Mathematics of Energy and Climate Change. CIM Series in Mathematical Sciences, pp. 377–388. Springer, Berlin (2015)
16. Steutel, F.W., Van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**, 893–899 (1979)
17. Tsay, R.S.: Time series model specification in the presence of outliers. J. Am. Stat. Assoc. **81**, 132–141 (1986)
18. Tsay, R.S.: Model checking via parametric bootstraps in time series analysis. J. R. Stat. Soc. Ser. C Appl. Stat. **41**, 1–15 (1992)
19. Weiß, C.H.: Controlling correlated processes of Poisson counts. Qual. Reliab. Eng. Int. **23**, 741–754 (2007)

Surveillance in Discrete Time Series

Maria da Conceição Costa, Isabel Pereira, and Manuel G. Scotto

Abstract The analysis of low integer-valued time series is an area of growing interest as time series of counts arising from many different areas have become available in the last three decades. Statistical quality control, computer science, economics and finance, medicine and epidemiology and environmental sciences are just some of the fields that we can mention to point out the wide variety of contexts from which discrete time series have emerged.

In many of these areas it is not just the statistical modelling of count data that matters. For instance, in environmental sciences or epidemiology, surveillance and risk analysis are critical and timely intervention is mandatory in order to ensure safety and public health. Actually, a major issue in the analysis of a large variety of random phenomena relates to the ability to detect and warn the occurrence of a catastrophe or some other event connected with an alarm system.

In this work, the principles for the construction of optimal alarm systems are discussed and their implementation is described. As there is no unifying approach to the modelling of all integer-valued time series, we will focus our attention in the class of observation-driven models. The implementation of the optimal alarm system will be described in detail for a particular non-linear model in this class, the INteger-valued Asymmetric Power ARCH, or, in short, INAPARCH(p, q).

M. C. Costa (✉) · I. Pereira
Departamento de Matemática and CIDMA, University of Aveiro, Aveiro, Portugal
e-mail: lopescosta@ua.pt; isabel.pereira@ua.pt

M. G. Scotto
Departamento de Matemática and CEMAT, Instituto Superior Técnico, University of Lisbon, Lisboa, Portugal
e-mail: manuel.scotto@tecnico.ulisboa.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_14

1 Introduction

An **alarm system** is an algorithm which, based on current information, predicts whether a level-crossing event is going to occur at a specified time in the future. As to remind that level-crossings do sometimes have very drastic consequences, the designation **catastrophe** is commonly used. Considering level-crossing events, we can distinguish between an exceedance and an up-crossing event. An exceedance is a one-dimensional level-crossing event where some critical level or threshold u is exceeded by a process at one single time point. An up-crossing is a two-dimensional level-crossing event involving two adjacent time points: the process is below the critical threshold at the first time point and above the threshold at the second time point. Throughout this work a catastrophe will thus be considered as the up-crossing event $C_{t,j} = \{X_{t+j-1} \leq u < X_{t+j}\}$ for some $j \in \mathbf{N}$ and some real u. At each moment, the algorithm of the alarm system signals whether or not a catastrophe is bound to happen j time steps ahead. An alarm is a **false alarm** if, after an alarm signal, no catastrophe occurs at the specified time; a catastrophe is said to be **undetected** if the catastrophe occurs without the previous alarm signalling. The success of the alarm system is measured by its false alarm rate and by its detection probability and the definition of *optimal* alarm involves a compromise between these two characteristics, referred to as the **operating characteristics** of the alarm system.

Lindgren [7] and de Maré [5] set the principles for the construction of optimal alarm systems. Establishing the analogy between alarm systems and hypothesis testing, [5] developed a general context optimal alarm system based on a likelihood-ratio argument. The alarm problem can be thought of as an hypothesis test where one has to choose whether to give an alarm or not. de Maré [5] showed that the Neyman-Pearson lemma gives a condition for this test to be optimal. Lindgren [7] restated this condition, giving an explicit formulation of the optimal alarm system in terms of the pair predicted value/predicted growth rate, for a Gaussian stationary process. The optimal alarm system is bound to give an alarm when the prediction exceeds a variable alarm level that adjusts according to the expected growth rate of the process. The optimal alarm condition is then, fundamentally, an alarm region (or decision boundary) that is defined by the likelihood ratio between predicted value and growth rate. Further developments on the construction of optimal alarm systems, recent applications and some alternative approaches on the analysis of risk can be found in [3].

The remaining part of this introductory section briefly presents basic definitions and the theoretical fundamentals of the method, for reader's convenience. In Sect. 2 the Integer-valued APARCH model is presented and the construction of the optimal alarm system is carried out for the particular INAPARCH(1,1) case. In Sect. 3 an application to the number of transactions in stocks is presented. The conditional maximum likelihood (CML) estimation procedure is used to model two real data series concerning the number of transactions per minute of two different stocks and the application of the optimal alarm system is illustrated. Section 4 concludes the

paper with a few remarks we would like to point out, some open questions and ideas for further work.

1.1 Optimal Alarm Systems: Basic Definitions

Let $(Y_t)_{t\in\mathbf{N}}$ be a discrete time stochastic process with parameter space $\Theta \subset \mathbf{R}^k$, for some fixed $k \in \mathbf{N}$. The time sequel $\{1, 2, \ldots, t-1, t, t+1, \ldots\}$ is divided into three sections, $\{1, 2, \ldots, t-q\}$, $\{t-q+1, \ldots, t\}$, and $\{t+1, \ldots\}$, namely, the past, the present and the future. For some $q > 0$, the sets $D_t = \{Y_1, \ldots, Y_{t-q}\}$, $\mathbf{Y_2} = \{Y_{t-q+1}, \ldots, Y_t\}$ and $\mathbf{Y_3} = \{Y_{t+1}, \ldots\}$ represent, respectively, the data or informative experience, the present experiment and the future experiment, at time point t.

Definition 1 The catastrophe[1] is defined as the up-crossing event of the fixed level u, at time point $t+j$, for some $j \in \mathbf{N}$ and for some real u:

$$C_{t,j} = \{Y_{t+j-1} \leq u < Y_{t+j}\}.$$

Definition 2 Any event $A_{t,j}$ in the σ-field generated by $\mathbf{Y_2}$, predictor of $C_{t,j}$, will be an event predictor or alarm.

It is said that an alarm is given at time t for the catastrophe $C_{t,j}$, if the observed value of $\mathbf{Y_2}$ belongs to the predictor event or alarm region. In addition, the alarm is said to be correct if the event $A_{t,j}$ is followed by the event $C_{t,j}$. Thus, the probability of correct alarm will be defined as the probability of catastrophe conditional on the alarm being given. Conversely, a false alarm is defined as the occurrence of $A_{t,j}$ without $C_{t,j}$. If an alarm is given when the catastrophe occurs, it is said that the catastrophe is detected and the probability of detection will be defined as the probability of an alarm being given conditional on the occurrence of the catastrophe.

Definition 3 The alarm region $A_{t,j}$ is said to have size $\alpha_{t,j}$ if $\alpha_{t,j} = P(A_{t,j}|D_t)$.

Note that $\alpha_{t,j}$ can be understood as the proportion of time spent in the alarm state.

Definition 4 The alarm region $A_{t,j}$ is optimal of size $\alpha_{t,j}$ if

$$P(A_{t,j}|C_{t,j}, D_t) = \sup_{B\in\sigma_{\mathbf{Y_2}}} P(B|C_{t,j}, D_t), \tag{1}$$

where the supreme is taken over all sets $B \in \sigma_{\mathbf{Y_2}}$ such that $P(B|D_t) = \alpha_{t,j}$.

In other words, Definition 4 states that the alarm region $A_{t,j}$ of size $\alpha_{t,j}$ is optimal, if it has the highest detection probability, among all regions with the same alarm size.

[1] A catastrophe is generally defined as any event of interest in the σ-field generated by $\mathbf{Y_3}$.

Definition 5 An optimal alarm system of size $\alpha_{t,j}$ is a family of alarm regions $(A_{t,j})$ in time, satisfying (1).

The following lemma enables to obtain the optimal alarm region as a ratio of two conditional probabilities, which is very useful when one turns into the practical construction of the alarm system.

Lemma 1 *Let $p(\mathbf{y_2}|D_t)$ and $p(\mathbf{y_2}|C_{t,j}, D_t)$ be the predictive density of $\mathbf{Y_2}$ and the predictive density of $\mathbf{Y_2}$ conditional on the event $C_{t,j}$, respectively. Then, the alarm system $(A_{t,j})$ with alarm region given by*

$$A_{t,j} = \left\{ \mathbf{y_2} \in \mathbf{R}^q : \frac{p(\mathbf{y_2}|C_{t,j}, D_t)}{p(\mathbf{y_2}|D_t)} \geq k_{t,j} \right\},$$

or, equivalently,

$$A_{t,j} = \left\{ \mathbf{y_2} \in \mathbf{R}^q : \frac{P(C_{t,j}|\mathbf{y_2}, D_t)}{P(C_{t,j}|D_t)} \geq k_{t,j} \right\},$$

for a fixed $k_{t,j}$ such that $P(\mathbf{Y_2} \in A_{t,j}|D_t) = \alpha_{t,j}$ is optimal of size $\alpha_{t,j}$.

If (Y_t) is an integer-valued process, simple adaptations of the previous lemma are required. In the discrete case, $p(\mathbf{y_2}|D_t)$ represents the predictive probability of $\mathbf{Y_2}$ and $p(\mathbf{y_2}|C_{t,j}, D_t)$, the predictive probability of $\mathbf{Y_2}$ conditional on the event $C_{t,j}$. In this case, $\mathbf{y_2} \in \mathbf{N}_0^q$, also. This lemma ensures that the alarm region defined above renders the highest detection probability. Moreover, to enhance the fact that the optimal alarm system depends on the choice of $k_{t,j}$, it is important to stress that, due to the fact that $P(C_{t,j}|D_t)$ does not depend on $\mathbf{y_2}$, the alarm region can be rewritten in the form

$$A_{t,j} = \{\mathbf{y_2} \in \mathbf{R}^q : P(C_{t,j}|\mathbf{y_2}, D_t) \geq k\}, \tag{2}$$

where $k = k_{t,j} P(C_{t,j}|D_t)$ is chosen in some optimal way to accommodate conditions over the operating characteristics of the alarm system.

Definition 6 The following probabilities are called the operating characteristics of an alarm system:

1. $P(A_{t,j}|D_t)$—Alarm size,
2. $P(C_{t,j}|A_{t,j}, D_t)$—Probability of correct alarm,
3. $P(\underline{A}_{t,j}|C_{t,j}, D_t)$—Probability of detecting the event,
4. $P(\overline{C}_{t,j}|A_{t,j}, D_t)$—Probability of false alarm,
5. $P(\overline{A}_{t,j}|C_{t,j}, D_t)$—Probability of undetected event.

The choice of k will depend on a compromise between maximizing the probabilities of correct alarm and of detecting the event. As it is not possible, in general, to maximize both alarm characteristics simultaneously, some criteria must be found in order that the alarm system achieves a satisfactory behaviour. Several criteria have

already been proposed in the literature and this issue will be addressed further on, when dealing with the application of the alarm system to a particular situation.

2 Optimal Alarm Systems: Application to the INAPARCH (1,1) Model

The INteger-valued Asymmetric Power ARCH model, or, in short, INAPARCH(p, q) is a non-linear model in the class of observation driven models for time series of counts. It is the integer-valued counterpart for the APARCH representation for the volatility introduced by Ding et al. [6] and is able to accommodate asymmetric responses relative to the mean of the process. Asymmetric responses on the volatility are also commonly observed in the analysis of time series representing the number of intra-day transactions is stocks, (see [2], e.g.) and this feature could not be addressed by any of the other non-linear models in the class previous to the development of the INAPARCH(p, q). The probabilistic properties and asymptotic theory related to maximum likelihood estimation for this model have already been addressed by the authors and can be found in [4]. Regarding reader's convenience the INAPARCH(1, 1) process is defined as follows. It is an integer-valued process (Y_t) such that

$$\begin{aligned} &Y_t|F_{t-1} \sim Po(\lambda_t) \\ &\lambda_t^{\delta} = \omega + \alpha(|Y_{t-1} - \lambda_{t-1}| - \gamma(Y_{t-1} - \lambda_{t-1}))^{\delta} + \beta\lambda_{t-1}^{\delta},\ t \in \mathbf{Z} \end{aligned}$$

with $\omega > 0, \alpha \geq 0, \beta \geq 0, |\gamma| < 1$ and $\delta \geq 0$.

The application to the INAPARCH(1, 1) model will be done for the particular case $q = 1$ and $j = 2$. Thus, the time sequel is divided in the following manner:

$$D_t = \{y_1, y_2, \ldots, y_{t-1}\} \qquad \mathbf{y_2} = \{y_t\} \qquad \mathbf{y_3} = \{y_{t+1}, y_{t+2}, \ldots\}.$$

The event of interest or the catastrophe is defined as the up-crossing of some fixed level u two steps ahead,

$$C_{t,2} = \{(y_{t+1}, y_{t+2}) \in \mathbf{N}^2 : y_{t+1} \leq u < y_{t+2}\}.$$

The optimal alarm region of size α_2 is given by

$$A_{t,2} = \left\{ y_t \in \mathbf{N} : \frac{P(C_{t,2}|y_t, D_t)}{P(C_{t,2}|D_t)} \geq k_{t,2} \right\} = \left\{ y_t \in \mathbf{N} : P(C_{t,2}|y_t, D_t) \geq k \right\},$$

where $k = k_{t,2}P(C_{t,2}|D_t)$. The first step in the construction of the alarm system consists of the calculation of both probabilities: the probability of catastrophe

conditional on D_t and $\mathbf{y_2}$, i.e., $P(C_{t,2}|y_t, D_t)$, and the probability of catastrophe conditional on D_t, $P(C_{t,2}|D_t)$. Indeed

$$\begin{aligned} P(C_{t,2}|y_t, D_t) &= P(Y_{t+1} \leq u < Y_{t+2}|y_t, D_t) \\ &= \sum_{y_{t+1}=0}^{u} p(y_{t+1}|y_t) \left(1 - \sum_{y_{t+2}=0}^{u} p(y_{t+2}|y_{t+1})\right) \\ &= \sum_{y_{t+1}=0}^{u} \frac{e^{-\lambda_{t+1}} \lambda_{t+1}^{y_{t+1}}}{(y_{t+1})!} \left(1 - \sum_{y_{t+2}=0}^{u} \frac{e^{-\lambda_{t+2}} \lambda_{t+2}^{y_{t+2}}}{(y_{t+2})!}\right) \end{aligned}$$

and

$$\begin{aligned} P(C_{t,2}|D_t) &= P(Y_{t+1} \leq u < Y_{t+2}|D_t) \\ &= \sum_{y_{t+1}=0}^{u} p(y_{t+1}|y_{t-1}) \left(1 - \sum_{y_{t+2}=0}^{u} p(y_{t+2}|y_{t+1})\right) \\ &= \sum_{y_t} \frac{e^{-\lambda_t} \lambda_t^{y_t}}{(y_t)!} \sum_{y_{t+1}=0}^{u} \frac{e^{-\lambda_{t+1}} \lambda_{t+1}^{y_{t+1}}}{(y_{t+1})!} \left(1 - \sum_{y_{t+2}=0}^{u} \frac{e^{-\lambda_{t+2}} \lambda_{t+2}^{y_{t+2}}}{(y_{t+2})!}\right). \end{aligned}$$

Having calculated these probabilities it is then possible to explicit all the operating characteristics.

1. Alarm size
   Since $\mathbf{y_2} = \{y_t\}$, the alarm size is simply

$$\alpha_{t,2} = P(A_{t,2}|D_t) = \sum_{y_t \in A_{t,2}} P(Y_t = y_t|D_t) = \sum_{y_t \in A_{t,2}} \frac{e^{-\lambda_t} \lambda_t^{y_t}}{(y_t)!}, \tag{3}$$

   with $A_{t,2}$ being the alarm region which depends on the choice of $k_{t,2}$.
2. Probability of correct alarm

$$\begin{aligned} P(C_{t,2}|A_{t,2}, D_t) &= \frac{P(C_{t,2} \bigcap A_{t,2}|D_t)}{P(A_{t,2}|D_t)} = \frac{P(Y_{t+1} \leq u < Y_{t+2}, Y_t \in A_{t,2}|D_t)}{P(Y_t \in A_{t,2}|D_t)} \\ &= \frac{\sum_{y_t \in A_{t,2}} p(y_t|y_{t-1}) P(C_{t,2}|y_t, D_t)}{\sum_{y_t \in A_{t,2}} p(y_t|y_{t-1})} \end{aligned}$$

and given $P(C_{t,2}|y_t, D_t)$ it follows that

$$P(C_{t,2}|A_{t,2}, D_t) = \sum_{y_t \in A_{t,2}} \left[\frac{e^{-\lambda_t}\lambda_t^{y_t}}{(y_t)!} \sum_{y_{t+1}=0}^{u} \frac{e^{-\lambda_{t+1}}\lambda_{t+1}^{y_{t+1}}}{(y_{t+1})!} \times \left(1 - \sum_{y_{t+2}=0}^{u} \frac{e^{-\lambda_{t+2}}\lambda_{t+2}^{y_{t+2}}}{(y_{t+2})!}\right)\right] \left[\sum_{y_t \in A_{t,2}} \frac{e^{-\lambda_t}\lambda_t^{y_t}}{(y_t)!}\right]^{-1}.$$

3. Probability of detecting the event

$$P(A_{t,2}|C_{t,2}, D_t) = \frac{P(A_{t,2} \bigcap C_{t,2}|D_t)}{P(C_{t,2}|D_t)} = \frac{P(Y_t \in A_{t,2}, Y_{t+1} \leq u < Y_{t+2}|D_t)}{P(C_{t,2}|D_t)}.$$

Once again, the numerator in this expression is the same as the numerator in the expression for the probability of correct alarm, and, given the probability of catastrophe, $P(C_{t,2}|D_t)$, the above expression can be rewritten as

$$= \sum_{y_t \in A_{t,2}} \left[\frac{e^{-\lambda_t}\lambda_t^{y_t}}{(y_t)!} \sum_{y_{t+1}=0}^{u} \frac{e^{-\lambda_{t+1}}\lambda_{t+1}^{y_{t+1}}}{(y_{t+1})!} \left(1 - \sum_{y_{t+2}=0}^{u} \frac{e^{-\lambda_{t+2}}\lambda_{t+2}^{y_{t+2}}}{(y_{t+2})!}\right)\right] \times \left[\sum_{y_t} \frac{e^{-\lambda_t}\lambda_t^{y_t}}{(y_t)!} \sum_{y_{t+1}=0}^{u} \frac{e^{-\lambda_{t+1}}\lambda_{t+1}^{y_{t+1}}}{(y_{t+1})!} \left(1 - \sum_{y_{t+2}=0}^{u} \frac{e^{-\lambda_{t+2}}\lambda_{t+2}^{y_{t+2}}}{(y_{t+2})!}\right)\right]^{-1}.$$

4. Probability of false alarm

$$P(\overline{C_{t,2}}|A_{t,2}, D_t) = 1 - P(C_{t,2}|A_{t,2}, D_t).$$

5. Probability of undetected event

$$P(\overline{A_{t,2}}|C_{t,2}, D_t) = 1 - P(A_{t,2}|C_{t,2}, D_t).$$

3 Real Data Example

The application of the alarm system was done to two time series of count data generated from stock transactions, namely, the tick-by-tick data for Glaxosmithkline and Astrazeneca for one trading day. Data was downloaded from www.dukascopy.com and treated as explained in [4]. Each series consists of 501 observations and contains a reasonably high number of zeros. The CML estimation procedure considering the INAPARCH(1,1) model was applied (see [4] for details) and the

application of the alarm system was done using the CML estimates. All necessary programs and routines were developed in Matlab. The analysis was done for the time instants $t = 450$ to $t = 460$. A preliminary study was carried out in order to choose the fixed value u. The probabilities $P(C_{t,2}|y_t, D_t)$ and $P(C_{t,2}|D_t)$ and also the alarm region were calculated for different values of u, for all the time instants mentioned. As a result of this preliminary study and in order to have reasonable probabilities of catastrophe, two different values of u were chosen for each data series: the 39th percentile ($Q_{0.39}$) and the 50th percentile ($Q_{0.50}$). It is worth mentioning that as these time series have many zero counts the probability of catastrophe for higher percentiles is very low. Hence, the fixed levels u considered in this application cannot be understood as a catastrophe in the sense that it should be related to a relatively rare event, but it is simply a fixed level for which the probability of up-crossing is not negligible.

In order to obtain the optimal alarm region for each case, it is necessary to obtain the alarm region for several values of k, according to expression (2). For each value of k, the alarm size $\alpha_{t,2}$, the probability of correct alarm $P(C_{t,2}|A_{t,2}, D_t)$, and the probability of detecting the event $P(A_{t,2}|C_{t,2}, D_t)$ are then calculated. For every fixed value of k, the alarm region has to be obtained through a systematic search in a three-dimensional region for $\{y_t, y_{t+1}, y_{t+2}\}$. We considered y_t taking all the integer values from 0 to 150 and determined, for each value of y_t, if $P(C_{t,2}|y_t, D_t)$ exceeds or not k. This procedure is repeated for all the values of k tested. The step and range of variation in k were chosen for each case in order to have as many different situations as possible. Although the analysis was done from $t = 450$ to $t = 460$ for both time series, to illustrate the procedure described, Tables 1 and 2 show the operating characteristics just for time points $t = 456$ and $t = 458$ for Astrazeneca and Glaxosmithkline time series, respectively. The alarm system shows the same behaviour for both time series not only in what concerns the general tendencies of the operating characteristics but also in what concerns the comparison of the level crossings of the 39 and 50th percentiles. Generally speaking we can say that, comparing the level crossings of both percentiles, alarm size always starts at lower values for the 50th percentile than for the 39th percentile, for corresponding time instants. Another general conclusion is that for the level crossing of the 39th percentile, the probability of detection is similar to the alarm size, having the same variation with k. In the case of the 50th percentile, although the probability of detection is slightly higher than the alarm size, the variation with k is also similar. It is not surprising that the probability of detection has the same behaviour as the alarm size, because as the alarm size decreases with the increase in k, the number of alarms decreases, leading directly to a lower probability of detecting the event. On the other hand, as k increases, the probability of the alarm being correct increases. This behaviour is not also unexpected: as the number of alarms decreases, the probability of false alarm also decreases, and, consequently, the probability of the alarm being correct is expected to increase.

As is obvious from the remarks above it is not possible to maximize simultaneously $P(C_{t,2}|A_{t,2}, D_t)$ and $P(A_{t,2}|C_{t,2}, D_t)$. A compromise must be reached between these operating characteristics by a proper choice of k. Several criteria

Table 1 Operating characteristics at time point $t = 456$ for the Astrazeneca series

$u = Q_{0.39} = 19$				$u = Q_{0.50} = 25$			
$t = 456$				$t = 456$			
$P(C_{t,2}\|y_t, D_t) = 0.1464$				$P(C_{t,2}\|y_t, D_t) = 0.0250$			
k	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$	k	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$
0.1520	0.5094	0.1572	0.5270	0.0325	0.3961	0.0418	0.5101
0.1530	0.4373	0.1581	0.4547	0.0425	0.1186	0.0552	0.2018
0.1540	0.3734	0.1588	0.3902	0.0525	0.0534	0.0647	0.1063
0.1550	0.3734	0.1588	0.3902	0.0625	0.0227	0.0744	0.0519
0.1560	0.3300	0.1592	0.3456	0.0725	0.0088	0.0844	0.0228
0.1570	0.2673	0.1599	0.2813	0.0825	0.0031	0.0945	0.0089
0.1580	0.2229	0.1604	0.2352	0.0925	0.0017	0.0994	0.0053
0.1590	0.2229	0.1604	0.2352	0.1025	5.2170×10^{-4}	0.1087	0.0017
0.1600	0.1306	0.1612	0.1385	0.1125	1.4272×10^{-4}	0.1170	5.1454×10^{-4}
0.1610	0.0738	0.1615	0.0784	0.1225	1.7229×10^{-5}	0.1260	6.6871×10^{-5}

Table 2 Operating characteristics at time point $t = 458$ for the Glaxosmithkline series

$u = Q_{0.39} = 13$				$u = Q_{0.50} = 18$			
$t = 458$				$t = 458$			
$P(C_{t,2}\|y_t, D_t) = 0.1884$				$P(C_{t,2}\|y_t, D_t) = 0.0977$			
k	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$	k	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$
0.1962	0.7517	0.1979	0.7584	0.0754	0.3925	0.0862	0.4485
0.1964	0.7256	0.1980	0.7324	0.0854	0.1697	0.0960	0.2160
0.1966	0.7256	0.1980	0.7324	0.0954	0.0709	0.1062	0.0998
0.1968	0.7256	0.1980	0.7324	0.1054	0.0276	0.1159	0.0424
0.1970	0.6852	0.1980	0.6917	0.1154	0.0093	0.1257	0.0156
0.1972	0.6852	0.1980	0.6917	0.1254	0.0028	0.1355	0.0050
0.1974	0.6075	0.1981	0.6136	0.1354	0.0014	0.1401	0.0027
0.1976	0.5504	0.1982	0.5561	0.1454	1.6640×10^{-4}	0.1532	3.3802×10^{-4}
0.1978	0.5504	0.1982	0.5561	0.1554	3.4203×10^{-5}	0.1610	7.3031×10^{-5}
0.1980	0.4764	0.1982	0.4814	0.1654	6.2986×10^{-6}	0.1676	1.4003×10^{-6}
0.1982	0.3864	0.1983	0.3905	0.1754	1.5678×10^{-7}	0.1778	3.6973×10^{-7}

have already been proposed in the literature. For instance, [1] suggested that k should be chosen so that the alarm size is about twice the probability of having a catastrophe given the past values of the process, $P(C_{t,2}|D_t) \simeq \frac{1}{2}P(A_{t,2}|D_t)$, meaning that in this situation the system spends twice the time in the alarm state than in the catastrophe region. The first criterion used in this application is a variation of the former. Since the alarm size is given by $P(A_{t,2}|D_t)$ and, as was seen above, the probability of detecting the event has the same behaviour of the alarm size with the variation in k, taking also similar values, we decided to substitute the alarm size with the detection probability. Moreover, we also found that the probability of correct alarm is always of the same order of the probability of catastrophe given past values of the process, $P(C_{t,2}|D_t)$: the difference between these two probabilities never exceeds 0.02. As such, we also substituted $P(C_{t,2}|D_t)$ by $P(C_{t,2}|A_{t,2}, D_t)$, the probability of correct alarm. Therefore, our Criterion 1 relates directly to operating characteristics and is $P(A_{t,2}|C_{t,2}, D_t) \simeq 2P(C_{t,2}|A_{t,2}, D_t)$.

Another criterion found in the literature is the one suggested by Svensson et al. [8], in which k should be chosen so that the probability of correct alarm and the probability of detecting the event are approximately equal. Our Criterion 1 is already related with these two operating characteristics. Also, because the probability of detection is directly dependent on the alarm size, it can be chosen to be as high as desired. Thus, it seems wise to look for the best set of operating characteristics in a different perspective, looking towards minimizing the number of false alarms, which is the same as maximizing the probability of the alarm being correct. As the probability of the alarm being correct increases, the detection probability decreases and, in order not to have too small detection probability we state the Criterion 2 as: Maximum $P(C_{t,2}|A_{t,2}, D_t)$, as long as $P(A_{t,2}|C_{t,2}, D_t) \geq 0.001$.

The online prediction is illustrated in Tables 3 and 4. The informative experience evolves as the time instant varies from $t = 450$ to $t = 460$ and D_t is updated at each time instant. The probability of catastrophe given the past experience, the alarm region and respective operating characteristics are presented, for each criteria. The analysis was done for the level crossings of both 39th and 50th percentiles but only the results for the first case are shown here. Table 3 refers to the fixed level crossing $u = Q_{0.39} = 19$ for the Astrazeneca series and Table 4 to the fixed level crossing $u = Q_{0.39} = 13$ for the Glaxosmithkline series. One general remark regarding the online prediction system is that Criterion 2, which tends to minimize the number of false alarms, is always satisfied for a higher value of k, when compared with Criterion 1. As was already discussed from previous results, a higher value of k implies smaller alarm size and smaller probability of detection, which in turn results in greater probability of correct alarm. This observation is not surprising since a greater probability of correct alarm is actually the main goal of Criterion 2.

In order to test the alarm system, three extra values of both time series were simulated: $(\mathbf{y_2}, \mathbf{y_3}) = (y_t, y_{t+1}, y_{t+2})$. This procedure was repeated 100,000 times with the same informative experience, D_t, for each series. Considering the alarm regions obtained before for $u = Q_{0.39}$ and for $u = Q_{0.50}$ and for the two criteria already mentioned, it was observed for each of the 100,000 samples whether an alarm was given or not and whether a catastrophe occurred or not. The

Table 3 Astrazeneca series: operating characteristics at different time points, with Criteria 1 and 2, considering $u = Q_{0.39} = 19$

Criterion	t	$P(C_{t,2}\|D_t)$	k	Alarm region	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$
1	450	0.0655	0.0855	$\{0, \ldots, 6\} \cup \{22, \ldots, 47\}$	0.0853	0.1037	0.1350
	451	0.0604	0.0804	$\{0, \ldots, 5\} \cup \{21, \ldots, 48\}$	0.1060	0.0944	0.1656
	452	0.0381	0.0581	$\{0, \ldots, 3\} \cup \{20, \ldots, 53\}$	0.0721	0.0715	0.1353
	453	0.0388	0.0588	$\{0, \ldots, 3\} \cup \{20, \ldots, 53\}$	0.0744	0.0721	0.1381
	454	0.0575	0.0775	$\{0, \ldots, 5\} \cup \{21, \ldots, 49\}$	0.0973	0.0924	0.1564
	455	0.1085	0.1185	$\{0, \ldots, 11\} \cup \{23, \ldots, 41\}$	0.2127	0.1323	0.2594
	456	0.1520	0.1560	$\{5, \ldots, 15\} \cup \{25, \ldots, 31\}$	0.3300	0.1592	0.3456
	457	0.1156	0.1256	$\{0, \ldots, 11\} \cup \{23, \ldots, 39\}$	0.2318	0.1372	0.2751
	458	0.1345	0.1420	$\{0, \ldots, 13\} \cup \{24, \ldots, 36\}$	0.2818	0.1498	0.3138
	459	0.1485	0.1525	$\{1, \ldots, 15\} \cup \{25, \ldots, 32\}$	0.3267	0.1576	0.3467
	460	0.1094	0.1194	$\{0, \ldots, 11\} \cup \{23, \ldots, 41\}$	0.2148	0.1330	0.2611
2	450	0.0655	0.1455	$\{30, \ldots, 39\}$	9.2971×10^{-4}	0.1518	0.0022
	451	0.0604	0.1404	$\{29, \ldots, 41\}$	0.0013	0.1455	0.0032
	452	0.0381	0.1181	$\{28, \ldots, 45\}$	5.5679×10^{-4}	0.1283	0.0019
	453	0.0388	0.1188	$\{28, \ldots, 45\}$	5.9269×10^{-4}	0.1287	0.0020
	454	0.0575	0.1375	$\{29, \ldots, 41\}$	0.0011	0.1443	0.0028
	455	0.1085	0.1585	$\{30, \ldots, 34\}$	0.0074	0.1603	0.0109
	456	0.1520	0.1610	$\{9, \ldots, 12\} \cup \{28, 29\}$	0.0738	0.1615	0.0784
	457	0.1156	0.1606	$\{31, 32\}$	0.0042	0.1616	0.0058
	458	0.1345	0.1595	$\{0, \ldots, 5\} \cup \{29, \ldots, 32\}$	0.0310	0.1611	0.0371
	459	0.1485	0.1615	$\{8, 9\} \cup \{29\}$	0.0232	0.1617	0.0252
	460	0.1094	0.1594	$\{30, \ldots, 33\}$	0.0073	0.1604	0.0107

Table 4 Glaxosmithkline series: operating characteristics at different time points, with Criteria 1 and 2, considering $u = Q_{0.39} = 13$

Criterion	t	$P(C_{t,2}\|D_t)$	k	Alarm region	α_2	$P(C_{t,2}\|A_{t,2}, D_t)$	$P(A_{t,2}\|C_{t,2}, D_t)$
1	450	0.1725	0.1755	$\{0, \ldots, 8\} \cup \{16, \ldots, 33\}$	0.3495	0.1822	0.3691
	451	0.1688	0.1718	$\{0, \ldots, 8\} \cup \{16, \ldots, 35\}$	0.3423	0.1794	0.3638
	452	0.1364	0.1424	$\{0, \ldots, 6\} \cup \{15, \ldots, 44\}$	0.2763	0.1528	0.3094
	453	0.1789	0.1789	$\{0, \ldots, 9\} \cup \{16, \ldots, 31\}$	0.4200	0.1859	0.4365
	454	0.1969	0.1979	$\{9, \ldots, 12\} \cup \{17, 18\}$	0.3786	0.1982	0.3811
	455	0.1878	0.1888	$\{0, \ldots, 9\} \cup \{17, \ldots, 27\}$	0.3556	0.1938	0.3669
	456	0.1882	0.1892	$\{0, \ldots, 9\} \cup \{17, \ldots, 27\}$	0.3565	0.1940	0.3675
	457	0.1961	0.1981	$\{13, \ldots, 17\}$	0.4761	0.1982	0.4814
	458	0.1962	0.1982	$\{13, \ldots, 16\}$	0.3864	0.1983	0.3905
	459	0.1969	0.1979	$\{10, \ldots, 13\} \cup \{17, 18\}$	0.4389	0.1982	0.4418
	460	0.1951	0.1963	$\{1, \ldots, 10\} \cup \{17, \ldots, 21\}$	0.3963	0.1974	0.4010
2	450	0.1725	0.1965	$\{23, \ldots, 27\}$	0.0094	0.1978	0.0107
	451	0.1688	0.1958	$\{23, \ldots, 28\}$	0.0084	0.1974	0.0099
	452	0.1364	0.1904	$\{25, \ldots, 34\}$	7.1587×10^{-4}	0.1940	0.0010
	453	0.1789	0.1969	$\{22, \ldots, 25\}$	0.0197	0.1979	0.0218
	454	0.1969	0.1982	$\{10, 11\} \cup \{18\}$	0.1820	0.1983	0.1832
	455	0.1878	0.1978	$\{0, 1\} \cup \{21, 22\}$	0.0324	0.1982	0.0342
	456	0.1882	0.1982	$\{0\} \cup \{21, 22\}$	0.0327	0.1982	0.0345
	457	0.1961	0.1983	$\{14\} \cup \{16\}$	0.1964	0.1983	0.1987
	458	0.1962	0.1982	$\{13, \ldots, 16\}$	0.3864	0.1983	0.3905
	459	0.1969	0.1982	$\{11, 12\} \cup \{17\}$	0.2279	0.1983	0.2295
	460	0.1951	0.1981	$\{4, \ldots, 7\} \cup \{19, 20\}$	0.1098	0.1982	0.1116

Table 5 Results for the Astrazeneca series, with $u = Q_{0.39} = 19$

Time instant	Criterion	Alarms		Catastrophes	
		False	Total	Detected	Total
$t = 456$	1	21,011 (0.6369)	32,992	11,981 (0.3755)	31,906
	2	4381 (0.5886)	7443	3062 (0.0948)	32,315
$t = 457$	1	17,464 (0.7505)	23,271	5807 (0.3105)	18,705
	2	249 (0.5818)	428	179 (0.0095)	18,761
$t = 458$	1	19,618 (0.6958)	28,193	8575 (0.3523)	24,340
	2	1820 (0.5938)	3065	1245 (0.0504)	24,713
$t = 459$	1	20,963 (0.6449)	32,508	11,545 (0.3798)	30,396
	2	1417 (0.5984)	2368	951 (0.0313)	30,389
$t = 460$	1	16,254 (0.7655)	21,233	4979 (0.2914)	17,089
	2	464 (0.6097)	761	297 (0.0170)	17,433

Percentages in parenthesis

Table 6 Results for the Glaxosmithkline series, with $u = Q_{0.39} = 13$

Time instant	Criterion	Alarms		Catastrophes	
		False	Total	Detected	Total
$t = 454$	1	20,607 (0.5452)	37,794	17,187 (0.3786)	45,399
	2	9873 (0.5449)	18,119	8246 (0.1819)	45,340
$t = 457$	1	25,776 (0.5397)	47,761	21,985 (0.4496)	48,898
	2	10,609 (0.5401)	19,641	9032 (0.1848)	48,867
$t = 458$	1	21,062 (0.5429)	38,795	17,733 (0.3638)	48,742
	2	21,062 (0.5429)	38,795	17,733 (0.3638)	48,742
$t = 459$	1	24,280 (0.5499)	44,152	19,872 (0.4338)	45,814
	2	12,447 (0.5510)	22,589	10,142 (0.2198)	46,145
$t = 460$	1	22,415 (0.5663)	39,583	17,168 (0.4169)	41,183
	2	5918 (0.5408)	10,944	5026 (0.1226)	41,006

Percentages in parenthesis

operating characteristics can then be estimated with these counts. This procedure was repeated for several time instants and results for the fixed level crossing $u = Q_{0.39}$ are presented in Tables 5 and 6 for the Astrazeneca and Glaxosmithkline series, respectively. The time instants were chosen for their better set of operating characteristics and particularly for the higher values of $P(C_{t,2}|A_{t,2}, D_t)$.

Regarding these results a few conclusions can be outlined. First of all, the results obtained from the application overestimate the operating characteristic of the probability of correct alarm (given as $1 - P(\overline{C_{t,2}}|A_{t,2}, D_t)$, in Tables 5 and 6), whose theoretical value (in Tables 3 and 4) is always around a half of the estimated one. Alarm size and probability of detection are the operating characteristics better estimated with this application. Particularly, the alarm size (not shown directly in Tables 5 and 6, but easily obtainable dividing the total number of alarms by the number of samples, 100,000) always follows the theoretical value up to the second

or third decimal place. Overall, Criterion 1 seems to provide better estimates of the operating characteristics than Criterion 2.

4 Conclusion

In this work the implementation of an optimal alarm system was carried out for the first time for the INAPARCH(1,1) model. It was possible to demonstrate online prediction, which contributes to the minimization of the number of false alarms with the constant update of the informative experience. An application was done with reasonable estimation of the theoretical operating characteristics. This work also establishes new criteria for the optimization of the operating characteristics. Regarding this new criteria, we cannot conclude which one leads to better results. That is something that must be chosen in agreement with the particular application. If one is interested in having a very small number of false alarms, then Criterion 2 should be chosen. If the risk analysis situation demands for a higher probability of detection instead, then Criterion 1 should be preferable, as this criterion looks for the alarm region for which the detection probability is approximately twice the probability of the alarm being correct.

An overall remark we feel necessary is that a theoretical probability of correct alarm that does not exceed 20% may not be enough in many situations. We believe, from previous experience in the application of optimal alarm systems to real-valued time series, that this result is related to the very nature of these particular data sets with small counts and a significant number of zeros. As such, future work involves application of optimal alarm systems to other real data time series, also exhibiting a significant number of zero counts. We would like to explore if rare events can actually be detected in this kind of data series.

Acknowledgements This work was supported in part by the Portuguese Foundation for Science and Technology (FCT—Fundação para a Ciência e a Tecnologia) through CIDMA—Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

References

1. Antunes, M., Amaral-Turkman, M., Turkman, K.: A Bayesian approach to event prediction. J. Time Ser. Anal. **24**, 631–646 (2003)
2. Brännäs, K., Quoreshi, A.M.M.S.: Integer-valued moving average modelling of the number of transactions in stocks. Appl. Financ. Econ. **20**, 1429–1440 (2010)
3. Costa, M.C.: Optimal alarm systems and its application to financial time series. PhD Thesis in Mathematics, University of Aveiro, Portugal (2014). http://hdl.handle.net/10773/12872
4. Costa, M.C., Scotto, M.G., Pereira, I.: Integer-valued APARCH processes. In: Rojas, I., Pomares, H. (eds.) Time Series Analysis and Forecasting, pp. 189–202. Springer International Publishing, Berlin (2016)

5. de Maré, J.: Optimal prediction of catastrophes with applications to Gaussian processes. Ann. Probab. **8**, 841–850 (1980)
6. Ding, Z., Engle, R.F., Granger, C.W.J.: A long memory property of stock market returns and a new model. J. Emp. Financ. **1**, 83–106 (1993)
7. Lindgren, G.: Optimal prediction of level crossings in Gaussian processes and sequences. Ann. Probab. **13**, 804–824 (1985)
8. Svensson, A., Holst, J., Lindquist, R., Lindgren, G.: Optimal prediction of catastrophes in autoregressive moving average processes. J. Time Ser. Anal. **17**, 511–531 (1996)

On the Maxima of Integer Models Based on a New Thinning Operator

Sandra Dias and Maria da Graça Temido

Abstract This paper introduces and studies a non-negative integer-valued process referred to as Ψ-INARMA(1,1), an extension of the geometric ARMA(1,1) process, introduced by McKenzie (Adv Appl Probab 18:679–705, 1986). The Ψ-INARMA(1,1) process is obtained by replacing the binomial thinning operator, proposed in Steutel and van Harn (Ann. Probab. 7:893–899, 1979), by a generalized thinning operator, introduced in Aly and Bouzar (REVSTAT Stat J 6:101–121, 2005). We prove its strictly stationarity and specify its asymptotic independence and local dependence behaviour. As a consequence, we conclude that the sequence of maxima converges in distribution to a discrete Gumbel distribution, when the sequence of innovations belongs to Anderson's class (J Appl Probab 7:99–113, 1970).

1 Introduction

The analysis of non-negative integer-valued time series has received increasing attention of the probabilistic and statistical literature, during the last three decades. This fact is due to the wide applicability of the underlying models in many different areas, where count data arises. In particular the amount of new integer-valued models proposed and studied in the last few years illustrates a great interest in this subject. Some examples of such series include applications in medicine [2], environmental processes [19] and alarm systems [18], among others. However, for

S. Dias (✉)
Pole CMAT-UTAD and CEMAT/Department of Mathematics, School of Sciences and Technology, University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal
e-mail: sdias@utad.pt

M. d. G. Temido
CMUC/Department of Mathematics, Faculty of Sciences and Technology, University of Coimbra, Coimbra, Portugal
e-mail: mgtm@mat.uc.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_15

almost all those models very little is known about their extremal behaviour, because many integer-valued distributions do not belong to the domain of attraction of any max-stable distribution. In this context, in the literature of extremes, we refer to [6, 10, 11] and [12, 13].

The approach, followed by many authors in order to obtain models for integer-valued or count data, consists in replacing the usual multiplication in real classical models by a random thinning operator, which maintains the discrete pattern of the process. This procedure, initially based on the binomial thinning operator, introduced by Steutel and van Harn [20], leads to the first class of INARMA models proposed by McKenzie [15, 17] and Al-Osh and Alzaid [3, 4].

In this paper, we propose a model based on a thinning operator whose details we describe in what follows.

Given an integer random variable (r.v.) Z and $\eta \in]0, 1[$, [5] introduced the thinning operator $\odot_\Psi$, which assigns to the pair (η, Z) the r.v.

$$\eta \odot_\Psi Z \equiv Y_1 + Y_2 + \cdots + Y_Z,$$

where $\{Y_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables (r.v.'s), independent of Z. The probability generating function (p.g.f.) of $\{Y_n\}$, denoted by $\Psi_t(z)$, $t = -\ln \eta$, belongs to the family defined by

$$\Psi_{t_1+t_2}(z) = \Psi_{t_1}(\Psi_{t_2}(z)), \quad |z| \leq r_Y$$

with $\Psi_t(0) \neq 0$, that is, $P(Y = 0) > 0$. In this work, r_Y represents the convergence ratio of the p.g.f. of the r.v. Y. The solution of the previous functional equation is (besides the identity function) $\Psi_t(z) = g^{-1}(g(z) \pm t)$, where g is a strictly increasing function (see [1]). In particular, we have the specific family of probability generating functions (p.g.f.'s)

$$\Psi_t^{(\theta)}(z) = 1 - \frac{\overline{\theta} e^{-\overline{\theta} t}(1 - z)}{\overline{\theta} + \frac{\theta}{2}(1 - e^{-\overline{\theta} t})(1 - z)}, \tag{1}$$

for $|z-1| < 2\overline{\theta}/(\theta(1-e^{-\overline{\theta} t}))$, with $t \geq 0$, $\theta \in [0, 1[$ and $\overline{\theta} = 1-\theta$. This particular family of p.g.f.'s is associated with a mixture of an integer-valued distribution function (d.f.) over $\mathbb{N}_0$ with the Dirac d.f. over $\{0\}$. This kind of mixtures renders appropriate to model zero-inflated count data, for example data associated with the phenomena with structural zeros as well as zeros resulting in the non-occurrence of the underlying event. If $\theta = 0$, then $\Psi_t^{(0)}(z) = 1 - e^{-t} + e^{-t} z$, that is the r.v.'s of $\{Y_n\}$ have a Bernoulli distribution. In this case $\odot_\Psi$ coincides with the well-known binomial thinning operator, denoted here by $\star$.

Aly and Bouzar [5] studied the Ψ-INAR(1) process described by the equation $X_n = \eta \odot_\Psi X_{n-1} + \epsilon_n$, where $0 < \eta < 1$ and $\{\epsilon_n\}$ is a sequence of i.i.d. integer r.v.'s, independent from the r.v.'s of the sequence $\{Y_n\}$.

McKenzie [16] introduced the Geometric ARMA(1,1) process defined by $X_n = \beta \star Z_n + V_n W_{n-1}$, with $W_n = \eta \star W_{n-1} + U_n Z_n$, where $\{Z_n\}$, $\{U_n\}$ and $\{V_n\}$

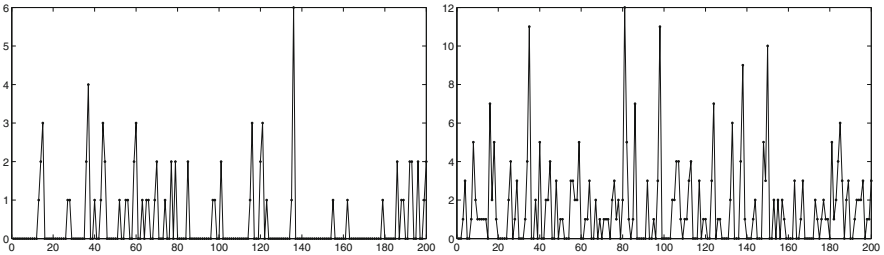


Fig. 1 Two sample paths of 200 observations simulated from a Ψ-INARMA(1,1) process with $\eta = \beta = 0.5$, $\theta = 0.1$ and Z_n has a geometric distribution with parameter 0.7 (left) and 0.4 (right)

are sequences of i.i.d. r.v.'s, $\{U_n\}$ and $\{V_n\}$ both have a Bernoulli distribution (with parameters $1-\eta$ and $1-\beta$, respectively) and W_0 is independent of all other r.v.'s.

In this work we consider an extension of McKenzie's process in the form

$$X_n = \beta \odot_\Psi Z_n + V_n W_{n-1}, \text{ where } W_n = \eta \odot_\Psi W_{n-1} + U_n Z_n,$$

with $\{Z_n\}$, $\{U_n\}$ and $\{V_n\}$ under the same assumptions, which we denote by Ψ-INARMA(1,1) process. In Fig. 1 we include two sample paths from this process, when the family Ψ_t^θ given by (1) is taken. We observe that if the common distribution of $\{Y_n\}$ is zero-inflated then the distribution of $\{W_n\}$ and $\{X_n\}$ also tends to be zero-inflated when an appropriate distribution for $\{Z_n\}$ is considered. This pattern (zero-inflated and not) is illustrated in the sample paths of Fig. 1.

Looking for a well-defined limit in distribution for the sequence of maxima of the process under consideration is the aim of this paper. We start by proving some useful properties of the operated variable $\eta \odot_\Psi Z$, that will be used henceforth. This is the content of Sect. 2. In Sect. 3 we present the proof of the strict stationarity of the process and Sect. 4 contains the main result of this work, which is a generalization of the well-known Leadbetter's Extremal Types Theorem. Then we analyse the asymptotic independence and local dependence of the process driven by conditions $D_{k_n}(u_n)$ and $D'_{k_n}(u_n)$, introduced in [22], where $\{k_n\}$ is a non-decreasing sequence of integers such that

$$\lim_{n\to+\infty} \frac{k_{n+1}}{k_n} = r > 1. \tag{2}$$

We now recall such conditions. Let $\{k_n\}$ be an increasing sequence of positive integers satisfying (2) and $\{u_n\}$ a real sequence. The sequence $\{X_n\}$ satisfies condition $D_{k_n}(u_n)$ if for any integers $1 \le i_1 < \ldots < i_p < j_1 < \ldots < j_q \le k_n$, with $j_1 - i_p > \ell_n$ and $A_j := \{X_j \le u_n\}$, we have

$$\left|P\Big(\bigcap_{s=1}^{p} A_{i_s}, \bigcap_{m=1}^{q} A_{j_m}\Big) - P\Big(\bigcap_{s=1}^{p} A_{i_s}\Big)P\Big(\bigcap_{m=1}^{q} A_{j_m}\Big)\right| \leq \alpha_{n,\ell_n},$$

where $\lim_{n\to+\infty} \alpha_{n,\ell_n} = 0$, for some sequence $\ell_n = o_n(k_n)$.

Taking into consideration strictly stationary processes, [22] proved that, under $D_{k_n}(u_n)$, the limit in distribution for $M_{k_n} = \max(X_1, X_2, \ldots, X_{k_n})$, under linear normalization, whenever it exists, is max-semistable. Following [9] we say that a d.f G on $\mathbb{R}$ is max-semistable if there are reals $r > 1$, $\gamma > 0$ and β such that $G(x) = G^r(x/\gamma + \beta)$, $x \in \mathbb{R}$, or equivalently, if there is a sequence of i.i.d. r.v.'s with d.f. F and two real sequences $\{a_n > 0\}$ and $\{b_n\}$ such that $\lim_{n\to+\infty} F^{k_n}(x/a_n + b_n) = G(x)$, for each continuity point of G, with $\{k_n\}$ satisfying (2). The analytical expressions of these d.f.'s can be found in [9].

Furthermore, condition $D'_{k_n}(u_n)$ holds, if there exists a sequence of positive integers $\{s_n\}$ such that $k_n/s_n \to +\infty$, $s_n\alpha_{n,\ell_n} \to 0$, $n \to +\infty$, and

$$\lim_{n\to+\infty} k_n \sum_{j=2}^{[k_n/s_n]} P\left(X_1 > u_n, X_j > u_n\right) = 0. \tag{3}$$

These conditions are adaptations to the max-semistable context of the well-known Leadbetter's conditions $D(u_n)$ and $D'(u_n)$ [14]. So the limit in distribution of the maximum M_{k_n} of strictly stationary processes which satisfy these two conditions is equal to that of the i.i.d. associated process. Then, taking the results of [22] also into account, we present the following lemma.

Lemma 1 *Let $\{X_n\}$ be a strictly stationary process and $\{k_n\}$ a positive integer-valued sequence satisfying (2). Under $D_{k_n}(u_n)$ and $D'_{k_n}(u_n)$ for some real sequence $\{u_n\}$, the sequences $\{P(M_{k_n} \leq u_n)\}$ and $\{F_X^{k_n}(u_n)\}$, when convergent, have the same limit which is max-semistable.*

Regarding the marginal distribution of the process $\{X_n\}$, we assume that $\{Z_n\}$ belongs to Anderson's class [6] that is, to the class of d.f.'s F satisfying $(1 - F(n-1))/(1 - F(n)) \to r > 1$, $n \to +\infty$. More specifically, to a subclass consisting of d.f.'s that satisfy

$$1 - F(z) \sim A[z]^{\xi} r^{-[z]}, \; z \to +\infty, \tag{4}$$

where $\xi \in \mathbb{R}$, $A > 0$ and $r > 1$, which will be denoted by $\mathscr{C}_{\mathscr{A}}(r)$. Under this hypothesis, we prove that the same happens with $\{W_n\}$ and $\{X_n\}$. Moreover, [21] proved that if $\{Z_n\}$ belongs to Anderson's class, then there is a sequence $\{k_n\}$ satisfying (2), such that the sequence $\{F^{k_n}(x+b_n)\}$ converges to the discrete Gumbel d.f., given by $G(x) = \exp(-r^{-[x]})$, $x \in \mathbb{R}$, which is max-semistable.

Once established the assumptions of Lemma 1, we conclude that the sequence of maxima of the Ψ-INARMA(1,1) process is attracted in distribution to a discrete Gumbel d.f.

2 Properties of the Operated Variable

In this section we characterize the p.g.f., $P_{\eta \odot_\Psi Z}$, assuming that the p.g.f. Ψ_t, associated with the random operator $\odot_\Psi$, has a convergence ratio greater than one. From the second order Taylor's expansion in the neighbourhood of the point $z = 1$, we have

$$\begin{aligned}\Psi_t(z) &= \Psi_t(1) + \Psi_t^{'}(1)(z-1) + \frac{\Psi_t^{''}(\vartheta)}{2}(z-1)^2 \\ &= 1 + E(Y)(z-1) + \frac{\Psi_t^{''}(\vartheta)}{2}(z-1)^2,\end{aligned}$$

where $|\vartheta - 1| < |z-1|$ and $\vartheta := \vartheta(z)$. Taking $h = z - 1$ and $\xi := \frac{\Psi_t^{''}(\vartheta)}{2}$, we obtain

$$\Psi_t(1+h) = 1 + E(Y)h + \xi h^2 = 1 + E(Y) f(h),$$

with $f(h) = h(1 + \xi h/E(Y))$. In [5] it is established that $E(Y) = \eta^{\delta_\Psi}$, with $\delta_\Psi = -\ln \Psi_1^{'}(1)$, so we get $E(\eta \odot_\Psi Z) = \eta^{\delta_\Psi} E(Z)$.

Lemma 2 *Consider the operated r.v. $X = \eta \odot_\Psi Z$ and suppose that P_Z has convergence ratio $r_Z > 1$.*

1. *If $1 + E(Y) f(h) < r_Z$, then*
   a. $P_X(1+h) = P_Z(\Psi_t(1+h)) = 1 + E(Y)E(Z)h(1 + o_h(1))$, $h \to 0$;
   b. $P_X(1+h) \leq (1 + C_1 E(Y) f(h))^2$, *where C_1 is a constant dependent on r_Z and $E(Z)$.*
2. $E((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}) = P_Z(1 + (\eta_1\eta_2)^{\delta_\Psi} f_1(h) f_2(h) + \eta_1^{\delta_\Psi} f_1(h) + \eta_2^{\delta_\Psi} f_2(h))$, *where $f_i(h) = h(1 + \xi h/E(Y^{(i)}))$, $i \in \{1,2\}$, and $1 + (\eta_1\eta_2)^{\delta_\Psi} f_1(h) f_2(h) + \eta_1^{\delta_\Psi} f_1(h) + \eta_2^{\delta_\Psi} f_2(h) < r_Z$.*

Proof

1. a. Let S_Z be the support of Z. We have

$$\begin{aligned}P_X(1+h) &= E((1+h)^X) = E(E((1+h)^X | Z)) \\ &= \sum_{k \in S_Z} \prod_{i=1}^{k} E((1+h)^{Y_i}) P(Z = k)\end{aligned}$$

$$= \sum_{k \in S_Z} (\Psi_t(1+h))^k \, P(Z=k)$$
$$= E\left[(\Psi_t(1+h))^Z\right] = P_Z(\Psi_t(1+h)), \tag{5}$$

where

$$\begin{aligned}\sum_{k \in S_Z} (\Psi_t(1+h))^k \, P(Z=k) &= \sum_{k=1}^{+\infty} (1+E(Y)f(h))^k \, P(Z=k) \\ &= (1+E(Y)f(h))P(Z=1) \\ &\quad + \sum_{k=2}^{+\infty} (1+kE(Y)f(h))P(Z=k) \\ &\quad + \sum_{k=2}^{+\infty} \sum_{j=2}^{k} C_j^k \, (E(Y)f(h))^j \, P(Z=k) \quad (6) \\ &= 1+E(Y)f(h)E(Z) \\ &\quad + \sum_{k=2}^{+\infty} \sum_{j=2}^{k} C_j^k \, (E(Y)f(h))^j \, P(Z=k).\end{aligned}$$

On the other hand, as $C_{j+2}^k = \frac{k(k-1)}{(j+2)(j+1)} C_j^{k-2}$, we get

$$\begin{aligned}\sum_{j=2}^{k} C_j^k \, (E(Y)f(h))^j &= (E(Y)f(h))^2 \sum_{j=0}^{k-2} C_{j+2}^k \, (E(Y)f(h))^j \\ &\leq (E(Y)f(h))^2 \, k^2 \, (1+E(Y)f(h))^{k-2},\end{aligned}$$

whereby the series in (6) does not exceed

$$(E(Y)f(h))^2 \sum_{k=2}^{+\infty} k^2 \, (1+E(Y)f(h))^{k-2} \, P(Z=k). \tag{7}$$

Since $1 + E(Y)f(h) < r_Z$, by D'Alembert criterion for the convergence of positive series, we conclude that this last series in (7) is convergent.

Finally, due to (5), (6) and (7), it follows that

$$P_X(1+h) = 1+E(Y)E(Z)f(h)\left(1+\frac{\sum_{k=2}^{+\infty}\sum_{j=2}^{k} C_j^k\,(E(Y)f(h))^j\,P(Z=k)}{E(Y)E(Z)f(h)}\right)$$

with

$$\begin{aligned} 0 &\le \frac{\sum_{k=2}^{+\infty}\sum_{j=2}^{k} C_j^k\,(E(Y)f(h))^j\,P(Z=k)}{E(Y)E(Z)f(h)} \\ &\le \frac{(E(Y)f(h))^2\sum_{k=2}^{+\infty} k^2\,(1+E(Y)f(h))^{k-2}\,P(Z=k)}{E(Y)E(Z)f(h)} \\ &\le C_1 f(h) \to 0,\ h \to 0^+, \end{aligned}$$

where C_1 is a positive constant. As $f(h) \sim h$, $h \to 0^+$, we complete the proof of 1.a.

b. As the series in (7) is convergent we have

$$\begin{aligned} P_X(1+h) &\le 1 + E(Y)E(Z)f(h) + C(E(Y)f(h))^2 \\ &\le (1 + C_1 E(Y)f(h))^2, \end{aligned}$$

where $C_1 = \max\{\sqrt{C}/E(Z), 1\}$.

2. Assuming that the two counting sequences $\{Y_j^{(1)}\}$ and $\{Y_j^{(2)}\}$ are independent, we get

$$\begin{aligned} E\left((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}\right) &= E\left(E\left((1+h)^{\eta_1 \odot_\Psi Z + \eta_2 \odot_\Psi Z}\right)|Z\right) \\ &= E\left(E\left((1+h)^{\eta_1 \odot_\Psi Z}\right)E\left((1+h)^{\eta_2 \odot_\Psi Z}\right)|Z\right) \\ &= \sum_{i=1}^{+\infty}(\Psi_{-\ln\eta_1}(1+h))^k(\Psi_{-\ln\eta_2}(1+h))^k P(Z=k) \\ &= \sum_{i=1}^{+\infty}(1+\eta_1^{\delta_\Psi} f_1(h))^k(1+\eta_2^{\delta_\Psi} f_2(h))^k P(Z=k). \end{aligned}$$

□

3 Strictly Stationarity of the Process

Since the p.g.f.'s of W_n and $U_n Z_n$, denoted here by P_{W_n} and P_{UZ}, respectively, satisfy $P_{W_n}(z) = P_{W_{n-1}}(\Psi_t(z))P_{UZ}(z)$, $n \in \mathbb{N}$, with $t = -\ln \eta$, assuming that the p.g.f. $\Psi_t(z)$ satisfies $\Psi_{t_1}(\Psi_{t_2}(z)) = \Psi_{t_1+t_2}(z)$, $|z| \leq r_Y$, the process $\{W_n\}$ has the representation

$$W_n \stackrel{d}{=} \eta^k \odot_\Psi W_{n-k} + \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-i} Z_{n-i}, \ \forall\, n \in \mathbb{N}, \ \forall\, k \geq 1.$$

Then, for all $n \in \mathbb{N}$,

$$X_n \stackrel{d}{=} \beta \odot_\Psi Z_n + V_n \left(\eta^k \odot_\Psi W_{n-1-k} + \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i} \right).$$

We prove the strict stationarity of the process following some of the arguments of [7] and [8].

Proposition 1 *If $E(Z) < +\infty$, then the sequence $\{X_n\}$ is strictly stationary.*

Proof Let us remember that if $\{Q_n\}$ is a sequence of r.v.'s with finite mean and $\sum_{n=1}^{+\infty} E(|Q_n|)$ is convergent then the series $\sum_{n=1}^{+\infty} Q_n$ is almost surely (a.s.) absolutely convergent. As $\sum_{i=0}^{+\infty} E(\eta^i \odot_\Psi U_{n-i} Z_{n-i}) = E(UZ) \sum_{i=0}^{+\infty} \eta^{i\delta_\Psi} < +\infty$, we get

$$W_n^{(k)} = \sum_{i=0}^{k-1} \eta^i \odot_\Psi U_{n-i} Z_{n-i} \xrightarrow[k]{q.c.} W_n' := \sum_{i=0}^{\infty} \eta^i \odot_\Psi U_{n-i} Z_{n-i}.$$

On the other hand, due to the fact that the r.v.'s of the sequence $\{U_n Z_n\}$ are i.i.d., the p.g.f.'s of the vectors $(W_n^{(k)}, W_{n+1}^{(k)}, \ldots, W_{n+t}^{(k)})$ and $(W_{n+\ell}^{(k)}, W_{n+1+\ell}^{(k)}, \ldots, W_{n+t+\ell}^{(k)})$ coincide, for any $\ell > 1$, and so $\{W_n^{(k)}\}$ is strictly stationary.

As the almost surely convergence of a vector is equivalent to the almost surely convergence of its margins, we prove that those two vectors converge almost surely to $(W_n', W_{n+1}', \ldots, W_{n+t}')$ and $(W_{n+\ell}', W_{n+1+\ell}', \ldots, W_{n+t+\ell}')$. As almost surely convergence implies convergence in distribution and the limit is unique, we conclude that the vectors $(W_n', W_{n+1}', \ldots, W_{n+t}')$ and $(W_{n+\ell}', W_{n+1+\ell}', \ldots, W_{n+t+\ell}')$ are identically distributed. Since $\eta^k \odot_\Psi W_{n-k} \xrightarrow[k]{a.s.} 0$,

we deduce that $W_n \stackrel{d}{=} \sum_{i=0}^{+\infty} \eta^i \odot_\Psi U_{n-i} Z_{n-i}$ as well as

$$X_n \stackrel{d}{=} \beta \odot_\Psi Z_n + V_n \sum_{i=0}^{+\infty} \eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i},$$

due to the independence of sequences $\{Z_n\}$ and $\{V_n\}$. Then it is proven that $\{W_n\}$ and $\{X_n\}$ are strictly stationary. □

4 Limit Distribution of the Maximum

Let us begin by characterizing the tails of the marginal distribution of the process in study. The next lemma, due to [11], is a key result for this work.

Lemma 3 *Let Y_1 and Y_2 be independent r.v.'s. If $Y_1 \in \mathscr{C}_{\mathscr{A}}(r_{Y_1})$ and Y_2 has finite p.g.f. for some $z > r_{Y_1}$, then $Y_1 + Y_2 \in \mathscr{C}_{\mathscr{A}}(r_{Y_1})$, with A replaced by $AE((r_{Y_1})^{Y_2})$.*

The tail marginal distribution of $\{W_n\}$ and $\{X_n\}$ is related to the distribution of the innovations $\{Z_n\}$ by the following result. Before, we should clarify that if the d.f. of any r.v. Z belongs to Anderson's class, then r is the convergence ratio of P_Z because

$$\frac{P(Z=z)}{P(Z=z+1)} = \frac{\frac{1-F_Z(z-1)}{1-F_Z(z)} - 1}{1 - \frac{1-F_Z(z+1)}{1-F_Z(z)}} \to r,\ z \to +\infty.$$

Theorem 1 *If the margins of $\{Z_n\}$ belong to $\mathscr{C}_{\mathscr{A}}(r_Z)$, then F_{W_n} and F_{X_n} belong to the same class with*

$$P(W_n > z) \sim A^*[z]^\xi r_Z^{-[z]},\ n \to +\infty,$$

and

$$P(X_n > z) \sim A'[z]^\xi r_Z^{-[z]},\ n \to +\infty,$$

where $A^* = AE(r_Z^{\sum_{i=1}^{+\infty} \eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i}})$ and $A' = A^* E\left((\Psi_{-\ln\beta}(r_Z))^Z\right)$.

Proof Since we have

$$X_n \stackrel{d}{=} \beta \odot_\Psi Z_n + V_n U_{n-1} Z_{n-1} + V_n \sum_{i=1}^{+\infty} \eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i}$$

and $V_n U_{n-1} Z_{n-1} \in \mathscr{C}_{\mathscr{A}}(r_Z)$, in order to apply Lemma 3, we first prove that the r.v. $\beta \odot_\Psi Z_n + V_n \sum_{i=1}^{+\infty} \eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i}$ has finite p.g.f., for some $z > r_Z$. Choose $M \geq 1$ such that for $i > M$, $1 + \eta^{i\delta_\Psi} f(h) < r_Z < 1 + h$. Indeed, by Lemma 2, we deduce

$$\begin{aligned}\prod_{i=M}^{+\infty} E\left((1+h)^{\eta^i \odot_\Psi U_{n-1-i} Z_{n-1-i}}\right) &< \prod_{i=M}^{+\infty} \left(1 + C_1 \eta^{i\delta_\Psi} f(h)\right)^2 \\ &\leq \exp\left(\sum_{i=M}^{+\infty} \ln\left(1 + C_1 \eta^{i\delta_\Psi} f(h)\right)^2\right) \\ &\leq \exp\left(\sum_{i=M}^{+\infty} 2C_1 \eta^{i\delta_\Psi} f(h)\right) < +\infty.\end{aligned}$$

As $\beta \odot_\Psi Z_n$ also has finite p.g.f., for some $z > r_Z$, Lemma 3 establishes the conclusion.

The approximation for $P(W_n > z)$ follows trivially. □

Remark 1 We observe that the double inequality $1 + \eta^{i\delta_\Psi} f(h) < r_Z < 1 + h$ is satisfied by a large spectrum of values of (θ, η, r_Z). For instance, for $(\theta, \eta, r_Z) = (0.5, 0.5, e)$ or $(0.8, 0.4, 1.34)$ we can choose $1 + h = 2.75$ (or $1 + h = 1.35$) and obtain $1 + \eta^{i\delta_\Psi} f(h) < 2.64$ (or $1 + \eta^{i\delta_\Psi} f(h) < 1.33$), respectively. Similar comments can be stated concerning the inequality $1 + \theta^{2i+j-1} f^2(h) + \theta^i (1 + \theta^{j-1}) f(h) < r_Z$, which is an assumption of Lemma 2 and is applied in the proof of the next theorem. Indeed, for the same values of (θ, η, r_Z) we can choose $1+h = 1.5$ (or $1 + h = 1.18$) and get, respectively, the upper bounds 1.77 and 1.32 for $1 + \theta^{2i+j-1} f^2(h) + \theta^i (1 + \theta^{j-1}) f(h)$, both less than the considered values of r_Z.

The next theorem is the expected generalization of the Leadbetter's Extremal Types Theorem, when a $\Psi - INARMA(1, 1)$ process is considered.

Theorem 2 *If $Z \in \mathscr{C}_{\mathscr{A}}(r_Z)$, then exist $\{b_n\}$, with $b_n \in \mathbb{N}$, and $\{k_n\}$ satisfying (2) such that the strictly stationary sequence $\{X_n\}$ satisfies $D_{k_n}(x+b_n)$ and $D'_{k_n}(x+b_n)$. Therefore*

$$P(M_{k_n} \leq x + b_n) \to \exp(-r_Z^{-[x]}),\ n \to +\infty, \forall\, x \in \mathbb{R}.$$

Proof Consider the positive integer values $i_1, \ldots, i_p, j_1, \ldots, j_q$ and the positive integer-valued sequence $\{\ell_n\}$ specified in the definition of $D_{k_n}(x + b_n)$. Let be

$u_n = x + b_n$ and

$$X_j^* = \beta \odot_\Psi Z_j + V_j \sum_{i=1}^{\ell_n - 1} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i}.$$

Use the notations $A_j := \{X_j \leq u_n\}$ and $A_j^* := \{X_j^* \leq u_n\}$. As $A_j \subseteq A_j^*$ and $(X_{i_1}^*, \ldots, X_{i_p}^*)$ and $(X_{j_1}^*, \ldots, X_{j_q}^*)$ are independent, we have, with $\varepsilon_n > 0$,

$$\begin{aligned} P\left(\bigcap_{s=1}^{p} A_{i_s}, \bigcap_{t=1}^{q} A_{j_t}\right) &\leq P\left(\bigcap_{s=1}^{p} A_{i_s}^*\right) P\left(\bigcap_{t=1}^{q} A_{j_t}^*\right) \\ &\leq P\left(\bigcap_{s=1}^{p} \{X_{i_s} \leq u_n + \varepsilon_n\}\right) P\left(\bigcap_{t=1}^{q} \{X_{j_t} \leq u_n + \varepsilon_n\}\right) \\ &\quad + 3k_n P\left(V_1 \sum_{i=\ell_n}^{+\infty} \eta^i \odot_\Psi U_{-i} Z_{-i} > \varepsilon_n\right) \end{aligned} \tag{8}$$

where, by Markov's inequality, the last term does not exceed

$$3(1-\beta)k_n \frac{E\left(\sum_{i=\ell_n}^{+\infty} \eta^i \odot_\Psi U_{-i} Z_{-i}\right)}{\varepsilon_n} = 3(1-\beta)E(UZ)\frac{k_n}{\varepsilon_n}\frac{\eta^{\delta_\Psi \ell_n}}{1-\eta^{\delta_\Psi}}. \tag{9}$$

Taking $\ell_n = [k_n^\alpha]$, $\varepsilon_n = k_n^{-\beta}$, with $\alpha \in]0, 1[$ and $\beta > 0$, we get $\frac{k_n}{\varepsilon_n} \eta^{\delta_\Psi \ell_n} \to 0$, $n \to +\infty$. The mutual inequality of (8) is obtained similarly.

To prove that $D'_{k_n}(u_n)$ occurs, let us begin by splitting the sum of its definition into two sums in accordance with $j \leq \gamma_n - 1$ and $j \geq \gamma_n$. Since $X_j \leq X_j | \{V_j = 1\}$, $j \geq 1$, it holds

$$\begin{aligned} X_1 + X_j \leq T_j := \; & \beta \odot_\Psi Z_1 + \beta \odot_\Psi Z_j + \eta^{j-2} \odot_\Psi U_1 Z_1 \\ & + \sum_{i=1}^{j-2} \eta^{i-1} \odot_\Psi U_{j-i} Z_{j-i} \\ & + \sum_{i=0}^{+\infty} (\eta^i \odot_\Psi U_{-i} Z_{-i} + \eta^{i+j-1} \odot_\Psi U_{-i} Z_{-i}). \end{aligned}$$

Then, by Markov's inequality, it follows that

$$\begin{aligned} k_n \sum_{j=1}^{\gamma_n - 1} P(X_1 > u_n, X_j > u_n) &\leq k_n \gamma_n \max_{j \geq 2} P(X_1 + X_j > 2u_n) \\ &\leq k_n \gamma_n \max_{j \geq 2} P((1+h)^{T_j} > (1+h)^{2u_n}) \quad (10) \\ &\leq k_n \gamma_n (1+h)^{-2u_n} \max_{j \geq 2} E((1+h)^{T_j}), \end{aligned}$$

where

$$\begin{aligned} E((1+h)^{T_j}) \;=\; & P_Z(\Psi_{-\ln\beta}(1+h)) P_Z(\Psi_{-\ln\beta}(1+h)\Psi_{-\ln\eta^{j-2}}(1+h)) \\ & \times \prod_{i=1}^{j-2} P_{UZ}\left(\Psi_{-\ln\eta^{i-1}}(1+h)\right) \prod_{i=0}^{+\infty} P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right). \end{aligned}$$

We only prove the convergence of the last product operator, given the similarity of the convergence of the other factors. Let $\theta := \eta^{\delta_\Psi}$. Consider h such that

$$\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h) = 1 + \theta^{2i+j-1} f^2(h) + \theta^i (1+\theta^{j-1}) f(h) < r_Z.$$

Using the arguments of [11] (page 372–373) and applying properties 1.b and 2 from Lemma 2, we get

$$\begin{aligned} & P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right) \\ &= P_{UZ}\left(1 + \theta^{2i+j-1} f^2(h) + \theta^i (1+\theta^{j-1}) f(h)\right) \\ &\leq P_{UZ}\left(1+\theta^i f(h)\right)\left(1 + C_1 \theta^i \theta^{j-1} f(h)\right)\left(1 + C_2 \theta^{2i+j-1} f^2(h)\right) \\ &\leq \left(1+C_3\theta^i f(h)\right)^2 \left(1 + C_1 \theta^i \theta^{j-1} f(h)\right)\left(1 + C_2 \theta^{2i+j-1} f^2(h)\right), \end{aligned}$$

whereby

$$\begin{aligned} & \prod_{i=0}^{+\infty} P_{UZ}\left(\Psi_{-\ln\eta^{i+j-1}}(1+h)\Psi_{-\ln\eta^{i}}(1+h)\right) \\ & \quad \leq \exp\left((2C_3 + C_1\theta^{j-1}) f(h) \sum_{i=0}^{+\infty} \theta^i + C_2 \theta^{j-1} f^2(h) \sum_{i=0}^{+\infty} \theta^{2i} \right) \end{aligned}$$

which is uniformly bounded in j. Now consider $b_n = n$, $k_n = [\frac{1}{A} n^{-\xi} r_Z^n]$, $s_n = [k_n^\alpha]$, with $\alpha \in]0,1[$, $\gamma_n = [(\frac{k_n}{s_n})^\mu]$, with $\mu \in]0,1[$, and $(1+h)^2 = r_Z^\phi$, with $\phi \in]1,2[$, so that $\mu(1-\alpha) < \phi - 1$. Then, due to (10), the sum correspondent to $j \leq \gamma_n - 1$ converges to zero, when $n \to +\infty$. Moreover, we have

$$k_n \sum_{j=\gamma_n}^{k_n/s_n} P(X_1 > u_n, X_j > u_n) \leq \frac{k_n^2}{s_n} P(X_1 > u_n)$$
$$\times P\left(\beta \odot_\Psi Z_j + U_{j-1} Z_{j-1} + \sum_{i=1}^{\gamma_n} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i} > u_n - \varepsilon\right) \quad (11)$$
$$+ \frac{k_n^2}{s_n} P\left(\sum_{i=\gamma_n+1}^{+\infty} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i} > \varepsilon\right).$$

Since $U_{j-1} Z_{j-1} \in \mathscr{C}_{\mathscr{A}}(r_Z)$, applying once again Lemma 3, we conclude that the r.v. $X_j^{**} := \beta \odot_\Psi Z_j + U_{j-1} Z_{j-1} + \sum_{i=1}^{\gamma_n} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i}$ belongs to the same class. Then we have $k_n P(X_1 > u_n) \to r_Z^{-[x]}$, $n \to +\infty$, as well as $k_n P(X_j^{**} > u_n - \varepsilon) \to r_Z^{-[x-\varepsilon]}$, $n \to +\infty$. Dividing by s_n we deduce that the first sum of the second member of (11) goes to zero, when $n \to +\infty$. On the other hand, Markov's inequality enables the assertion that the second sum does not exceed

$$\frac{k_n^2}{s_n} \frac{E\left(\sum_{i=\gamma_n+1}^{+\infty} \eta^i \odot_\Psi U_{j-1-i} Z_{j-1-i}\right)}{\varepsilon} \leq \frac{k_n^2}{s_n} \frac{\theta^{\gamma_n}}{\varepsilon(1-\theta)} \to 0, \; n \to +\infty. \quad (12)$$

This finalizes the proof. □

Acknowledgements The work of the first author was partially supported by Portuguese Funds through FCT—Fundação para a Ciência e a Tecnologia, within the Project UID/MAT/00013/2013 and UID/MAT/04621/2013. The work of the second author was partially supported by the Centre for Mathematics of the University of Coimbra—UID/MAT/00324/2013, funded by the Portuguese Government through FCT/MEC and co-funded by the European Regional Development Fund through the Partnership Agreement PT2020.

References

1. Aczél, J., Dhombres, J.: Functional Equations in Several Variables. Cambridge University Press, Cambridge (1989)
2. Al-Osh, M.A.: The impact of missing data in a generalized integer-valued autoregression model for count data. J. Biopharm. Stat. **19**, 1039–1054 (2009)

3. Al-Osh, M.A., Alzaid, A.A.: First-order integer-valued autoregressive (INAR(1)) process. J. Time Ser. Anal. **8**, 261–275 (1987)
4. Al-Osh, M.A., Alzaid, A.A.: Integer-valued moving average (INMA) process. Stat. Pap. **29**, 281–300 (1988)
5. Aly, E.A., Bouzar, N.: On a class of $\mathbb{Z}_+-$valued autoregressive moving average (ARMA) processes. REVSTAT Stat. J. **6**, 101–121 (2005)
6. Anderson, C.W.: Extreme value theory for a class of discrete distribution with applications to some stochastic processes. J. Appl. Probab. **7**, 99–113 (1970)
7. Ferland, R., Latour, A., Oraichi, D.: Integer-valued GARCH process. J. Time Ser. Anal. **27**, 923–942 (2006)
8. Gonçalves, E., Mendes-Lopes, N., Silva, F.: Infinitely divisible distributions in integer-valued garch models. J. Time Ser. Anal. **36**, 503–527 (2015)
9. Grinevich, I.V.: Max-semistable limit laws under linear and power normalizations. Theory Probab. Appl. **38**, 640–650 (1992)
10. Hall, A.: Maximum term of a particular autoregressive sequence with discrete margins. Commun. Stat. Theory Methods **25**, 721–736 (1996)
11. Hall, A.: Extremes of integer-valued moving average models with exponential type tails. Extremes **6**, 361–379 (2003)
12. Hall, A., Temido, M.G.: On the maximum term of MA and Max-AR models with margins in Anderson's class. Theory Probab. Appl. **51**, 291–304 (2007)
13. Hall, A., Temido, M.G.: On the max-semistable limit of maxima of stationary sequences with missing values. J. Stat. Plann. Inference **3**, 875–890 (2009)
14. Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Processes. Springer, Berlin (1983)
15. McKenzie, E.: Some simple models for discrete variate time series. J. Am. Water Resour. Assoc. **21**, 645–650 (1985)
16. McKenzie, E.: Auto regressive-moving-average processes with negative binomial and geometric marginal distribution. Adv. Appl. Probab. **18**, 679–705 (1986)
17. McKenzie, E.: Some ARMA models for dependent sequences of poisson counts. Adv. Appl. Probab. **4**, 822–835 (1988)
18. Monteiro, M., Pereira, I., Scotto, M.G.: Optimal alarm systems for count processes. Commun. Stat. Theory Methods **37**, 3054–3076 (2008)
19. Scotto, M.G., Weiß, C.H., Silva, M.E., Pereira, I.: Bivariate binomial autoregressive models. J. Multivar. Anal. **125**, 233–251 (2014)
20. Steutel, F.W., van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**, 893–899 (1979)
21. Temido, M.G.: Domínios de atracção de funções de distribuição discretas. In: Carvalho, L. et al. (eds.) Novos Rumos em Estatística, pp. 415–426. Edições SPE, Lisboa (2002)
22. Temido, M.G., Canto e Castro, L.: Max-semistable laws in extremes of stationary random sequences. Theory Probab. Appl. **47**, 365–374 (2003)

Exact and Approximate Probabilities for the Null Distribution of Bartels Randomness Test

Ayana Mateus and Frederico Caeiro

Abstract In this work we revisit the statistical properties of the Bartels randomness test. The exact distribution of the statistic, under the randomization hypothesis, can only be obtained when the sample size (n) is small, since it requires the full set of permutations of the first n positive integers. Here, we present the exact null distribution without ties, for samples of size $10 \leq n \leq 17$, extending the results available in the literature. Since the null distribution is asymptotically normally distributed, but at a slow rate, Bartels concluded that the null distribution is well approximated by a Beta distribution, for samples of size $10 \leq n \leq 100$. We present a new approximation, based on the Edgeworth series, for the null distribution of the Bartels randomness statistic. The precision of this new approximation is also discussed.

1 Introduction

We shall consider and study the statistical properties of the Bartels nonparametric randomness test [3]. This test is the rank version of the ratio test for randomness developed by von Neumann [12] and is a linear transformation of the rank serial correlation coefficient introduced by Wald and Wolfowitz [13].
Let $(X_1, X_2, \ldots, X_n)$ be a sample from a population with continuous distribution and $R_i = \text{rank}(X_i)$, $i = 1, \ldots, n$, the rank of the i-th observation. Then, Bartels test statistic (or the Rank Version of von Neumann's ratio, RVN) for testing randomness is given by

$$RVN = \frac{\sum_{i=1}^{n-1}(R_i - R_{i+1})^2}{\sum_{i=1}^{n}(R_i - \bar{R})^2}, \tag{1}$$

A. Mateus (✉) · F. Caeiro
Faculdade de Ciências e Tecnologia and CMA, Universidade Nova de Lisboa, Caparica, Portugal
e-mail: amf@fct.unl.pt; fac@fct.unl.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_16

with $\bar{R} = \sum_{i=1}^{n} R_i/n$. The support of RVN is asymptotically the interval $(0, 4)$. By using a two-tailed test, the null hypothesis of randomness is tested against either a positive or negative serial correlation. But if we consider the existence of positive (negative) serial correlation in the alternative hypothesis, the null hypothesis is rejected if RVN is very small (large). If the sample has no ties, all ranks are distinct and we have $\sum_{i=1}^{n}(R_i - \bar{R})^2 = n(n^2 - 1)/12$. Consequently the numerator (NM) of RVN in (1) is an equivalent test statistic. The computation of the exact null distribution of NM is a very computational intensive process, except for small values of n, since it requires the full set of $n!$ permutations of $(1, 2, \ldots, n)$. Selected values of the exact null distribution of NM, when the sample has no ties, can be found in Bartels [3], for $4 \leq n \leq 10$. Bartels also derived the exact expression for the first four moments of RVN, in (1), and concluded that RVN converges in distribution slowly to the normal distribution. To be able to use the test of randomness for moderated sample sizes ($10 \leq n \leq 100$), Bartels also considered an approximation based on the Beta distribution (see [2, 3, 6, 8] for further details).

The remainder of this paper is organized as follows. In Sect. 2 we provide the exact null distribution of NM for samples of size $10 \leq n \leq 17$. In Sect. 3, after describing the approximations to the exact null distribution of RVN provided by the normal and the Beta distributions, we consider a new approximation based on the Edgeworth series. Next we evaluate the performance of all approximations to the exact null distribution function of RVN and present the main conclusions of this work.

2 Exact Probabilities for Bartels Statistic Test

As far as we know, the largest sample size for which the exact null distribution of NM is available is $n = 10$ and the results are presented in Bartels paper. To obtain the exact null distribution, we implemented a computer program in C programming language and use it to compute the null distribution of NM up to $n = 17$. In Tables 1, 2, and 3, we present values of the exact tail probabilities $LT = P(NM \leq x)$ and $RT = P(NM \geq y)$ for samples of size $10 \leq n \leq 17$. Those values can be used to obtain critical values. Due to the extensive numerical data, we only present selected values of both tail probabilities. For each n, we present select values of LT and RT in $[0.005, 0.01]$ and the two probabilities adjacent to the previous interval. Missing values associated with probabilities in $[0.005, 0.01]$ can be computed via interpolation. The complete null distribution for $n \leq 17$ can be obtained from the authors.

Table 1 Exact tail probabilities for the Bartels *NM* statistic ($10 \leq n \leq 13$)

n	x	LT	x	LT	x	LT	x	LT	y	RT	y	RT	y	RT	y	RT
10	51	0.0050	65	0.0157	81	0.0405	97	0.0841	228	0.1023	243	0.0550	259	0.0240	275	0.0082
	52	0.0054	67	0.0178	83	0.0448	99	0.0906	229	0.0990	245	0.0504	261	0.0215	277	0.0071
	53	0.0060	69	0.0204	85	0.0493	101	0.0985	231	0.0914	247	0.0456	263	0.0189	279	0.0060
	55	0.0072	71	0.0233	87	0.0543	102	0.1017	233	0.0846	249	0.0413	265	0.0166	281	0.0051
	57	0.0085	73	0.0260	89	0.0601			235	0.0778	251	0.0372	267	0.0145	282	0.0045
	59	0.0100	75	0.0290	91	0.0652			237	0.0719	253	0.0336	269	0.0127		
	61	0.0117	77	0.0328	93	0.0711			239	0.0658	255	0.0300	271	0.0111		
	63	0.0136	79	0.0365	95	0.0775			241	0.0605	257	0.0270	273	0.0097		
11	72	0.0048	91	0.0153	111	0.0383	131	0.0787	301	0.1011	319	0.0565	339	0.0257	359	0.0095
	73	0.0051	93	0.0170	113	0.0414	133	0.0837	302	0.0976	321	0.0528	341	0.0236	361	0.0085
	75	0.0060	95	0.0188	115	0.0449	135	0.0893	303	0.0950	323	0.0490	343	0.0215	363	0.0075
	77	0.0068	97	0.0207	117	0.0483	137	0.0948	305	0.0897	325	0.0456	345	0.0196	365	0.0067
	79	0.0077	99	0.0228	119	0.0521	138	0.0977	307	0.0842	327	0.0421	347	0.0178	367	0.0059
	81	0.0087	101	0.0249	121	0.0558	139	0.1007	309	0.0793	329	0.0391	349	0.0162	369	0.0052
	83	0.0099	103	0.0273	123	0.0601			311	0.0741	331	0.0360	351	0.0146	370	0.0048
	85	0.0110	105	0.0297	125	0.0644			313	0.0696	333	0.0333	353	0.0132		
	87	0.0124	107	0.0325	127	0.0689			315	0.0649	335	0.0305	355	0.0118		
	89	0.0137	109	0.0352	129	0.0735			317	0.0608	337	0.0281	357	0.0106		
12	99	0.0048	121	0.0136	145	0.0331	169	0.0677	387	0.1014	409	0.0580	433	0.0278	457	0.0112
	100	0.0050	123	0.0148	147	0.0354	171	0.0715	388	0.0989	411	0.0548	435	0.0259	459	0.0102
	101	0.0053	125	0.0160	149	0.0377	173	0.0751	389	0.0969	413	0.0518	437	0.0242	461	0.0094
	103	0.0059	127	0.0174	151	0.0403	175	0.0793	391	0.0922	415	0.0488	439	0.0225	463	0.0086
	105	0.0065	129	0.0187	153	0.0427	177	0.0833	393	0.0880	417	0.0461	441	0.0210	465	0.0079

(continued)

Table 1 (continued)

n	x	LT	x	LT	x	LT	x	LT	y	RT	y	RT	y	RT	y	RT
	107	0.0072	131	0.0203	155	0.0455	179	0.0876	395	0.0837	419	0.0434	443	0.0194	467	0.0071
	109	0.0079	133	0.0218	157	0.0482	181	0.0919	397	0.0797	421	0.0409	445	0.0181	469	0.0065
	111	0.0088	135	0.0235	159	0.0512	183	0.0966	399	0.0757	423	0.0384	447	0.0167	471	0.0059
	113	0.0096	137	0.0252	161	0.0542	184	0.0986	401	0.0720	425	0.0361	449	0.0155	473	0.0053
	115	0.0105	139	0.0271	163	0.0575	185	0.1011	403	0.0682	427	0.0338	451	0.0143	474	0.0051
	117	0.0115	141	0.0290	165	0.0607			405	0.0647	429	0.0317	453	0.0132	475	0.0048
	119	0.0125	143	0.0310	167	0.0642			407	0.0612	431	0.0296	455	0.0121		
13	133	0.0048	161	0.0138	191	0.0336	221	0.0687	488	0.1011	517	0.0555	547	0.0261	577	0.0103
	134	0.0050	163	0.0148	193	0.0353	223	0.0718	489	0.0994	519	0.0530	549	0.0247	579	0.0096
	135	0.0053	165	0.0157	195	0.0372	225	0.0748	491	0.0956	521	0.0506	551	0.0233	581	0.0090
	137	0.0057	167	0.0168	197	0.0392	227	0.0780	493	0.0920	523	0.0483	553	0.0220	583	0.0084
	139	0.0062	169	0.0178	199	0.0412	229	0.0813	495	0.0884	525	0.0461	555	0.0208	585	0.0078
	141	0.0067	171	0.0190	201	0.0433	231	0.0847	497	0.0850	527	0.0439	557	0.0196	587	0.0072
	143	0.0073	173	0.0202	203	0.0455	233	0.0881	499	0.0816	529	0.0418	559	0.0185	589	0.0067
	145	0.0078	175	0.0214	205	0.0477	235	0.0917	501	0.0784	531	0.0398	561	0.0174	591	0.0062
	147	0.0085	177	0.0227	207	0.0501	237	0.0953	503	0.0752	533	0.0378	563	0.0163	593	0.0057
	149	0.0091	179	0.0241	209	0.0524	239	0.0991	505	0.0722	535	0.0359	565	0.0153	595	0.0053
	151	0.0098	181	0.0255	211	0.0550	240	0.1009	507	0.0691	537	0.0342	567	0.0144	596	0.0051
	153	0.0105	183	0.0270	213	0.0575			509	0.0663	539	0.0324	569	0.0135	597	0.0049
	155	0.0113	185	0.0285	215	0.0602			511	0.0634	541	0.0307	571	0.0126		
	157	0.0121	187	0.0301	217	0.0629			513	0.0607	543	0.0291	573	0.0118		
	159	0.0129	189	0.0318	219	0.0658			515	0.0580	545	0.0276	575	0.0111		

Table 2 Exact tail probabilities for the Bartels *NM* statistic ($n = 14, 15$)

n	x	LT	x	LT	x	LT	x	LT	y	RT	y	RT	y	RT	y	RT
14	174	0.0048	209	0.0139	245	0.0331	281	0.0673	605	0.1009	639	0.0565	675	0.0271	711	0.0110
	175	0.0050	211	0.0147	247	0.0346	283	0.0698	606	0.0993	641	0.0544	677	0.0258	713	0.0104
	177	0.0054	213	0.0155	249	0.0361	285	0.0723	607	0.0978	643	0.0524	679	0.0247	715	0.0099
	179	0.0057	215	0.0163	251	0.0377	287	0.0748	609	0.0947	645	0.0505	681	0.0236	717	0.0093
	181	0.0061	217	0.0171	253	0.0393	289	0.0774	611	0.0918	647	0.0486	683	0.0225	719	0.0088
	183	0.0065	219	0.0180	255	0.0409	291	0.0801	613	0.0888	649	0.0467	685	0.0215	721	0.0083
	185	0.0069	221	0.0190	257	0.0426	293	0.0829	615	0.0860	651	0.0449	687	0.0205	723	0.0078
	187	0.0074	223	0.0199	259	0.0444	295	0.0857	617	0.0832	653	0.0431	689	0.0195	725	0.0074
	189	0.0079	225	0.0209	261	0.0462	297	0.0885	619	0.0805	655	0.0415	691	0.0186	727	0.0069
	191	0.0083	227	0.0220	263	0.0481	299	0.0915	621	0.0778	657	0.0398	693	0.0177	729	0.0065
	193	0.0089	229	0.0230	265	0.0500	301	0.0945	623	0.0752	659	0.0382	695	0.0168	731	0.0061
	195	0.0094	231	0.0242	267	0.0520	303	0.0976	625	0.0726	661	0.0366	697	0.0160	733	0.0057
	197	0.0100	233	0.0253	269	0.0540	304	0.0990	627	0.0702	663	0.0351	699	0.0152	735	0.0054
	199	0.0105	235	0.0265	271	0.0561	305	0.1007	629	0.0677	665	0.0337	701	0.0144	737	0.0051
	201	0.0112	237	0.0277	273	0.0582			631	0.0654	667	0.0323	703	0.0137	738	0.0049
	203	0.0118	239	0.0290	275	0.0604			633	0.0631	669	0.0309	705	0.0130		
	205	0.0125	241	0.0303	277	0.0626			635	0.0609	671	0.0296	707	0.0123		
	207	0.0132	243	0.0317	279	0.0650			637	0.0586	673	0.0283	709	0.0116		
15	224	0.0050	265	0.0137	307	0.0318	349	0.0639	739	0.1006	779	0.0570	821	0.0279	863	0.0118
	225	0.0051	267	0.0143	309	0.0330	351	0.0659	740	0.0992	781	0.0552	823	0.0269	865	0.0113
	227	0.0054	269	0.0149	311	0.0342	353	0.0679	741	0.0980	783	0.0535	825	0.0259	867	0.0108
	229	0.0057	271	0.0156	313	0.0355	355	0.0699	743	0.0955	785	0.0518	827	0.0250	869	0.0103
	231	0.0060	273	0.0163	315	0.0367	357	0.0720	745	0.0929	787	0.0502	829	0.0240	871	0.0099
	233	0.0063	275	0.0170	317	0.0380	359	0.0741	747	0.0905	789	0.0486	831	0.0231	873	0.0094

(continued)

Table 2 (continued)

n	x	LT	x	LT	x	LT	x	LT	y	RT	y	RT	y	RT	y	RT
	235	0.0067	277	0.0177	319	0.0394	361	0.0763	749	0.0881	791	0.0471	833	0.0222	875	0.0090
	237	0.0070	279	0.0185	321	0.0408	363	0.0785	751	0.0857	793	0.0455	835	0.0214	877	0.0085
	239	0.0074	281	0.0193	323	0.0422	365	0.0808	753	0.0833	795	0.0441	837	0.0205	879	0.0081
	241	0.0078	283	0.0201	325	0.0436	367	0.0831	755	0.0811	797	0.0426	839	0.0197	881	0.0078
	243	0.0082	285	0.0209	327	0.0451	369	0.0854	757	0.0788	799	0.0412	841	0.0189	883	0.0074
	245	0.0086	287	0.0217	329	0.0466	371	0.0878	759	0.0766	801	0.0398	843	0.0182	885	0.0070
	247	0.0090	289	0.0226	331	0.0482	373	0.0902	761	0.0745	803	0.0385	845	0.0174	887	0.0067
	249	0.0095	291	0.0235	333	0.0498	375	0.0927	763	0.0724	805	0.0372	847	0.0167	889	0.0063
	251	0.0099	293	0.0245	335	0.0514	377	0.0952	765	0.0703	807	0.0359	849	0.0160	891	0.0060
	253	0.0104	295	0.0254	337	0.0531	379	0.0978	767	0.0683	809	0.0347	851	0.0154	893	0.0057
	255	0.0109	297	0.0264	339	0.0548	380	0.0990	769	0.0663	811	0.0335	853	0.0147	895	0.0054
	257	0.0114	299	0.0274	341	0.0565	381	0.1004	771	0.0643	813	0.0323	855	0.0141	897	0.0052
	259	0.0120	301	0.0285	343	0.0583			773	0.0624	815	0.0312	857	0.0135	898	0.0050
	261	0.0125	303	0.0296	345	0.0602			775	0.0606	817	0.0301	859	0.0129	899	0.0049
	263	0.0131	305	0.0307	347	0.0620			777	0.0587	819	0.0290	861	0.0124		

Table 3 Exact tail probabilities for the Bartels *NM* statistic ($n = 16, 17$)

n	x	LT	x	LT	x	LT	x	LT	y	RT	y	RT	y	RT	y	RT
16	282	0.0049	331	0.0136	381	0.0317	431	0.0638	891	0.1005	939	0.0564	989	0.0275	1039	0.0116
	283	0.0051	333	0.0141	383	0.0326	433	0.0654	892	0.0993	941	0.0549	991	0.0266	1041	0.0112
	285	0.0053	335	0.0146	385	0.0336	435	0.0671	893	0.0983	943	0.0535	993	0.0258	1043	0.0108
	287	0.0055	337	0.0152	387	0.0347	437	0.0688	895	0.0961	945	0.0520	995	0.0250	1045	0.0104
	289	0.0058	339	0.0157	389	0.0357	439	0.0705	897	0.0940	947	0.0507	997	0.0242	1047	0.0100
	291	0.0061	341	0.0163	391	0.0368	441	0.0723	899	0.0919	949	0.0493	999	0.0234	1049	0.0096
	293	0.0063	343	0.0169	393	0.0379	443	0.0741	901	0.0898	951	0.0480	1001	0.0227	1051	0.0092
	295	0.0066	345	0.0175	395	0.0390	445	0.0759	903	0.0878	953	0.0467	1003	0.0220	1053	0.0088
	297	0.0069	347	0.0181	397	0.0402	447	0.0778	905	0.0858	955	0.0454	1005	0.0212	1055	0.0085
	299	0.0072	349	0.0188	399	0.0413	449	0.0797	907	0.0838	957	0.0442	1007	0.0205	1057	0.0082
	301	0.0075	351	0.0194	401	0.0425	451	0.0816	909	0.0818	959	0.0430	1009	0.0199	1059	0.0078
	303	0.0078	353	0.0201	403	0.0438	453	0.0835	911	0.0799	961	0.0417	1011	0.0192	1061	0.0075
	305	0.0082	355	0.0208	405	0.0450	455	0.0855	913	0.0780	963	0.0406	1013	0.0186	1063	0.0072
	307	0.0085	357	0.0215	407	0.0463	457	0.0875	915	0.0762	965	0.0394	1015	0.0179	1065	0.0069
	309	0.0089	359	0.0223	409	0.0476	459	0.0896	917	0.0744	967	0.0383	1017	0.0173	1067	0.0066
	311	0.0092	361	0.0230	411	0.0489	461	0.0917	919	0.0726	969	0.0372	1019	0.0167	1069	0.0064
	313	0.0096	363	0.0238	413	0.0503	463	0.0938	921	0.0708	971	0.0361	1021	0.0161	1071	0.0061
	315	0.0100	365	0.0246	415	0.0517	465	0.0959	923	0.0691	973	0.0351	1023	0.0156	1073	0.0058
	317	0.0104	367	0.0254	417	0.0531	467	0.0981	925	0.0674	975	0.0341	1025	0.0150	1075	0.0056
	319	0.0108	369	0.0262	419	0.0545	468	0.0991	927	0.0658	977	0.0331	1027	0.0145	1077	0.0053
	321	0.0113	371	0.0271	421	0.0560	469	0.1003	929	0.0641	979	0.0321	1029	0.0140	1079	0.0051
	323	0.0117	373	0.0280	423	0.0575			931	0.0625	981	0.0311	1031	0.0135	1080	0.0050
	325	0.0122	375	0.0288	425	0.0590			933	0.0609	983	0.0302	1033	0.0130		
	327	0.0126	377	0.0298	427	0.0606			935	0.0594	985	0.0293	1035	0.0125		
	329	0.0131	379	0.0307	429	0.0622			937	0.0578	987	0.0284	1037	0.0121		

(continued)

Table 3 (continued)

n	*x*	LT	*x*	LT	*x*	LT	*x*	LT	*y*	RT	*y*	RT	*y*	RT	*y*	RT
17	350	0.0050	406	0.0132	466	0.0313	526	0.0641	1062	0.1005	1120	0.0552	1180	0.0264	1240	0.0108
	351	0.0051	409	0.0139	469	0.0326	529	0.0663	1063	0.0997	1123	0.0534	1183	0.0253	1243	0.0103
	352	0.0052	412	0.0145	472	0.0339	532	0.0684	1066	0.0968	1126	0.0516	1186	0.0243	1246	0.0098
	355	0.0055	415	0.0152	475	0.0352	535	0.0707	1069	0.0941	1129	0.0499	1189	0.0233	1249	0.0094
	358	0.0058	418	0.0159	478	0.0366	538	0.0729	1072	0.0913	1132	0.0481	1192	0.0224	1252	0.0089
	361	0.0061	421	0.0167	481	0.0380	541	0.0753	1075	0.0888	1135	0.0465	1195	0.0215	1255	0.0085
	364	0.0065	424	0.0174	484	0.0394	544	0.0776	1078	0.0861	1138	0.0449	1198	0.0205	1258	0.0080
	367	0.0068	427	0.0182	487	0.0409	547	0.0801	1081	0.0836	1141	0.0433	1201	0.0197	1261	0.0076
	370	0.0072	430	0.0190	490	0.0424	550	0.0825	1084	0.0811	1144	0.0418	1204	0.0188	1264	0.0072
	373	0.0076	433	0.0199	493	0.0440	553	0.0851	1087	0.0787	1147	0.0403	1207	0.0181	1267	0.0069
	376	0.0080	436	0.0208	496	0.0456	556	0.0876	1090	0.0763	1150	0.0388	1210	0.0173	1270	0.0065
	379	0.0084	439	0.0217	499	0.0473	559	0.0903	1093	0.0740	1153	0.0374	1213	0.0165	1273	0.0062
	382	0.0089	442	0.0226	502	0.0489	562	0.0929	1096	0.0717	1156	0.0360	1216	0.0158	1276	0.0058
	385	0.0094	445	0.0236	505	0.0507	565	0.0957	1099	0.0695	1159	0.0347	1219	0.0151	1279	0.0055
	388	0.0098	448	0.0246	508	0.0524	568	0.0985	1102	0.0673	1162	0.0334	1222	0.0144	1282	0.0052
	391	0.0104	451	0.0256	511	0.0543	569	0.0995	1105	0.0652	1165	0.0321	1225	0.0138	1284	0.0050
	394	0.0109	454	0.0267	514	0.0562	570	0.1004	1108	0.0631	1168	0.0309	1228	0.0131	1285	0.0049
	397	0.0114	457	0.0278	517	0.0581			1111	0.0611	1171	0.0297	1231	0.0125		
	400	0.0120	460	0.0289	520	0.0601			1114	0.0591	1174	0.0286	1234	0.0119		
	403	0.0126	463	0.0301	523	0.0621			1117	0.0572	1177	0.0275	1237	0.0114		

3 Approximations to the Distribution Function of RVN and Conclusions

Since several approximation methods for the distribution function (df) of random variables involve matching a finite set of moments, we first present the first four moments of RVN available in Bartels:

$$\begin{aligned}
\mu &= 2 \\
\sigma^2 &= \frac{4(n-2)(5n^2-2n-9)}{5n(n+1)(n-1)^2} \\
\mu_3 &= \frac{96(n-4)(n+2)(n+4)}{35n(n+1)^2(n-1)^3} \\
\mu_4 &= \frac{48k}{175n^3(n+1)^3(n-1)^4} \quad \text{where} \\
k &= 175n^8 - 931n^7 + 1090n^6 + 3146n^5 \\
&\quad -10445n^4 + 761n^3 + 34380n^2 + 7104n - 17640
\end{aligned} \tag{2}$$

Those moments were computed by using the formulas developed by Young [14] for the null distribution of $1 - (1/2)RVN$.
For large sample sizes, the limit distribution is:

$$\frac{RVN - 2}{\sqrt{\frac{4(n-2)(5n^2-2n-9)}{5n(n+1)(n-1)^2}}} \xrightarrow[n\to\infty]{d} N(0,1), \tag{3}$$

but the rate of convergence to the normal distribution is slow. In order to have simple and accurate approximations to the exact null distribution function of RVN, required to compute for example the p-value of the test, we shall study the following approximations: the beta distribution, the normal distribution, and the Edgeworth series. All the approximations were implemented in R [10] and the computer code is available in the Appendix. Bartels already considered that the null distribution of $RVN/4$ can be approximated by a Beta distribution with density

$$f(x) = \frac{x^{p-1}(1-x)^{q-1}}{\mathscr{B}(p,q)}, \quad 0 < x < 1$$

where $\mathscr{B}(.,.)$ is the beta function and $p = q = \frac{5n(n+1)(n-1)^2}{2(n-2)(5n^2-2n-9)} - \frac{1}{2}$. By using these expressions for p and q, the first two moments of the exact and the approximated distribution are equal. Critical values, computed with the approximation provided

by the Beta distribution, for samples of size $10 \leq n \leq 100$, are available in [3]. The approximation to the null distribution of RVN by the normal law follows from the result in (3). This test, with the approximations mentioned above, can be performed by the use of the following R packages: EnvStats [9], lawstat [7], and randtests [4]. The new approximation based on the Edgeworth series [1, 11], using the first four moments, is given by

$$P(RVN \leq x) \approx \Phi(z) - \phi(z) \left\{ \frac{\gamma_1}{6}(z^2 - 1) + \frac{(\gamma_2 - 3)}{24}(z^3 - 3z) + \frac{\gamma_1^2}{72}(z^5 - 10z^3 + 15z) \right\}$$

where $z = (x - \mu)/\sigma$, ϕ and Φ are, respectively, the density and distribution functions of the standard Normal distribution and $\gamma_1 = \mu_3/\sigma^3$ and $\gamma_2 = \mu_4/\sigma^4$ are, respectively, the third and fourth standardized moments (coefficients of skewness and kurtosis). In addition, we can apply the continuity correction (cc) to all approximations, since the test statistic RVN is discrete. In other words, the continuity correction factor 0.5 is applied to the value of NM and x in $P(RVN \leq x)$ should be modified to $x + 0.5 \times 12/(n(n^2 - 1))$.

Next we study the precision of the different approximations. In Fig. 1, we present an illustration of the error of the beta and the Edgerworth approximations for samples of size 5 and 10, i.e., the values of the $error_{x \in S} = (\tilde{F}(x) - F(x))$ where $\tilde{F}$ and F are, respectively, the approximate and exact null df of RVN (the exact df of RVN is obtained straightforwardly from the exact df of NM) and S denotes the support of RVN. We also studied the approximations with a continuity correction factor. Figure 1 suggests that the continuity correction factor usually reduces the absolute error.

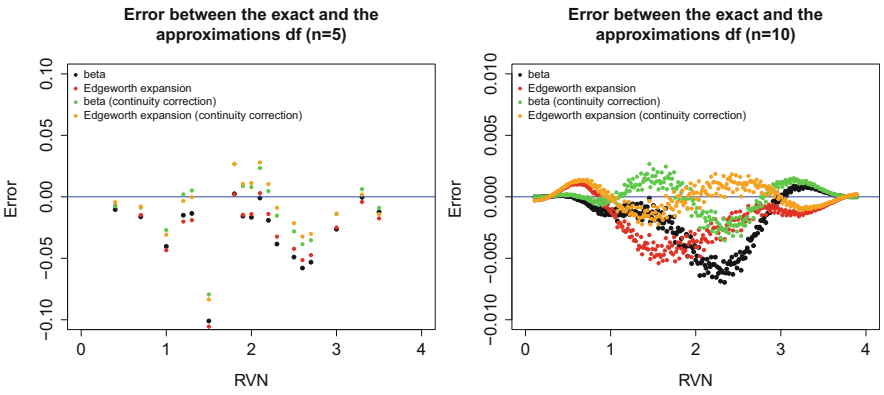


Fig. 1 Error of the beta and the Edgerworth approximations for samples size of $n = 5, 10$

Table 4 Maximum absolute error of the distribution function approximations

n	5	10	11	12	13	14	15	16	17
Beta	0.1010	0.0070	0.0055	0.0046	0.0040	0.0035	0.0031	0.0028	0.0026
Beta (cc)	0.0795	0.0035	0.0028	0.0025	0.0023	0.0021	0.0019	0.0018	0.0017
Normal	0.1206	0.0148	0.0126	0.0112	0.0101	0.0092	0.0084	0.0078	0.0072
Normal (cc)	0.0984	0.0117	0.0100	0.0092	0.0085	0.0078	0.0073	0.0068	0.0064
Edgeworth	0.1056	0.0054	0.0039	0.0032	0.0026	0.0021	0.0018	0.0015	0.0013
Edgeworth (cc)	0.0835	0.0022	0.0016	0.0013	0.0011	0.0009	0.0007	0.0007	0.0006
n	25	30	40	50	75	100			
Beta	0.0016	0.0014	0.0009	0.0008	0.0006	0.0004			
Beta (cc)	0.0012	0.0011	0.0008	0.0007	0.0005	0.0004			
Normal	0.0047	0.0038	0.0028	0.0022	0.0014	0.0011			
Normal (cc)	0.0044	0.0036	0.0027	0.0021	0.0014	0.0011			
Edgeworth	0.0006	0.0004	0.0003	0.0001	0.0001	0.0002			
Edgeworth (cc)	0.0003	0.0002	0.0002	0.0001	0.0001	0.0002			

To evaluate the precision of the normal, the beta, and the Edgeworth approximations, for finite samples of size n, we computed the maximum absolute error of the approximations, i.e., the values of $\epsilon = \max_{x \in S} \left| \tilde{F}(x) - F(x) \right|$ (computed with and without continuity correction in the approximation $\tilde{F}(x)$). In Table 4 we present the values of ϵ for different sample sizes. Note that for sample sizes greater or equal to 25 the exact df was not available and was obtained by computer simulation, based on 5×10^7 samples. The standard error of the probabilities $\tilde{F}(x)$ is smaller than 7.1×10^{-5}.

From Table 4, we can conclude that for moderate sample sizes, the beta and the Edgeworth approximations provide accurate approximations for the exact null distribution of RVN. The best results are underlined and mainly obtained with the Edgeworth approximation with continuity correction. The limit normal distribution presents greater values of ϵ than the beta and Edgeworth approximations, even for large sample sizes. This result is not surprising since it is known that the convergence in distribution to normality of this kind of statistics is quite slow[5]. We do not need to use these approximations for small sample sizes since we provide the exact null distribution for samples of sizes $10 \leq n \leq 17$. For moderate and large sample sizes we advise the use of the Edgeworth approximation with continuity correction.

Acknowledgements This research is partially supported by National Funds through *FCT*—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through project PEst-OE/MAT/UI0297/2013 (Centro de Matemática e Aplicações).

Appendix

R function for the evaluation of the first four moments (mean, variance, skewness, and kurtosis)

```
# Mean, Variance, skewness and kurtosis coefficients of
# the Bartels Statistic RVN
mbartels <- function(n){
  mu <- 2
  sigma2<- 4*(n-2)*(5*n^2-2*n-9)/(5*n*(n+1)*(n-1)^2)
  sigma <- sqrt(sigma2)
  gm1 <- 96*(n-4)*(n+2)*(n+4)/(35*n*(n+1)^2*(n-1)^3)/sigma^3
  k <- 175*n^8-931*n^7+1090*n^6+3146*n^5-10445*n^4+761*n^3+
    34380*n^2+7104*n-17640
  gm2 <- 48*k/(175*n^3*(n+1)^3*(n-1)^4)/sigma^4
  return(c(mu, sigma2, gm1, gm2))
}
```

R function for the evaluation of the approximations to the null Distribution Function of the Bartels Statistic

```
# Approximation to the null Distribution Function
# of the Bartels Statistic
```

```
pbartels <- function(x, n, approx = "edge", cc=T){
    if (approx == "normal"){
      if (cc) x <- x+.5/(n*(n^2-1)/12)
      y <- mbartels(n)
      z <- (x-y[1])/sqrt(y[2])
      pp <- pnorm(z)
    }
    # compute approximate $F(x)$ using the Edgerworth expansion
    if (approx == "edge"){
      if (cc) x <- x+.5/(n*(n^2-1)/12)
      y <- mbartels(n)
      z <- (x-y[1])/sqrt(y[2])
      pp <- pnorm(z)-dnorm(z)*(y[3]*(z^2-1)/6)
      pp <- pp-dnorm(z)*((y[4]-3)*(z^3-3*z)/24+(y[3]^2)*
      (z^5-10*z^3+15*z)/72)
    }
    # compute F(x) with the beta approximation
    if (approx == "beta"){
      if (cc) x <- x+.5/(n*(n^2-1)/12)
      btp <- (5*n*(n+1)*(n-1)^2)/(2*(n-2)*(5*n^2-2*n-9))-1/2
      pp <- pbeta(x/4, shape1=btp, shape2=btp)
    }
  return(pp)
}
```

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Applied Mathematics Series, vol. 55, 10th edn. National Bureau of Standards, US Governement Printing Office, Washington, DC (1972).
2. Bagdonavičius, V., Kruopis, J., Nikulin, M.: Nonparametric Tests for Complete Data. Wiley, London (2010)
3. Bartels, R.: The rank version of von Neumann's ratio test for randomness. J. Am. Stat. Assoc. **77**(377), 40–46 (1982)
4. Caeiro, F., Mateus, A.: randtests: Testing randomness in R. R package version 1.0. http://CRAN.R-project.org/package=randtests (2014)
5. David, F.N., Fix, E.: Randomization and the serial correlation coefficient. In: David, F.N. (ed.) Research Papers in Statistics, pp.461–468. Wiley, London (1966)
6. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference, 5th edn. CRC Press, Boca Raton (2010)
7. Gastwirth, J.L., Gel, Y.R., Wallace Hui, W. L., Lyubchich, V., Miao, W., Noguchi, K.: lawstat: Tools for biostatistics, public policy, and law. R package version 3.0. http://CRAN.R-project.org/package=lawstat (2015)
8. Madansky, A.: Prescriptions for Working Statisticians. Springer Texts in Statistics. Springer, New York (1988)
9. Millard, S.P.: EnvStats: package for environmental statistics, including US EPA guidance. R package version 2.1.1. http://CRAN.R-project.org/package=EnvStats (2016)
10. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (2013)
11. Remillard, B.: Statistical Methods for Financial Engineering. CRC Press, Boca Raton (2013)

12. von Neumann, J.: Distribution of the ratio of the mean square successive difference to the variance. Ann. Math. Stat. **12**(4), 367–395 (1941)
13. Wald, A., Wolfowitz, J.: An exact test for randomness in the non-parametric case based on serial correlation. Ann. Math. Stat. **14**(4), 378–388 (1943)
14. Young, L.C.: On randomness in ordered sequences. Ann. Math. Stat. **12**(3), 293–300 (1941)

Gamma-Series Representations for the Sum of Independent Gamma Random Variables and for the Product of Independent Beta Random Variables

Filipe J. Marques

Abstract In this work it is shown that using well-known series expansions it is possible to represent a single gamma distribution and also the logarithm of a single beta distribution, as an infinite mixture of gamma distributions. Then, using these representations, it is possible to derive simple gamma-series representations for the distribution of the sum of independent gamma random variables and for the sum of independent logbeta random variables, which by simple transformation may be used to represent also the distribution of the product of independent beta random variables. These representations may be used to develop accurate asymptotic approximations for corresponding distributions.

1 Introduction

The gamma and beta distributions are widely used in most different applications, and as such, applications involving sums or products of these distributions may arise naturally in the multivariate context. For example, the distribution of the sum of independent gamma random variables is very important in problems related to wireless communications [1, 3, 15]. On the other hand, the distribution of many likelihood ratio test statistics used in multivariate analysis may be represented as a product of independent beta random variables [12] and as such this distribution is of crucial importance for those who need to apply testing procedures in multivariate analysis. Given their huge importance in many statistical procedures there is a vast literature on the distribution of the sum of independent gamma random variables and on the product of independent beta random variables; in what follows, we have selected only some of those results. Concerning the sum of independent gamma random variables, there is one first result, for a particular case, given in exercises

F. J. Marques (✉)
Universidade NOVA de Lisboa (UNL) and Centro de Matemática e Aplicações (CMA), Faculdade de Ciências e Tecnologia, Caparica, Portugal
e-mail: fjm@fct.unl.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_17

12 and 13 of [9], for the sum of independent exponential random variables (with different parameters); also for a particular case in [2, 5] the authors developed results for the sum of integer gamma random variables, and finally in [14] a result is derived, for the general case, based on an infinite mixture of gamma distributions. For the distribution of the product of independent beta random variables there are also many results that may be referred, from Tukey and Wilks [17] to Tang and Gupta [16] and Moschopoulos [13], this last reference is about representations for the distribution of a class of likelihood ratio statistics but the results obtained may be used to address the product of independent beta random variables, and more recently [7]. Most of the results are based on the H or Meijer G functions or infinite mixtures representations, which in our days may still be difficult to implement or to use in modern softwares. On the other hand, some of the known approximations are based on a single χ^2 distribution [18], on Box method [4], which is usually presented in the form of mixtures of gamma distributions, on saddle point approximations [8], and on Edgeworth expansions or Cornish-Fisher series approximations. However, most of these approximations reveal lack of precision in extreme cases, for example when one has a large number of variables. More recently, there are some advanced results that can be used to improve these approximations, the so-called near-exact approximations [6, 7].

Our aim in this work is to show how it is possible to obtain, in a very simple way, gamma-series representations for the density and cumulative distribution functions of a gamma random variable and of a logbeta random variable. These gamma-series representations are obtained, mainly, by using the binomial and exponential expansions and are derived in such a way that it is possible to choose the rate parameter of all the gamma distributions involved in these representations. This feature is of extreme importance since it will allow us to obtain single gamma-series representations for the sum of independent gamma random variables and for the sum of independent logbeta random variables. Using these representations it is possible to derive simple and accurate asymptotic approximations for these distributions by simple truncation of the series and/or equating moments.

2 The Sum of Independent Gamma Random Variables

Let $Y_i \sim \Gamma(r_i, \lambda_i)$, $i = 1, \ldots, p$ be p independent gamma random variables, with density function of Y_i given by

$$f_{Y_i}(y) = \frac{\lambda_i^{r_i}}{\Gamma(r_i)} y^{r_i - 1} \exp\{-\lambda_i y\}, \quad y > 0, \quad r_i > 0, \quad \lambda_i > 0, \quad i = 1, \ldots, p$$

which, for $\delta > 0$ and $\delta \neq \lambda_i$ for all i, may still be represented as

$$\begin{aligned} f_{Y_i}(y) &= \frac{\lambda_i^{r_i}}{\Gamma(r_i)} y^{r_i-1} \exp\{-\lambda_i y\} \frac{\exp\{(\lambda_i-\delta)y\}}{\exp\{(\lambda_i-\delta)y\}} \\ &= \frac{\lambda_i^{r_i}}{\Gamma(r_i)} y^{r_i-1} \exp\{-\delta y\} \sum_{j=0}^{\infty} \frac{(\delta-\lambda_i)^j}{j!} y^j \\ &= \sum_{j=0}^{\infty} \frac{(\delta-\lambda_i)^j}{j!} \frac{\lambda_i^{r_i}}{\Gamma(r_i)} \frac{\Gamma(r_i+j)}{\delta^{r_i+j}} \frac{\delta^{r_i+j}}{\Gamma(r_i+j)} y^{r_i+j-1} \exp\{-\delta y\} \\ &= \sum_{j=0}^{\infty} p_{i,j} f_{X_{i,j}}(y) \end{aligned}$$

which represents a mixture of gamma distributions, $X_{i,j} \sim \Gamma(r_i + j, \delta)$, with weights given by

$$p_{i,j} = \frac{1}{j!} \left(\frac{\delta-\lambda_i}{\delta}\right)^j \left(\frac{\lambda_i}{\delta}\right)^{r_i} \frac{\Gamma(r_i+j)}{\Gamma(r_i)}. \tag{1}$$

As one may observe from the above expressions the parameter $\delta \neq \lambda_i$, for all i, may be the same in all the series representations of the densities of Y_i, and thus we have gamma-series representations, which are infinite mixtures of $\Gamma(r_i + j, \delta)$ for each of the Y_i, $i = 1, \ldots, p$. This is an essential feature which together with the appealing properties of the mixtures allows us, for example, to obtain a representation for the sum of independent gamma random variables in the form of a single gamma-series representation.

The characteristic function of $W = \sum_{i=1}^{p} Y_i$ with $Y_i \overset{ind}{\sim} \Gamma(r_i, \lambda_i)$, $i = 1, \ldots, p$ may be written as

$$\Phi_W(t) = \prod_{i=1}^{p} \left\{ \sum_{j=0}^{\infty} p_{i,j} \Phi_{X_{i,j}}(t) \right\}, \quad t \in \mathbb{R} \tag{2}$$

for $\delta > 0$ and $\delta \neq \lambda_i$, for all i, with $p_{i,j}$ in (1) and where $\Phi_{X_{i,j}}(t)$ is the characteristic function of $X_{i,j} \sim \Gamma(r_i + j, \delta)$ which is given by

$$\Phi_{X_{i,j}}(t) = \left(\frac{\delta}{\delta - \mathrm{i}t}\right)^{r_i+j}.$$

with $\mathrm{i} = \sqrt{-1}$. As already mentioned, given the fact that all $X_{i,j}$ have the same rate parameter and using simple results available for the product of series (please see

[10]), the characteristic function in (2) may be written as

$$\Phi_W(t)=\underbrace{\sum_{j_0=0}^{\infty}\left\{\sum_{j_1=0}^{j_0}\sum_{j_2=0}^{j_1}\cdots\sum_{j_{p-1}=0}^{j_{p-2}}\left\{\prod_{i=1}^{p}p_{i,j_{p-i}-j_{p-i+1}}\right\}\right\}}_{p^{\star}_{j_0}}\left(\frac{\delta}{\delta-\mathrm{i}t}\right)^{\sum_{i=1}^{p}r_i+j_0} \tag{3}$$

with $j_p = 0$, and which may still be simplified as

$$\Phi_W(t) = \sum_{j_0=0}^{\infty} p^{\star}_{j_0}\left(\frac{\delta}{\delta-\mathrm{i}t}\right)^{\sum_{i=1}^{p}r_i+j_0} \tag{4}$$

with $p^{\star}_{j_0}$ in (3) and which is a gamma-series representation, corresponding to an infinite mixture of gamma distributions, $\Gamma\left(\sum_{i=1}^{p} r_i + j_0, \delta\right)$, with weights $p^{\star}_{j_0}$. As a result, the density and cumulative distribution functions of $W = \sum_{i=1}^{p} Y_i$ are, respectively, given by

$$f_W(w) = \sum_{j_0=0}^{\infty} p^{\star}_{j_0} f_{\Gamma\left(\sum_{i=1}^{p} r_i+j_0,\delta\right)}(w), \quad w > 0,$$

$$F_W(w) = \sum_{j_0=0}^{\infty} p^{\star}_{j_0} F_{\Gamma\left(\sum_{i=1}^{p} r_i+j_0,\delta\right)}(w), \quad w \in \mathbb{R}$$

where $f_{\Gamma\left(\sum_{i=1}^{p} r_i+j_0,\delta\right)}$ and $F_{\Gamma\left(\sum_{i=1}^{p} r_i+j_0,\delta\right)}$ are, respectively, the density and cumulative distribution functions of a random variable with gamma distribution, $\Gamma\left(\sum_{i=1}^{p} r_i+j_0,\delta\right)$.

Although there is a high computational investment required for the implementation of these last expressions, it is possible to use them to obtain accurate approximations. We should also note that the parameter δ can be chosen in order to increase the speed of convergence of the series representations. Based on our empirical knowledge we propose the use of δ equal to the rate parameter of a mixture of two gamma distributions which match the first four moments of the exact distribution of W. The results in [11] also suggest the replacement of the original weights by the ones obtained by the method of matching moments, this procedure leads to a considerable reduction of the number of terms in the mixture. There are other more elaborate techniques that will not be subject to our attention in this work and that may be used to obtain even more accurate approximations, for example the so-called near-exact distributions (please see, [11]). In [14] it is

also presented, using a different approach, a gamma-series representation for the sum of independent gamma random variables. We should note that our results are equivalent to the ones in [14] if in expression (2.9), in [14], we consider $\beta_1 = \delta$. In [14] the weights of the mixture are defined using recurrence formulas.

Let us consider a simple illustration with two scenarios:

(i) $r_i = \left\{\frac{5}{7}, 3, \frac{8}{5}\right\}$ and $\lambda_i = \left\{\frac{8}{5}, \frac{3}{4}, \frac{1}{3}\right\}$

(ii) $r_i = \left\{10, \frac{11}{3}, \frac{9}{5}, \frac{15}{2}\right\}$ and $\lambda_i = \left\{\frac{4}{3}, \frac{5}{3}, \frac{15}{2}, \frac{14}{4}\right\}$.

The approximating probability density function is given by

$$f_W^\star(w) = \sum_{j=0}^{m^\star} \pi_j^\star f_{\Gamma\left(\sum_{i=1}^{p} r_i + j, \delta\right)}(w), \quad w > 0$$

where δ is equal to the rate parameter of a mixture of two gamma distributions which match the first four moments of the exact distribution of W, that is δ is obtained as solution of

$$\frac{\partial^h}{\partial t^h} \Phi_W(t)\Big|_{t=0} = \frac{\partial^h}{\partial t^h} \left\{l(\delta)^{r_1}(\delta - \mathrm{i}t)^{-r_1} + (1-l)(\delta)^{r_2}(\delta - \mathrm{i}t)^{-r_2}\right\}\Big|_{t=0},$$
$$h = 1, \ldots, 4$$

and the new weights, $\pi_j^\star$, are determined in order to ensure the matching of the first $m^\star$ exact moments, that is by solving the system of equations

$$\frac{\partial^h}{\partial t^h} \Phi_W(t)\Big|_{t=0} = \frac{\partial^h}{\partial t^h} \Phi_W^\star(t)\Big|_{t=0}, h = 1, \ldots, m^\star \tag{5}$$

with Φ_W in (4),

$$\Phi_W^\star(t) = \sum_{j=0}^{m^\star} \pi_j^\star \left(\frac{\delta}{\delta - \mathrm{i}t}\right)^{\sum_{i=1}^{p} r_i + j}$$

and

$$\pi_{m^\star}^\star = 1 - \sum_{j=0}^{m^\star - 1} \pi_j^\star .$$

In Fig. 1, we present the smooth empirical probability density function (solid line) obtained from a simulated data of dimension 1,000,000, the approximating

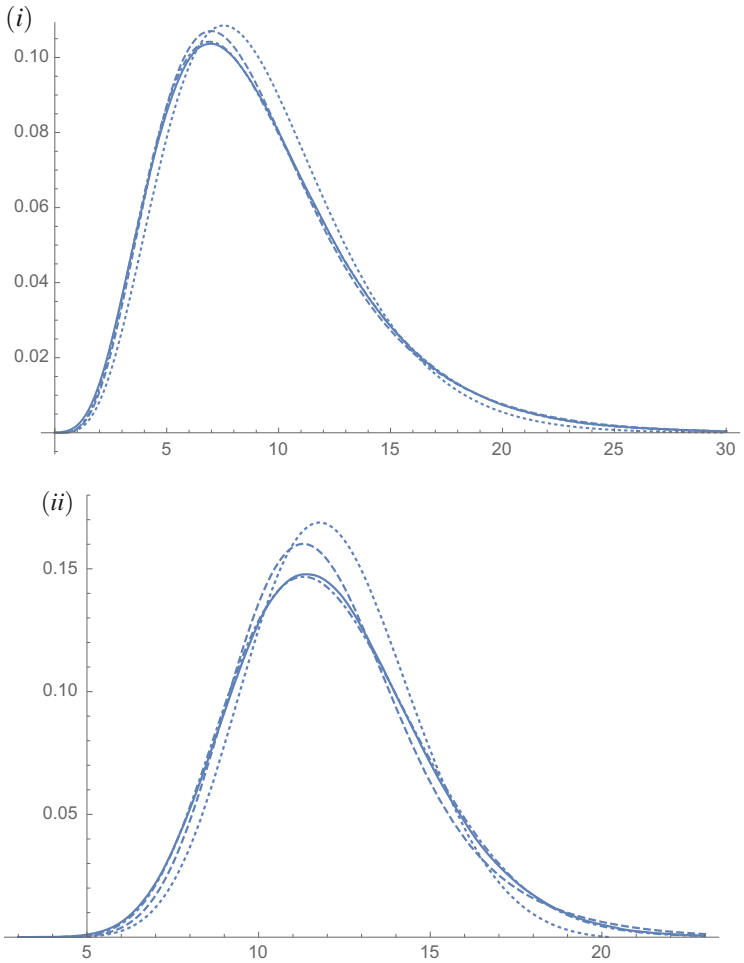


Fig. 1 The smooth empirical probability density function (solid line) and the approximating probability density functions for $m^\star = 1, 2$, and 4, respectively the dotted, dashed, and dot-dashed lines, for scenarios (i) and (ii)

probability density function, $f_W^\star$, for $m^\star = 1, 2$, and 4 (respectively, the dotted, dashed and dot-dashed lines). As it is possible to observe from Fig. 1, by matching four moments, which means that the mixture will have five terms, one obtains a fair agreement between the empirical and the approximating distribution in the two scenarios considered. Clearly, for other cases, it may be necessary to match more moments or to consider more advanced techniques such as near-exact distributions [11] in order to have good approximations.

3 The Sum of Independent Logbeta Random Variables and the Product of Independent Beta Random Variables

Using a similar approach to the one used in Sect. 2, we show that it is also possible to obtain a gamma-series representation for a single logbeta random variable, and that this representation may be used to derive a gamma-series representation for the sum of independent logbeta random variables. Clearly, these results may also be used to obtain a representation for the product of independent beta random variables. Thus, let $Y_i \sim \text{Beta}(a_i, b_i)$, $i = 1, \ldots, p$ be p independent beta random variables, then the density function of Y_i is given by

$$f_{Y_i}(y) = \frac{1}{B(a_i, b_i)}(1-y)^{b_i-1}\, y^{a_i-1}, \;\; 0 < y < 1, \;\; a_i > 0, \;\; b_i > 0\,.$$

We say that the random variable $W_i = -\log Y_i$ has a logbeta distribution with parameters $a_i > 0$ and $b_i > 0$, and we denote this fact by $W_i \sim \text{Logbeta}(a_i, b_i)$. The probability density function of W_i is given by

$$f_{W_i}(w) = \frac{1}{B(a_i, b_i)} \exp\{-a_i w\}(1-\exp\{-w\})^{b_i-1}, \;\; w > 0\,.$$

If we expand the factor $(1-\exp\{-w\})^{b_i-1}$ we obtain

$$f_{W_i}(w) = \sum_{j=0}^{\infty} \frac{(-1)^j \binom{b_i-1}{j}}{B(a_i, b_i)} \exp\{-(a_i+j)w\}, \;\; w > 0,$$

where

$$\binom{b_i-1}{j} = \frac{\Gamma(b_i)}{\Gamma(j+1)\Gamma(b_i-j+1)}\,.$$

From this last expression we obtain a well-known result which mentions that a single logbeta random variable may be represented as a mixture of exponential distributions with parameters $a_i + j$ and weights

$$\frac{(-1)^j \binom{b_i-1}{j}}{B(a_i, b_i)(a_i+j)}\,.$$

Now, if we consider a similar procedure to the one used for the sum of independent gamma random variables we have, for $\delta > 0$ and $\delta \neq j + a_i$ for all j and all i,

$$
\begin{aligned}
f_{W_i}(w) &= \sum_{j=0}^{\infty} \frac{(-1)^j \binom{b_i - 1}{j}}{B(a_i, b_i)} \exp\{-(a_i + j)w\} \frac{\exp\{(a_i + j)w - \delta w\}}{\exp\{(a_i + j)w - \delta w\}} \\
&= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{(-1)^j \binom{b_i - 1}{j}^{(-j-a_i+\delta)^k}}{B(a_i, b_i) k!} w^k \exp\{-\delta w\} \\
&= \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \frac{(-1)^j \binom{b_i - 1}{j}^{(-j-a_i+\delta)^k \Gamma(k+1)}}{B(a_i, b_i) k! \delta^{k+1}} \frac{\delta^{k+1}}{\Gamma(k+1)} w^k \exp\{-\delta w\} \\
&= \sum_{k=0}^{\infty} \left\{ \sum_{j=0}^{\infty} \frac{(-1)^j \binom{b_i - 1}{j}^{(-j-a_i+\delta)^k \Gamma(k+1)}}{B(a_i, b_i) k! \delta^{k+1}} \right\} \frac{\delta^{k+1}}{\Gamma(k+1)} w^k \exp\{-\delta w\} \\
&= \sum_{k=0}^{\infty} p_{i,k} f_{X_k}(w)
\end{aligned}
$$

which represents a mixture of gamma distributions with $X_k \sim \Gamma(k + 1, \delta)$ and weights given by

$$
p_{i,k} = \sum_{j=0}^{\infty} \frac{(-1)^j \binom{b_i - 1}{j}^{(-j-a_i+\delta)^k \Gamma(k+1)}}{B(a_i, b_i) k! \delta^{k+1}}. \tag{6}
$$

Again, for all W_i, we have a gamma-series representation with the same rate parameter δ, and this will allow us to obtain a representation for the sum of independent logbeta random variables in the form of a single gamma-series representation. Clearly, by simple transformation, we may derive a series representation for the product of independent beta random variables.

The characteristic function of $W_i = -\log Y_i$ is given by

$$
\Phi_{W_i}(t) = \sum_{k=0}^{\infty} p_{i,k} \Phi_{X_k}(t), \quad t \in \mathbb{R}
$$

with $p_{i,k}$ in (6) and where $\Phi_{X_k}(t)$ is the characteristic function of $X_k \sim \Gamma(k+1, \delta)$, for $\delta > 0$, which is given by

$$\Phi_{X_k}(t) = \left(\frac{\delta}{\delta - \mathrm{i}t}\right)^{k+1}, \quad t \in \mathbb{R}.$$

The characteristic function of $W = \sum_{i=1}^{p} W_i$ with $W_i \overset{ind}{\sim} \text{logBeta}(a_i, b_i)$, $i = 1, \ldots, p$ may be written as

$$\Phi_W(t) = \underbrace{\sum_{k_0=0}^{\infty} \left\{ \sum_{k_1=0}^{k_0} \sum_{k_2=0}^{k_1} \cdots \sum_{k_{p-1}=0}^{k_{p-2}} \left\{ \prod_{i=1}^{p} p_{i,k_{p-i}-k_{p-i+1}} \right\} \right\}}_{p^{\star}_{k_0}} \left(\frac{\delta}{\delta - \mathrm{i}t}\right)^{p+k_0} \tag{7}$$

with $k_p = 0$, which, in a simplified way, can be written as follows:

$$\Phi_W(t) = \sum_{k_0=0}^{\infty} p^{\star}_{k_0} \left(\frac{\delta}{\delta - \mathrm{i}t}\right)^{p+k_0}$$

which is a gamma-series representation, corresponding to an infinite mixture of gamma distributions, $\Gamma(p + k_0, \delta)$. As a result, the density and cumulative distribution functions of $W = \sum_{i=1}^{p} W_i$, are, respectively, given by

$$f_W(w) = \sum_{k_0=0}^{\infty} p^{\star}_{k_0} f_{\Gamma(p+k_0,\delta)}(w), \quad w > 0,$$

$$F_W(w) = \sum_{k_0=0}^{\infty} p^{\star}_{k_0} F_{\Gamma(p+k_0,\delta)}(w), \quad w \in \mathbb{R}$$

where $f_{\Gamma(p+k_0,\delta)}(w)$ and $F_{\Gamma(p+k_0,\delta)}(w)$ are, respectively, the density and cumulative distribution functions of a random variable with gamma distribution $\Gamma(p + k_0, \delta)$. We should note that, similar to what happens in Sect. 2, for the sum of independent gamma distributions, in these last representations the parameter δ can be chosen in order to increase the speed of convergence of the series and to improve the approximation obtained by truncation of the series and by matching a given number of the exact moments. By simple transformation the density and cumulative distribution functions of the product of independent beta

random variables, $Y = \prod_{i=1}^{p} Y_i$, are given by

$$f_Y(y) = \sum_{k_0=0}^{\infty} p_{k_0}^{\star} f_{\Gamma(p+k_0,\delta)}(-\log(y))\frac{1}{y}, \quad 0 < y < 1,$$

$$F_Y(y) = 1 - \sum_{k_0=0}^{\infty} p_{k_0}^{\star} F_{\Gamma(p+k_0,\delta)}(-\log(y)), \quad y \in \mathbb{R}.$$

Let us consider the following two scenarios:

(i) $a_i = \left\{\frac{3}{2}, \frac{8}{3}\right\}$ and $b_i = \left\{\frac{7}{5}, 2\right\}$

(ii) $a_i = \left\{\frac{13}{2}, \frac{15}{3}, \frac{9}{5}\right\}$ and $b_i = \left\{\frac{6}{5}, \frac{3}{2}, 1\right\}$.

The approximating probability density functions, for $W = \sum_{i=1}^{p} W_i$ and for $Y = \prod_{i=1}^{p} Y_i$, are, respectively, given by

$$f_W^{\star}(w) = \sum_{k=0}^{m^{\star}} \pi_k^{\star} f_{\Gamma(p+k,\delta)}(w), \quad w > 0$$

$$f_Y^{\star}(y) = \sum_{k=0}^{m^{\star}} \pi_k^{\star} f_{\Gamma(p+k,\delta)}(-\log(y))\frac{1}{y}, \quad 0 < y < 1.$$

In the above expressions, the parameter δ and the weights $\pi_k^{\star}$ are determined by a method analogous to the one used in Sect. 2. In Fig. 2 we present the smooth empirical probability density function (solid line) obtained from a simulated data of dimension 1,000,000, the approximating probability density function, $f_Y^{\star}$, of the product of independent beta random variables, for $m^{\star} = 1, 2$, and 4 (respectively, the dotted, dashed, and dot-dashed lines) in scenario (i), and the approximating probability density function, $f_W^{\star}$, of the sum of independent logbeta random variables, for $m^{\star} = 1, 2$, and 4 (respectively, the dotted, dashed, and dot-dashed lines) in scenario (ii). Similar to what happens in Sect. 2 one may observe, from Fig. 2, that when $m^{\star} = 4$, which corresponds to a case with five terms in the mixture, one obtains a better fit in both scenarios considered. However, we should mention that this feature may not happen in other scenarios for which the matching of more moments and/or the use of more technical and elaborated approximations as the ones developed in [7] may be needed. We should also note that, in Fig. 2, for $m^{\star} = 1$, the density in some intervals has negative values, this is due to the fact that, for the cases considered, when only one moment is matched, we have a mixture of two Gamma distributions, where one of the weights is negative and the other weight is bigger than one, and this leads to this strange behavior.

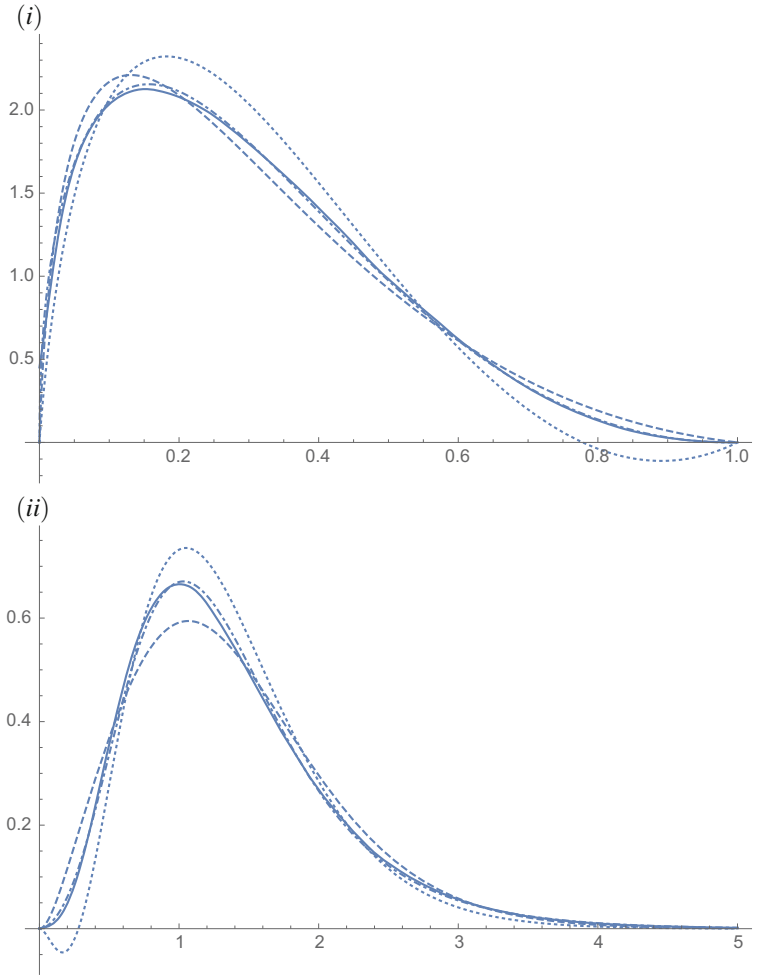


Fig. 2 The smooth empirical probability density function (solid line) and the approximating probability density functions for $m^\star = 1, 2$, and 4, respectively, the dotted, dashed, and dot-dashed lines, for the product of independent beta random variables in scenarios (i) and for the sum of independent logbeta random variables in scenario (ii)

4 Conclusions

Using the binomial and exponential expansions, simple gamma-series representations were obtained for the density of a gamma random variable and for the density of a logbeta random variable. In the gamma-series representations, all the gamma distributions have the same rate parameter which makes possible to derive gamma-series representations for the sum of independent gamma distributions and for the sum of independent logbeta distributions. By simple transformation it was also

possible to obtain a gamma-series representation for the product of independent beta random variables. The illustrations provided suggest that the truncation of the series and determination of the weights by matching a given number of the exact moments is an effective technique for the development of simple and accurate approximations for the distribution of the sum of independent gamma distributions and for the sum of independent logbeta distributions.

Acknowledgements This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

1. Alouini, M.-S., Abdi, A., Kaveh, M.: Sum of gamma variates and performance of wireless communication systems over Nakagami-fading channels. IEEE Trans. Veh. Technol. **50**, 1471–1480 (2001)
2. Amari, S.V., Misra, R.B.: Closed-form expressions for distribution of sum of exponential random variables. IEEE Trans. Reliab. **46**, 519–522 (1997)
3. Ansari, I.S., Yilmaz, F., Alouini, M.-S., Kucur, O.: New results on the sum of Gamma random variates with application to the performance of wireless communication systems over Nakagami-m fading channels. Trans. Emerg. Telecommun. Technol. **28**, e2912 (2014). https://doi.org/10.1002/ett.2912
4. Box, G.E.P.: A general distribution theory for a class of likelihood criteria. Biometrika **36**, 317–346 (1949)
5. Coelho, C.A.: The generalized integer Gamma distribution–a basis a logarithmized Gamma distribution and its characteristic for distributions in multivariate statistics. J. Multivar. Anal. **64**, 86–102 (1998)
6. Coelho, C.A.: The generalized near-integer Gamma distribution: a basis for 'near-exact' approximations to the distributions of statistics which are the product of an odd number of independent Beta random variables. J. Multivar. Anal. **89**, 191–218 (2004)
7. Coelho, C.A., Alberto, R.P.: On the distribution of the product of independent beta random variables applications. Technical Report, CMA 12 (2012)
8. Daniels, G.H.: Saddle point approximations in statistics. Ann. Math. Stat. **25**, 631–650 (1954)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 2. Wiley, Hoboken (1971)
10. Luke, Y.L.: The Special Functions and Their Approximations. Academic Press, London (1969)
11. Marques, F.J., Coelho, C.A.: A note on the distribution of the linear combination of independent Gamma random variables. Am. Inst. Phys. Conf. Proc. **1618**, 527–530 (2014)
12. Marques, F.J., Coelho, C.A., Arnold, B.C.: A general near-exact distribution theory for the most common likelihood ratio test statistics used in multivariate analysis. TEST **20**, 180–203 (2010)
13. Moschopoulos, P.G.: The distribution of the sum of independent gamma random variables. Ann. Inst. Stat. Math. **37**, 541–544 (1985)
14. Moschopoulos, P.G.: New representations for the distribution function of a class of likelihood ratio criteria. J. Stat. Res. **20**, 13–20 (1986)
15. Nakagami, M.: The m-distribution general formula of intensity distribution of rapid fading. In: Statistical Methods in Radio Wave Propagation, pp. 3–36. Pergamon, Oxford (1960)
16. Tang, J., Gupta, A.K.: Exact distribution of certain general test statistics in multivariate analysis. Aust. J. Stat. **28**, 107–114 (1986)

17. Tukey, J.W., Wilks, S.S.: Approximation of the distribution of the product of beta variables by a single beta variable. Ann. Math. Stat. **17**, 318–324 (1946)
18. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)

Likelihood Ratio Tests for Equality of Mean Vectors with Circular Covariance Matrices

Carlos A. Coelho

Abstract While the likelihood ratio test for the equality of mean vectors, when the covariance matrices are assumed to be only positive-definite, is a common test in multivariate analysis, similar likelihood ratio tests are not available in the literature when the covariance matrices are assumed to have some common given structure. In this compact paper the author deals with the problem of developing likelihood ratio tests for the equality of mean vectors when the covariance matrices are assumed to have a circular or circulant structure. The likelihood ratio statistic is obtained and its exact distribution is expressed in terms of products of independent Beta random variables. Then, it is shown how for some particular cases it is possible to obtain very manageable finite form expressions for the probability density and cumulative distribution functions of this distribution, while for the other cases, given the intractability of the expressions for these functions, very sharp near-exact distributions are developed. Numerical studies show the extreme closeness of these near-exact distributions to the exact distributions.

1 Introduction

The likelihood ratio test for the equality of mean vectors, when the covariance matrices are assumed to be just positive-definite is a well-known test in multivariate analysis, and the distribution of the associated test statistic has been extensively studied, often associated with the one-way MANOVA model [9, Chap. 9], [2, Chap. 8], [12, Chap. 10], [6, 11].

However, similar tests for the cases where some common given structure is assumed for the covariance matrices are not available in the literature. Let $\underset{\sim}{X}_k$ $(k = 1, \ldots, m)$ be p-variate random vectors, with expected value and

C. A. Coelho (✉)
Mathematics Department (DM) and Centro de Matemática e Aplicações (CMA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
e-mail: cmac@fct.unl.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_18

covariance matrix, respectively

$$E(\underline{X}_k) = \underline{\mu}_k \quad \text{and} \quad Cov(\underline{X}_k) = \Sigma_k \, .$$

In this paper we will be interested in the test to the hypotheses

$$\underline{\mu}_1 = \cdots = \underline{\mu}_k = \cdots = \underline{\mu}_m, \tag{1}$$

and

$$\underline{\mu}_1 = \cdots = \underline{\mu}_k = \cdots = \underline{\mu}_m = \underline{0}_{p\times 1}, \tag{2}$$

when $\Sigma_1, \ldots, \Sigma_k, \ldots, \Sigma_m$ are assumed to be equal and have a common circular structure. Initially the idea was to address: (1) the circular or circulant structure, (2) the compound symmetric structure, and (3) the spherical structure, but due to space limitations only the structure in (1) will be addressed. Tests when assuming one of the other structures will be addressed in a later paper.

The interest in these tests comes from the fact that such covariance structures are quite common or quite commonly assumed for covariance matrices in many situations, and, in case the assumption of such structures for the covariance matrices is correct, then not accounting for them when carrying out the tests for the mean vectors will lead to losses in power. The circular or circulant covariance structure is commonly assumed for real circulant stationary processes [14], cyclic designs [7], and in serially correlated time series [1], as well as in a wealth of other applications [8]. Therefore it is of interest to investigate tests for equality of mean vectors when assuming this structure for the covariance matrices. Since in these tests one assumes the equality of the covariance matrices, likelihood ratio tests for the equality of covariance matrices when assuming one of these structures are also of interest and will be addressed in a later publication.

2 Tests for the Equality and Simultaneous Nullity of Mean-Vectors When the Covariance Matrices Are Circular

2.1 Likelihood Ratio Test Statistic for the Equality of Mean Vectors

Let us assume that $\underline{X}_k \sim N_p\big(\underline{\mu}_k, \Sigma_k\big)$, $k = 1, \ldots, q$, where Σ_k are assumed to be equal and circular or circulant.

The $p \times p$ matrix Σ is said to be circular or circulant if $\Sigma = \Sigma_{cp}$, with

$$\Sigma_{cp} = [\sigma_{ij}], \quad i, j = 1, \ldots, p, \quad \text{where} \quad \sigma_{i,i+k} = \sigma_{i+k,i} = Cov(X_i, X_{i+k}) = \sigma_0^2 \rho_k, \tag{3}$$

with $\rho_0 = Corr(X_i, X_i) = 1$ and $\rho_k = \rho_{p-k} = Corr(X_i, X_{i+k})$, for $i = 1, \ldots, p-1$ and $k = 1, \ldots, p-i$. For example, for $p = 5$ and $p = 6$ we have, respectively,

$$\Sigma_{c5} = \sigma_0^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_2 & \rho_1 & 1 \end{bmatrix}, \qquad \Sigma_{c6} = \sigma_0^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_1 & \rho_2 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}.$$

Let us further suppose that we have a sample of size $n_k > p$ from $\underline{X}_k$ ($k = 1, \ldots, q$) and that these q samples are independent, with $n = \sum_{k=1}^{q} n_k$.

Then the $(2/n)$-th power of the likelihood ratio test (LRT) statistic to test the null hypothesis

$$\begin{aligned} H_0 : & \underline{\mu}_1 = \ldots \underline{\mu}_q \\ & \text{assuming } \Sigma_1 = \cdots = \Sigma_q (= \Sigma_{cp} \text{ non-specified}) \end{aligned} \tag{4}$$

where Σ_{cp} represents a circular matrix of order p, is

$$\Lambda = \prod_{j=1}^{p} \frac{v_j^*}{v_j^{**}}, \tag{5}$$

where, for $m = \lfloor p/2 \rfloor$,

$$v_j^* = \begin{cases} a_{jj}^{**}, & j = 1 \text{ and } j = 1 + m \text{ if } p \text{ is even} \\ \left(a_{jj}^{**} + a_{p-j+2,p-j+2}^{**}\right)/2, & j = 2, \ldots, p-m, m+2, \ldots, p, \end{cases} \tag{6}$$

and

$$v_j^{**} = \begin{cases} c_{jj}^{**}, & j = 1 \text{ and } j = 1 + m \text{ if } p \text{ is even} \\ \left(c_{jj}^{**} + c_{p-j+2,p-j+2}^{**}\right)/2, & j = 2, \ldots, p-m, m+2, \ldots, p, \end{cases} \tag{7}$$

with $v_j^* = v_{p-j+2}^*$ and $v_j^{**} = v_{p-j+2}^{**}$ for $j = 2, \ldots, p-m$. In (6) and (7) a_{jj}^{**} and c_{jj}^{**} represent the j-th diagonal elements, respectively, of

$$A^{**} = UAU' \quad \text{and} \quad C^{**} = U(A+B)U', \tag{8}$$

where

$$A = \sum_{k=1}^{q} (n_k - 1) S_k \quad \text{and} \quad B = \sum_{k=1}^{q} n_k \left(\overline{X}_k - \overline{X}\right)\left(\overline{X}_k - \overline{X}\right)' \tag{9}$$

are, respectively, the "within" and "between" sum of squares and sum of products matrices, with S_k and $\underline{X}_k$, respectively, the sample covariance matrix and mean vector of the k-th sample and

$$\underline{\overline{X}} = \frac{1}{n} \sum_{k=1}^{q} n_k \, \underline{X}_k \, . \tag{10}$$

The matrix U in (8) is an orthogonal symmetric matrix with running element

$$u_{ij} = \frac{1}{\sqrt{p}} \left\{ \cos\left(2\pi(i-1)(j-1)/p\right) + \sin\left(2\pi(i-1)(j-1)/p\right) \right\}, \tag{11}$$

for $i, j \in \{1, \ldots, p\}$.

There are a number of different ways in which the LRT statistic in (5) may be obtained. Likely, the easiest one is to derive it from the LRT statistic used to test the equality of mean vectors under the simple assumption of positive definiteness of the covariance matrices. We will use a similar statistic, to which we will now have to add the fact that the matrices Σ_k are assumed to be circular, by adequately computing the MLEs of the covariance matrices involved.

It is indeed not too hard to obtain explicit expressions for the MLEs of Σ_{cp} in (4) both under the null hypothesis and under the alternative hypothesis. If we take into account which are the elements in Σ_{cp} that are equal to $\sigma_0^2 \rho_k$, for $k = 0, 1, \ldots, m$, and that there are $2p$ of each of these elements, except for $k = 0$ and, when p is even, also for $k = m$, in which cases there are p of these elements, by the invariance property of the MLEs, the MLE of Σ_{cp}, under the alternative hypothesis

$$\begin{aligned} H_1 : \exists j, j' \in \{1, \ldots, q\} : \underline{\mu}_j \neq \underline{\mu}_{j'} \\ \text{assuming } \Sigma_1 = \cdots = \Sigma_q (= \Sigma_{cp} \text{ non-specified}), \end{aligned} \tag{12}$$

is $A^* = [a^*_{ij}]$, with

$$a^*_{j+k,\,j} = a^*_{j,\,j+k} = \widehat{\sigma_0^2 \rho_k}\big|_{H_1} = \frac{1}{2p} \sum_{j=1}^{p} \left(a_{j,\,mod^*(j+k,p)} + a_{mod^*(j+k,p),\,j} \right), \tag{13}$$

for $k = 0, \ldots, p-1$ and $j = 1, \ldots, p-k$, where a_{ij} represents the running element of the matrix A in (9), and

$$mod^*(a, b) = \begin{cases} mod(a, b), & mod(a, b) \neq 0 \\ b, & mod(a, b) = 0, \end{cases}$$

while the MLE of Σ_{cp} under the null hypothesis in (4) is $C^* = [c_{ij}^*]$, with

$$c_{j+k,j}^* = c_{j,j+k}^* = \widehat{\sigma_0^2 \rho_k}\big|_{H_0} = \frac{1}{2p}\sum_{j=1}^{p}\big(a_{j,mod^*(j+k,p)} + b_{j,mod^*(j+k,p)} + a_{mod^*(j+k,p),j} + b_{mod^*(j+k,p),j}\big), \tag{14}$$

for $k = 0, \ldots, p-1$ and $j = 1, \ldots, p-k$, where a_{ij} and b_{ij} represent the running elements of the matrices A and B in (9).

Then, the LRT statistic to test H_0 in (4) may be written as

$$\Lambda = \frac{|A^*|}{|C^*|} \tag{15}$$

where

$$\begin{aligned} |A^*| &= a_{11}^{**}\left(a_{1+m,1+m}^{**}\right)^{(p+1)\perp\!\!\!\perp 2}\prod_{j=2}^{p-m}\left(\frac{a_{jj}^{**} + a_{p-j+2,p-j+2}^{**}}{2}\right)^2 \\ &= v_1^*\left(v_{1+m,1+m}^*\right)^{(p+1)\perp\!\!\!\perp 2}\prod_{j=2}^{p-m} v_j^* \end{aligned} \tag{16}$$

and

$$\begin{aligned} |C^*| &= c_{11}^{**}\left(c_{1+m,1+m}^{**}\right)^{(p+1)\perp\!\!\!\perp 2}\prod_{j=2}^{p-m}\left(\frac{c_{jj}^{**} + c_{p-j+2,p-j+2}^{**}}{2}\right)^2 \\ &= v_1^{**}\left(v_{1+m,1+m}^{**}\right)^{(p+1)\perp\!\!\!\perp 2}\prod_{j=2}^{p-m} v_j^{**} \end{aligned} \tag{17}$$

with

$$(p+1) \perp\!\!\!\perp 2 = mod(p+1, 2) = \begin{cases} 1, \text{ for odd } p+1 \\ 0, \text{ for even } p+1, \end{cases}$$

and, as in (6) and (7), a_{jj}^{**} and c_{jj}^{**} represent, respectively, the j-th diagonal element of the matrices A^{**} and C^{**} in (8), so that Λ in (15) may be written as in (5).

We may note that v_j^* and v_j^{**}, defined in (6) and (7), are the MLEs of the eigenvalues δ_j in (18), respectively under H_1 in (12) and H_0 in (4).

2.1.1 Characterization of the Exact Distribution

In order to obtain the distribution of the LRT statistic Λ in (5) one only has to notice that

$$U\Sigma_{cp}U' = \Psi = diag(\delta_1, \delta_2, \ldots, \delta_p), \tag{18}$$

where $\delta_j = \delta_{p-j+2}$ for $j = 2, \ldots, p-m$, so that, since we know that the matrices A and B are independent, with

$$A \sim W_p(n-q, \Sigma_{cp}) \quad \text{and} \quad B \sim W_p(q-1, \Sigma_{cp}), \tag{19}$$

we have A^{**} and $B^{**} = UBU'$ independent, with

$$A^{**} \sim W_p(n-q, \Psi) \quad \text{and} \quad B^{**} \sim W_p(q-1, \Psi), \tag{20}$$

so that

$$C^{**} = A^{**} + B^{**} \sim W_p(n-1, \Psi)\,. \tag{21}$$

As such, we know that the diagonal elements of A^{**} are independent, as well as the diagonal elements of B^{**} and C^{**}, with

$$\frac{a_{jj}^{**}}{\delta_j} \sim \chi^2_{n-q}, \qquad \frac{b_{jj}^{**}}{\delta_j} \sim \chi^2_{q-1}, \qquad \frac{c_{jj}^{**}}{\delta_j} = \frac{a_{jj}^{**}}{\delta_j} + \frac{b_{jj}^{**}}{\delta_j} \sim \chi^2_{n-1}, \tag{22}$$

for $j = 1, \ldots, p$.

But then, from (16), (17), and (22) we see that

$$\Lambda \overset{d}{\equiv} Y_1 \left(Y^*\right)^{(p+1)\perp\!\!\!\perp 2} \prod_{j=2}^{p-m} Y_j^2 \tag{23}$$

where all r.v.'s are independent, with

$$Y_1 \overset{d}{\equiv} Y^* \sim Beta\left(\frac{n-q}{2}, \frac{q-1}{2}\right) \quad \text{and} \quad Y_j \sim Beta\,(n-q, q-1)\,. \tag{24}$$

2.1.2 Exact Distribution for an Odd Number of Samples

As such, using

$$\tfrac{\Gamma(n+a)}{\Gamma(a)} = \prod_{k=0}^{n-1}(a+k), \tag{25}$$

for any complex a and integer n, the h-th moment of Λ, for odd q, may be written as

$$
\begin{aligned}
E\left(\Lambda^h\right) &= \left(\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-q}{2}\right)}\frac{\Gamma\left(\frac{n-q}{2}+h\right)}{\Gamma\left(\frac{n-1}{2}+h\right)}\right)^{1+(p+1)\amalg 2}\prod_{j=2}^{p-\lfloor p/2\rfloor}\frac{\Gamma(n-1)}{\Gamma(n-q)}\frac{\Gamma(n-q+2h)}{\Gamma(n-1+2h)} \\
&= \left\{\prod_{k=0}^{\frac{q-3}{2}}\left(\frac{n-q}{2}+k\right)^{1+(p+1)\amalg 2}\left(\frac{n-q}{2}+k+h\right)^{-(1+(p+1)\amalg 2)}\right\} \\
&\quad\times\left\{\prod_{j=2}^{p-\lfloor p/2\rfloor}\prod_{k=0}^{q-2}(n-q+k)(n-q+k+2h)^{-1}\right\} \\
&= \prod_{\ell=1}^{q-1}\left(\frac{n-q+\ell-1}{2}\right)^{r_\ell}\left(\frac{n-q+\ell-1}{2}+h\right)^{-r_\ell},
\end{aligned}
$$

valid for any real or complex $h > -(n-q)/2$, and where for $\ell = 1, \ldots, q-1$,

$$
r_\ell = \left\lfloor\frac{p}{2}\right\rfloor + 1 - (1+(p+1)\amalg 2)\,((\ell-1)\amalg 2) = \begin{cases}\lfloor p/2\rfloor+1, & \text{odd } \ell \\ p-\lfloor p/2\rfloor-1, & \text{even } \ell\,.\end{cases} \tag{26}
$$

So that, for $W = -\log\Lambda$, we have

$$
\Phi_W(t) = E\left(e^{\mathrm{i}tW}\right) = E\left(\Lambda^{-\mathrm{i}t}\right) = \prod_{\ell=1}^{q-1}\left(\frac{n-q+\ell-1}{2}\right)^{r_\ell}\left(\frac{n-q+\ell-1}{2}-\mathrm{i}t\right)^{-r_\ell}
$$

which is the c.f. of a Generalized Integer Gamma (GIG) distribution of depth $q-1$, with rate parameters $(n-q+\ell-1)/2$ and shape parameters r_ℓ $(\ell = 1, \ldots, q-1)$, thus yielding the exact distribution of Λ in (5) and (23) as an Exponentiated GIG (EGIG) distribution of depth $q-1$ with rate parameters $(n-q+\ell-1)/2$ and shape parameters r_ℓ $(\ell = 1, \ldots, q-1)$, given by (26), with probability density function (p.d.f.) and cumulative distribution function (c.d.f.) given, respectively, by

$$
f_\Lambda(z) = f^{EGIG}\left(z \,\middle|\, \left\{r_\ell\right\}_{\ell=1:q-1}; \left\{\tfrac{n-q+\ell-1}{2}\right\}_{\ell=1:q-1}; q-1\right)
$$

and

$$
F_\Lambda(z) = F^{EGIG}\left(z \,\middle|\, \left\{r_\ell\right\}_{\ell=1:q-1}; \left\{\tfrac{n-q+\ell-1}{2}\right\}_{\ell=1:q-1}; q-1\right).
$$

See [4] for a reference on the GIG distribution and [3] for a reference on the EGIG distribution and the expressions of their p.d.f.'s and c.d.f.'s.

2.1.3 Near-Exact Distribution for an Even Number of Samples

For even q, using a similar technique to the one used in the previous subsection, one may write

$$
\begin{aligned}
E\left(\Lambda^h\right) &= \left(\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}+h\right)}{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}+h\right)}\,\frac{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-q}{2}+h\right)}{\Gamma\left(\frac{n-q}{2}\right)\Gamma\left(\frac{n-2}{2}+h\right)}\right)^{1+(p+1)\amalg 2} \\
&\quad \times \prod_{j=2}^{p-\lfloor p/2\rfloor} \frac{\Gamma(n)\,\Gamma(n-q+2h)}{\Gamma(n-q)\,\Gamma(n+2h)} \\
&= \left(\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\,\frac{\Gamma\left(\frac{n-2}{2}+h\right)}{\Gamma\left(\frac{n-1}{2}+h\right)}\right)^{1+(p+1)\amalg 2} \\
&\quad \times \left\{\prod_{k=0}^{\frac{q-4}{2}} \left(\frac{n-q}{2}+k\right)^{1+(p+1)\amalg 2} \left(\frac{n-q}{2}+k+h\right)^{-(1+(p+1)\amalg 2)}\right\} \\
&\quad \times \left\{\prod_{j=2}^{p-\lfloor p/2\rfloor}\prod_{k=0}^{q-2} (n-q+k)(n-q+k+2h)^{-1}\right\} \\
&= \left(\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}+h\right)}{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}+h\right)}\right)^{1+(p+1)\amalg 2} \prod_{\ell=1}^{q-1}\left(\frac{n-q+\ell-1}{2}\right)^{r_\ell}\left(\frac{n-q+\ell-1}{2}+h\right)^{-r_\ell}
\end{aligned}
$$

where, for $\ell = 1, \ldots, q-1$,

$$
r_\ell = \begin{cases} \lfloor p/2\rfloor + 1, & \text{odd } \ell,\ \ell \neq q-1 \\ p - \lfloor p/2\rfloor - 1, & \text{even } \ell \text{ and } \ell = q-1\,. \end{cases} \tag{27}
$$

Thus, for even q we may write

$$
\Phi_W(t) = \underbrace{\left(\frac{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n-2}{2}-\mathrm{i}t\right)}{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}-\mathrm{i}t\right)}\right)^{1+(p+1)\amalg 2}}_{\Phi_{W,1}(t)} \underbrace{\prod_{\ell=1}^{q-1}\left(\frac{n-q+\ell-1}{2}\right)^{r_\ell}\left(\frac{n-q+\ell-1}{2}-\mathrm{i}t\right)^{-r_\ell}}_{\Phi_{W,2}(t)}
$$

where we will leave $\Phi_{W,2}(t)$ unchanged and will asymptotically approximate $\Phi_{W,1}(t)$ by

$$
\widetilde{\Phi}_1(t) = \sum_{k=0}^{m^*} \pi_k (\lambda^*)^{1/2(1+(p+1)\amalg 2)} (\lambda^* - \mathrm{i}t)^{-1/2(1+(p+1)\amalg 2)},
$$

which is the c.f. of a finite mixture of $\Gamma(1/2, \lambda^*)$ or $\Gamma(1, \lambda^*)$ distributions, according to p being odd or even, and where λ^* is the rate parameter in

$$\Phi^*(t) = \theta(\lambda^*)^{r_1}(\lambda^* - \mathrm{i}t)^{-r_1} + (1-\theta)(\lambda^*)^{r_2}(\lambda^* - \mathrm{i}t)^{-r_2}$$

which will be computed together with θ, r_1 and r_2 in such a way that

$$\frac{\partial^h}{\partial t^h}\Phi^*(t)\bigg|_{t=0} = \frac{\partial^h}{\partial t^h}\Phi_{W,1}(t)\bigg|_{t=0}, \quad h = 1, \ldots, 4\,.$$

The weights π_k, $k = 0, \ldots, m^* - 1$, will then be computed in such a way that

$$\frac{\partial^h}{\partial t^h}\widetilde{\Phi}_1(t)\bigg|_{t=0} = \frac{\partial^h}{\partial t^h}\Phi_{W,1}(t)\bigg|_{t=0}, \quad h = 1, \ldots, m^* - 1,$$

with $\pi_{m^*} = 1 - \sum_{k=0}^{m^*-1} \pi_k$.

This approach is based on the fact that $\Phi_{W,1}(t)$ is the characteristic function (c.f.) of a single $Logbeta\left(\frac{n-2}{2}, \frac{1}{2}\right)$ distribution, for odd p, or the c.f. of a sum of two independent $Logbeta\left(\frac{n-2}{2}, \frac{1}{2}\right)$ distributions, for even p, and on the results in Section 5 of [15] which show that we can, for increasing values of a, replace asymptotically a $Logbeta(a, b)$ distribution by an infinite mixture of $\Gamma(b + k, a)$ $(k = 0, 1, \ldots)$ distributions. Then, using a somewhat heuristic approach, we truncate this mixture to a finite one and define λ^* and the weights π_k as above. This is an approach which, as it will be seen shortly ahead, gives in practice extremely good results.

By proceeding this way we will obtain

$$\Phi^*_W(t) = \widetilde{\Phi}_1(t)\,\Phi_{W,2}(t)$$

as a near-exact c.f. for W, which will yield as near-exact distributions for W mixtures with m^*+1 components, each of which is either a GIG or a Generalized Near-Integer Gamma (GNIG) distribution [5] of depth q, according to p being even or odd, with rate parameters r_ℓ $(\ell = 1, \ldots, q-1)$ and a q-th one either equal to 1 or 1/2, according to p being even or odd, and rate parameters $(n-q+\ell-1)/2$ $(\ell = 1, \ldots, q-1)$ and λ^*.

This gives near-exact distributions for Λ with p.d.f.

$$f_\Lambda(z) = \sum_{k=0}^{m^*} f^{GNIG}\left(\log z \,\middle|\, \left\{r_\ell\right\}_{\ell=1:q-1}, r; \left\{\tfrac{n-q+\ell-1}{2}\right\}_{\ell=1:q-1}, \lambda^*;\ q\right)\tfrac{1}{z}$$

and c.d.f.

$$F_\Lambda(z) = \sum_{k=0}^{m^*}\left(1 - F^{GNIG}\left(\log z \,\middle|\, \left\{r_\ell\right\}_{\ell=1:q-1}, r; \left\{\tfrac{n-q+\ell-1}{2}\right\}_{\ell=1:q-1}, \lambda^*;\ q\right)\right),$$

where

$$r = \begin{cases} 1, & \text{even } p \\ 1/2, & \text{odd } p, \end{cases}$$

being the case that when $r = 1$, the GNIG distribution becomes indeed a GIG distribution. See [5] and [11] for the expressions of $f^{GNIG}(\cdot)$ and $F^{GNIG}(\cdot)$, the p.d.f. and c.d.f. of the GNIG distribution.

This yields both for W and Λ very manageable distributions, which will match the first m^* exact moments of W and which may be shown to lie very close to the exact distribution.

In order to evaluate the proximity of these near-exact distributions to the exact distribution we will use the measure

$$\Delta = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| \frac{\Phi_W(t) - \Phi_W^*(t)}{t} \right| dt \tag{28}$$

with

$$\Delta \geq \max_{w>0} \left| F_W(w) - F_W^*(w) \right| \qquad \text{and} \qquad \Delta \geq \max_{0<z<1} \left| F_{\Lambda^*}(z) - F_{\Lambda^*}^*(z) \right|,$$

and where $\Phi_W(t)$ and $\Phi_W^*(t)$ represent, respectively, the exact and the near-exact characteristic functions of W and $F_W(\cdot)$ and $F_W^*(\cdot)$ the corresponding cumulative distribution functions.

In Table 1 we may analyze the values of the measure Δ in (28) for different sample sizes and different values of p and q, with smaller values of Δ showing a better agreement between the near-exact and the corresponding exact distribution. We may see how all the near-exact distributions exhibit extremely low values of the measure Δ and how they display a clear asymptotic behavior not only for increasing sample sizes but also for increasing values of p, the number of variables involved, as well as for increasing values of q, the number of samples involved. Noticeably, even for very small sample sizes the near-exact distributions exhibit extremely low values of Δ. As expected, near-exact distributions with higher values of m^* show lower values of the measure Δ, given that the near-exact distributions match the first m^* exact moments of W.

2.2 *Likelihood Ratio Test Statistic for the Simultaneous Nullity of Mean Vectors*

Let us consider a similar setting to the one in the previous Sect. 2.1. In the present section we are concerned with testing the hypothesis

$$\begin{aligned} H_0 : & \underline{\mu}_1 = \ldots \underline{\mu}_q = \underline{0}_{p\times 1} \\ & \text{assuming } \Sigma_1 = \cdots = \Sigma_q (= \Sigma_{cp} \text{ non-specified}) . \end{aligned} \tag{29}$$

Table 1 Values of the measure Δ for near-exact distributions that match the first m^* exact moments of W, for increasing values of p and q and samples of size $n = (p+1)q + 0, 100, 500$

			m^*			
p	q	n	2	4	10	20
3	6	24	1.05×10^{-8}	1.70×10^{-11}	4.65×10^{-18}	7.15×10^{-26}
		124	1.21×10^{-11}	6.93×10^{-16}	1.48×10^{-26}	5.01×10^{-41}
		524	3.65×10^{-14}	1.15×10^{-19}	4.10×10^{-34}	8.57×10^{-55}
10	6	66	5.96×10^{-11}	4.45×10^{-15}	3.56×10^{-25}	1.60×10^{-38}
		166	1.42×10^{-12}	1.65×10^{-17}	5.22×10^{-30}	5.39×10^{-47}
		566	1.02×10^{-14}	1.01×10^{-20}	1.99×10^{-36}	1.03×10^{-58}
30	6	186	1.78×10^{-13}	5.91×10^{-19}	5.52×10^{-33}	3.10×10^{-52}
		286	3.16×10^{-14}	4.41×10^{-20}	3.10×10^{-35}	2.47×10^{-56}
		686	9.44×10^{-16}	2.28×10^{-22}	8.32×10^{-40}	1.07×10^{-64}
3	16	64	2.92×10^{-11}	2.25×10^{-15}	2.05×10^{-25}	1.15×10^{-38}
		164	7.59×10^{-13}	9.60×10^{-18}	3.85×10^{-30}	5.56×10^{-47}
		564	5.64×10^{-15}	6.18×10^{-21}	1.58×10^{-36}	1.14×10^{-58}
3	36	144	2.85×10^{-13}	1.82×10^{-18}	1.21×10^{-31}	1.67×10^{-49}
		244	3.99×10^{-14}	9.72×10^{-20}	3.50×10^{-34}	3.84×10^{-54}
		644	9.18×10^{-16}	3.44×10^{-22}	4.42×10^{-39}	3.92×10^{-63}

Following a similar approach to the one used in the previous section, we may obtain the $(2/n)$-th power of the LRT statistic to test this null hypothesis as

$$\Lambda = \prod_{j=1}^{p} \frac{v_j^*}{v_j^{***}}, \tag{30}$$

where v_j^* are given by (6), and

$$v_j^{***} = \begin{cases} c_{jj}^{***}, & j = 1 \text{ and } j = 1+m \text{ if } p \text{ is even} \\ \left(c_{jj}^{***} + c_{p-j+2,p-j+2}^{***}\right)/2, & j = 2, \ldots, p-m, m+2, \ldots, p, \end{cases} \tag{31}$$

with c_{jj}^{***} representing the j-th diagonal element of

$$C^{***} = U(A + B^*)U', \tag{32}$$

where

$$B^* = \sum_{k=1}^{q} n_k \underline{\overline{X}}_k \underline{\overline{X}}_k' \,. \tag{33}$$

While the MLE of Σ_{cp} under the alternative hypothesis is the same as it is in the previous section, the MLE of Σ_{cp} under the null hypothesis in (29) is now $\widetilde{C}^* = \left[\tilde{c}_{ij}^*\right]$, with the elements $\tilde{c}_{ij}^*$ given by (14), with b_{ij} replaced by b_{ij}^*, the running element of the matrix B^* in (33). As such, the LRT statistic to test H_0 in (29) may be written as

$$\Lambda = \frac{|A^*|}{|\widetilde{C}^*|} \tag{34}$$

where $|A^*|$ is given by (16) and

$$\begin{aligned} |\widetilde{C}^*| &= c_{11}^{***} \left(c_{1+m,1+m}^{***}\right)^{mod(p+1,2)} \prod_{j=2}^{p-m} \left(\frac{c_{jj}^{***} + c_{p-j+2,p-j+2}^{***}}{2}\right)^2 \\ &= v_1^{***} \left(v_{1+m,1+m}^{***}\right)^{mod(p+1,2)} \prod_{j=2}^{p-m} v_j^{***} \end{aligned} \tag{35}$$

where c_{jj}^{***} is the j-th diagonal element of the matrix C^{***} in (32), so that Λ in (34) may be written as in (30).

We may note that v_j^{***} and v_j^*, defined in (6) and (31), are the MLEs of the eigenvalues δ_j in (18), respectively under H_0 in (29) and under the alternative hypothesis, where the matrices Σ_k are still assumed to be circular.

2.2.1 Characterization of the Exact Distribution

In order to obtain the distribution of the LRT statistic Λ in (30) and (34) one only has to notice that now the matrices A and B^* are independent, with

$$A \sim W_p(n-q, \Sigma_{cp}) \quad \text{and} \quad B^* \sim W_p(q, \Sigma_{cp}),$$

so that A^{**} in (8) and $B^{***} = U B^* U'$ are independent, with

$$A^{**} \sim W_p(n-q, \Psi) \quad \text{and} \quad B^{***} \sim W_p(q, \Psi),$$

for Ψ given by (18), and thus,

$$C^{***} = A^{**} + B^{***} \sim W_p(n, \Psi)\,.$$

Therefore, the diagonal elements of A^{**} are independent, as well as the diagonal elements of B^{***} and C^{***}, with

$$\frac{a_{jj}^{**}}{\delta_j} \sim \chi_{n-q}^2, \qquad \frac{b_{jj}^{***}}{\delta_j} \sim \chi_q^2, \qquad \frac{c_{jj}^{***}}{\delta_j} = \frac{a_{jj}^{**}}{\delta_j} + \frac{b_{jj}^{***}}{\delta_j} \sim \chi_n^2\,, \tag{36}$$

so that, from (16), (34), (35), and (36) we see that

$$\Lambda \overset{d}{\equiv} Y_1 \left(Y^*\right)^{(p+1)\perp\!\!\!\perp 2} \prod_{j=2}^{p-m} Y_j^2 \tag{37}$$

where all r.v.'s are independent, with

$$Y_1 \overset{d}{\equiv} Y^* \sim Beta\left(\frac{n-q}{2}, \frac{q}{2}\right) \quad \text{and} \quad Y_j \sim Beta\left(n-q, q\right) .$$

2.2.2 Exact Distribution for an Even Number of Samples

Using (25) and an approach in all similar to the one used in Sect. 2.1.2, we may write for even q the h-th moment of Λ as

$$E(\Lambda^h) = \prod_{\ell=1}^{q} \left(\frac{n-q+\ell-1}{2}\right)^{r_\ell} \left(\frac{n-q+\ell-1}{2}+h\right)^{-r_\ell}$$

with r_ℓ still given by (26), now for $\ell = 1, \ldots, q$. This yields for Λ its exact distribution as an EGIG distribution of depth q, with rate parameters $(n-q+\ell-1)/2$ and shape parameters r_ℓ $(\ell = 1, \ldots, q)$, with p.d.f. and c.d.f. similar to the ones in Sect. 2.1.2 with $q-1$ replaced by q.

2.2.3 Near-Exact Distribution for an Odd Number of Samples

Using a similar technique to the one used in the previous section and also in Sect. 2.1.3, we may write for odd q, the h-th moment of Λ as

$$E(\Lambda^h) = \underbrace{\left(\frac{\Gamma\left(\frac{n}{2}\right)\ \Gamma\left(\frac{n-1}{2}+h\right)}{\Gamma\left(\frac{n-1}{2}\right)\ \Gamma\left(\frac{n}{2}+h\right)}\right)^{1+(p+1)\perp\!\!\!\perp 2}}_{\Phi_{W,1}(t)} \underbrace{\prod_{\ell=1}^{q} \left(\frac{n-q+\ell-1}{2}\right)^{r_\ell} \left(\frac{n-q+\ell-1}{2}+h\right)^{-r_\ell}}_{\Phi_{W,2}(t)}$$

where now, for $\ell = 1, \ldots, q$,

$$r_\ell = \begin{cases} \lfloor p/2 \rfloor + 1, & \text{odd } \ell, \ell \neq q \\ p - \lfloor p/2 \rfloor - 1, & \text{even } \ell \text{ and } \ell = q \, . \end{cases}$$

Then, taking a similar approach to the one used in Sect. 2.1.3, we will obtain as near-exact distribution for Λ a mixture of m^*+1 exponentiated GNIG distributions of depth $q+1$, with shape parameters r_ℓ $(\ell = 1, \ldots, q)$ and $r = 1/2$ and rate

parameters $(n-q+\ell-1)/2$ and λ^*, for odd p, or a mixture of EGIG distributions of the same depth and with similar parameters, except for the shape parameter r which will have the value 1, for even p.

The performance of these near-exact distributions will be in all similar to that of the near-exact distributions in Sect. 2.1.3, with values of Δ of similar magnitudes for similar values of n, p, and q, with the only difference that now the values of q will be odd.

3 Concluding Remarks

The results obtained enable a quite simple implementation of both tests addressed, which are the test of equality of mean vectors and the test of simultaneous nullity of mean vectors, when the covariance matrices are assumed equal and with a circular symmetric structure. The fact that it was possible to express the exact distribution of the likelihood ratio statistics associated with these tests as EGIG distributions when q, the number of populations or samples involved, is odd in the test of equality of mean vectors, or when it is even in the test of simultaneous nullity of mean vectors, enables an extremely easy computation of quantiles and p-values from the c.d.f. of this distribution, which may be easily obtained from the c.d.f. of the GIG distribution. Mathematica® modules for the computation of the p.d.f. and the c.d.f. of the GIG distribution are available from a file placed on the web-page https://sites.google.com/site/nearexactdistributions/GIG-dist. For the construction of the near-exact p.d.f.'s and c.d.f.'s, the reader can find the Mathematica® modules for the p.d.f. and c.d.f. of the GNIG distribution on the web-page https://sites.google.com/site/nearexactdistributions/GNIG-dist.

A question that may arise in the mind of the reader interested in applying the tests addressed may then be: "but how can one test for circularity of the covariance matrices?". The LRT for circularity of the covariance matrix was developed by Olkin and Press [13] and for odd p, that is, for an odd number of variables involved, Marques and Coelho [10] have shown the exact distribution of the likelihood ratio statistic to be a GIG distribution. The same authors also developed near-exact distributions for the likelihood ratio statistic of this test for even p. However, in order to duly implement any of the two tests addressed in this paper, one would then still need to test the equality of the covariance matrices, assuming their circular structure. The LRT for such hypothesis, as well as near-exact distributions for the associated statistic are being developed by the same authors and are expected to be published shortly.

However, the use of this newly developed tests seems to entail rather slim systematic gains in power, when compared with the results obtained by using the common tests for positive-definite covariance matrices. But, the author believes that

by using the "trimmed" versions

$$\prod_{j=1}^{\lceil (p+1)/2 \rceil} \frac{v_j^*}{v_j^{**}} \quad \text{and} \quad \prod_{j=1}^{\lceil (p+1)/2 \rceil} \frac{v_j^*}{v_j^{***}}$$

instead of the statistics in (5) and (30), very large gains in power may be obtained. This is intended for future research, which due to space limitations is not undertaken here.

References

1. Anderson, T.W.: The Statistical Analysis of Time Series. Wiley, New York (1972)
2. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, Hoboken (2003)
3. Arnold, B.C., Coelho, C.A., Marques, F.J.: The distribution of the product of powers of independent uniform random variables – a simple but useful tool to address and better understand the structure of some distributions. J. Multivar. Anal. **113**, 19–36 (2013)
4. Coelho, C.A.: The generalized integer Gamma distribution – a basis for distributions in multivariate statistics. J. Multivar. Anal. **64**, 86–102 (1998)
5. Coelho, C.A.: The generalized near-integer Gamma distribution: a basis for 'near-exact' approximations to the distribution of statistics which are the product of an odd number of independent Beta random variables. J. Multivar. Anal. **89**, 191–218 (2004)
6. Coelho, C.A., Arnold, B.C., Marques, F.J.: Near-exact distributions for certain likelihood ratio test statistics. J. Stat. Theory Pract. **4**, 711–725 (2010)
7. John, J.A.: Cyclic Designs. Chapman and Hall, London (1987)
8. Khattree, R.: Multivariate statistical inference involving circulant matrices: a review. In: Gupta, A.K., Girko, V.L. (eds.) Multidimensional Statistical Analysis and Theory of Random Matrices, Proceedings of the Sixth Lukacs Symposium, pp. 101–110. VSP, The Nederlands (1996)
9. Kshirsagar, A.M.: Multivariate Analysis. Marcel Dekker, New York (1972)
10. Marques, F.J., Coelho, C.A.: Obtaining the exact and near-exact distributions of the likelihood ratio statistic to test circular symmetry through the use of characteristic functions. Comput. Stat. **28**, 2091–2115 (2013)
11. Marques, F.J., Coelho, C.A., Arnold, B.C.: A general near-exact distribution theory for the most common likelihood ratio test statistics used in multivariate analysis. TEST **20**, 180–203 (2011)
12. Muirhead, R.J.: Aspects of Multivariate Statistical Theory, 2nd edn. Wiley, Hoboken (2005)
13. Olkin, I., Press, S.J.: Testing and estimation for a circular stationary model. Ann. Math. Stat. **40**, 1358–1373 (1969)
14. Pollock, D.S.G.: Circulant matrices and time-series analysis. Int. J. Math. Educ. Sci. Technol. **33**, 213–230 (2002)
15. Tricomi, F.G., Erdélyi, A.: The asymptotic expansion of a ratio of Gamma functions. Pac. J. Math. **1**, 133–142 (1951)

Optimal Estimators in Mixed Models with Orthogonal Block Structures

Dário Ferreira, Sandra S. Ferreira, Célia Nunes, and João T. Mexia

Abstract Mixed models whose variance–covariance matrices are the positive definite linear combinations of pairwise orthogonal orthogonal projection matrices have orthogonal block structure. Here, we will obtain uniformly minimum-variance unbiased estimators for the relevant parameters when normality is assumed and we show that those for estimable vectors are, in general, uniformly best linear unbiased estimators. This is, they are best linear unbiased estimators whatever the variance components.

1 Introduction

A linear mixed model is an extended multivariate linear regression method of analysis for fixed and random effects. That kind of model is used for statistical modeling in a wide variety of fields. There have been extensive studies in estimation in mixed models, see, for example, [2, 16] and [15] or, in more recent years, [8] and [5].

In what follows, let us consider mixed models

$$\boldsymbol{Y} = \sum_{i=0}^{w} \boldsymbol{X}_i \boldsymbol{\beta}_i \tag{1}$$

D. Ferreira (✉) · S. S. Ferreira · C. Nunes
Department of Mathematics and Center of Mathematics and Applications, University of Beira Interior, Covilhã, Portugal
e-mail: dario@ubi.pt; sandraf@ubi.pt; celian@ubi.pt

J. T. Mexia
Center of Mathematics and Its Applications, Faculty of Science and Technology, New University of Lisbon, Lisbon, Portugal
e-mail: jtm@fct.unl.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_19

where $\boldsymbol{Y}$ is a vector of n random variables, $\boldsymbol{\beta}_0$ is a fixed vector, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_w$ are random and uncorrelated vectors, with null mean vectors and variance–covariance matrices $\sigma_1^2 \boldsymbol{I}_{c_1}, \ldots, \sigma_w^2 \boldsymbol{I}_{c_w}$, with $\boldsymbol{I}_{c_i}$ the $c_i \times c_i$ identity matrix and, possibly, $c_w = n$. Thus, these models will have mean vectors

$$E(\boldsymbol{Y}) = \boldsymbol{X}_0 \boldsymbol{\beta}_0 \tag{2}$$

and variance–covariance matrices

$$Var(\boldsymbol{Y}) = \sum_{i=1}^{w} \sigma_i^2 \boldsymbol{M}_i \tag{3}$$

where $\boldsymbol{M}_i = \boldsymbol{X}_i \boldsymbol{X}_i^\top$, $i = 1, \ldots, w$. We are interested in the case in which matrices $\boldsymbol{M}_i$, $i = 1, \ldots, w$, commute.

2 Estimators

When $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w$ commute they belong, see [10], to a commutative Jordan algebra, CJA, $\mathscr{A}$. This algebra will be a linear space constituted by symmetric matrices that commute and containing the squares of its matrices, see [11] and [12]. Moreover, $\mathscr{A}$ will have a unique basis, $pb(\mathscr{A}) = \left\{ \boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_w \right\}$, whose matrices are pairwise orthogonal orthogonal projection matrices so that

$$\boldsymbol{M}_i = \sum_{j=1}^{m} b_{i,j} \boldsymbol{Q}_j, \quad i = 1, \ldots, w, \tag{4}$$

see [17]. Therefore the model has variance–covariance matrices

$$\boldsymbol{V}(\boldsymbol{\sigma}^2) = \sum_{i=1}^{w} \sigma_i^2 \boldsymbol{M}_i = \sum_{j=1}^{m} \gamma_j \boldsymbol{Q}_j = \boldsymbol{V}(\boldsymbol{\gamma}) \tag{5}$$

with $\boldsymbol{\gamma} = \boldsymbol{B}^\top \boldsymbol{\sigma}^2$, where $\boldsymbol{B} = [b_{i,j}]$. When $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w$ is a basis for $\mathscr{A}$, the family $M = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w\}$ will be called perfect, see [6], and matrix $\boldsymbol{B}$ will be invertible with $m = w$. This is important for what follows since otherwise there would be linear restrictions on the $\gamma_1, \ldots, \gamma_w$. Then we may assume that $\boldsymbol{\gamma} \in \Gamma$, with Γ an open set in

$$R_>^w = \{\boldsymbol{v} : v_j > 0, j = 1, \ldots, w\} \tag{6}$$

so that, see [19], the model has orthogonal block structure, OBS. We must point out that this is not the original definition of OBS given by Nelder [13, 14], since we now assume the variance–covariance matrices to be positive definite, which is necessary to have densities. These models play an important role in design of experiments, see [7, 9] and [1], and in the theory of randomized block designs, see [3] and [4].

Let us now consider matrices $\boldsymbol{A}_j$, $j = 1, \ldots, w$, such that their row vectors constitute an orthonormal basis for the range space of $\boldsymbol{Q}_j$, $R(\boldsymbol{Q}_j)$, $j = 1, \ldots, w$. Then we may define the sub-models

$$\boldsymbol{Y}_j = \boldsymbol{A}_j \boldsymbol{Y}, \quad j = 1, \ldots, w, \tag{7}$$

which will have mean vector $\boldsymbol{\mu}_j = \boldsymbol{X}_{0,j}\boldsymbol{\beta}_0$, with

$$\boldsymbol{X}_{0,j} = \boldsymbol{A}_j \boldsymbol{X}_0, \quad j = 1, \ldots, w, \tag{8}$$

and variance–covariance matrix $\gamma_j \boldsymbol{Q}_j$, $j = 1, \ldots, w$. Then

$$\boldsymbol{P}_j = \boldsymbol{X}_{0,j}\boldsymbol{X}_{0,j}^{+}, \quad j = 1, \ldots, w, \tag{9}$$

where $+$ denotes MOORE-PENROSE inverse, will be the orthogonal projection matrix on the range space of $\boldsymbol{X}_{0,j}$, $R(\boldsymbol{X}_{0,j})$, $j = 1, \ldots, w$. Moreover, with $g_j = rank(\boldsymbol{Q}_j)$, $j = 1, \ldots, w$,

$$\boldsymbol{P}_j^c = \boldsymbol{I}_{g_j} - \boldsymbol{P}_j, \quad j = 1, \ldots, w, \tag{10}$$

will be the orthogonal projection matrices on the orthogonal complement of $R(\boldsymbol{X}_{0,j})$.

We put $p_j = rank(\boldsymbol{P}_j)$ and $p_j^c = rank(\boldsymbol{P}_j^c)$, $j = 1, \ldots, w$. Then, since $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_w$ are pairwise orthogonal orthogonal projection matrices, we have

$$\begin{aligned} \boldsymbol{X}_0 &= \sum_{j=1}^{w} \boldsymbol{Q}_j \boldsymbol{X}_0 = \sum_{j=1}^{w} \boldsymbol{A}_j^{\top} \boldsymbol{X}_{0,j} \\ &= \sum_{j=1}^{w} \boldsymbol{A}_j^{\top} \boldsymbol{P}_j \boldsymbol{X}_{0,j} = \sum_{j=1}^{w} (\boldsymbol{A}_j^{\top} \boldsymbol{P}_j \boldsymbol{A}_j) \boldsymbol{X}_0. \end{aligned} \tag{11}$$

With

$$\dot{\boldsymbol{Q}}_j = \boldsymbol{A}_j^{\top} \boldsymbol{P}_j \boldsymbol{A}_j, \quad j = 1, \ldots, w, \tag{12}$$

and

$$\bar{\boldsymbol{T}} = \sum_{j=1}^{w} \dot{\boldsymbol{Q}}_j, \tag{13}$$

we have

$$\boldsymbol{X}_0 = \bar{\boldsymbol{T}} \boldsymbol{X}_0. \tag{14}$$

So, since $\bar{\boldsymbol{T}}$ is an orthogonal projection matrix, with $\boldsymbol{T} = \boldsymbol{X}_0\boldsymbol{X}_0^+$ the orthogonal projection matrix on $\Omega = R(\boldsymbol{X}_0)$, and $\bar{\Omega} = R(\bar{\boldsymbol{T}})$ we have

$$\Omega \subset \bar{\Omega}, \tag{15}$$

as well as

$$\boldsymbol{T}\bar{\boldsymbol{T}} = \bar{\boldsymbol{T}}\boldsymbol{T} = \boldsymbol{T}, \tag{16}$$

so that

$$\boldsymbol{X}_0^+\boldsymbol{T} = \boldsymbol{X}_0^+\boldsymbol{T}\bar{\boldsymbol{T}} = \boldsymbol{X}_0^+\bar{\boldsymbol{T}}, \tag{17}$$

since

$$\boldsymbol{X}_0^+\boldsymbol{T} = \boldsymbol{X}_0^+\boldsymbol{X}_0\boldsymbol{X}_0^+ = \boldsymbol{X}_0^+. \tag{18}$$

Thus the least square estimator

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{X}_0^+\boldsymbol{T}\boldsymbol{Y}, \tag{19}$$

of $\boldsymbol{\beta}$ may be rewritten as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{X}_0^+\bar{\boldsymbol{T}}\boldsymbol{Y}. \tag{20}$$

Besides this, whenever $p_j^c < g_j$, we have the unbiased estimators

$$\widetilde{\boldsymbol{\gamma}}_j = \frac{\boldsymbol{Y}_j^\top \boldsymbol{P}_j^c \boldsymbol{Y}_j}{p_j^c} = \frac{S_j - \boldsymbol{Y}_j^\top \boldsymbol{P}_j \boldsymbol{Y}_j}{g_j - p_j}, \tag{21}$$

where $S_j = \|\boldsymbol{Y}_j\|^2$, $j = 1, \ldots, w$, and we also have

$$\widetilde{\boldsymbol{\sigma}}^2 = (\boldsymbol{B}^\top)^{-1}\widetilde{\boldsymbol{\gamma}}. \tag{22}$$

Let the row vectors of $\boldsymbol{W}_j$ constitute an orthogonal basis for $R(\boldsymbol{X}_{0,j})$, then, with $\boldsymbol{Z}_j = \boldsymbol{W}_j\boldsymbol{Y}_j$ and $\boldsymbol{\eta}_j = \boldsymbol{W}_j\boldsymbol{\mu}_j$, $j = 1, \ldots, w$, we have

$$\boldsymbol{Y}_j^\top \boldsymbol{P}_j \boldsymbol{Y}_j = \|\boldsymbol{Z}_j\|^2. \tag{23}$$

We now establish

Theorem 1 *If the model has OBS and* $rank(\boldsymbol{X}_0) = \sum_{j=1}^{w} p_j$ *then, for* $p_j^c < g_j$, $\widetilde{\boldsymbol{\beta}}$ *and* $\widetilde{\boldsymbol{\gamma}}_j$, $j = 1, \ldots, w$, *will be minimum variance unbiased estimators,*

UMVUE, when normality is assumed, while $\widetilde{\boldsymbol{\beta}}$ *will be uniformly best linear unbiased estimator, UBLUE, in general.*

Proof The first part of the thesis follows from, when normality is assumed, $\boldsymbol{Y}$ having density

$$n(\boldsymbol{y}) = \frac{e^{-\frac{1}{2}\sum_{j=1}^{w}\left(\vartheta_j S_j - 2\boldsymbol{\xi}_j^{\top}\boldsymbol{Z}_j + \frac{1}{\vartheta_j}\|\boldsymbol{\xi}_j\|^2\right)}}{\prod_{j=1}^{w}\left(\frac{2\pi}{\vartheta_j}\right)^{\frac{g_j}{2}}} \tag{24}$$

with $\vartheta_j = \gamma_j^{-1}$ and $\boldsymbol{\xi}_j = \frac{1}{\gamma_j}\boldsymbol{\eta}_j$, $j = 1, \ldots, w$, so the $(S_j, \boldsymbol{Z}_j)$, $j = 1, \ldots, w$, will constitute a sufficient statistic. Moreover, with $k = \sum_{j=1}^{w} p_j$, the parameter space will be $\Gamma \times R^k$, which will be open and so, see [18, p. 31], our estimators are derived from complete and sufficient statistics.

The second part of the thesis follows from the variance covariance of linear estimators $H\boldsymbol{Y}$ of $\boldsymbol{\beta}$ depending only on H and $\boldsymbol{V}(\boldsymbol{Y})$ and not on the distribution of $\boldsymbol{Y}$. Thus the proof is complete. □

Thus, if a model has OBS and $rank(\boldsymbol{X}_0) = \sum_{j=1}^{w} p_j$ then, see [19], it is an orthogonal model. A necessary and sufficient condition for a model with OBS to be orthogonal is that $\boldsymbol{T}$ commutes with the $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_w$, which is equivalent to $\boldsymbol{T}$ commuting with the $\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_w$.

3 Final Remarks

We considered mixed models with orthogonal block structure and showed that, when normality is assumed, our treatment leads to complete sufficient statistics and thus to UMVUE estimators for both variance components and estimable vectors. Besides this, we also showed that the estimators for estimable vectors are also UBLUE in general, since they only depend on the algebraic structure of the models.

Acknowledgements This work was partially supported by national funds of FCT—Foundation for Science and Technology under UID/MAT/00212/2013 and UID/MAT/00297/2013.

References

1. Bailey, R.A., Ferreira, S.S., Ferreira, D., Nunes, C.: Estimability of variance components when all model matrices commute. Linear Algebra Appl. **492**, 144–160 (2016)
2. Brown, H., Prescott, R.: Applied Mixed Models in Medicine. Wiley, New York (1999)
3. Calinski, T., Kageyama, S.: Block Designs: A Randomization Approach.Volume I: Analysis. Springer, New York (2000)

4. Calinski, T., Kageyama, S.: Block Designs: A Randomization Approach.Volume II: Analysis. Springer, New York (2003)
5. Demidenko, E.: Mixed Models: Theory and Applications with R. Wiley Series in Probability and Statistics. Wiley, New York (2013)
6. Ferreira, S.S., Ferreira, D., Fernandes, C., Mexia, J.T.: Orthogonal mixed models and perfect families of symmetric matrices. In: 56-th Session of the International Statistical Institute, ISI, p. 291 (2007)
7. Houtman, A., Speed, T.: Balance in designed experiments with orthogonal block structure. Ann. Stat. **11**(4), 1069–1085 (1983)
8. McCulloch, C.E., Searle, S.R., Neuhaus, J.M.: Generalized, Linear and Mixed Models. Wiley Series in Probability and Statistics. Wiley, New York (2008)
9. Mejza, S.: On some aspects of general balance in designed experiments. Statistica **52**, 263–278 (1992)
10. Mexia, J.T., Vaquinhas, R., Fonseca, M., Zmyślony, R.: COBS: segregation, matching, crossing and nesting. In: Latest Trends on Applied Mathematics, Simulation, Modeling, 4th International Conference on Applied Mathematics, Simulation, Modelling (ASM'10), pp. 249–255 (2010)
11. Michalski, A., Zmysloni, R.: Testing hypothesis for variance components in mixed linear models. Statistics **27**, 297–310 (1996)
12. Michalski, A., Zmysloni, R.: Testing hypothesis for linear functions of parameters in mixed linear models. Tatra Mt. Math. Publ. **17**, 103–110 (1999)
13. Nelder, J.A.: The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. Proc. R. Soc. Lond. Ser. A **273**, 147–162 (1965)
14. Nelder, J.A.: The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. Proc. R. Soc. Lond. Ser. A **273**, 163–178 (1965)
15. Pinheiro, J.C., Bates, D.M.: Mixed-Effects Models in S and S-PLUS. Springer, New York (2000)
16. Rao, C.R., Kleffe, J.: Estimation of Variance Components and Applications. North-Holland, Amsterdam (1988)
17. Seely, J.: Quadratic subspaces and completeness. Ann. Stat. **42**(2), 710–721 (1971)
18. Silvey, S.D.: Statistical Inference. Chapman & Hall, London (1965)
19. VanLeeuwen, D.M., Seely, J.F., Birkes, D.S.: Sufficient conditions for orthogonal designs in mixed linear models. J. Stat. Plan. Inference **73**, 373–389 (1998)

Constructing Random Klein Surfaces Without Boundary

Antonio F. Costa and Eran Makover

Abstract We describe a method to construct some non-compact Klein surfaces without boundary from 3-regular oriented graphs with a bicoloration of edges. These Klein surfaces are in fact Belyi Klein surfaces where we pinch the preimages of branched points. We can complete such surfaces to compact ones and these surfaces are the random Klein surfaces (without boundary). The consideration of random 3-regular graphs with orientation and a bicoloration of edges give us a method for computing the probability of satisfying a given geometrical property for random Klein surfaces. The relation between random Klein surfaces and random Riemann surfaces allows to claim new properties for random Klein surfaces.

1 Introduction

In [4] and using a probabilistic model, the authors obtain information about geometrical properties of typical Riemann surfaces of large genera (see also [3, 6] and the recent work [8] by B. Petri). For this purpose they construct Riemann surfaces from 3-regular graphs with orientation. Random oriented 3-regular graphs with big number of vertices provide surfaces of large genera and an experimental model to claim important geometrical properties of such surfaces. For instance, the existence of a low bound for systoles when the genus tends to infinity.

In this article we adapt a similar procedure for Klein surfaces without boundary.

Klein surfaces are surfaces with a dianalytic structure instead as an analytic one. In dianalytic structures the changes of charts may be analytic or antianalytic (see [1]). These surfaces were introduced by Klein for the study of real algebraic curves

A. F. Costa (✉)
Departamento de Matemáticas Fundamentales, UNED, Madrid, Spain
e-mail: acosta@mat.uned.es

E. Makover
Department of Mathematics, Central Connecticut State University, New Britain, CT, USA
e-mail: makovere@ccsu.edu

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_20

(there is a functorial equivalence between the category of Klein surfaces and the category of real algebraic curves). Klein surfaces may be non-orientable and with boundary, in this work we shall restrict our attention to the case of surfaces without boundary.

First we shall describe a method to construct some non-compact Klein surfaces without boundary from 3-regular oriented graphs with a bicoloration of edges and we shall call such surfaces open random Klein surfaces. The open random Klein surfaces are in fact Belyi Klein surfaces where we pinch the preimages of branched points (see Theorem 2). Using the completion method of Brooks [2] we can complete such surfaces to compact ones and these surfaces will be the random Klein surfaces. The consideration of random 3-regular graphs with orientation and a bicoloration of edges give us a method for computing the probability of satisfying a given geometrical property for random Klein surfaces. The relation between random Klein surfaces and random Riemann surfaces allows to claim interesting properties for random Klein surfaces (see Sect. 6).

2 Combinatorial Model

In this section we shall define a combinatorial object that we shall use to define the random Klein surfaces without boundary: oriented 3-regular graphs with a bicolouration of edges. An oriented 3-regular graph (G, O) is a pair consisting of a 3-regular (pseudo-multi) graph G (multiple edges and loops are allowed) and an orientation O of G. Let us recall the definition of orientation for a 3-regular graph G. Let $v \in V(G)$ and e_1^v, e_2^v, e_3^v be the three adjacent to v edges. There are two cyclic orderings (orientations) of $\{e_1^v, e_2^v, e_3^v\}$: o_1^v and o_2^v, where

$$(e_1^v, e_2^v, e_3^v) \in o_1^v, (e_1^v, e_3^v, e_2^v) \in o_2^v.$$

Let $o_v = \{o_1^v, o_2^v\}$ be the set of orientations in v. An orientation of G is a map $O : V(G) \to \{o_v : v \in V(G)\}$. Finally a bicoloration of the edges of G is a map:

$$\varepsilon : E(G) \to \{0, 1\}.$$

We shall denote (G, O, ε), where (G, O) is an oriented 3-regular graph and ε is a bicoloration of the edges of G.

3 Construction of Random Klein Surfaces Without Boundary

Now we describe the method of construction of an open Klein surface without boundary from (G, O, ε). First we construct $\gamma_v : \{e_1^v, e_2^v, e_3^v\} \to \{0, 1, \infty\}$ such that

$$\left(\gamma_v^{-1}(0), \gamma_v^{-1}(1), \gamma_v^{-1}(\infty)\right) \in O(v), \text{ for each } v \in V(G).$$

Now, for each $v \in V(G)$, we consider a copy of the ideal triangle with vertices $0, 1, \infty$: T_v. Let v, w be two adjacent vertices of G and $e \in E(G)$ connecting v with w. Let $e \in E(G)$. If $\varepsilon(e) = 0$, we identify the side opposed to $\gamma_v(e)$ of T_v with the opposed side to $\gamma_v(e)$ of T_w by the identity map $T_v \to T_w$ or using a power of the rotation $z \to \frac{1}{z-1}$. This corresponds to make the identification between edges given by the map that sends the midpoint of the edge of T_v to the midpoint of the edge of T_w and preserves the orientation of T. If $\varepsilon(e) = 1$, we identify the side opposed to $\gamma_v(e)$ of T_v with the side opposed $\gamma_v(e)$ of T_w using one of the reflections:

$$z \to \frac{1}{\overline{z}}; \quad z \to 1 - \overline{z},$$

or one of the above reflections composed with a power of the rotation $z \to \frac{1}{z-1}$. In this case the identification preserves the midpoints but reverses the orientation from T_v to T_w. With the above construction we obtain a triangular map on a finite area open Klein surface $S^O(G, O, \varepsilon)$ without boundary that is orientable or not. The surfaces $S^O(G, O, \varepsilon)$ obtained from all (G, O, ε) are the open random Klein surfaces.

4 Probability Measure

For each n let $\mathscr{G}_{2n}$ denote the finite set of 3-regular multigraphs with $2n$ vertices. We begin with a graph consisting only of $2n$ vertices $\{v_1, \ldots, v_{2n}\}$. We consider a set W of $6n$ elements partitioned in $2n$ subsets labelled $v_1, \ldots, v_{2n}$ of 3 points each one. A perfect matching of the elements of W into $3n$ sets of two elements is called a pairing. From each pairing we construct a multigraph with n vertices in the following way: if $\{x, y\}$ is a set of pairing, $x \in v_i$, $y \in v_j$, we add to the graph an edge joining the vertex v_i with the vertex v_j. If $\mathscr{P}_W$ is the set of pairings of W, we have a surjective map ϕ from $\mathscr{P}_W$ to the $\mathscr{G}_{2n}$. Giving to all partitions the same probability and using the map ϕ we have a probability measure on $\mathscr{G}_{2n}$. For each element in $\mathscr{G}_{2n}$ we consider all the orientations and bicolorations of edges with uniform probability. This procedure produces a probability space for the set of 3-regular oriented and bicolored multigraphs (note that $\mathscr{G}_{2n}$ is not a uniform space, see [9]).

Remark 1 The graph G obtained in this way may be non-connected, but

$$\lim_{n\to\infty} \text{Prob}_n[(G, O, \varepsilon) \text{ when } G \text{ is a simple graph, } G \text{ connected}] \to 1$$

(see [9] section 2.6).

Note There is other method to obtain a probability measure on the set of open random Klein surfaces. The open random Klein surfaces without boundary are orbifold

coverings of the orbifold $\mathbf{H}/\langle\Lambda\rangle$, where Λ is the hyperbolic crystallographic group generated by the reflections in the sides of the hyperbolic triangle T in Sect. 3 and $\mathbf{H}$ is the complex upper half-plane. Such coverings of $2n$ sheets are given by three permutations that are products of n disjoint 2-cycles. Then each such covering is given by a set of three partitions of $\{1, \ldots, 2n\}$ in sets of two elements. This procedure gives a probability measure on the set of open random Klein surfaces that is different from the one described before.

5 Belyi Klein Surfaces and Random Riemann Surfaces

A compactification $S^C(G, O, \varepsilon)$ of an open random Klein surface $S^O(G, O, \varepsilon)$ is called random Klein surface. In this way we obtain a very important family of surfaces: Belyi Klein surfaces without boundary. Recall the following theorem by Kock and Singerman (see [7]).

Theorem 1 *A Klein surface S admits a Belyi map, i.e. an orbifold covering $\beta : S \to \Delta$, where Δ is a hyperbolic compact triangle, if and only if S is isomorphic to the compactification of the quotient surface $\mathbf{H}/L$ for some surface (non-cocompact) subgroup L of finite index of the extended modular group Γ^* acting on the upper half-plane $\mathbf{H}$.*

A Klein surface admitting a Belyi map is called a Belyi Klein surface. Note that if S is a Belyi Klein surface with non-empty boundary, S can be represented by an algebraic curve defined in $\overline{\mathbf{Q}} \cap \mathbf{R}$, but this result is unknown in the case of non-orientable Klein surfaces without boundary, precisely the surfaces considered in the present work. We have the following theorem that is the analogous to Lemma 2.1 in [5]:

Theorem 2 *The surface S is a Belyi Klein surface without boundary if and only if S is a random Klein surface, i.e. the compactification of an open random Klein surface.*

Proof Let S be a Belyi Klein surface without boundary. By Theorem 3.2 in [7] S is the compactification of $\mathbf{H}/L$ for some surface subgroup L of finite index of the extended modular group Γ^*. Note that Γ^* is a triangular group of signature $(2, 3, \infty)$. There is an orbifold covering $\beta : \mathbf{H}/L \to \Delta$ where Δ is a hyperbolic triangle of angles $\pi/2, \pi/3$ and a vertex in the infinity line with angle 0. The lifting of the side of Δ from the vertex with angle $\pi/3$ to the vertex with angle $\pi/2$, produces a 3-regular graph G in $\mathbf{H}/L$ and the lifting of the orientation in Δ gives an orientation on G. Now consider a spanning tree T for G and let R be the edges of G that are not in T. The surface S is obtained from a hyperbolic polygon, where T is embedded, pairwise identifying sides that are traversed by the edges in R. Then we give coloration 0 to the sides in T and the sides of R traversing sides of the hyperbolic polygon to be identified by orientation preserving isometries. We give coloration 1 to the sides of R through sides of the hyperbolic polygon to be

identified by orientation reversing isometries. The oriented 3-regular graph with this edge coloration produces the surface $\mathbf{H}/L$, then S is a random Klein surface.

Let now $S^O(G, O, \varepsilon)$ be an open random Klein surface. Since O is a orientation of the graph G we can construct $\gamma_v : \{e_1^v, e_2^v, e_3^v\} \to \{1, 2, 3\}$ such that

$$\left(\gamma_v^{-1}(1), \gamma_v^{-1}(2), \gamma_v^{-1}(3)\right) \in O(v),$$

for each $v \in V(G)$. An orbifold covering of the hyperbolic triangle $\Delta(2, 3, \infty)$ is given by its monodromy $\omega : \pi_1 O(2, 3, \infty) \to \Sigma_h$, where $\pi_1 O(2, 3, \infty)$ is the orbifold fundamental group, h is the degree of the covering and Σ_h is the group of permutations of $\{1, \ldots, h\}$. If m is the number of vertices of G, we consider $h = 6m$ and we label the elements of $\{1, \ldots, h\}$ by $\{(i, j, k) : i \in \{1, \ldots, m\}, j \in \{1, 2, 3\}, k \in \{0, 1\}\}$. The group $\pi_1 O(2, 3, \infty)$ has a presentation:

$$\left\langle c_1, c_2, c_3 : c_i^2 = (c_1 c_2)^2 = (c_2 c_3)^3 = 1 \right\rangle.$$

We define

$$\omega(c_1)(i, j, k) = (i', j', k')$$

if there is an edge e of G joining the vertices v_i with $v_{i'}$, $\gamma_{v_i}(e) = j$ and $\gamma_{v_i'}(e) = j'$, $k = k'$ if $\varepsilon(e) = 0$ and $k = (k' + 1)\text{mod}2$ if $\varepsilon(e) = 1$;

$$\omega(c_2)(i, j, k) = (i, j, (k + 1)\text{mod}2),$$

and

$$\omega(c_2)(i, j, k) = (i, (j + 1)\text{mod}3, (k + 1)\text{mod}2).$$

The covering of Δ with monodromy ω is precisely $S^O(G, O, \varepsilon)$, then $S^O(G, O, \varepsilon)$ has a Belyi surface as compactification.

The following proposition tells us that with our method and for big random Klein surfaces the non-orientability is the usual situation:

Proposition 1 $\lim_{n\to\infty} \text{Prob}_n[S^{O,C}(G, O, \varepsilon)$ *is orientable*$] \to 0$.

Proof Consider a spanning tree T of G and let R be the sides of G that are not in T. For each bicoloration of the sides of T there is only a coloration of the elements of R producing an orientable surface $S^O(G, O, \varepsilon)$. Since the number of elements of R is $n + 1$, the probability of having an orientable random surface $\to 0$ when $n \to \infty$.

6 From Random Klein Surfaces to Random Riemann Surfaces

If (G, O) is an oriented 3-regular graph we shall denote $S^O(G, O)$ to the open random Riemann surface constructed from (G, O) following the construction in [7], note that $S^O(G, O) = S^O(G, O, \varepsilon_0)$ where ε_0 is the trivial 0 coloration of the sides of G. The usual way relating Klein with Riemann surfaces is using the complex double, see, for instance, [1]. Given a non-orientable Klein surface K without boundary, a complex double of K is a Riemann surface S such that $S \to K$ is a two-fold unbranched covering. Assume that $S^O(G, O, \varepsilon)$ is non-orientable. The complex double $DS^O(G, O, \varepsilon)$ of $S^O(G, O, \varepsilon)$ is the finite area non-compact random Riemann surface $S^O(G', O')$, where (G', O') is a two-fold covering of (G, O). The two-fold covering $c : G' \to G$ is given as follows: for each cycle γ of G the lift $c^{-1}(\gamma)$ has two components if and only if $\Sigma\varepsilon(e_i)\text{mod}(2) = 0$, where e_i are the edges of the cycle γ. From the case of open random Klein surfaces follows the relation between compact non-orientable random Klein surfaces without boundary and compact random Riemann surfaces using complex doubles: Let $S^C(G, O, \varepsilon)$ be a compactification of $S^O(G, O, \varepsilon)$. There is a compactification $S^C(G', O')$ of the complex double $S^O(G', O')$ of $S^O(G, O, \varepsilon)$ and a two-fold covering $p : S^C(G', O') \to S^C(G, O, \varepsilon)$. Using this relation we can control the hyperbolic metric of the compactification $S^C(G, O, \varepsilon)$ from the metric in $S^O(G, O, \varepsilon)$ as in the case of random Riemann surfaces (see 3.2 of [4] or [2]).

In addition to the complex double there is other natural way of associating a random Riemann surface with a random Klein surface: given a random Klein surface $S^O(G, O, \varepsilon)$ simply consider $S^O(G, O) = S^O(G, O, \varepsilon_0)$. Note that $S^O(G, O, \varepsilon)$ and $S^O(G, O)$ have many common properties, in fact all the properties coming from properties of the graph G as the length of closed geodesics. As a difference a left-hand-turns path is not the boundary of a region containing a cusp in $S^O(G, O, \varepsilon)$, when the sum $\Sigma\varepsilon(e_i)\text{mod}(2) = 1$, where e_i are the edges of the path. In this case to obtain the boundary of a region containing a cusp it is necessary to consider two left-hand-turns paths. In any case from Theorem 2.1A of [4] there is some L such that, when $n \to \infty$, $\text{Prob}_n[S^O(G, O, \varepsilon)$ has cusps of length $\geq L] \to 1$, where we are considering the probability measure described in Sect. 4. Also as a consequence of this relation there exists a constant C such that

$$\lim_{n\to\infty} \text{Prob}_n\big[\text{syst}(S^C(G, O, \varepsilon)) \geq C\big] \to 1.$$

As conclusion, with the model presented in this article we open the possibility of the study of geometric properties of random Klein surfaces without boundary, these surfaces correspond to a special type of real algebraic curves.

Acknowledgements The first author was partially supported by the Spanish research project MTM2014-55812-P.

References

1. Alling, N., Greenleaf, N.: Klein surfaces and real algebraic function fields. Bull. Am. Math. Soc.**75**(4), 869–872 (1969)
2. Brooks, R.: Platonic surfaces. Comment. Math. Helv. **74**(1), 156–170 (1999)
3. Brooks, R.: A statistical model of Riemann surfaces. In: Complex Analysis and Dynamical Systems. Contemporary Mathematics, vol. 364, pp. 15–25. American Mathematical Society, Providence (2004)
4. Brooks, R., Makover, E.: Random construction of Riemann surfaces. J. Differ. Geom. **68**(1), 121–157 (2004)
5. Gamburd, A.: Poisson-Dirichlet distribution for random Belyi surfaces. Ann. Probab. **34**(5), 1827–1848 (2006)
6. Gamburd, A., Makover, E.: On the genus of a random Riemann surface. In: Complex Manifolds and Hyperbolic Geometry (Guanajuato, 2001). Contemporary Mathematics, vol. 311, pp. 133–140. American Mathematical Society, Providence (2002)
7. Köck, B., Singerman, D.: Real Belyi theory. Q. J. Math. **58**(4), 463–478 (2007)
8. Petri, B.: Random regular graphs and the systole of a random surface. J. Topol. **10**(1), 211–267 (2017)
9. Wormald, N.C.: Models of random regular graphs. In: Surveys in Combinatorics, 1999 (Canterbury). London Mathematical Society Lecture Note Series, vol. 267, pp. 239–298. Cambridge University Press, Cambridge (1999)

Performance Analysis of a GPS Equipment

M. Filomena Teodoro, Fernando M. Gonçalves, and Anacleto Correia

Abstract In emerging economies the easiest way to ensure the geodetic support still is the static relative positioning (SRP) using a single reference station. This technique provides surveyors the ability to determine the 3D coordinates of a new point with centimeter-level accuracy. The objective of this work is to evaluate GPS SRP regarding accuracy, as the equivalent of a real time kinematic (RTK) network and to address the practicality of using either a continuously operating reference stations (CORS) or a passive control point for providing accurate positioning control. The precision of an observed 3D relative position between two global navigation satellite systems (GNSS) antennas, and how it depends on the distance between these antennas and on the duration of the observing session, was studied. We analyze the performance of the software for each of the six chosen ranges of length in each of the four scenarios created, considering different intervals of observation time. An intermediate inference level technique (Tamhane and Dunlop, Statistics and data analysis: from elementary to intermediate, Prentice Hall, New Jersey, 2000), an analysis of variance, establishes the evidence of relation between observing time and baseline length.

M. F. Teodoro (✉)
CINAV, Center of Naval Research, Naval Academy, Almada, Portugal

CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Lisboa, Portugal
e-mail: maria.alves.teodoro@marinha.pt

F. M. Gonçalves
NGI, Nottingham Geospatial Institute, University of Nottingham, Nottingham, UK

A. Correia
CINAV, Center of Naval Research, Naval Academy, Almada, Portugal
e-mail: cortez.correia@marinha.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_21

1 Introduction

RTK networks are common in Europe but this is not the case in emerging economies where huge construction projects are running requiring geodetic support. In such cases, the easiest way to ensure that kind of support still is the SRP using a single reference station. This technique provides surveyors the ability to determine the 3D coordinates of a new point with centimeter-level accuracy relative to a control point located several hundred kilometers away, which in turn can be associated with another GNSS receiver of a CORS operated by some institution.

Today the global navigation satellite systems play a fundamental role in the way that surveyors measure positional coordinates. It is now possible to determine the 3D coordinates of a new point with centimeter-level accuracy relative to a control point located several hundred kilometers away, which in turn can be associated with another GNSS receiver of a CORS operated by some institution. Examples of such networks are the ordnance survey (OS) Network across the UK [14] or, globally, the International GNSS Service Network [7].

With the implementation of real time networks (RTN), particularly across Europe and North America, the way surveyors work has dramatically changed over the last few years. Certainly, the growth of RTN will continue and it is expected that in the near future the work taking place in areas covered by these infrastructures will be dominated by RTK techniques. However, in other regions of the world which can become, or already are, of interest for scientific or industry projects, this type of infrastructure does not exist. Consequently, "old" methods such as the static observation and post processing continue to play a prominent role in GNSS surveying in order to provide accurate position solutions without support of network corrections. Furthermore, when surveyors decide which GNSS methods to use, they must consider several aspects of a project. Besides specific requirements from clients, other important factors to be considered are budget, schedule, accuracy, and control over how data is managed.

In this research the coordinates of the OS active stations were used as "true" values to address the practicality of using either a CORS or a passive control point for providing accurate positioning control and, implicitly, the performance of the software used. The precision of an observed 3D relative position between two GNSS antennas, and how it depends on the distance between these antennas and on the duration of the observing session, was studied. These results were attained through using commercial software LGO to process 105 single baselines, ranging from 61 to 898 km, according to observing sessions of varying lengths. ABEP was used as a reference station, with fixed coordinates, and the values obtained for the rover stations compared with those provided by OS. Also, to address the differences between using broadcast or precise ephemerides and computing the tropospheric effects or for simply applying a tropospheric model, the data processing was repeated for all different strategies.

Generally results show, whatever the strategy followed, that the length of the baseline matters, regarding the rate of successful baselines processed for a priori

given values of 1D (ellipsoidal height accuracy) and 2D (compound of longitude and latitude accuracy). While distance matters, under the conditions of this experiment, the results also indicate that the duration of the observing session does not present the same pattern for 1D and 2D. In addition to the length of the baseline and the duration of the observing session, positioning precision depends on several other factors, including the methodology and the software used for processing GPS data, in this case the LGO. Biases associated with meteorological effects (ionosphere and troposphere) also play an important role in the total error budget of positioning precision.

This work investigates the performance of commercial software LGO when processing baselines in static mode. The parameter to be tested is the time of observation needed to achieve a given accuracy (1D and 2D) for a set of ranges of baseline lengths. Four different scenarios were created, as follows:

- Broadcast ephemerides and Hopfield model (BH);
- Broadcast ephemerides and Computing the troposphere (BC);
- Precise ephemerides and Hopfield model (PH);
- Precise ephemerides and Computing the troposphere (PC).

Summarizing, the present work is comprised of introduction and conclusion sections, a section with background information, another describing the data and the methodology adopted and two sections containing specific tests and results.

2 GNSS Overview

In this section, we provide an introduction of GPS, the navigation system used in this research. As there are a number of relevant references available, e.g. [4, 9, 10], only a very brief discussion on the basics of the system will be given, with particular emphasis on the parts which are relevant to observation modeling of systematic biases and errors affecting GPS measurements. The various types of GPS observables of interest on baseline determination in SRP are also described, as are some of their possible combinations. The possible usefulness of Precise Ephemerides, in terms of the increased accuracy in long baselines, is also evaluated.

There are numerous sources of measurement errors that influence GPS performance. Both observables types, code and phase, are affected by many systematic biases and errors, different in their source and suitable method of treatment. The most important of these biases and errors are briefly reviewed here. The orbital errors and tropospheric effects will be discussed later with more detail.

Finally, because tropospheric delay is a dominant factor for the relative positioning accuracy in GPS/GNSS long baselines, as the LGO strategy using the "ionospherically-free observable" almost removes all first-order ionospheric biases, a description of the different strategies available to mitigate tropospheric biases is also provided. Differences between the Hopfield model and tropospheric computing techniques are highlighted.

2.1 Systematic Biases and Errors

There are numerous sources of measurement errors that influence GPS performance. Both observables types, code and phase, are affected by many systematic biases and errors, different in their source and suitable method of treatment. The most important of these biases and errors are briefly reviewed here. The orbital errors and tropospheric effects will be discussed later with more detail.

The sum of all systematic biases and errors contributing to the measurement error is referred to as a range bias. Bingley in [2] argues that this bias is caused by a physical phenomenon, as is the case, for example, in ionospheric or tropospheric delays, and error is the quantity remaining after the bias has been mitigated to some extent, which is the case, for example, for errors in broadcast ephemerides. According to the same author, the systematic biases and errors affecting GPS measurements can be grouped into three main categories: satellite related, atmospheric related, and station related.

2.2 Satellite Related Biases and Errors

Satellite related biases consist of biases of satellite ephemerides (orbital errors), satellite clock offsets and satellite antenna phase centers, as the selective availability (SA) was internationally terminated by the US Government in May 1, 2000.

The error in satellite coordinates is the difference between the predicted and the "true" satellite position. The predicted position is estimated by the Master Control Stations (MCS), using data collected by Master Stations (MS), and uploaded to the satellites, which in turn broadcast that information to users through the navigation message. The predicted satellite position is currently on the order of 1 m [2]. Besides broadcast ephemerides, precise ephemerides are available from IGS [7], providing an accuracy of 2.5 cm in their rapid and final format.

Although, precise as they are, satellites clocks are not perfect. The satellite clock error is defined as the difference between satellite clock time and true GPS time. The MCS computes and broadcasts to the users the parameters to correct the satellite clock error, according to the equation in [4, p. 52].

Because GPS orbit is calculated with respect to the satellite' center of mass but the observation refers to the antenna phase center (point of transmission), which are not coincident, the offset between these two centers has to be known. In addition to this, at the point of transmission, the electrical center is not the geometrical center. By applying Phase Center Offsets (PCO) and Phase Center Variations (PCV) corrections it is possible to relate the measurements consistently to the satellite's center of mass. In [11] the author states that in global networks absolute PCVs have to be taken into account due to the fact that the GPS satellites are normally seen at different elevations from the ends of a baseline.

2.3 Atmospheric Related Biases and Errors

Atmospheric biases are due to ionospheric and tropospheric delays. The ionospheric bias is caused by the propagation of the GPS signals in the ionosphere, which is the region of the atmosphere between about 50 and 1000 km above the Earth surface. Within this region ions and free electrons, originating in sun radiation, are present in quantities that affect the propagation of electromagnetic signals. In the GPS case, the code (pseudo-range) is delayed and the carrier phase is advanced. Because this is a dispersive medium at GPS frequencies, i.e., the propagation speed depends on the carrier frequency, resolution of ionospheric delays can be accomplished by using a dual-frequency receiver. However, according to Wells in [18], during a high solar activity cycle (e.g., solar maximum between 2011 and 2013) and in mid afternoons this technique may not be adequate for certain applications. The ionospheric delay depends on the Total Electron Content (TEC) along the signal path and on the frequency used [6]. The ionospheric bias may range from 5 (at night, the satellite at the zenith) to 150 m (at midday and the satellite at low elevation) [18].

The troposphere is the lowest atmosphere layer, from the Earth's surface to 50 km. The tropospheric delay is caused by the refraction of the GPS signal in this layer. This bias depends on parameters such as the temperature, humidity, and pressure. It varies with the height of the station. Unlike the ionosphere, this is a non-dispersive medium for GPS frequencies, that is, the delay is independent from the carrier frequency, so that dual-frequency receivers cannot be used to eliminate it. In GPS case both pseudo-range and carrier phase will experiment the same delay. Usually, the tropospheric bias is broken in two components [5]:

A hydrostatic component, including about 80–90% of the error and highly predictable, according to atmospheric pressure and temperature, and satellite' elevation angle.
A wet component, including about 10–20% of the error, is more difficult to predict, due to variations of the partial water vapor on the atmosphere.

A number of studies have been performed to create tropospheric models to mitigate the influence of this bias, among them the Hopfield model, used in this research. The hydrostatic, or dry, component can be precisely described by these models with an accuracy of $\pm 1\%$, while the wet component can be modeled by surface weather data to within 3–4 cm [18]. Besides using models, usually based on meteorological parameters, other approaches to determine the wet component include direct measurement with water vapor radiometers and the use of a station-dependent zenith scale factor for each satellite pass [12].

2.4 Station Related Biases and Errors

Station related biases and errors to be considered include those related to the equipment (receiver clock offset and receiver antenna phase centers) and location of the station (multipath and geophysical phenomena).

The receiver clock error is the difference between the time maintained by the receiver clock and the reference GPS time. Differencing observations between satellites can eliminate the receiver dock error. This procedure is based on the assumption that the clock bias is independent at each measurement time. "In the case of relative positioning, such as static positioning using carrier phase, the receiver clock offsets are eliminated, with the assumption that a receiver appears to make observations to all satellites at the same time" [2].

Receiver Antenna Phase centers biases are the compound of PCOs and PCVs. PCO is the offset between the point of reception (mean phase center), at the antenna, and the physical Antenna Reference Point (ARP). This offset is constant, for each antenna and frequency, whereas PCVs vary depending on the direction (azimuth and elevation of the satellite) and frequency of the transmitting signal. According to [2], if not accounted for, the bias due to these variations can reach several centimeters in the observed carrier phase for some types of antenna. For high accurate applications, in static positioning using carrier phase, these biases have to be mitigated. Some procedures should then be followed, in order to eliminate or reduce this type of bias, such as the use of similar antennas (choke ring, if possible), directed north on both sides of the baseline. Nevertheless, even in the case of similar antennas being used, models for receiver antenna phase centers (particularly PCVs) must be applied for baselines greater than 100 km [2].

Multipath is the phenomena whereby a signal arrives at a receiver from more than one path because of the reflections during the signal propagation. As the bias due to multipath is wavelength dependent, code and carrier phase are affected in different ways. Pseudo-range multipath can reach up to one chip length of the PRN codes (293 m for C/A code and 29.3 m for P code). Carrier phase measurements are not free from multipath either, although the effect is about two orders of magnitude smaller than in pseudo-ranges (e.g. 5 cm for L1), it contributes to the phase measurement noise [3]. Furthermore, because multipath affects L1 and L2 signals differently, this can cause problems during cycle slip detection and correction [2]. For static positioning using carrier phase, as demonstrated in [13], multipath signature can be detected through the analysis of strong correlation present in the adjustment' residuals of two consecutive sidereal days, due to the geometry repetition of satellite-antenna-reflector.

Applying models for geophysical phenomena, such as the solid earth tides (SET) or ocean tide loading (OTL) is important when striving for centimeter-level accuracies using carrier phase relative positioning over long baselines lengths.

3 Data and Methodology

OS' active stations were used to investigate the relation between time of observation and length of the baseline. A total of 105 baselines were processed using LGO, separated into six range groups ($R_i,\ i = 1, \ldots, 6$) according to their lengths in kilometers:

- $R_1 = [000 - 100] \rightarrow$(5 baselines)
- $R_2 = [100 - 200] \rightarrow$(14 baselines)
- $R_3 = [200 - 300] \rightarrow$(27 baselines)
- $R_4 = [300 - 400] \rightarrow$(29 baselines)
- $R_5 = [400 - 500] \rightarrow$(14 baselines)
- $R_6 = [500 - 900] \rightarrow$ (16 baselines)

All the stations are permanent stations of clear sky visibility and with low multipath conditions. The quality of the data is therefore expectantly high. Day 13/06/2013 of receiver independent exchange (RINEX) data of GPS week 1744 was downloaded from the data archive of the active GPS network of Ordnance Survey (OS Net) for each of the 106 stations [15]. These RINEX data include phase measurement of the carrier waves L_1 and L_2, P_1, P_2 and C/A pseudo-range code at a 30 s interval.

For this experiment, 24 h of dual-frequency GPS carrier phase observations for each of 105 baselines formed by ABEP, chosen as reference station, and all the other active stations, designated as rover, from OS Network were used. These 105 baselines range in length from 61 to 898 km and correspond to all active stations considered "healthy" on 13/06/2013. The data for each baseline comprised the same 24-h session that was further subdivided into periods of time of 1, 2, 3, 4, 6, 8, 12, and 24 h as follows, where the two first digits represent the beginning of the observation period and the last two the end:

- 1 h periods: [0001], [0607], [1213], [1819];
- 2 h periods: [0002], [0608], [1214], [1820];
- 3 h periods: [0003], [0609], [1215], [1821];
- 4 h periods: [0004], [0408], [0812], [1216], [1620], [2024];
- 6 h periods: [0006], [0612], [1218], [1824];
- 8 h periods: [0008], [0816], [1624];
- 12 h periods: [0012], [1224];
- 24 h period: [0024].

The division of time in this way was done in order to evaluate the performance of the software for different lengths of observation time.

The criteria followed to select the reference station were primarily based on location. Thus ABEP, on the west coast of England, was chosen, because of its high altitude and location, providing a well-distributed range of radial vectors to all the other active stations, either in latitude and longitude. Its 3D positional coordinates were fixed to the official values adopted by OS.

In order to evaluate at what range of baseline lengths the use of precise ephemerides becomes worthwhile, both results using broadcast and precise ephemerides are presented as well. The corresponding SP3 files were downloaded from the data archive of IGS [8]. These include precise ephemerides at a sampling interval of 15 min and the high-rate precise satellite clocks with a sampling of 30 s.

Hence, the four different scenarios can be compared as follows:

- Direct comparison of the results obtained using the broadcast ephemerides and the precise ephemerides (BH versus PH and BC versus PC);
- Direct comparison of the results obtained using Hopfield model and computing the troposphere (BH versus BC and PH versus PC).

At starting points 1D, 2D, and 3D accuracy criteria were established for each baseline, as only successful processed baselines are of interest for this research. The chosen values were set to 1D and 2D accuracies to be better than 3 cm and 3D better than 4.5 cm. These are realistic values, as the OS active stations have 1D accuracy of about 2 cm in magnitude and close to 1 cm in 2D. Therefore, assuming the 3 cm as 1D and 2D threshold seems to be reasonable due to the fact that this tolerance allows for the "absorption" of errors inherent to the coordinates of the stations. Despite how perfectly the baseline was calculated an error of up to 4 cm in height and 2 cm in plan could arise due to the uncertainty associated with the coordinates.

The published coordinates of each of these stations (in Cartesian format on the header of the corresponding RINEX file) are assumed as "true" and used to compute the errors (1D, 2D, and 3D) in the solutions processed by LGO.

Figure 1 presents the percentage of successful baselines per range in 1D (black) and 2D (blue). There is a clear trend for fewer successful baselines as the length increases, either in 1D or 2D.

In Fig. 2, where the results are organized in a detailed form (each rectangle contains the eight box-plots, relatively to four strategies for 1D and 2D per range length), the trend is evident in all cases. In Table 1 we present the percentage of successful baselines per range in 1D (black) and 2D (blue) and per strategy. It is also easily to detect such behavior. That is a clear trend for a lower quantity of successful baselines as the length increases, regardless of the strategy adopted, either in 1D or 2D.

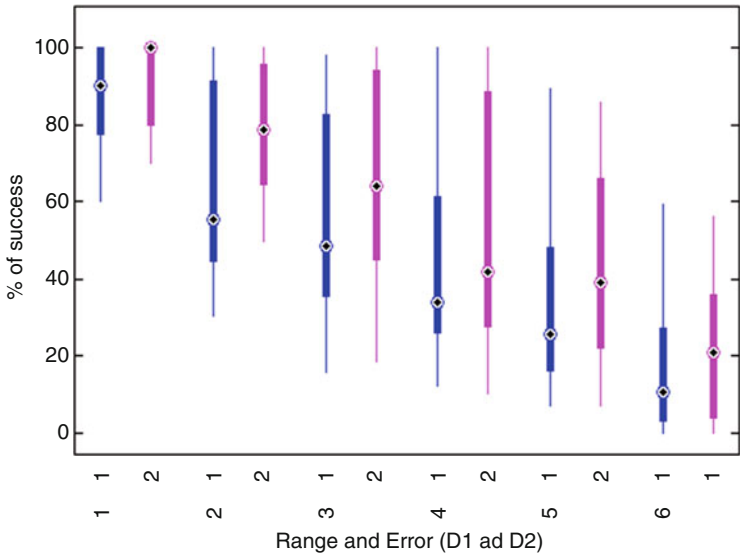


Fig. 1 Percentage of successful baselines distinct ranges (R_i, $i = 1, \ldots, 6$) in 1D (in blue) and 2D (in magenta)

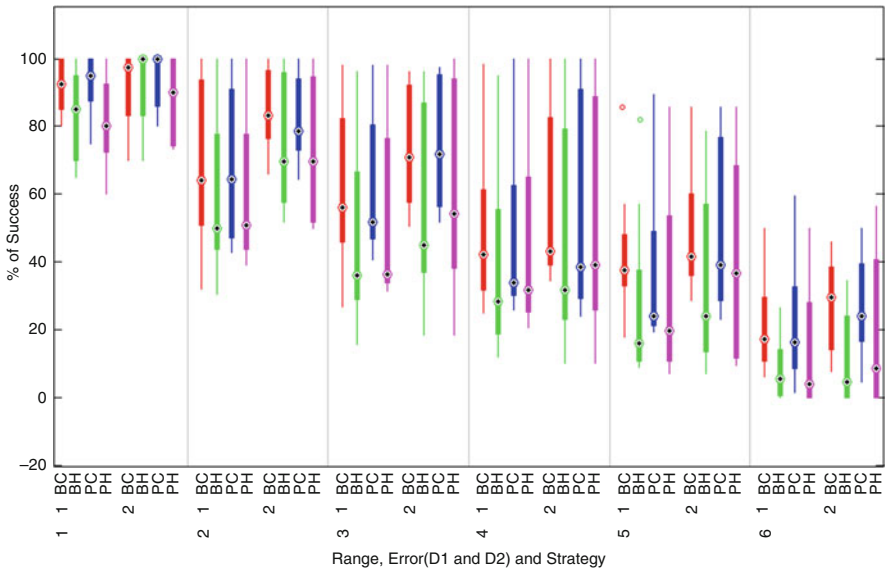


Fig. 2 Percentage of successful baselines distinct ranges (R_i, $i = 1, \ldots, 6$) in 1D (rectangle left side) and 2D (rectangle right side) for strategies BH (in red), BC (in green), PH (in blue), and PC (in magenta)

Table 1 Averaged percentage of successful baselines in 1D and 2D for total (T) and distinct ranges ($R_i,\ i = 1, \ldots, 6$)

Strategies	Processed	Pass	T	R_1	R_2	R_3	R_4	R_5	R_6
BH	2940	880	29.932	77.143	53.571	37.434	27.094	13.520	1.339
	2940	1756	59.728	97.145	83.418	71.296	60.099	44.643	20.537
BC	2940	1379	46.905	82.857	70.663	57.804	38.424	34.949	22.321
	2940	1875	63.776	98.571	82.908	76.190	65.394	49.235	25.000
PH	2940	840	28.571	72.857	50.510	37.302	25.739	11.990	0.223
	2940	2055	69.898	97.143	84.184	81.746	73.153	59.694	32.143
PC	2940	1206	41.020	84.286	66.582	54.365	28.818	27.296	16.518
	2940	2087	70.986	100.000	84.439	83.466	73.522	59.184	35.045

Strategies BH, BC, PH, and PC. Percentage of success for 1D in black; for 2D in blue
Four strategies considering 1D and 2D

In a preliminary approach, it was found that the different ranges led to significantly different results. Were used parametric tests to compare proportions (t-test). With some small samples in certain ranges, were also applied some nonparametric tests that allow us to compare location measures, or a chi-square test and a Kruskal-Wallis to evaluate if the proportions of success are the same in the different ranges; a chi-square independence test was also used to evaluate the relation between the proportion of success and range. In Table 2 are the p-values obtained when the differences of the proportions of success for different ranges and strategies are tested.

In general, different strategies conduce to similar results: almost all comparisons have the same conclusion—the proportions of success in different ranges are not equal except when the ranges are sequential of each other. Also were performed similar tests comparing different strategies considering the same range. Generally, the proportions of success for the same range, but with different strategies conducted to significant tests, meaning that there is statistical evidence of different proportions of success per different strategies for same range. These conclusions are visible in Fig. 2.

We also performed an analysis of variance with four factors (parametric and non-parametric approach). The results of such analysis are similar for both cases: generally each factor is significant, meaning that the probability for success is not equal for each level of the considered factor. The intersection of each factor with the other considering secondary and third level intersections was not significant. The resume of that analysis can be found in Fig. 3. In this study we have merged the classes with smaller exposure time. The level 3 of factor duration means exposure time until 3 h.

We also performed the Scheffe's S procedure, derived from F Distribution. This technique provides a simultaneous confidence level for comparisons for all linear combinations of means, namely for comparisons of simple differences of pairs. Figure 4 illustrates such comparisons for all levels of each factor. The conclusions are similar.

Table 2 Tests for difference of success probability for distinct ranges

Strategies	Ranges	R_2	R_3	R_4	R_5	R_6
BH	R_1	0.3204	0.0679	0.0204	0.0073	0.0008
		0.1129	0.3127	0.6257	0.9886	0.5567
	R_2		0.3271	0.0988	0.0200	0.0007
			0.1115	0.0096	0.0126	0.0004
	R_3			0.4097	0.0745	0.0006
				0.2769	0.1261	0.0099
	R_4				0.2767	0.0052
					0.4756	0.0827
	R_5					0.2140
						0.3991
BC	R_1	0.5652	0.2053	0.0267	0.0367	0.0065
		0.1864	0.0291	0.0029	0.0032	0.0000
	R_2		0.4100	0.0395	0.0530	0.0054
			0.6076	0.1981	0.0545	0.0005
	R_3			0.1452	0.1585	0.0158
				0.3743	0.0934	0.0005
	R_4				0.8250	0.2491
					0.3190	0.0060
	R_5					0.4492
						0.1698
PH	R_1	0.3640	0.1159	0.0357	0.0122	0.0017
		0.3059	0.1534	0.0383	0.0237	0.002
	R_2		0.4222	0.1208	0.0229	0.0008
			0.8434	0.3924	0.1420	0.0019
	R_3			0.3533	0.0538	0.0003
				0.4418	0.1514	0.0009
	R_4				0.2541	0.0033
					0.3896	0.0063
	R_5					0.1900
						0.1278
PC	R_1	0.4018	0.1237	0.0048	0.0116	0.0018
		0.1266	0.0278	0.0028	0.0064	0.0000
	R_2		0.4451	0.0168	0.0320	0.0034
			0.9360	0.3946	0.1339	0.0033
	R_3			0.0502	0.0844	0.0071
				0.3645	0.1125	0.0012
	R_4				0.9174	0.3316
					0.3598	0.0110
	R_5					0.4812
						0.1845

Strategies BH, BC, PH, and PC. P-values for 1D in black; P-values for 2D in blue
Four strategies considering 1D and 2D

Analysis of Variance

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Range	196068.7	5	39213.7	102.22	0
Strategy	9323.3	3	3107.8	8.1	0
PrecisionD1D2	2649.9	1	2649.9	6.91	0.0089
Duration	37693.4	5	7538.7	19.65	0
Error	141552.6	369	383.6		

Constrained (Type III) sums of squares.

Fig. 3 Analysis of Variance (four factors)—Percentage of successful baselines distinct ranges (R_i, $i = 1, \ldots, 6$); Duration of Exposure; Strategies BH, BC, PH, PC; Precision (1D and 2D)

4 Discussion of Results, Conclusions, and on Going Work

This work studies the relation for single baselines between lengths ranges and between the different ranges and the observation time required to obtain high-accurate positioning, using commercial software LGO. A brief analysis for different amplitudes of time interval of exposure, considering the four strategies is reproduced partially in this paper. The results are valid for this specific software and under the conditions of the experiments. Four different strategies were established and evaluated through the processing of a total of 11,760 baselines. The data processing and testing used several options concerning the best thresholds for accuracy. The LGO results were compared with the published coordinates by Ordnance Survey and the baselines passing the accuracy criteria were isolated. The division of time in this way was done in order to evaluate the performance of the software for different lengths of observation time. It revealed that the largest amplitude of time exposure interval, the bigger percentage of success.

Clearly was shown the dependence of success in 1D regarding the baseline length. No matter the strategy adopted, broadcast or precise ephemerides, Hopfield model or computing the troposphere, the rate of successful baselines processed decreases as the baseline length increases, following a linear trend. Generally, when looking at the range 1 to range 3 baseline length classes, BC performance is slightly better than PC but it is absolutely certain that computing the troposphere leads to higher rates of success for these three classes (BC vs BH and PC vs PH). Using LGO to process individual longer baselines (range 4 to range 6 classes), without any kind of redundancy, represents a risk, as the percentage of success is always less than 50 %.

A preliminary experiment shows that to obtain high accurate relative positioning 3D coordinates for long baselines in static mode with LGO at least 4 h of observation are recommended. Therefore, it is important to give, in a short time, a special focus to periods of this magnitude and over. These cover the whole day in nonoverlapping periods, whereas for the 1, 2 and 3 h intervals only representative samples were chosen. It is still need to analyze the results from similar lengths but at different times of the day experiencing diverse atmospheric conditions. Other tests and

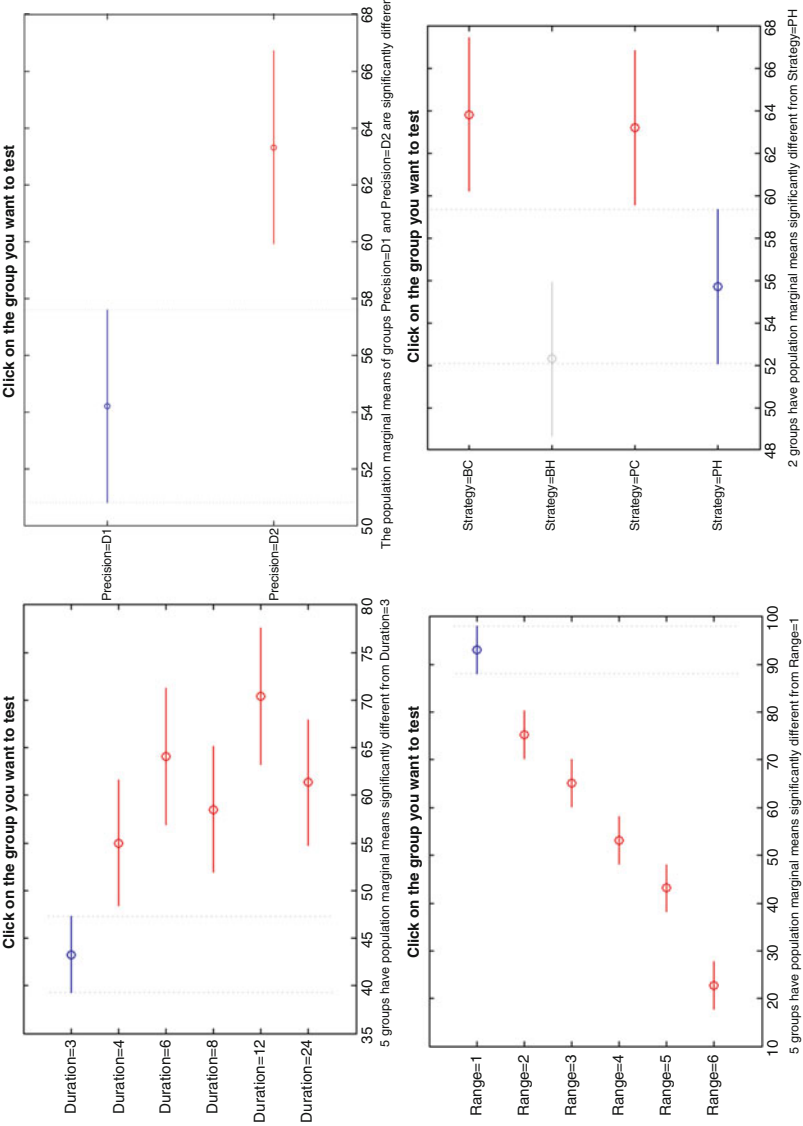


Fig. 4 Scheffe simultaneous mean percentage of successful baselines confidence intervals for the each factor levels—Duration of Exposure (top-left); Precision 1D and 2D (top-right); distinct ranges R_i, $i = 1, \ldots, 6$ (bottom-left); Strategies BH, BC, PH, PC (bottom-right)

techniques, inquiring about the significance of the hour of the day, the amplitude of time interval of exposure, considering the four strategies.

An Analysis of Variance with several factors [16] (range, strategies, amplitude of interval time of exposure) was applied. Another possible approach is to model the data by General Linear Models [1, 17]. Such statistical approach details will be found in a future continuation of this manuscript.

Acknowledgements This work was supported by Portuguese funds through the *Center for Computational and Stochastic Mathematics* (CEMAT), *The Portuguese Foundation for Science and Technology* (FCT), University of Lisbon, Portugal, project UID/Multi/04621/2013, and *Center of Naval Research* (CINAV), Naval Academy, Portuguese Navy, Portugal.

References

1. Anderson, T.W.: An Introduction to Multivariate Analysis. Wiley, New York (2003)
2. Bingley, R.M.: GNSS principles and observables: systematic biases and errors. Short Course, University of Nottingham, Nottingham Geospatial Institute (2013)
3. Georgiadou, Y., Kleusberg, A.: Multipath effects in static and kinematic GPS surveying. In: Global Positioning System: An Overview. International Association of Geodesy Symposia, vol. 102, pp. 82–89. Springer, Heidelberg (2002)
4. Hofmann-Wellenhof, B., Lichtenegger, H., Wasle, H.: GNSS-Global Navigation Satellite Systems: GPS, GLONASS, Galileo, and More. Springer Verlag-Wien, New York (2008)
5. Hopfield, H.S.: Tropospheric effect on electromagnetically measured range: prediction from surface weather data. Radio Sci. **6**(3), 357–367 (1971)
6. Hoque, M.M., Jakowski, N.: Ionospheric propagation effects on GNSS signals and new correction approaches. In: Shuanggen, J. (ed.) Global Navigation Satellite Systems: Signal, Theory and Applications, pp. 381–405. InTech, Rijeka (2012). https://doi.org/10.5772/30090
7. IGS: International GNSS Service. http://igscb.jpl.nasa.gov/ (2016). Accessed 19 Aug 2016
8. IGS: International GNSS Service. http://igscb.jpl.nasa.gov/components/prods_cb.html (2016). Accessed 19 Aug 2016
9. Kaplan, E.D., Hegarty, C.J.: Understanding GPS: Principles and Applications. Artech House, Norwood (2006)
10. Leick, A.: GPS Satellite Surveying. Wiley, New Jersey (2004)
11. Mader, G.L.: GPS antenna calibration at the national geodetic survey. GPS Solutions **3**(1), 50–58 (1999)
12. Mendes, V.B.: Modeling the neutral-atmosphere propagation delay in radiometric space techniques. PhD thesis, University of New Brunswick (1999)
13. Meng, X.: Modeling the neutral-atmosphere propagation delay in radiometric space techniques. PhD thesis, University of Nottingham (2002)
14. OS Net Business and Government. Ordnance survey. http://www.ordnancesurvey.co.uk/oswebsite/products/os-net/index.html (2016). Accessed 19 Aug 2016
15. OS Net Business and Government. Ordnance survey. http://www.ordnancesurvey.co.uk/gps/os-net-rinex-data/ (2016). Accessed 19 Aug 2016
16. Tamhane, A.C., Dunlop, D.D.: Statistics and Data Analysis: From Elementary to Intermediate. Prentice Hall, New Jersey (2000)
17. Turkman, M.A., Silva, G.: Modelos Lineares Generalizados da Teoria a Prática. Sociedade Portuguesa de Estatística, Lisboa (2000)
18. Wells, D.E., et al.: Guide to GPS Positioning. Canadian GPS Associates, Fredericton. http://plan.geomatics.ucalgary.ca/papers/guide_to_gps_positioning_book.pdf (1986). Accessed 1 Oct 2016

Multivariate Generalized Birnbaum-Saunders Models Applied to Case Studies in Bio-Engineering and Industry

Víctor Leiva and Carolina Marchant

Abstract Birnbaum-Saunders models are receiving considerable attention in the literature. Multivariate regression models are a useful tool in the multivariate analysis, which takes into account the correlation between variables. Diagnostic analysis is an important aspect to be considered in the statistical modeling. In this work, we formulate a statistical methodology based on multivariate generalized Birnbaum-Saunders regression models and their diagnostics. We implement the obtained results in the `R` software, which are illustrated with two real-world multivariate data sets related to case studies in bio-engineering and industry to show their potential applications.

1 Introduction

The univariate Birnbaum-Saunders (BS) distribution is unimodal, positively skewed, and has two parameters that modify its shape and scale. The BS distribution has been widely studied because of its good properties, its relation with the normal distribution, and its diverse applications. A logarithmic version of the BS (log-BS) distribution is often used to formulate regression models. For more details of the BS distribution, see the recent book by Leiva [11]. The family of elliptically contoured (EC) distributions has the Laplace, logistic, normal and Student-t (hereafter called "t") cases as some of its members. The interested reader is referred to [4–7, 10, 12] for more details on EC distributions and their modeling. EC distributions provide a flexible framework to model extreme cases which are often found in data following distributions of heavy-tails. Particularly, some authors [10] suggested the t distribution as an alternative to the normal distribution, permitting atypical

V. Leiva (✉)
School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
e-mail: victor.leiva@pucv.cl; www.victorleiva.cl

C. Marchant
Faculty of Basic Sciences, Universidad Católica del Maule, Talca, Chile

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_22

data to be accommodated. Robust estimation of parameters is reached when the t distribution is used. This distribution has shown to provide robust estimation in several regression models. Indeed, the t distribution includes an additional parameter which allows the kurtosis to be modeled in a flexible way. This flexibility of the t distribution permits outliers to be accommodated suitably, which is the main difference between the t and normal distributions, because in the normal distribution its kurtosis is fixed. Therefore, the maximum likelihood estimators under normality are sensitive to outliers, even when the number of them is small, for example, less than 1% [10]. Note that [3] proposed a generalized version of the BS (GBS) distribution based on the EC family of distributions. Univariate log-linear regression models for GBS distributions were studied by [1, 14, 19]. For multivariate versions of GBS distributions and their modeling, diagnostics and applications based on a logarithmic version of the GBS (log-GBS) distribution, the interested reader is referred to [8, 15–17].

The objectives of this work are (1) to derive a methodology based on multivariate GBS regression models and their diagnostics, as well as (2) to illustrate this methodology with two real-world multivariate data sets related to bio-engineering and industry to show their potential applications. The numerical results were obtained with a programming code implemented in the R software (www.R-project.org).

The contents of the work are organized as follows. In Sect. 2, we formulate the multivariate GBS log-linear regression models. Also, we discuss the maximum likelihood method for estimating the corresponding parameters. In Sect. 3, we derive diagnostics for multivariate GBS log-linear regression models considering local influence, as well as global influence by the Mahalanobis distance (MD). In Sect. 4, we summarize the proposed methodology based on multivariate GBS regression models. In Sects. 5 and 6, we conduct two case studies with real-world multivariate data sets. Finally, in Sect. 7, we discuss some conclusions and future research related to the topic of this work.

2 Multivariate GBS Regression Models

Consider the multivariate GBS log-linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{Y} = (Y_{ij}) \in \mathbb{R}^{n\times m}$ is the log-response matrix, and $\boldsymbol{X} = (x_{is}) \in \mathbb{R}^{n\times p}$ represents the model matrix of rank p containing the values of p covariates. Here, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linked by a coefficient matrix $\boldsymbol{\beta} = (\beta_{sj}) = \left(\underline{\beta}_1, \ldots, \underline{\beta}_m\right) \in \mathbb{R}^{p\times m}$ to be estimated, whereas $\boldsymbol{E} = (\varepsilon_{ij}) \in \mathbb{R}^{n\times m}$ is the error matrix. In addition, in the model given in (1), let $\underline{Y}_i^\top$, $\underline{x}_i^\top$, and $\underline{\varepsilon}_i^\top$ be the ith rows of $\boldsymbol{Y}$, $\boldsymbol{X}$, and $\boldsymbol{E}$, respectively. Thus, we can write

$$\underline{Y}_i = \underline{\mu}_i + \underline{\varepsilon}_i = \boldsymbol{\beta}^\top \underline{x}_i + \underline{\varepsilon}_i, \quad i = 1, \ldots, n,$$

where $\underline{\varepsilon}_1, \ldots, \underline{\varepsilon}_n$ are independently and identically log-GBS$_m\big(\alpha \underline{1}_{m\times 1}, \underline{0}_{m\times 1}, \boldsymbol{\Psi}, g^{(m)}\big)$ distributed, with $\underline{1}_{m\times 1}$ being a vector of ones for the indicated dimension, $\underline{0}_{m\times 1}$ being a vector of zeros, $\boldsymbol{\Psi} = (\varrho_{rs}) \in \mathbb{R}^{m\times m}$ being the correlation matrix, and $g^{(m)}$ being the multivariate EC density generator. For details about the m-variate GBS and log-GBS distributions, denoted by GBS$_m$ and log-GBS$_m$ respectively, see [15, 17] and the references therein.

Let $\boldsymbol{Y} = (\underline{Y}_1, \ldots, \underline{Y}_n)^\top$ be a sample from a multivariate log-GBS distribution with $\mathrm{E}[\underline{Y}_i] = \boldsymbol{\beta}^\top \underline{x}_i$ (multivariate GBS log-linear regression structure), and $\boldsymbol{y} = \big(\underline{y}_1, \ldots, \underline{y}_n\big)^\top$ their observations. Then, the log-likelihood function for $\underline{\theta} = \big(\alpha, \mathrm{vec}(\boldsymbol{\beta})^\top, \mathrm{svec}(\boldsymbol{\Psi})^\top\big)^\top$, with "vec" and "svec" denoting the vectorization and vectorization of a symmetric matrix, respectively, is given by

$$\ell(\underline{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i(\underline{\theta}) = \sum_{i=1}^{n} \log\left(f_{\mathrm{EC}_m}\left(\underline{\phi}_i; \boldsymbol{\Psi}, g^{(m)}\right)\right) + \sum_{i=1}^{n}\sum_{j=1}^{m} \log(\xi_{ij}), \tag{2}$$

where f_{EC_m} is the probability density function of an m-variate EC distribution [6] and $\underline{\phi}_i = (\phi_{i1}, \ldots, \phi_{im})^\top$, with

$$\phi_{ij} = \frac{2}{\alpha}\sinh\left(\frac{y_{ij} - \mu_{ij}}{2}\right), \quad \xi_{ij} = \frac{2}{\alpha}\cosh\left(\frac{y_{ij} - \mu_{ij}}{2}\right),$$

and $\mu_{ij} = \underline{\beta}_j^\top x_i$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$. The log-likelihood function defined in (2) can be specified for different EC distributions based on the probability density functions defined in Table 1. We compare the BS and BS-t models, both particular cases of GBS models. Results for other members of the EC family, as the BS-Laplace, BS-Cauchy, BS-power-normal, and BS-logistic models, are directly obtained from the methodology proposed in this work. We focus on the BS-t model due to its interesting robustness property in the parameters estimation.

3 Diagnostics

Global influence in the multivariate regression model defined in (1) can be assessed by the MD expressed as

$$\mathrm{MD}_i = \underline{\phi}_i^\top \boldsymbol{\Psi}^{-1} \underline{\phi}_i, \quad i = 1, \ldots, n, \tag{3}$$

where $\underline{\phi}_i$ is given in (2). Based on [9, 10]: (1) $\mathrm{MD}_i \sim \chi^2(m)$, that is, the MD defined in (3) follows the central χ^2 distribution with m degrees of freedom, when $g^{(m)}$ is the multivariate normal density generator, and (2) $\mathrm{MD}_i/m \sim \mathscr{F}(m, \nu)$, that is, it is related to the central $\mathscr{F}$ distribution with m degrees of freedom in the numerator and ν in the denominator, if $g^{(m)}$ is the multivariate t density generator, for $i = 1, \ldots, n$.

Table 1 Normalizing constant ($c^{(m)}$), density generator ($g^{(m)}(u)$) for $u > 0$ and probability density function ($f_{\mathrm{EC}_m(u)} = c^{(m)}|\boldsymbol{\Psi}|^{-\frac{1}{2}}\, g^{(m)}(u)$), for the indicated distribution

Distribution	$c^{(m)}$	$g^{(m)}(u)$	$f_{\mathrm{EC}_m(u)}$
Normal	$(2\pi)^{-\frac{m}{2}}$	$\exp\left(\frac{-u}{2}\right)$	$(2\pi)^{-\frac{m}{2}}\,\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}\,\exp\left(\frac{-u}{2}\right)$
t	$\dfrac{\Gamma\left(\frac{\nu+m}{2}\right)}{(\nu\pi)^{\frac{m}{2}}\Gamma\left(\frac{\nu}{2}\right)}$	$\left(1+\dfrac{u}{\nu}\right)^{-\frac{(\nu+m)}{2}}$	$\dfrac{\Gamma\left(\frac{\nu+m}{2}\right)\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}}{(\nu\pi)^{\frac{m}{2}}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\dfrac{u}{\nu}\right)^{-\frac{(\nu+m)}{2}}$ $\nu > 0$
Symmetric Pearson type VII	$\dfrac{\Gamma(\xi)}{\Gamma\left(\frac{2\xi-m}{2}\right)(\theta\pi)^{m/2}}$	$\left(1+\dfrac{u}{\theta}\right)^{-\xi}$	$\dfrac{\Gamma(\xi)\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}}{\Gamma\left(\frac{2\xi-m}{2}\right)(\theta\pi)^{\frac{m}{2}}}\left(1+\dfrac{u}{\theta}\right)^{-\xi}$ $\xi > m/2;\ \theta > 0$
Symmetric Kotz type	$\dfrac{\delta\Gamma\left(\frac{m}{2}\right)\lambda^{\frac{(2\eta+m-2)}{2\delta}}}{\pi^{\frac{m}{2}}\Gamma\left(\frac{2\eta+m-2}{2\delta}\right)}$	$u^{\eta-1}\exp(-\lambda\, u^{\delta})$	$\dfrac{\delta\Gamma\left(\frac{m}{2}\right)\lambda^{\frac{(2\eta+m-2)}{2\delta}}u^{\eta-1}\exp(-\lambda\, u^{\delta})}{\pi^{\frac{m}{2}}\Gamma\left(\frac{2\eta+m-2}{2\delta}\right)\lvert\boldsymbol{\Psi}\rvert^{\frac{1}{2}}}$ $\delta > 0;\ \lambda > 0;\ 2\eta + m > 2$
Laplace	$\dfrac{1}{2^{\frac{(m-2)}{2}}\pi^{\frac{m}{2}}\sigma^m\Gamma(\frac{m}{2})}$	$K\left(\frac{(2u)^{\frac{1}{2}}}{\sigma}\right)$	$\dfrac{\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}K\left(\frac{(2u)^{\frac{1}{2}}}{\sigma}\right)}{2^{\frac{m}{2}-1}\pi^{\frac{m}{2}}\sigma^m\Gamma(\frac{m}{2})}$ $\sigma > 0$
Symmetric logistic	$\dfrac{\pi^{\frac{m}{2}}\kappa}{\Gamma(\frac{m}{2})}$	$\dfrac{\exp(-u)}{(1+\exp(-u))^2}$	$\dfrac{\pi^{\frac{m}{2}}\kappa\lvert\boldsymbol{\Psi}\rvert^{-\frac{1}{2}}}{\Gamma(\frac{m}{2})}\dfrac{\exp(-u)}{(1+\exp(-u))^2}$

Where Γ and K denote, respectively, the gamma function and the modified Bessel function of the third kind, whereas the constant $\kappa = \int_0^\infty z^{\frac{m}{2}-1}(\exp(-z)/(1+\exp(-z))^2)\,\mathrm{d}z$

Consider the log-likelihood function $\ell(\underline{\theta})$ for the parameter $\underline{\theta}$ of the model defined in (1), which we call the non-perturbed model. In addition, consider a perturbation vector $\underline{w} \in \mathbb{R}^q$ in the model, for q being a generic value which can correspond to the sample size n or the number of responses m, and $\underline{w} \in \Omega$, with $\Omega \subset \mathbb{R}^q$ being a set of perturbations. Then, $\ell(\underline{\theta}|\underline{w})$ is the log-likelihood function of the perturbed model, with $\widehat{\underline{\theta}}_{\underline{w}}$ being the maximum likelihood estimate of $\underline{\theta}$ obtained from $\ell(\underline{\theta}|\underline{w})$. Furthermore, let $\underline{w}_0 \in \Omega \subset \mathbb{R}^q$ be a non-perturbation vector with $\underline{w}_0 = \underline{0}^\top_{q\times 1}$, or $\underline{w}_0 = \underline{1}^\top_{q\times 1}$, or a possible third choice, so that $\ell(\underline{\theta}) = \ell(\underline{\theta}|\underline{w}_0)$. Assuming that $\ell(\underline{\theta}|\underline{w})$ is a twice continuously differentiable function in a neighborhood of $\left(\widehat{\underline{\theta}}, \underline{w}_0\right)$, we compare the maximum likelihood estimates $\widehat{\underline{\theta}}$ and $\widehat{\underline{\theta}}_{\underline{w}}$ by the local influence method to investigate how inference is affected by the corresponding perturbation. The likelihood distance (LD) is given by

$$\mathrm{LD}(\underline{w}) = 2\left(\ell\left(\widehat{\underline{\theta}}\right) - \ell\left(\widehat{\underline{\theta}}_{\underline{w}}\right)\right), \tag{4}$$

which is used to detect the influence of $\underline{w}$. Large values of $\mathrm{LD}(\underline{w})$ in (4) indicate that $\widehat{\underline{\theta}}$ and $\widehat{\underline{\theta}}_{\underline{w}}$ differ considerably in relation to the contours of the non-perturbed log-likelihood function $\ell(\underline{\theta})$. We study the local behavior of the influence plot $a(\underline{w}) = (\underline{w}^\top, \mathrm{LD}(\underline{w}))^\top$ around $\underline{w}_0$. The direction in which the LD locally changes most rapidly is evaluated, that is, the maximum curvature of the surface $a(\underline{w})$. For $\mathrm{LD}(\underline{w})$ given in (4), this maximum curvature is read to be

$$C_{\max} = \max_{||\underline{d}||=1} C_{\underline{d}}, \tag{5}$$

where $C_{\underline{d}} = 2|\underline{d}^\top \boldsymbol{F} \underline{d}|$, with the matrix $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ and $\underline{d}$ being the unit-length direction vector. To compute $C_{\max}$ given in (5) and the corresponding direction vector $\underline{d}_{\max}$, we must calculate

$$\boldsymbol{F} = -\boldsymbol{\Delta}\big(\widehat{\underline{\theta}}, \underline{w}_0\big)^\top \ddot{\boldsymbol{\ell}}\big(\widehat{\underline{\theta}}\big)^{-1} \boldsymbol{\Delta}\big(\widehat{\underline{\theta}}, \underline{w}_0\big), \tag{6}$$

where $-\ddot{\boldsymbol{\ell}}(\widehat{\underline{\theta}}) \in \mathbb{R}^{p^*\times p^*}$ is the observed information matrix for the non-perturbed model and $\boldsymbol{\Delta}(\underline{\theta}, \underline{w}) \in \mathbb{R}^{p^*\times n}$ is a matrix partitioned accordingly for the perturbed model obtained from (1), called perturbation matrix, with elements defined as

$$\Delta_{ij} = \frac{\partial^2 \ell(\underline{\theta}|\underline{w})}{\partial \underline{\theta}_i \partial \underline{w}_j}, \quad i = 1, \ldots, n, \;\; j = 1, \ldots, p^*,$$

evaluated at $\underline{\theta} = \widehat{\underline{\theta}}$ and $\underline{w} = \underline{w}_0$, where $p^* = p\,m+l+1$, with $l = m(m-1)/2$. Then, $\underline{d}_{\max}$ is a unit-length eigenvector associated with the largest absolute eigenvalue $C_{\max}$ given in (5). If the absolute value of $\underline{d}_{\max_i}$ is large, it indicates that the case i is potentially influential. In addition to $\underline{d}_{\max_i}$, another direction of interest is $\underline{d}_i = \underline{e}_{in}$, which is related to the direction of the case i, where $\underline{e}_{in} \in \mathbb{R}^n$ is a vector of zeros and a one at the ith position. Thus, the normal curvature is $C_i(\underline{\theta}) = 2|f_{ii}|$, for $i = 1, \ldots, n$, where f_{ii} is the ith diagonal element of $\boldsymbol{F}$ given in (6), evaluated at $\underline{\theta} = \widehat{\underline{\theta}}$. The case i is considered as potentially influential if $C_i(\widehat{\underline{\theta}}) > 2\overline{C}(\widehat{\underline{\theta}})$, for $i = 1, \ldots, n$, where

$$\overline{C}\big(\widehat{\underline{\theta}}\big) = \frac{1}{n}\sum_{i=1}^{n} C_i\big(\widehat{\underline{\theta}}\big). \tag{7}$$

The diagnostic method defined in (7) is called total local influence.

By using the model formulated in (1) and its perturbed version, we determine normal curvatures for local influence. We compute the observed information matrix $-\ddot{\boldsymbol{\ell}}(\widehat{\underline{\theta}})$, find the perturbation matrix $\boldsymbol{\Delta}(\widehat{\underline{\theta}}, \underline{w}_0)$, and then obtain the eigenvector associated with the largest absolute eigenvalue of $\boldsymbol{F}$ given in (6) as a local influence measure. Next, we detail the perturbation matrices for different schemes.

For the scheme of case-weight (ca) perturbation, let $\underline{w} = (w_1, \ldots, w_n)^\top \in \mathbb{R}^n$ be the perturbation vector, where the w_is are positive values denoting the weight

corresponding to the case i, and $\ell_{\mathrm{ca}}(\underline{\theta}|\underline{w})$ is the perturbed log-likelihood function. Let $\underline{w}_0 = \underline{1}_{1\times n}^{\top}$ be the non-perturbation vector such that $\ell_{\mathrm{ca}}(\underline{\theta}|\underline{w}_0) = \ell(\underline{\theta})$. Then, the log-likelihood function for the perturbed model under this scheme is read to be

$$\ell_{\mathrm{ca}}(\underline{\theta}|\underline{w}) = \sum_{i=1}^{n} w_i \ell_i(\underline{\theta}), \tag{8}$$

with $\ell_i(\underline{\theta})$ defined from (2). Hence, we establish the matrix $\boldsymbol{\Delta}_{\mathrm{ca}}(\underline{\theta}, \underline{w})$ by taking the derivatives of $\ell_{\mathrm{ca}}(\underline{\theta}|\underline{w})$ given in (8) with respect to $\underline{\theta}$ and $\underline{w}$, evaluating it at $\underline{\theta} = \widehat{\underline{\theta}}$ and $\underline{w} = \underline{w}_0$. For details about these derivatives, see [16].

For the scheme of correlation matrix (cm) perturbation, let $\underline{w} = (w_1, \ldots, w_n)^{\top} \in \mathbb{R}^n - \{0\}$ be the perturbation vector and $\ell_{\mathrm{cm}}(\underline{\theta}|\underline{w})$ the corresponding perturbed log-likelihood function. Let $\underline{w}_0 = \underline{1}_{1\times n}^{\top}$ be the non-perturbation vector such that $\ell_{\mathrm{cm}}(\underline{\theta}|\underline{w}_0) = \ell(\underline{\theta})$. Then, the log-likelihood function for the perturbed model under this scheme is read to be

$$\ell_{\mathrm{cm}}(\underline{\theta}|\underline{w}) = \sum_{i=1}^{n} \left(\log\left(f_{\mathrm{EC}_m}\left(\underline{\phi}_i; w_i^{-1}\boldsymbol{\Psi}, g^{(m)}\right)\right) + \sum_{j=1}^{m} \log(\xi_{ij}) \right). \tag{9}$$

Again we establish the matrix $\boldsymbol{\Delta}_{\mathrm{cm}}(\underline{\theta}, \underline{w})$ by taking the derivatives now of $\ell_{\mathrm{cm}}(\underline{\theta}|\underline{w})$ given in (9) with respect to $\underline{\theta}$, and then with respect to $\underline{w}$, evaluating it at $\underline{\theta} = \widehat{\underline{\theta}}$ and $\underline{w} = \underline{w}_0$. For details about these derivatives, see [16].

In the scheme of covariate (co) perturbation, we replace the value of a continuous covariate x_{il} by $x_{il} + w_i$, where $x_{il} \in \mathbb{R}^n$ is the lth column of $\underline{x}_i$ and $\underline{w} = (w_1, \ldots, w_n)^{\top} \in \mathbb{R}^n$ is the perturbation vector. Here, $\underline{w}$ can be expressed as a proportional value to the standard deviation (SD) of the perturbed covariate and $\underline{w}_0 = \underline{0}_{1\times n}^{\top}$ is the non-perturbation vector such that $\ell_{\mathrm{co}}(\underline{\theta}|\underline{w}_0) = \ell(\underline{\theta})$. Then, the log-likelihood function for the perturbed model under this scheme is read to be

$$\ell_{\mathrm{co}}(\underline{\theta}|\underline{w}) = \sum_{i=1}^{n} \left(\log\left(f_{\mathrm{EC}_m}\left(\underline{\phi}_i(\underline{w}); \boldsymbol{\Psi}, g^{(m)}\right)\right) + \sum_{j=1}^{m} \log(\xi_{ij}(\underline{w})) \right). \tag{10}$$

Here, we establish $\boldsymbol{\Delta}_{\mathrm{co}}(\underline{\theta}, \underline{w})$ by taking the derivatives of $\ell_{\mathrm{co}}(\underline{\theta}|\underline{w})$ given in (10) with respect to $\underline{\theta}$ and $\underline{w}$, evaluating it at $\underline{\theta} = \widehat{\underline{\theta}}$ and $\underline{w} = \underline{w}_0$. For details about these derivatives, see [16].

In the scheme of response (re) perturbation, we replace $\underline{y}_i$ by $\underline{y}_i + \underline{w}_i$, with $\underline{w}_i = (w_{i1}, \ldots, w_{im})^{\top} \in \mathbb{R}^m$ denoting the corresponding perturbation to the case i. Here, $\underline{w}_i$ can be expressed as a proportional value to the SD of the response and $\underline{w}_0 = \underline{0}_{1\times m}^{\top}$ is the non-perturbation vector such that $\ell_{\mathrm{re}}(\underline{\theta}|\underline{w}_0) = \ell(\underline{\theta})$. Then, the log-likelihood function for the perturbed model under this scheme is read to be

$$\ell_{\mathrm{re}}(\underline{\theta}|\underline{w}) = \sum_{i=1}^{n} \left(\log\left(f_{\mathrm{EC}_m}\left(\underline{\phi}_i(\underline{w}); \boldsymbol{\Psi}, g^{(m)}\right)\right) + \sum_{j=1}^{m} \log(\xi_{ij}(\underline{w})) \right). \tag{11}$$

Again, $\boldsymbol{\Delta}_{\mathrm{re}}(\underline{\theta}, \underline{w})$ is obtained by taking the corresponding derivatives now of $\ell_{\mathrm{re}}(\underline{\theta}|\underline{w})$ given in (11) with respect to $\underline{\theta}$ and $\underline{w}$, and evaluating it at $\underline{\theta} = \widehat{\underline{\theta}}$ and $\underline{w} = \underline{w}_0$. For details about these derivatives, see [16].

4 Summary of the Proposed Methodology

The proposed statistical methodology based on multivariate GBS log-linear regression models and their diagnostics is summarized in Algorithm 1.

5 Case Study I: Bio-Engineering Data

In this case study, we investigate four types of bone densities: (1) bulk density, which considers the mass of the intact core, including fat and water; (2) dry density, which also considers the mass of the intact core, but excluding fat and water; (3) ash density, which is related to the mineral content into the mass of the core; and (4) computed tomography (CT) density, which is obtained from a calibration equation that is derived from known bone mineral content phantoms. Clinical CT scans are used to assess their application in inferring physical properties of human trabecular bone. The prediction of apparent density from ash density allows for estimation of

Algorithm 1 Methodology based on multivariate GBS regression models and their diagnostics

Step 1. Collect a data set of size n with m responses and p covariates for regression modeling.

Step 2. Make an exploratory data analysis based mainly on correlations to justify the use of multivariate distributions and regression models, as well as discarding possible multicollinearity problems among the covariates.

Step 3. Propose a multivariate GBS log-linear regression model for the data set collected in Step 1 based on the information obtained from Step 2.

Step 4. Estimate the model parameters using maximum likelihood and non-linear optimization methods, for example Newton and quasi-Newton methods [18].

Step 5. Fit the multivariate regression model to the data set collected in Step 1 using goodness-of-fit tools, for example, the MD and probability versus probability (PP) plots with acceptance bands based on Kolmogorov-Smirnov (KS) test for the transformed MD [2]. If the model fits adequately the data, then to go to Step 6. Otherwise, others models must be considered to describe the data.

Step 6. Use the local influence method in the fitted model given in Step 4 to identify potentially influential cases. Here, it is necessary to compute the observed information matrix $-\ddot{\boldsymbol{\ell}}(\widehat{\underline{\theta}})$ and obtain the perturbation matrix $\boldsymbol{\Delta}(\widehat{\underline{\theta}}, \underline{w}_0)$ for the four perturbation schemes.

Step 7. Carry out an analysis to evaluate whether removal of potentially influential cases detected in Step 6 produces inferential changes or not. If no inferential changes are detected, then to use the model obtained in Step 4 for prediction. Otherwise, potentially influential cases should be removed and a new estimated model should be obtained.

mechanical properties of bone, which can subsequently be used in a finite element analysis/model of bone [20]. The data set is presented in Table 5 of the appendix. In this table, "Age," "Gender," and "ID" correspond to age, gender, and identification of the bone core for a patient under study, respectively, whereas "Voxels" is the number of volumetric pixels (voxels), "HU mean," "HU SD," and "HU CV" are the mean, SD, and coefficient of variation for the voxel Hounsfield units (HUs) within a bone core. In addition, also in Tables 4 and 5, "ρ_{bulk}" is the bulk density, "ρ_{dry}" is the dry density, ρ_{ash} is the ash density, and "ρ_{ct}" is the CT density, all of these four densities measured over the specimen bulk volume. Next, we describe the data analytic following the steps of Algorithm 1.

Step 1 We consider as responses: (1) bulk density (T_1, in mg/cm^3) and (2) dry density (T_2, in mg/cm^3). The covariates that can affect these responses are: (1) CT density (X_1, in mg/cm^3) and (2) ash density (X_2, in mg/cm^3). We illustrate the proposed multivariate models with real-world bone density data associated with these variables. We work with the log-responses $Y_j = \log(T_j)$, for $j = 1, 2$.

Step 2 We make an exploratory data analysis computing correlations for Y_1, Y_2, X_1 and X_2. Figure 1 displays the scatter-plots for these variables and their corresponding correlations. From this figure, we detect that exist: (1) large correlations between (Y_1, Y_2), justifying the use of a multivariate distribution; (2) large correlations between (X_1, X_2), indicating a possible collinearity problem; and (3) medium correlations between (X_1, Y_1), and (X_1, Y_2), and large correlations between (X_2, Y_1) and (X_2, Y_2), which supports the elimination of X_1. This must be confirmed by the inferential analysis.

Step 3 We propose a multivariate regression model for describing (Y_1, Y_2) in function of X_2 (because X_1 is discarded due to collinearity—see Step 2). Therefore, the proposed multivariate log-linear regression model is given by

$$\underline{Y}_i = \boldsymbol{\beta}^\top \underline{x}_i + \underline{\varepsilon}_i, \quad i = 1, \ldots, 74,$$

where $\underline{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})^\top \sim \text{log-GBS}_2\big(\alpha \underline{1}_{2\times 1}, \underline{0}_{2\times 1}, \boldsymbol{\Psi}_{2\times 2}, g^{(2)}\big)$.

Step 4 We estimate the parameters of the multivariate BS and BS-t regression models via the maximum likelihood method using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) quasi-Newton method [18], which we have implemented in `R` code. Starting values, $\widehat{\underline{\theta}}^{(0)}$ say, used in the maximization procedure are: $\widehat{\alpha}^{(0)} = 0.077963$,

$$\widehat{\boldsymbol{\beta}}^{(0)} = \begin{pmatrix} 6.677099 & 4.682944 \\ 0.001157 & 0.004538 \end{pmatrix}, \quad \widehat{\boldsymbol{\Psi}}^{(0)} = \begin{pmatrix} 1.000000 & 0.378761 \\ 0.378761 & 1.0000 \end{pmatrix}.$$

In addition, we have used the value $\nu = 4$ for the t distribution and verified that it corresponds to the value that maximizes the log-likelihood function within a range of values for ν. For details about starting values for the maximum likelihood estimation procedure used in this illustration, see [16].

Table 2 displays the parameter estimates, the value of the maximized log-likelihood function, estimated asymptotic standard error of the corresponding

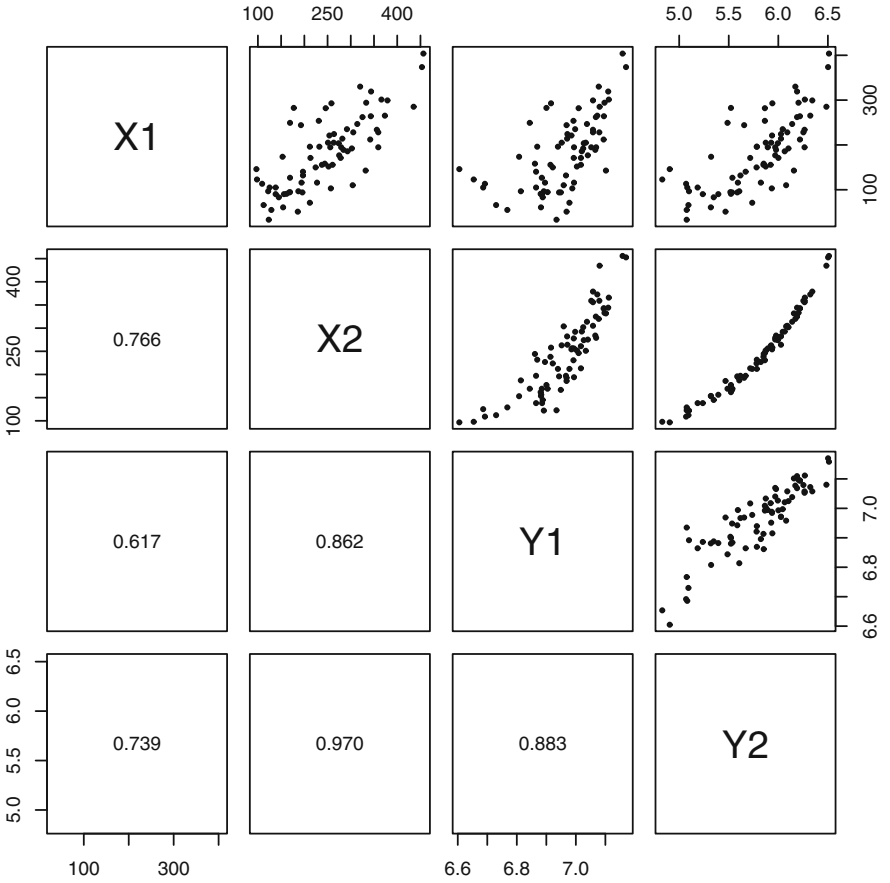


Fig. 1 Scatter-plots with their corresponding correlations for the indicated variable with bone density data

Table 2 Maximum likelihood estimate of the indicated parameter and model, with its corresponding estimated asymptotic standard error, p-value and log-likelihood function with bone density data

	BS_2 model			BS-t_2 model		
Parameter	Estimate	Standard error	p-value	Estimate	Standard error	p-value
ρ	0.377034	0.050354	<0.001	0.373313	0.059365	<0.001
β_{01}	6.676328	0.025193	<0.001	6.678921	0.029402	<0.001
β_{02}	4.679009	0.025101	<0.001	4.687253	0.032067	<0.001
β_{11}	0.001159	0.000097	<0.001	0.001151	0.000110	<0.001
β_{12}	0.004550	0.000097	<0.001	0.004526	0.000120	<0.001
α	0.072004	0.003628	<0.001	0.071836	0.004840	<0.001
Log-likelihood	165.3831	–	–	169.3681	–	–

maximum likelihood estimators for both models, and p-values of each t-test. As usual, we use the square root of the diagonal elements of the observed Fisher information inverse matrix to approximate the corresponding estimated standard errors [19]. From this table, and for a 5% significance level, we obtain the following conclusions: (1) estimated correlation from the BS_2 and BS-t_2 log-linear models results to be statistically significant, corroborating our conjecture from the exploratory analysis; and (2) the regression coefficients β_0 (constant term of the model) and β_1 (slope) must be considered in the prediction of T_1 and T_2 because β_0 and β_1 are statistically significant at 5%. We can also see that the value that maximizes the log-likelihood function is greater for the BS-t_2 model than for the BS model, indicating a better fit with the BS-t_2 model.

Step 5 As mentioned, m-variate log-GBS model checking can be conducted by using the MD. Here, this distance follows the $\chi^2(m = 2)$ or $2\mathcal{F}(m = 2, \nu = 4)$ distribution if $g^{(2)}$ is the bivariate normal or t_2 density generator, respectively. We substitute the maximum likelihood estimator of $\underline{\theta}$ in $\mathrm{MD}_i(\widehat{\underline{\theta}})$, which has asymptotically the same distribution of $\mathrm{MD}_i(\underline{\theta})$. We use the Wilson-Hilferty (WH) approximation for transforming this distance, which should follow now a normal distribution. Then, we check normality of the transformed distances with the WH approximation using goodness-of-fit techniques. Figure 2a, b shows the corresponding PP plots with acceptance bands for a significance level of 5%. From this figure, we detect that the BS-t_2 log-linear regression model provides a better fit than the BS_2 model, which is corroborated by the p-values 0.4068 and 0.0227, respectively, of the KS test associated with these PP plots. Therefore, we can conclude that the multivariate BS-t_2 log-linear regression model fits better the bone density data. The MD is a global influence measure to detect multivariate outliers. Figure 2c, d displays the index plots of this distance for the BS_2 and BS-t_2 log-linear regression models. In addition, Fig. 2e presents the plot of estimated weights versus MD_i for the BS-t_2 log-linear regression model, with $i = 1, \ldots, 74$. From Fig. 2c, d, note that the cases {#20, #48, #55, #69, #70} appear as possible multivariate outliers in the BS_2 model, but not in the BS-t_2 model. In Fig. 2e, observe that these cases have smaller weight in the BS-t_2 model than in the BS_2 model, which confirms the inherent robustness of the maximum likelihood procedure against possible outlying observations.

Step 6 In order to identify possible influential cases in the fitted models, we present some diagnostic graphs for total local influence (C_i). Figure 3 shows these plots under the case-weight, correlation, covariate and response perturbation schemes for $\widehat{\underline{\theta}}$. From this figure, note that the cases {#20, #48, #55, #69, #70} appear with a large influence in the BS_2 model, but not in the BS-t_2 model. These cases coincide with those detected by the MD in Step 5.

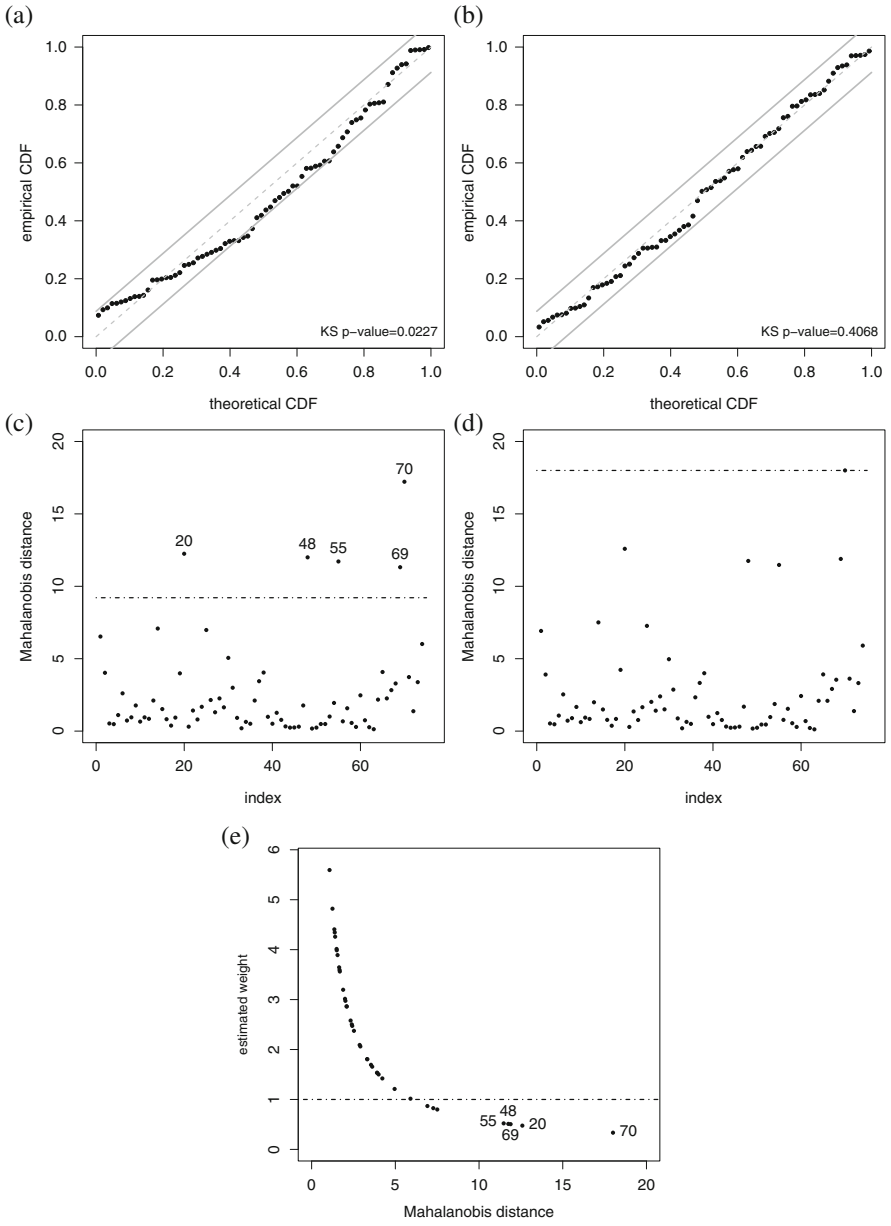


Fig. 2 PP plots with KS acceptance regions at 5% for transformed MDs in BS_2 (**a**) and BS-t_2 (**b**) models; index plots of MDs for the BS_2 (**c**) and BS-t_2 (**d**) models; and plot of estimated weights of MDs for BS-t_2 (**e**) and BS_2 models (straight line at a value equal to one) with bone density data

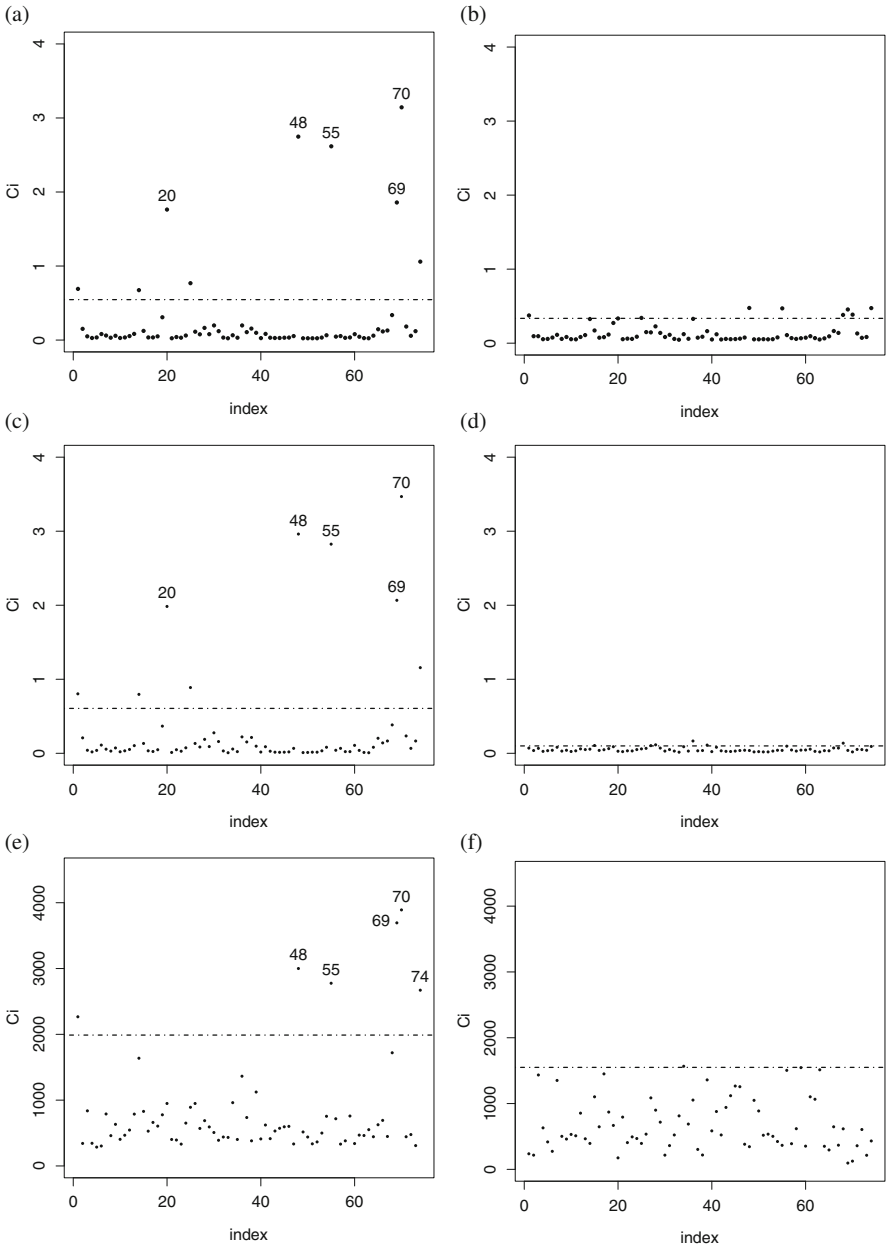


Fig. 3 Total local influence index plots of case-weight (**a**, **b**), correlation (**c**, **d**), covariate (**e**, **f**), and response (**g**, **h**) perturbations for $\widehat{\underline{\theta}}$ in the indicated model with bone density data

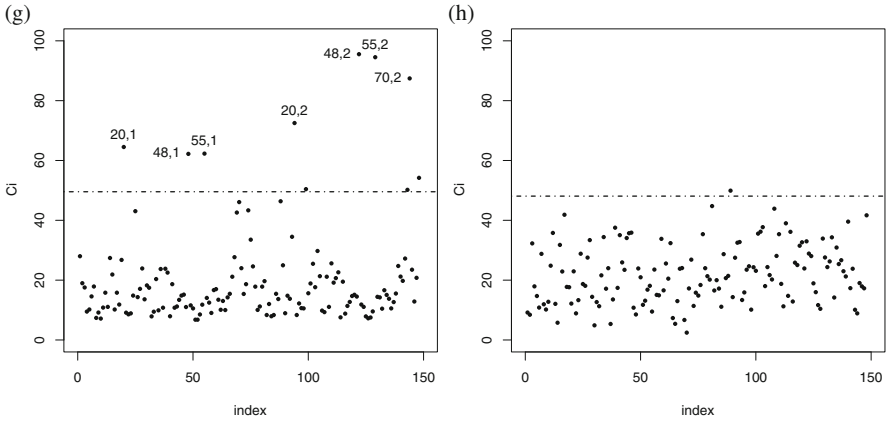


Fig. 3 (continued)

6 Case Study II: Industry Data

In this case study, we analyze die fracture, which is a typical metal fatigue caused by cyclic stress in the course of the service life cycle of dies (die lifetime). Although this fatigue can be mainly determined by die lifetime, other random variables can also be considered as responses to this fatigue. The purpose of this illustration is to model fatigue in a metal forming process. The data set is presented in Table 5 of the appendix. Next, we describe the data analytic following the steps of Algorithm 1.

Step 1 We consider as responses: (1) Von Mises stress (T_1, in N/mm^2) and (2) manufacturing force (T_2, in Newton –N–). The covariates that can affect these responses are: (1) friction coefficient (X_1, dimensionless) and (2) work temperature (X_2, in $^{\text{o}}$C).

Step 2 Figure 4 displays the scatter-plots for all log-responses and covariates, from which we detect that: (1) no correlations exist between (X_1, X_2), discarding any possible collinearity problem in our model; (2) large correlation between (Y_1, Y_2), justifying the use of a multivariate distribution; (3) small correlations between (X_1, Y_1) and (X_1, Y_2); (4) large correlations between (X_2, Y_1) and (X_2, Y_2).

Step 3 We propose a multivariate regression model for describing (Y_1, Y_2) in function of X_2 (because X_1 is discarded due to its small correlations with the responses—see Step 2). Therefore, the proposed multivariate log-linear regression model is given by

$$\underline{Y}_i = \boldsymbol{\beta}^\top \underline{x}_i + \underline{\varepsilon}_i, \quad i = 1, \ldots, 15,$$

where $\underline{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2})^\top \sim \text{log-GBS}_2\big(\alpha \underline{1}_{2\times 1}, \underline{0}_{2\times 1}, \boldsymbol{\Psi}_{2\times 2}, g^{(2)}\big)$.

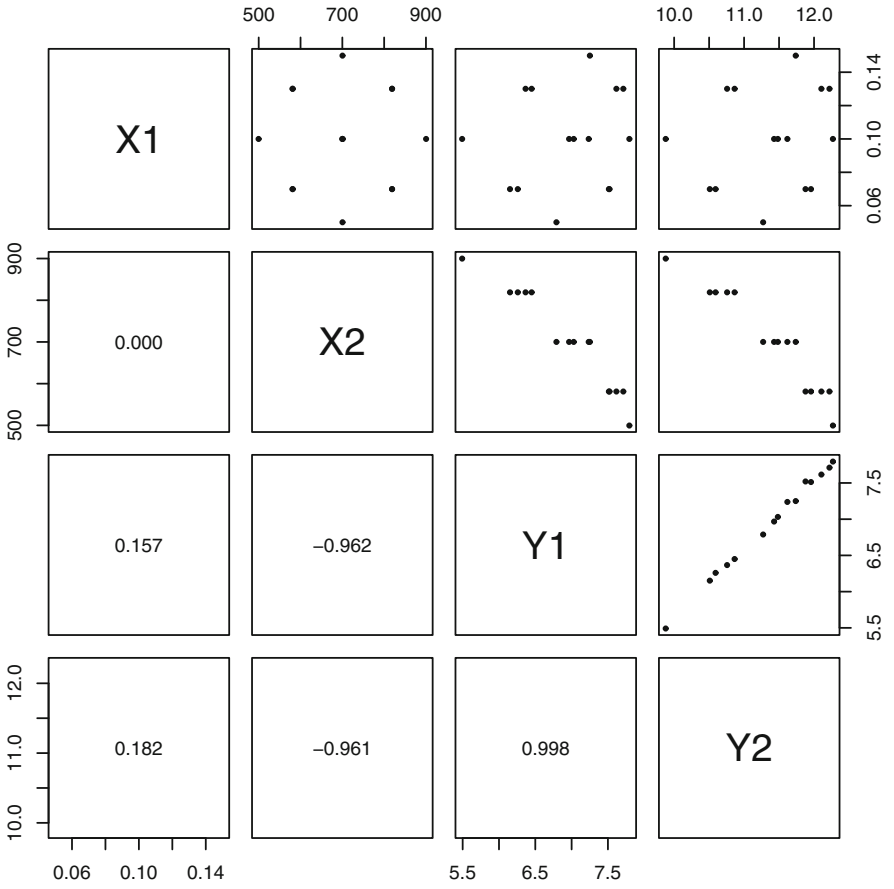


Fig. 4 Scatter-plots with their corresponding correlations for the indicated variables with die lifetime data

Step 4 We estimate the parameters of the multivariate BS and BS-t log-linear regression models via the maximum likelihood method and the BFGS approach. Starting values, $\widehat{\underline{\theta}}^{(0)}$ say, used in the maximization procedure are: $\widehat{\alpha}^{(0)} = 0.184808$,

$$\widehat{\boldsymbol{\beta}}^{(0)} = \begin{pmatrix} 10.825491 & 15.448092 \\ -0.005546 & -0.005823 \end{pmatrix}, \widehat{\boldsymbol{\Psi}}^{(0)} = \begin{pmatrix} 1.000000 & 0.972114 \\ 0.972114 & 1.000000 \end{pmatrix}.$$

Once again, we use the value $\nu = 4$ for the t distribution, which we verify that corresponds to the value that maximizes the log-likelihood function within a range of values for ν.

Table 3 displays the model parameter estimates, the estimated asymptotic standard errors of the corresponding maximum likelihood estimators for both models, and p-values of each t-test. From this table, and for a significance level of 1%,

Table 3 Maximum likelihood estimate of the indicated parameter and model with its corresponding estimated asymptotic standard error, p-value and log-likelihood function with die lifetime data

	BS_2 model			BS-t_2 model		
Parameter	Estimate	Standard error	p-value	Estimate	Standard error	p-value
ρ	0.972392	0.005219	<0.001	0.975965	0.004694	<0.001
β_{01}	10.897981	0.236175	<0.001	10.759766	0.234839	<0.001
β_{02}	15.524423	0.235865	<0.001	15.432361	0.242804	<0.001
β_{11}	−0.005647	0.000333	<0.001	−0.005438	0.000334	<0.001
β_{12}	−0.005930	0.000333	<0.001	−0.005779	0.000344	<0.001
α	0.147407	0.014813	<0.001	0.101474	0.009816	<0.001
Log-likelihood	26.9665	–	–	23.6383	–	–

we can obtain the following conclusions: (1) estimated correlation from the BS_2 and BS-t_2 log-linear models results to be statistically significant, corroborating our conjecture from the exploratory analysis; and (2) the regression coefficients β_0 and β_1 must be considered in the prediction of T_1 and T_2 because they are significant at 1%. We can also see that the value that maximizes the log-likelihood function is greater for the BS_2 model than for the BS-t_2 model, which is a conclusion different from the another case analysis, indicating a better fit of the BS_2 model.

Step 5 Figure 5a, b shows the corresponding PP plots for the transformed MDs with acceptance bands for a significance level of 5%. From this figure, we detect that the BS_2 model provides a better fit than the BS-t_2 model, which is corroborated by the p-values 0.2480 and 0.0736, respectively, of the KS test associated with these PP plots. Therefore, we can conclude that the BS_2 log-linear regression model fits better the die lifetime data. Figure 5c, d displays the index plots of this distance for the BS_2 and BS-t_2 log-linear regression models. In addition, Fig. 5e presents the plot of estimated weights versus MD_i for the BS-t_2 log-linear regression model, with $i = 1, \ldots, 15$. From Fig. 5c, d, note that the case #1 appears as possible multivariate outlier in the BS_2 model, but not in the BS-t_2 log-linear regression model. In Fig. 5e, observe that these cases have smaller weight in the BS-t_2 log-linear regression model than the BS_2 log-linear regression model.

Step 6 Figure 6 shows the index plots of C_i under the case-weight, correlation, covariate, and response perturbation schemes for $\widehat{\underline{\theta}}$. From this figure, note that the case #1 appears with a large influence in the BS_2 log-linear regression model under all perturbation schemes, but not in the BS-t_2 model. These cases coincide with those detected by the MD in Step 5. In addition, the case #14 appears with a large influence in the BS_2 log-linear regression model under case-weight and correlation perturbation schemes.

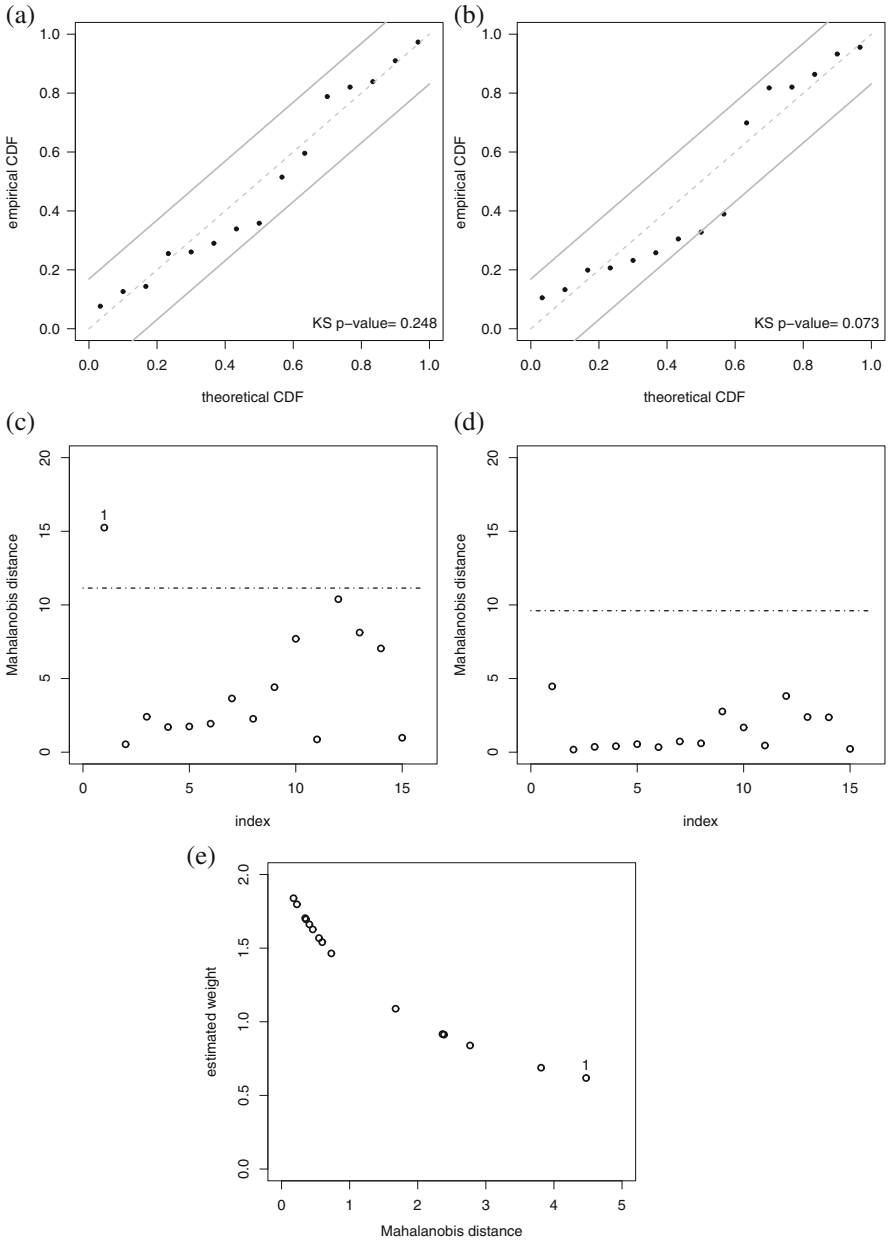


Fig. 5 PP plots with KS acceptance regions at 5% for transformed MDs with BS_2 (**a**) and BS-t_2 (**b**) models; index plots of MDs for the BS_2 (**c**) and BS-t_2 (**d**) models; and plot of estimated weights of MDs for the BS-t_2 model (**e**) and BS_2 model—straight line at a value equal to one—with die lifetime data

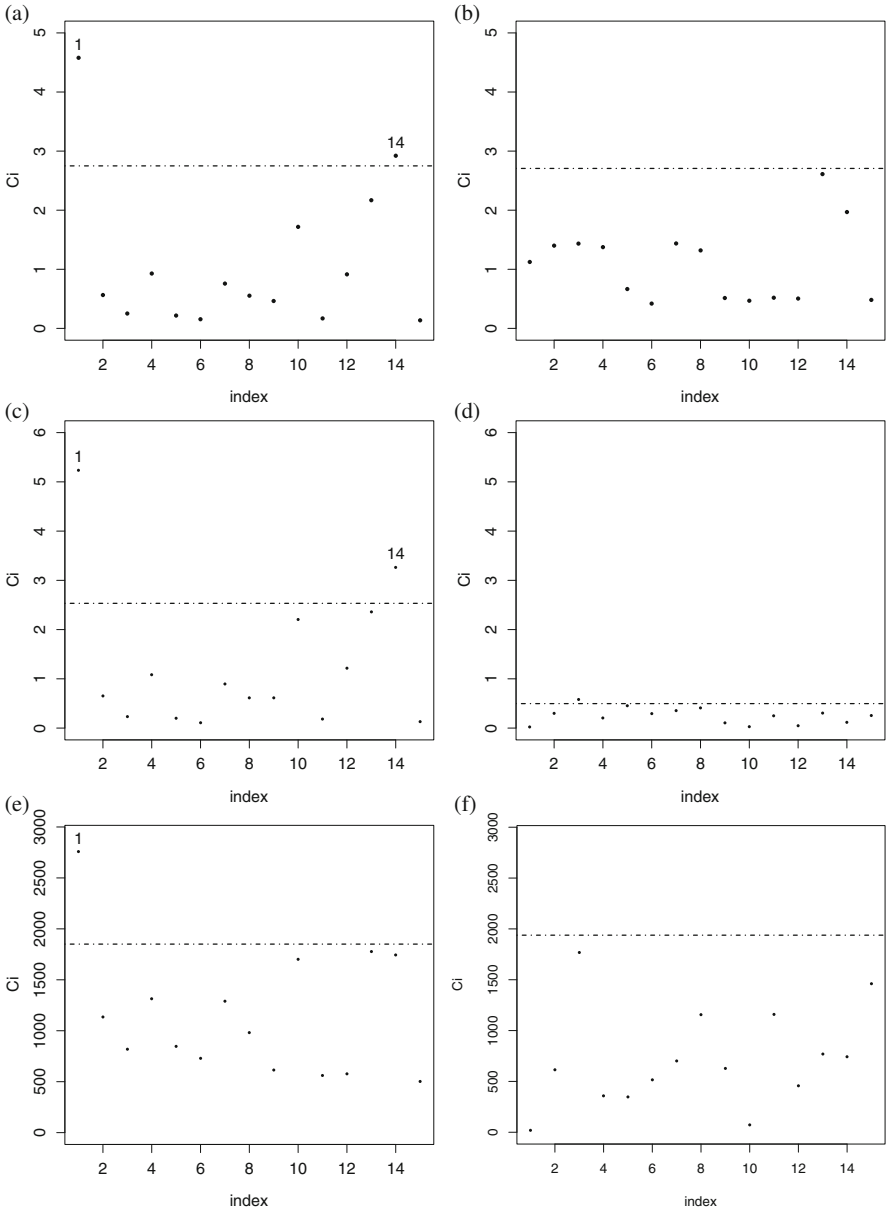


Fig. 6 Total local influence index plots of case-weight (**a**, **b**), correlation (**c**, **d**), covariate (**e**, **f**) and response (**g**, **h**) perturbations for $\widehat{\underline{\theta}}$ in the indicated model with die lifetime data

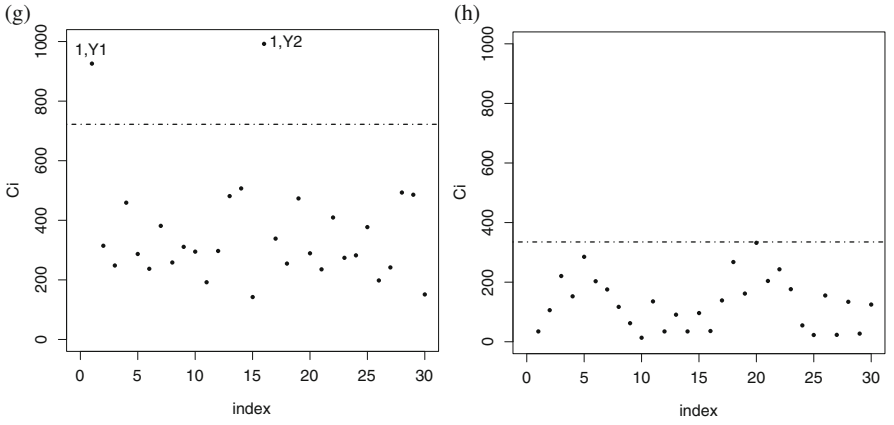


Fig. 6 (continued)

7 Conclusions and Future Research

We have derived a methodology based on generalized multivariate Birnbaum-Saunders log-linear regression models and their diagnostics. We have considered the Mahalanobis distance for evaluating the suitability of the distributional assumption by transforming this distance with the Wilson-Hilferty approximation and then by using goodness-of-fit techniques. In addition, the Mahalanobis distance has been employed as a global influence measure to detect multivariate outliers. Furthermore, we have developed local influence under perturbation schemes of case-weight, correlation matrix, response variable, and a continuous covariate. We have implemented the obtained results in the `R` software. These results have been applied to case studies in bio-engineering and industry to illustrate their good performance.

Some aspects to be studied in a future research are related to exploring the effect of considering different shape parameters in each response. In addition, heterogeneity presented in the observations can also be modeled. Furthermore, other estimation methods and robustness aspects may be studied. Moreover, a residual analysis should be derived for these regression models, whose appropriateness is usually detected by residuals. Also, random effects may be inserted in the modeling considered in this work. All of these aspects provide us future challenging issues to be analyzed.

Acknowledgements The authors thank the editors and reviewers for their constructive comments on an earlier version of this manuscript. This research work was partially supported by FONDECYT 1160868 grant from the Chilean government.

Appendix: Bio-Engineering and Industry Data Sets

See Tables 4 and 5.

Table 4 CT scan data to study the bone quality

							Density (mg/cm^3)			
Age	Gender	ID	Voxels	HU mean	HU SD	HU CV	ρ_{ct}	ρ_{bulk}	ρ_{dry}	ρ_{ash}
68	F	213	2311	144	153	1.060	105	801	161	125
68	F	231	1811	154	177	1.150	116	988	338	227
68	F	233	2214	260	262	1.010	195	1156	526	359
68	F	234	2447	327	205	0.626	235	1094	422	292
68	F	235	2026	272	227	0.832	205	1045	376	263
87	F	411	1902	230	132	0.573	156	1006	347	238
87	F	412	1902	329	121	0.368	228	1187	519	359
87	F	413	2059	401	132	0.330	282	1107	352	246
87	F	421	1376	104	122	1.173	71	1073	310	212
87	F	422	1791	232	115	0.497	156	1116	373	261
87	F	423	1862	284	123	0.433	195	1175	391	279
87	F	424	1761	213	111	0.520	143	1214	472	332
87	F	425	2267	349	123	0.352	244	1063	286	193
87	F	426	1614	88	89	1.010	55	869	160	129
87	F	427	1833	425	159	0.374	301	1226	527	366
86	F	511	4250	180	157	0.870	132	1061	275	197
86	F	512	3231	121	117	0.967	83	980	210	145
86	F	521	3047	129	140	1.080	95	996	249	169
86	F	522	3083	148	136	0.922	105	958	179	138
86	F	523	3618	45	95	2.130	33	1027	160	123
86	F	524	3750	131	136	1.040	94	1036	267	196
86	F	525	3623	140	142	1.020	103	1084	377	257
86	F	526	3256	232	179	0.770	171	1089	404	278
86	F	531	4972	149	167	1.120	116	1090	268	194
86	F	532	4740	128	155	1.210	97	984	164	122
86	F	534	5141	125	153	1.220	94	1042	253	167
86	F	535	4170	74	138	1.860	61	974	204	154
86	F	536	4413	116	151	1.300	90	978	188	138
81	F	611	2487	193	102	0.529	127	977	254	169
81	F	613	2891	324	116	0.360	224	1066	414	264
81	F	621	1448	211	110	0.521	140	958	290	197
81	F	622	2688	297	136	0.457	204	1126	401	274
81	F	624	3210	329	125	0.380	228	1162	442	305

(continued)

Table 4 (continued)

Age	Gender	ID	Voxels	HU mean	HU SD	HU CV	Density (mg/cm^3)			
							ρ_{ct}	ρ_{bulk}	ρ_{dry}	ρ_{ash}
77	F	712	3109	377	161	0.428	265	1179	556	373
77	F	713	2462	280	116	0.415	192	1124	446	302
77	F	714	2319	102	107	1.052	66	837	163	112
77	F	722	3817	361	160	0.442	254	1088	352	231
77	F	724	2922	234	136	0.580	158	955	348	244
77	F	726	3112	422	195	0.463	299	1162	568	379
79	M	311	1668	260	118	0.454	177	1142	391	275
79	M	312	1952	450	125	0.278	319	1224	488	344
79	M	321	1554	278	101	0.364	189	1171	394	284
79	M	322	1612	465	147	0.317	330	1185	479	320
79	M	323	1601	417	146	0.349	294	1208	494	333
79	M	324	1542	309	104	0.336	212	1205	501	342
79	M	325	1755	377	137	0.365	264	1206	501	343
79	M	326	1520	223	112	0.501	150	1014	324	224
79	M	327	1625	564	187	0.332	404	1285	672	456
79	M	331	1900	374	174	0.466	262	1174	487	325
79	M	332	1781	355	122	0.344	247	1139	464	314
79	M	333	1486	284	145	0.512	195	1090	363	256
79	M	334	1368	227	124	0.549	152	1100	355	252
79	M	335	1448	299	106	0.353	205	1134	355	251
79	M	336	1225	252	145	0.577	171	1115	304	213
79	M	337	1333	524	164	0.313	374	1300	667	453
66	M	811	2394	258	129	0.500	174	905	205	153
66	M	812	3029	310	162	0.522	214	1065	415	282
66	M	821	2270	273	93	0.339	186	1119	430	293
66	M	822	2207	337	98	0.290	234	1162	526	356
66	M	823	2307	287	99	0.384	196	962	325	232
66	M	824	2838	358	117	0.328	249	938	242	169
66	M	825	2168	403	91	0.227	283	992	250	177
66	M	826	1893	146	64	0.441	90	975	220	156
66	M	827	2001	320	124	0.387	221	1080	379	254
66	M	831	2045	287	120	0.418	196	1032	325	212
66	M	832	2029	147	87	0.589	91	973	250	162
66	M	833	885	93	53	0.573	51	1063	237	186
66	M	834	2500	177	84	0.477	113	806	159	109
66	M	835	3000	220	88	0.401	146	739	135	97
66	M	836	3334	190	75	0.396	123	776	125	98
66	M	837	2172	155	81	0.523	97	910	272	187
66	M	838	1942	172	88	0.514	110	1052	437	304
66	M	839	3428	416	182	0.439	293	1008	380	258
66	M	8310	3477	405	127	0.313	285	1188	656	435

Source: [20]

Table 5 Fatigue data for the indicated variable

Friction coefficient	Angle	Temperature	Von Misses stress	Maximum deformation	Manufacturing force	Lifetime
0.07	23.00	581.08	1850	1.260	144,000	6420
0.07	23.00	818.92	470	1.349	36,700	33,700
0.07	31.96	581.08	1830	1.532	156,000	9430
0.07	31.96	818.92	523	1.614	39,900	36,600
0.13	23.00	581.08	2030	1.801	181,000	12,100
0.13	23.00	818.92	581	1.824	46,900	32,000
0.13	31.96	581.08	2230	1.939	203,000	13,200
0.13	31.96	818.92	632	1.928	52,300	32,100
0.05	27.50	700.00	889	1.275	78,600	19,900
0.15	27.50	700.00	1410	1.921	125,000	15,000
0.10	20.00	700.00	1060	1.692	92,100	20,900
0.10	35.00	700.00	1390	1.888	111,000	21,200
0.10	27.50	500.00	2430	1.666	213,000	9170
0.10	27.50	900.00	243	1.685	19,500	74,800
0.10	27.50	700.00	1130	1.651	96,900	19,900

Source: [13]

References

1. Barros, M., Paula, G.A., Leiva, V.: A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. Lifetime Data Anal. **14**, 316–332 (2008)
2. Barros, M., Leiva, V., Ospina, R., Tsuyuguchi, A.: Goodness-of-fit tests for the Birnbaum-Saunders distribution with censored reliability data. IEEE Trans. Reliab. **63**, 543–554 (2014)
3. Díaz-García, J.A., Leiva, V.: A new family of life distributions based on elliptically contoured distributions. J. Stat. Plan. Inference **128**, 445–457 (2005)
4. Díaz-García, J.A., Leiva, V., Galea, M.: Singular elliptic distribution: density and applications. Commun. Stat. Theory Methods **31**, 665–681 (2002)
5. Díaz-García, J.A., Galea, M., Leiva, V.: Influence diagnostics for elliptical multivariate linear regression models. Commun. Stat. Theory Methods **32**, 625–641 (2003)
6. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman and Hall, London (1990)
7. Galea, M., Paula, G.A., Uribe-Opazo, M.A.: On influence diagnostic in univariate elliptical linear regression models. Stat. Pap. **44**, 23–45 (2003)
8. Garcia-Papani, F., Uribe-Opazo, M.A., Leiva, V., Aykroyd, R.G.: Birnbaum-Saunders spatial modelling and diagnostics applied to agricultural engineering data. Stoch. Env. Res. Risk A. **31**, 105–124 (2017)
9. Lange, K., Sinsheimer, J.: Normal/independent distributions and their applications in robust regression. J. Comput. Graph. Stat. **2**, 175–198 (1993)
10. Lange, K., Little, J., Taylor, M.: Robust statistical modeling using the t distribution. J. Am. Stat. Assoc. **84**, 881–896 (1989)
11. Leiva, V.: The Birnbaum-Saunders Distribution. Academic Press, New York (2016)
12. Leiva, V., Liu, S., Shi, L., Cysneiros, F.J.A.: Diagnostics in elliptical regression models with stochastic restrictions applied to econometrics. J. Appl. Stat. **43**, 627–642 (2016)

13. Lepadatu, D., Kobi, A., Hambli, R., Barreau, A.: Lifetime multiple response optimization of metal extrusion die. In: Proceedings of Annual Reliability and Maintainability Symposium, pp. 37–42. IEEE, Piscataway (2005)
14. Li, A., Chen, Z., Xie, F.: Diagnostic analysis for heterogeneous log-Birnbaum-Saunders regression models. Stat. Probab. Lett. **89**, 1690–1698 (2012)
15. Marchant, C., Leiva, V., Cysneiros, F.J.A.: A multivariate log-linear model for Birnbaum-Saunders distributions. IEEE Trans. Reliab. **65**, 816–827 (2016)
16. Marchant, C., Leiva, V., Cysneiros, F.J.A., Vivanco, J.F.: Diagnostics in multivariate generalized Birnbaum-Saunders regression models. J. Appl. Stat. **43**, 2829–2849 (2016)
17. Marchant, C., Leiva, V., Cysneiros, F.J.A., Liu, S.: A multivariate log-linear model for Birnbaum-Saunders distributions. J. Stat. Comput. Simul. **88**, 182–202 (2018)
18. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (2006)
19. Paula, G.A., Leiva, V., Barros, M., Liu, S.: Robust statistical modeling using the Birnbaum-Saunders-t distribution applied to insurance. Appl. Stoch. Model. Bus. Ind. **28**, 16–34 (2012)
20. Vivanco, J.F., Burgers, T.A., García, S., Crookshank, M., Kunz, M., MacIntyre, N.J., Harrison, M.M., Bryant, J.T., Sellens, R.W., Ploeg, H.L.: Estimating the density of femoral head trabecular bone from hip fracture patients using computed tomography scan data. Proc. Inst. Mech. Eng. H J. Eng. Med. **228**, 616–626 (2014)

Energy Prices Forecasting Using GLM

M. Filomena Teodoro, Marina A. P. Andrade, Eliana Costa e Silva, Ana Borges, and Ricardo Covas

Abstract The work described in this article results from a problem proposed by the company EDP—Energy Solutions Operator, in the framework of ESGI 119th, European Study Group with Industry, during July 2016. Markets for electricity have two characteristics: the energy is mainly no-storable and volatile prices at exchanges are issues to take into consideration. These two features, between others, contribute significantly to the risk of a planning process. The aim of the problem is the short-term forecast of hourly energy prices. In the present work, GLM is considered a useful technique to obtain a predictive model where its predictive power is discussed. The results show that in the GLM framework the season of the year, month, or winter/summer period revealed significant explanatory variables in the different estimated models. The in-sample forecast is promising, conducting to adequate measures of performance.

M. F. Teodoro (✉)
CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, Almada, Portugal

CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Lisboa, Portugal
e-mail: maria.alves.teodoro@marinha.pt

M. A. P. Andrade
ISCTE-IUL/ISTAR, Lisboa, Portugal
e-mail: marina.andrade@iscte.pt

E. C. e Silva · A. Borges
CIICESI/ESTG - P.Porto, Margaride, Felgueiras, Portugal
e-mail: eos@estg.ipp.pt; aib@estg.ipp.pt

R. Covas
CMA - Centro de Matemática e Aplicações, Universidade Nova de Lisboa, Caparica, Portugal
EDP - Energias de Portugal, Lisboa, Portugal

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_23

1 Introduction

The objective of the present work is the short term forecast of hourly energy prices. Electricity Price Forecasting (EPF) is a difficult purpose. A wide number of methods have been proposed to EFP. In [14] is described an almost complete review about the enormous quantity of available methods, analyzing their strengths and weaknesses. The author proposes the classification of such methods in four categories: multi-agent models, fundamental models, reduced-form models, statistical models and computational intelligence models.

Most of the statistical approaches consist in methods that forecast the current electricity price by using a mathematical combination of the previous prices and/or previous or current values of exogenous factors, such as consumption and production figures or weather variables (see [14] for further detail).

Statistical EPF models are mainly inspired from economics literature such as game theory models and time-series econometric models, as explained also by [8], where they present an extremely relevant summary of selected finance and econometrics inspired literature on spot electricity price forecasting (see Table 3 in [8]).

Considering the short-term forecasting in an EPF context, the more frequent techniques are the ones which take into account the autoregression and moving average models ARMA, that can be combined with the stationary form of time series, the ARIMA models. When seasonality is an important issue, the extended form of such models results in the SARIMA approach. The forecasting of ARMA-type models can be conducted via the Durbin-Levinson algorithm or the innovations algorithm, or by the Kalman filter for models in space state form. ARX, ARMAX, ARIMAX, and SARIMAX are the extension of these models when some exogenous factors [14] are considered (e.g., generation capacity, load profiles, and meteorological conditions).

Multivariate time series analysis is used when one wants to model and explain the interactions and co-movements among a group of time series variables. In this scope [3, 4, 11] have proposed some techniques: VAR, MAR, VARMA, GARCH, ARFNN (fusion of VAR and fuzzy neural networks), Extended Kalman Filter, Polynomial fitting. A vector autoregressive structure (VAR) approach has been recently proposed [14]. Temporal Distribution Extrapolation is another possible approach. It considers the kernel density estimation taking into account, for example, pseudo-points. It is a nonparametric technique which estimates the distribution of a random (univariate or multivariate) variable minimizing some measure. Quite interesting work is presented in [9, 12].

Another method that can be found in literature is the GLM approach. For example, a semi-parametric model for electricity spot prices [5] is built applying GLM where an unknown link function is estimated together with the linear part of the model, followed by a Principal Components Analysis and cross validation to reduce the dimensionality of the problem, avoiding the over-fitting. Also in a GLM setting [10], a Gausss-Laplacian mixture model was used as a basis for stochastic optimization of electricity market.

In 1972, was born the idea of GLM as a powerful method in Statistics, standardizing the different theoretical and applied points of view about all the structure of linear regression developed until then. Due to the large number of models, and simplicity of development associated with rapid computational analysis, the GLM have been playing an important role in statistical analysis. The idea is the establishment of a functional relation between the variable to predict (dependent variable) and a set of other exogenous variables (explanatory variables or covariates). This relation allows to predict the dependent variable. The dependent variables and the explanatory variables can be of any type: continuous, discrete, dichotomous, quantitative, qualitative, stochastic, non-stochastic. The response variable can also be a proportion, be positive, have a non-normal random component. In 1935, Bliss proposed the probit model to proportions; in 1944, Berkson developed the logistic regression, log-linear models for contingency tables were introduced by Birch at 1963. In 1972, Nelder and Wedderbrun proved that all these models are particular cases of a general family: the generalized linear models. In GLM, the random component belongs to exponential family and a transformation of expected value of response variable is related with explanatory variables. The simplest models, where the explanatory variables are nonrandom and the disturbances are Gaussian white noise, which are estimated by ordinary least squares, can be extended for more general models in which the disturbances are auto-correlated, heteroskedastic, not Gaussian, etc., or when some of the explanatory variables are stochastic. Recently, data mining methodology has increased its influence mostly by its fast computational performance. It does not mean that data mining shall replace the proven effective techniques such as GLM. The advantages of both techniques can be combined (see, e.g., [6]).

In the present work, GLM is considered a useful technique to obtain a predictive model where its predictive power is discussed.

The outline of this article is developed in four sections. In Sect. 2 are given more details on the challenge proposed by EDP and on the data provided. Will be presented a summary about exploratory analysis of the data sets provided by EDP and continues with the study on the co-variables that may predict the hourly prices pattern. In Sect. 3 is presented a GLM approach. Finally in Sect. 4 conclusions are drawn and suggestions for future work are pointed.

2 Exploratory Analysis

Taking into consideration the challenge proposed by EDP, the available data consists in the daily market electricity prices as a strip of prices (one for each hour of the day), all simultaneously observed once at a given time of each day:

$$Y_t = [y_{1t}, y_{2t}, \ldots, y_{nt}], \quad n = 1, \ldots, 24 \text{ (or 23 or 25)}, \quad t = 1, 2, \ldots.$$

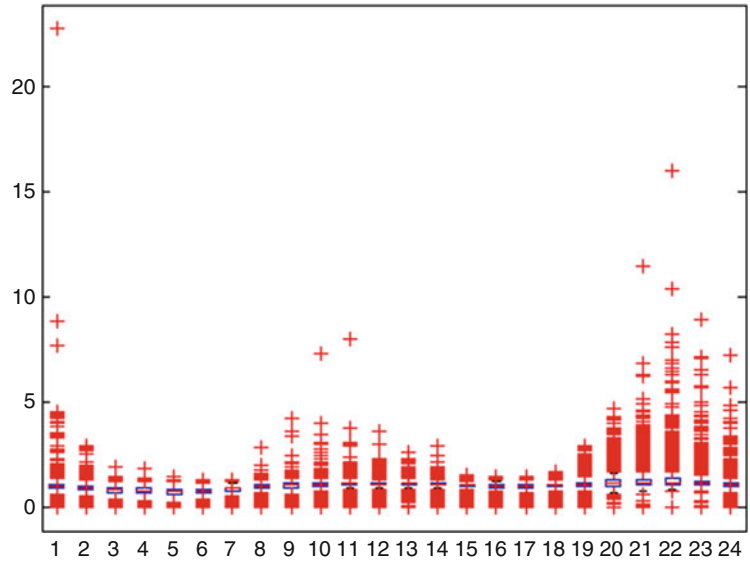


Fig. 1 Boxplot diagrams (rescaled data 01.01.2008–31.12.2016)

In the present work we consider the disaggregated data, i.e., hourly prices and average day price, from January 2008 to June 2016, in a total of 3102 observations of the 24 (23 or 25) h of the day.

In a preliminary exploratory analysis, the data originally provided consisted in a transformed ratio (in what follows named rescaled data) and revealed serious problems which can be visualized in the boxplot diagrams (Fig. 1). The rescaled data has different distributions and a great number of anomalies per hour. These details are also confirmed in Table 1 where some descriptive statistics and tests are summarized.

From Table 1, we can see the different patterns of dispersion (observe the standard deviation and inter-quartile range columns, respectively). Also we confirm that the data does not have normal distribution when we check the Kolmogorov-Smirnov and Jarcke and Bera normality tests.

Consequently, we consider a new data set with the real data. In a preliminary analysis, we have taken the period from 1st January 2008 to 31st December 2010, to exemplify some details and issues and to estimate the initial models considering several covariates of interest.

Since we have a huge dimensional data set, to compare graphically the rescaled data set and the real data set we restrict to the year 2010 graphics in Fig. 2. We can conclude that rescaled data present a huge quantity of "uncommon" observations each hour of the day with exception of hours 4, 5, and 6. The rescaled data also presents different patterns of dispersion. On the other hand, the real data displays unusual observations but in a fewer quantity than in rescaled data. The dispersion of real data presents more homogeneous patterns each hour.

Table 1 Descriptive summary (rescaled data 01.01.2008–31.12.2016)

Hora	Mean	Trimmean	Median	Std.	Iqr	Hora	Skewness	Kurtosis	*P*-value (KS)	*P*-value (JB)
1.0000	1.0212	0.9908	0.9900	0.5327	0.1800	1.0000	24.5154	927.2962	0	0.0010
2.0000	0.8695	0.8872	0.8900	0.2487	0.2000	2.0000	−0.0747	12.7788	0	0.0010
3.0000	0.7531	0.7943	0.8100	0.2483	0.2100	3.0000	−1.4003	5.3185	0	0.0010
4.0000	0.7114	0.7541	0.7800	0.2523	0.2300	4.0000	−1.2967	4.6556	0	0.0010
5.0000	0.6802	0.7230	0.7500	0.2504	0.2400	5.0000	−1.2724	4.2777	0	0.0010
6.0000	0.7107	0.7573	0.7800	0.2369	0.1900	6.0000	−1.6251	5.3662	0	0.0010
7.0000	0.8111	0.8594	0.8700	0.2211	0.1600	7.0000	−2.2851	8.3236	0	0.0010
8.0000	0.9488	0.9773	0.9900	0.2067	0.1600	8.0000	−1.8786	12.7847	0	0.0010
9.0000	0.9911	1.0163	1.0300	0.2457	0.1900	9.0000	0.0254	24.4068	0	0.0010
10.0000	1.0582	1.0666	1.0700	0.2596	0.1500	10.0000	4.8500	122.5760	0	0.0010
11.0000	1.0975	1.0988	1.1000	0.2322	0.1200	11.0000	8.7122	269.0530	0	0.0010
12.0000	1.0823	1.0896	1.0900	0.1724	0.1100	12.0000	−0.2557	39.8107	0	0.0010
13.0000	1.0955	1.0998	1.1000	0.1633	0.1200	13.0000	−1.5259	24.3516	0	0.0010
14.0000	1.0709	1.0807	1.0800	0.1597	0.1100	14.0000	−1.8895	28.5563	0	0.0010
15.0000	1.0096	1.0282	1.0300	0.1575	0.1000	15.0000	−3.4210	21.3683	0	0.0010
16.0000	0.9690	0.9973	1.0000	0.1774	0.1300	16.0000	−2.9380	15.0985	0	0.0010
17.0000	0.9547	0.9872	1.0000	0.1913	0.1500	17.0000	−2.6087	12.1901	0	0.0010
18.0000	0.9987	1.0209	1.0300	0.1843	0.1400	18.0000	−2.4575	13.7446	0	0.0010
19.0000	1.0861	1.0715	1.0600	0.2388	0.1700	19.0000	0.8480	13.3137	0	0.0010
20.0000	1.1944	1.1275	1.1000	0.3818	0.2500	20.0000	3.1875	20.0743	0	0.0010
21.0000	1.2651	1.1717	1.1500	0.4885	0.2575	21.0000	6.6494	88.1468	0	0.0010
22.0000	1.3302	1.2027	1.1700	0.6446	0.2400	22.0000	8.4558	126.6330	0	0.0010
23.0000	1.2139	1.1298	1.1100	0.4657	0.2000	23.0000	7.0455	75.4992	0	0.0010
24.0000	1.0760	1.0336	1.0200	0.3265	0.1700	24.0000	6.5014	82.1613	0	0.0010

Left: Mean, trimean, media, standard deviation, inter-quartile range. *Right:* Skewness, kurtosis, Kolmogorov-Smirnov, and Jarcke and Bera normality tests

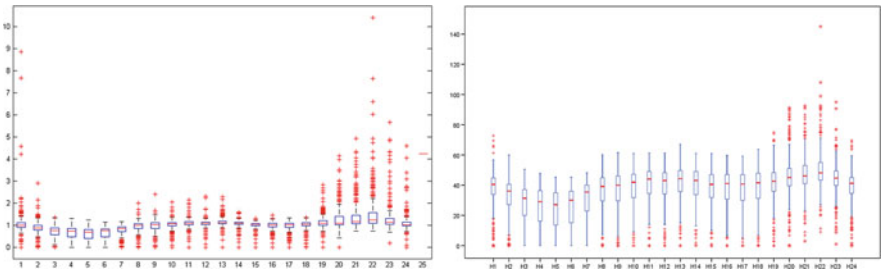


Fig. 2 Boxplot diagrams of rescaled (*left*) and real data (*right*). Time interval: 01.01.2010–31.01.2010

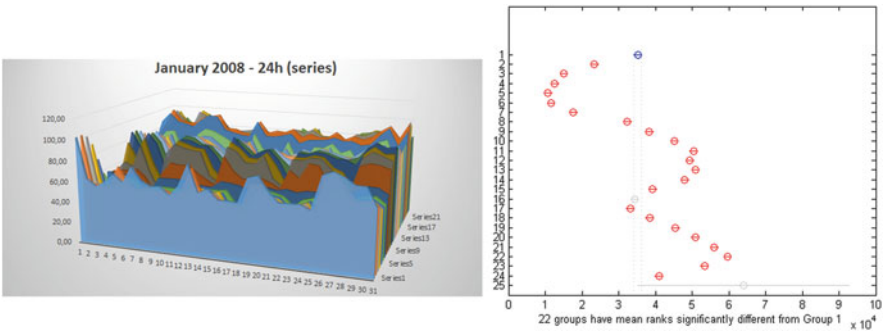


Fig. 3 Real data (01.01.2008–31.01.2008). Left: Different patterns per day. Right: Mean price per hour

Considering the real data, for example from January 2008, we found different patterns per day and per hour (see Fig. 3, left).

The same behavior was found in Fig. 3 (right), where, for example, we can see that 22 groups (hours) have mean ranks significantly different from group 1 (hour 1).

Electricity prices may be influenced by the present and past values of various exogenous factors, such as generation capacity, load profiles, and meteorological conditions [14]. In a preliminary stage we have selected defined and code the following candidates to co-variables: Day of the week—$C_1 = 0, 1, 2, 3, 4, 5, 6$ (Mon, ..., Sunday); Weekday/Saturday/Sunday—$C_2 = 0, 1, 2$; Weekday/Weekend—$C_3 = 0, 1$; Regular day/ holiday—$C_4 = 0, 1$; Season—$C_5 = 0, 1, 2, 3$ (Winter, Spring, Summer, Autumn); Month—$C_6 = 0, \ldots, 11$ (Jan, ..., Dec); Summer/Winter Hour—$C_7 = 0, 1$.

3 GLM Approach

3.1 General Linear Models

In the classical linear model, a vector X with p explanatory variables $X = (X_1, X_2, \ldots, X_p)$ can explain the variability of the variable of interest Y (response variable), where $Y = Z\beta + \epsilon$. Z is a specification matrix with size $n \times p$ (usually $Z = X$, considering a unitary vector in first column), β a parameter vector and ϵ a vector of random errors ϵ_i, independent and identical distributed to a reduced Gaussian.

The data are in the form $(y_i, x_i), i = 1, \ldots, n$, as a result of observation of (Y, X) n times. The response variable Y has expected value $E[Y|Z] = \mu$.

GLM is an extension of classical model where the response variable, following an exponential family distribution [13], does not need to be Gaussian. Another extension from the classical model is that the function which relates the expected value and the explanatory variables can be any differentiable function. Y_i has expected value $E[Y_i|x_i] = \mu_i = b'(\theta_i), i = 1, \ldots, n$.

It is also defined a differentiable and monotone link function g which relates the random component with the systematic component of response variable. The expected value μ_i is related with the linear predictor $\eta_i = z_i^T \beta_i$ using the relation

$$\mu_i = h(\eta_i) = h(z_i^T \beta_i), \qquad \eta_i = g(\mu_i) \tag{1}$$

where h is a differentiable function; $g = h^{-1}$ is the link function; β is a vector of parameter with size p (the same size of the number of explanatory variables); Z is a specification vector with size p.

There are different link functions in GLM. When the random component of response variable has a Poisson distribution, the link function is logarithmic and the model is log-linear. In particular, when the linear predictor $\eta_i = z_i^T \beta_i$ coincides with the canonical parameter θ_i, $\theta_i = \eta_i$, which implies $\theta_i = z_i^T \beta_i$, the link function is denominated as canonical link function. Sometimes, the link function is unknown, for example, in [5] the link function is estimated simultaneously with the linear component of the semi-parametric model for electricity spot prices. A detailed description of GLM methodology can be found in several references such as [7, 13].

3.2 Model Estimation

Initially, to estimate the model as described before, we considered the time interval from 01/01/2008 to 31/12/2010. The first approach using IBM SPSS Statistics (version 22) was performed with difficulty due to the high dimensionality of data. A question that arose was: "Can we reduce the number of components of Y_t?", e.g.,

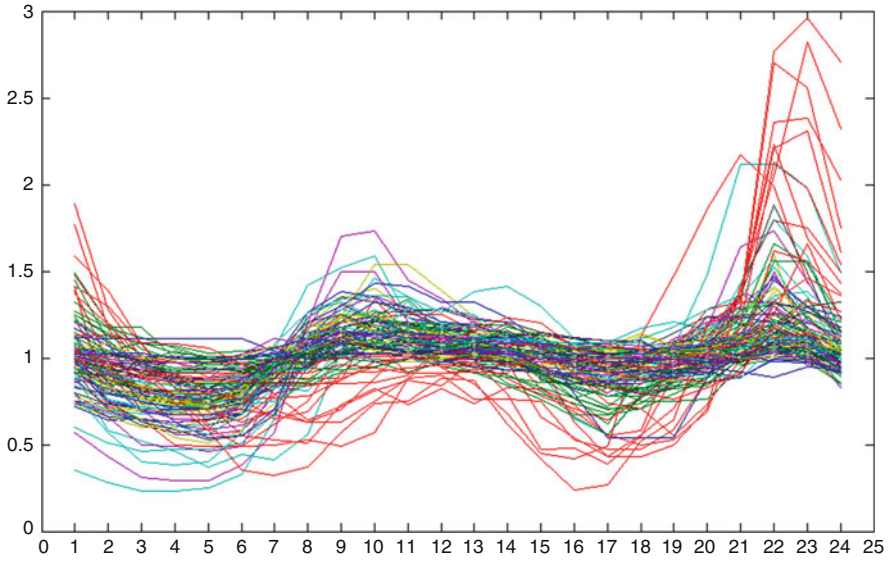


Fig. 4 Data representation—prices per hour (time interval from 01/01/2008 to 31/12/2010)

Are there significant differences between Y_i and Y_j, for $i \neq j$? To solve partially such issue, we try to reduce the 24 h of a day to fewer reference hours. First of all, an analysis of data plot per hour was performed. The graphical representation of data (see Fig. 4) shows similar behavior in some distinct time intervals. Identified such similar hours we merge them into a unique interval of similarity. In this way the dimension of data can be reduced, by taking the mean or median or other measure of response variable.

We have selected and defined some time intervals which conduced to the best model performance. In this way, it was reduced the dimension defining the following time intervals: aurora, lunch time, and dinner time. Aurora corresponds to the hours 3, 4, and 5 respectively. Lunch time merges the hours 11, 12, 13, and 14. Dinner time takes into account hours 17, 18, and 19. When the data is graphically overlapped for each hour in the defined time intervals (see Fig. 5) no significant differences were found.

We studied some possible explanatory variables which can contribute to the explication of energy price per hour. In a preliminary stage of the study, using the initial explanatory variables proposed in Sect. 2, an analysis of variance with second order interaction was performed. The best candidates to explanatory variables of a GLM model were chosen: C_1, C_4, C_5, C_6, C_7.

It was also considered the fare defined by EDP as possible explanatory variable but it was not significant.

The best models were obtained for log or square root link function. The diagnostic analysis and selection of the order of the models was done but we do not reproduce with detail such work. The significant explanatory variables were

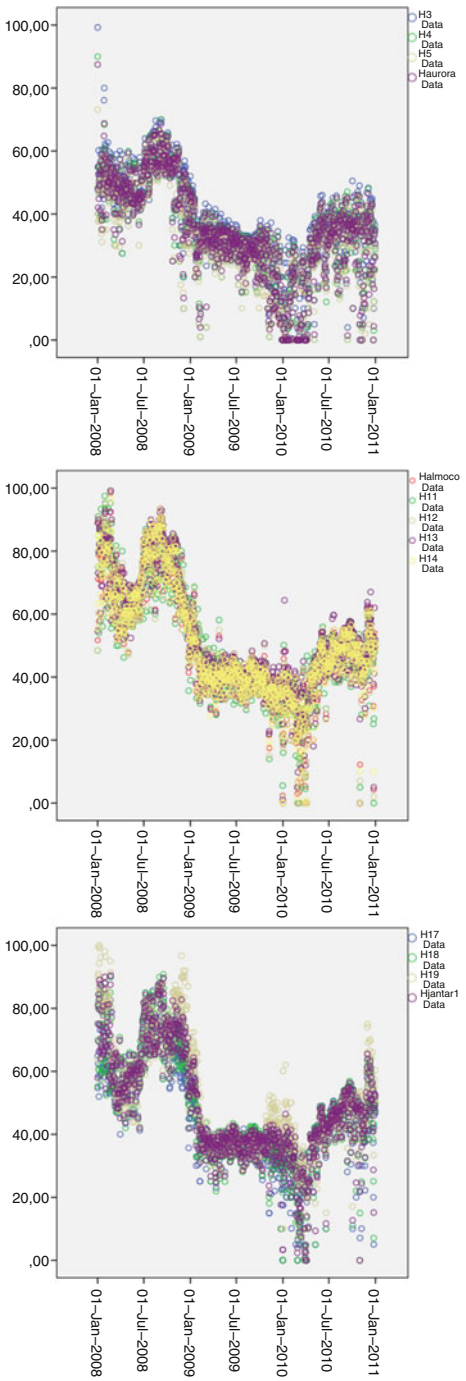


Fig. 5 Overlapped data: aurora time (top), lunch time (center), and dinner time (bottom). Time interval from 01/01/2008 to 31/12/2010

C_4, C_6, C_7, H_2, H_7, H_8, H_{16}, H_{20}, H_{22}, H_{23}, H_{24} and lunch time (link function: square root). When we consider the log as link function, the best explanatory variables were C_4, C_6, C_7, H_2, H_7, H_8, H_{16}, H_{20}, H_{22}, H_{23}, H_{24}. Notice that other transformations should be considered taking into account the time series nature of the data. Eventually, we could get models with better performance.

Considering the obtained results as indicators, we can conclude that some of the explanatory variables proposed initially were not relevant for dependent variable, such as EDP fares, Portuguese holidays (maybe the Iberian holidays can have some relevance, and not just the Portuguese ones). Also, some periods of time can be dropped off as relevant explanatory variables, such as dinner time or some others. The season, month, or winter/summer time period revealed significant explanatory variables in the different estimated models.

Using this preliminary model estimation as starting point, we repeated all estimation process considering a more recent sample so we could compare with the results published in [1]. The GLM model was estimated using hourly prices from 10/03/2014 to 29/5/2016. The remaining sample, from 30/05/2014 to 28/06/2016, was used to evaluate the forecasting performance of the selected model. To assess the in-sample prediction quality of the model, we use the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE).

Following the preliminary model estimation, in models formulation, we considered the response variable with a Gamma distribution and selected the link function with options: 1-log, 2-square root, 3-identity. We have selected as preliminary explanatory variables the same used earlier also considered in [1], where its done a VAR approach. There were estimated the model parameters and analyzed the suitability measures of estimates. The selection and validation of models such as selection of variables, diagnostics, residual analysis and interpretation was concluded. All models obtained good significant results in Likelihood Ratio Chi-Square test, Pearson Chi-Square test, etc. The best models in the sense of performance (estimation and forecasting) are the models with the identity link function. The model A (with higher dimensionality), where each hour of the day is considered, has lower performance in sense of residual analysis and forecasting than model B, where we consider the aurora time, lunch time and dinner time, and the remaining hours (lower dimensionality).

When we analyze the graphics in Fig. 6, we can conclude that model B presents better performance estimation than model A.

From Table 2 we can analyze the quality of prediction in-sample using the MAPE and RMSE. We can conclude that the forecasting quality is promising. In both models (A and B) the prediction performance measures are close, but model B gets better results. Notice that the RMSE values are in accordance with the results obtained using the VAR approach [1].

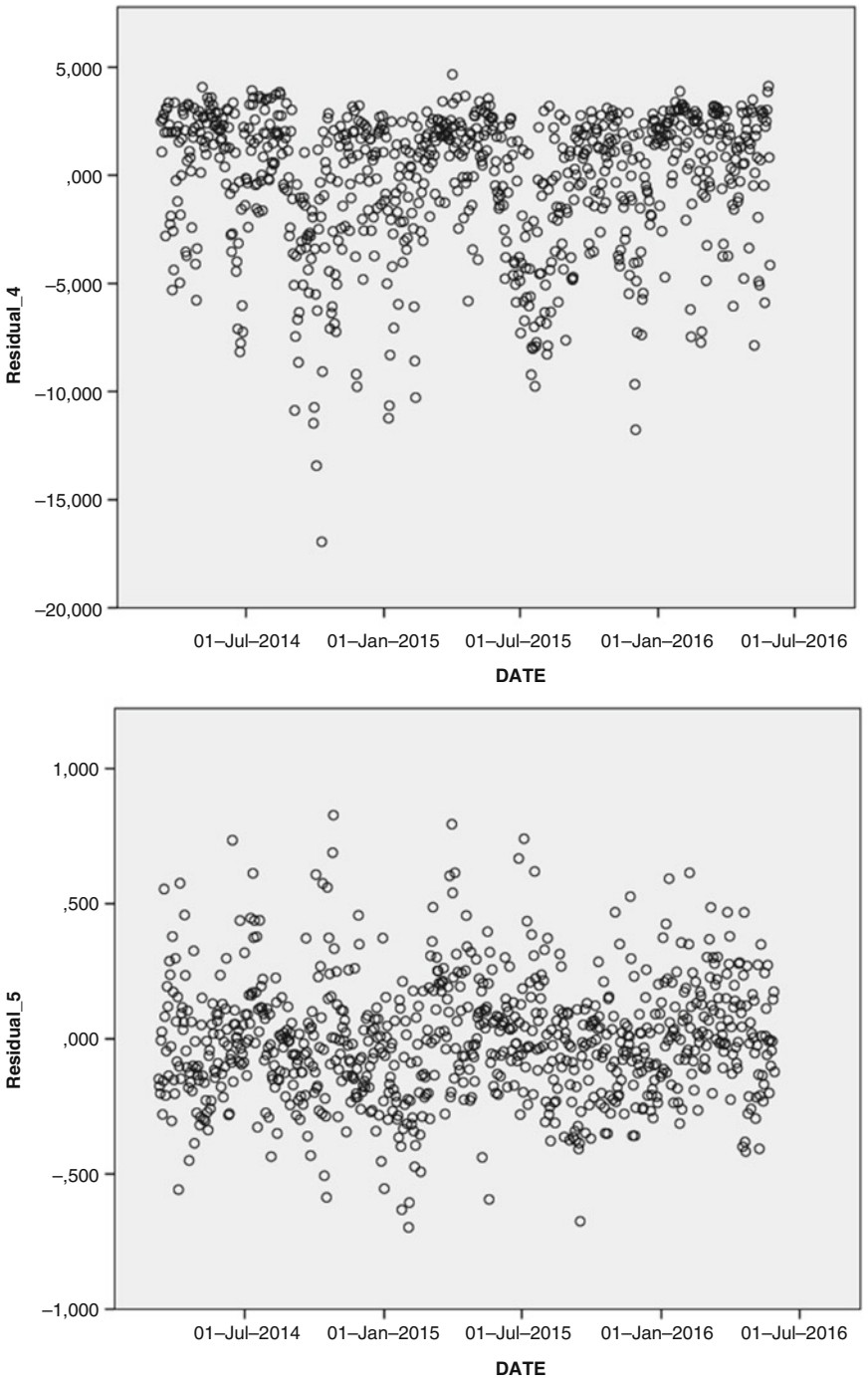


Fig. 6 Residuals representation (×0.01). Left: model A. Right: model B. Estimation period: 10/03/2014 to 29/5/2016

Table 2 MAPE and RMSE

	MAPE (%)	RMSE		MAPE (%)	RMSE		MAPE (%)	RMSE
30/05/2016	7.86	3.45	09/06/2016	7.62	3.85	19/06/2016	39.91	12.44
31/05/2016	5.75	2.61	10/06/2016	9.52	4.94	20/06/2016	7.24	4.03
01/06/2016	8.41	3.58	11/06/2016	14.66	7.08	21/06/2016	8.41	4.24
02/06/2016	8.16	3.76	12/06/2016	14.20	6.94	22/06/2016	9.28	4.09
03/06/2016	4.70	2.52	13/06/2016	8.38	3.83	23/06/2016	11.21	4.88
04/06/2016	3.48	1.79	14/06/2016	15.14	6.43	24/06/2016	10.20	4.43
05/06/2016	11.79	5.08	15/06/2016	15.94	6.19	25/06/2016	29.36	11.58
06/06/2016	4.04	2.07	16/06/2016	26.21	9.55	26/06/2016	55.75	14.17
07/06/2016	8.44	4.73	17/06/2016	20.42	10.47	27/06/2016	16.55	6.44
08/06/2016	4.75	2.91	18/06/2016	15.69	7.12	28/06/2016	18.41	10.18

Forecasting period: 30/05/2016 to 28/06/2016. Model B

4 Conclusions and Recommendations

The challenge proposed by EDP consisted in simulating electricity prices not only for risk measures purposes but also for scenario analysis in terms of pricing and strategy. Data concerning hourly electricity prices from 2008 to 2016 were provided by EDP.

The data were explored using different statistical software, namely IBM SPSS Statistics, Matlab, and R Statistical Software. In this work a GLM approach was considered. The different link functions and the identity case were performed. The season of the year, month, or winter/summer period revealed significant explanatory variables in the different estimated models. We got better results when is considered the reduced form of day hours (aurora time, lunch time, dinner time). From Table 2 we can analyze the quality of prediction in-sample by MAPE and RMSE. We can conclude that the forecasting quality is promising. When compared with multivariate approach using the VAR approach [1] for the same period (from 30/05/2016 to 28/06/2016) the RMSE values are in accordance with the RMSE computed using the VAR method. Although the forecast do not exactly replicate the real price the results are quite promising. The introduction of other co-variables, such as oil price, gas price, wind energy production, other meteorological variables, would certainly improve the model and the forecast. The GLM approach still needs to be improved in the sense of trying other link functions or some differentiation of data. Other methods should be explored. Longitudinal modeling is an approach which has not yet been addressed in Electricity Price Forecasting and deserves our future attention. Univariate time series is other possible future work.

EPF literature has mainly concerned on models that use information at daily level, however this particularly problem proposed is interested in forecasting intra-day prices using hourly data (disaggregated data), maybe it is necessary to consider models that explore the complex dependence structure of the multivariate price series. The problem of modeling distributional properties of energy prices can be classified in three main classes: reduced form models, forward price models, and hybrid price models [2]. Temporal Distribution Extrapolation is another possible idea for our future work.

Acknowledgements This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Portuguese Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2013.

References

1. E Silva, E.C., Borges, A., Teodoro, M.F., Andrade, M.A.P., Covas, R.: Time series data mining for energy prices forecasting: an application to real data. In: Madureira, A., Abraham, A., Gamboa, D., Novais, P. (eds.) Intelligent Systems Design and Applications. ISDA 2016. Advances in Intelligent Systems and Computing, vol. 557, pp. 649–658. Springer Science+Business Media, New York (2017)
2. Eydeland, A., Wolyniec, K.: Energy and Power Risk Management: New Developments in Modeling, Pricing, and Hedging. Wiley, New Jersey (2003)
3. Fu, Y., et al.: ARFNNs with SVR for prediction of chaotic time series with outliers. Expert Syst. Appl. **37**, 4441–4451 (2014)
4. Joens, S.S., et al.: A multivariate time series approach to modeling and forecasting demand in the emergency department. J. Biomed. Inform. **42**, 123–139 (2009)
5. Kocacevic, R.M., Wozabal, D.: A semiparametric model for electricity spot prices. IIE Trans. **46**(4), 344–356 (2014)
6. Kolyshkina, I., Wong, S., Lim, S.: Enhancing generalised linear models with data mining. Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.566.4377. Last viewed in 08/10/2016
7. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall, Londres (1989)
8. Murthy, G.G.P., Sedidi, V., Panda, A.K., Rath, B.N.: Forecasting electricity prices in deregulated wholesale spot electricity market-a review. Int. J. Energy Econ. Policy **4**(1), 32 (2014)
9. Poczos, B., et al.: Nonparametric kernel estimators for image classification. Technical report (2012)
10. Shenoy, S., Gorinevsky, D.: Gaussian-Laplacian mixture for electricity market. In: 2014 IEEE 53rd Annual Conference on Decision and Control (CDC), pp. 158–169. IEEE, New York (2015)
11. Su, L.: Prediction of multivariate chaotic time series with local polynomial fitting. Comput. Math. Appl. **59**(2), 737–744 (2010)
12. Taylor, J.W., Jeon, J.M.: Using conditional kernel density estimation for wind power density forecasting. J. Am. Stat. Assoc. **107**, 66–79 (2012)
13. Turkman, M.A., Silva, G.: Modelos Lineares Generalizados da Teoria á Prática. Sociedade Portuguesa de Estastística, Lisboa (2000)
14. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. **30**(4), 1030–1081 (2014)

Pseudo Maximum Likelihood and Moments Estimators for Some Ergodic Diffusions

Pedro Mota and Manuel L. Esquível

Abstract When $(X_t)_{t\geq 0}$ is an ergodic process, the density function of X_t converges to some invariant density as $t \to \infty$. We will compute and study some asymptotic properties of pseudo moments estimators obtained from this invariant density, for a specific class of ergodic processes. In this class of processes we can find the Cox-Ingersoll & Ross or Dixit & Pindyck processes, among others. A comparative study of the proposed estimators with the usual estimators obtained from discrete approximations of the likelihood function will be carried out.

1 Introduction

Ergodic diffusion processes like the Cox-Ingersoll & Ross [3], the geometric Ornstein-Uhlenbeck or Dixit & Pindyck [4] are widely used in the mathematical finance context, see [2] or [4].

Many times, for ergodic diffusions, the transition density is not known and the parameter estimation is made using approximations of the likelihood function based in some kind of discretization or using martingale estimating functions, see, for instance, [1, 5–7, 12]. In [10], a new parameter estimation technique was presented and applied to the stochastic processes satisfying the following stochastic differential equation,

$$dX_t = b(a - X_t)X_t^{\gamma}dt + \sigma\sqrt{X_t^{\gamma+1}}dB_t, \quad a, b, \sigma > 0, \gamma \geq 0, \tag{1}$$

and for the combination of parameters that makes this processes ergodic.

P. Mota (✉) · M. L. Esquível
Centro de Matemática e Aplicações (CMA), Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Caparica, Portugal
e-mail: pjpm@fct.unl.pt; mle@fct.unl.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_24

The idea was that if $(X_t)_{t\geq 0}$ is an ergodic process then as $t \to \infty$ the density function of X_t converges to the invariant density and then the process parameters can be estimated from the invariant density as if the observations, $X_1, \ldots, X_n$, of the process were independent and identically distributed (i.i.d.) random variables, all of them with the same invariant distribution.

Also in [10], pseudo maximum likelihood estimators were deduced from the invariant density and in the present work we will compute pseudo moments estimators and their asymptotic properties will be studied. In the final section a comparative study, through simulation, will be implemented to compare the pseudo moments estimators with the pseudo maximum likelihood estimators already mentioned and also with the usual estimators obtained from discrete approximations of the transition density.

2 Ergodicity

A continuous time diffusion process

$$dX_t = \mu(X_t; \boldsymbol{\theta})dt + \sigma(X_t; \boldsymbol{\theta})dB_t,$$

with state space $\mathbb{R}$, is said to be ergodic (see, for instance, [9]), if

$$S(x; \boldsymbol{\theta}) = \int_{x_0}^{x} \exp\left(-2\int_{x_0}^{y} \frac{\mu(v; \boldsymbol{\theta})}{\sigma^2(v; \boldsymbol{\theta})}dv\right) dy \to \pm\infty, \text{ as } x \to \pm\infty,$$

and

$$M(\boldsymbol{\theta}) = \int_{-\infty}^{+\infty} \frac{1}{\sigma^2(x; \boldsymbol{\theta})} \exp\left(2\int_{x_0}^{x} \frac{\mu(v; \boldsymbol{\theta})}{\sigma^2(v; \boldsymbol{\theta})}dv\right) dx < \infty,$$

with x_0 an interior point of the state space.
The invariant density is then given by

$$f_{\boldsymbol{\theta}}(x) = \frac{1}{M(\boldsymbol{\theta})\sigma^2(x; \boldsymbol{\theta})} \exp\left(2\int_{x_0}^{x} \frac{\mu(v; \boldsymbol{\theta})}{\sigma^2(v; \boldsymbol{\theta})}dv\right).$$

Theorem 1 *The processes satisfying the stochastic differential equation (1) are ergodic, when* $2ab > \sigma^2(\gamma + 1)$*, with invariant density,*

$$f_{(\alpha,\beta)}(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} \sim Gamma(\alpha, \beta), \text{ with } \alpha = \frac{2ab}{\sigma^2} - \gamma, \quad \beta = \frac{2b}{\sigma^2}.$$

Proof The processes have state space $]0, \infty[$ and with, $\boldsymbol{\theta} = (a, b, \gamma, \sigma)$, we have

$$S(x;\boldsymbol{\theta}) = \int_{x_0}^{x} s(y,\boldsymbol{\theta})dy = \int_{x_0}^{x} \exp\left(-2\int_{x_0}^{y} \frac{b(a-v)v^{\gamma}}{\sigma^2 v^{\gamma+1}}dv\right)dy$$
$$= x_0^{\frac{2ab}{\sigma^2}} e^{-\frac{2b}{\sigma^2}x_0} \int_{x_0}^{x} y^{-\frac{2ab}{\sigma^2}} e^{\frac{2b}{\sigma^2}y} dy \to \pm\infty, x \to +\infty, x \to 0,$$

and

$$M(\boldsymbol{\theta}) = \int_0^{+\infty} \frac{1}{\sigma^2 x^{\gamma+1}} \exp\left(2\int_{x_0}^{x} \frac{b(a-v)v^{\gamma}}{\sigma^2 v^{\gamma+1}}dv\right)dx$$
$$= \frac{x_0^{-\frac{2ab}{\sigma^2}} e^{\frac{2b}{\sigma^2}x_0}}{\sigma^2} \int_0^{\infty} x^{\frac{2ab}{\sigma^2}-\gamma-1} e^{-\frac{2b}{\sigma^2}x} dx < \infty, \text{ if } 2ab > \sigma^2(\gamma+1).$$

The invariant density is then given by

$$f_{\boldsymbol{\theta}}(x) = \frac{x_0^{-\frac{2ab}{\sigma^2}} e^{\frac{2b}{\sigma^2}x_0}}{\sigma^2} x^{\frac{2ab}{\sigma^2}-\gamma-1} e^{-\frac{2b}{\sigma^2}x} \left(\frac{x_0^{-\frac{2ab}{\sigma^2}} e^{\frac{2b}{\sigma^2}x_0}}{\sigma^2} \int_0^{\infty} x^{\frac{2ab}{\sigma^2}-\gamma-1} e^{-\frac{2b}{\sigma^2}x} dx\right)^{-1}$$
$$= \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} \sim Gamma(\alpha,\beta), \text{ with } \alpha = \frac{2ab}{\sigma^2} - \gamma, \ \beta = \frac{2b}{\sigma^2},$$

completing the proof.

3 Estimators and Consistency

If we are working with a strictly stationary ergodic process (for instance, if X_0 have already the invariant distribution), then for any $t > 0$ the random variable X_t will have the invariant distribution. In this framework we propose to deal with the observations of the process like if they were identically distributed with the invariant distribution and then use the invariant density for estimation purposes. From the previous section we know that the processes satisfying the stochastic differential equation (1) with $2ab > \sigma^2(\gamma+1)$ are ergodic with the invariant density $Gamma(\alpha,\beta)$, where $\alpha = \frac{2ab}{\sigma^2} - \gamma$, $\beta = \frac{2b}{\sigma^2}$.

In the following, let us suppose that we have observations $X_1, \ldots, X_n$ of the process, collected at equally spaced times $t_1 < \ldots < t_n$ and that γ and σ are known parameters, that is, the only parameters of interest for estimation purposes are a and b.

3.1 Pseudo Maximum Likelihood Estimators

We can compute pseudo maximum likelihood estimators, that is, defining the likelihood function

$$f_{X_1,\ldots,X_n}(\alpha,\beta;x_1,\ldots,x_n) := \prod_{i=1}^{n} f_{X_i}(\alpha,\beta;x_i)$$

just like in the case of i.i.d. observations and where f_{X_i} is the $Gamma(\alpha,\beta)$ density of Eq. (1).

Since

$$\forall i = 1,\ldots,n, \quad f_{X_i}(\alpha,\beta;x_i) = \frac{x_i^{\alpha-1}e^{-\beta x_i}\beta^{\alpha}}{\Gamma(\alpha)}$$

we get the likelihood function,

$$L(\alpha,\beta;X_1,\ldots,X_n) = \prod_{i=1}^{n} \frac{X_i^{\alpha-1}e^{-\beta X_i}\beta^{\alpha}}{\Gamma(\alpha)}$$

and the log-likelihood,

$$\log(L(\alpha,\beta;X_1,\ldots,X_n)) = (\alpha-1)\sum_{i=1}^{n}\log(X_i) - \beta\sum_{i=1}^{n}X_i + n\alpha\log(\beta) - n\log(\Gamma(\alpha)).$$

From differentiating the log-likelihood function and equating to zero, we get (with $\psi(.)$ the digamma function),

$$\frac{1}{n}\sum_{i=1}^{n}\log(X_i) + \log\left(\frac{2b}{\sigma^2}\right) - \psi\left(\frac{2\bar{X}_n b}{\sigma^2}\right) = 0 \tag{2}$$

with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and

$$a = \bar{X}_n + \frac{\sigma^2\gamma}{2b}$$

getting the estimator $\hat{b}_n$ of b as the solution of the Eq. (2) and the estimator of a, as

$$\hat{a}_n = \bar{X}_n + \frac{\sigma^2\gamma}{2\hat{b}_n}.$$

Theorem 2 *If* $2ab > \sigma^2(\gamma+1)$*, the pseudo maximum likelihood estimators* $\hat{a}_n$ *and* $\hat{b}_n$ *are almost sure (a.s.) consistent estimators for* a *and* b*.*

Proof The proof of the theorem can be found in [10].

3.2 Pseudo Moments Estimators

We can obtain the moments estimators for a and b, using the invariant gamma density and by solving the equations,

$$\begin{cases} \bar{X}_n &= \frac{\alpha}{\beta} \\ M_{2,n} &= \frac{\alpha+\alpha^2}{\beta^2} \end{cases},$$

with $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ the sample mean and $M_{2,n} = \frac{1}{n}\sum_{i=1}^n X_i^2$ the empirical second moment.

Solving these equations we get the moments estimators for the parameters a and b,

$$\tilde{a}_n = \bar{X}_n + \frac{M_{2,n} - \bar{X}_n^2}{\bar{X}_n}\gamma \quad \wedge \quad \tilde{b}_n = \frac{\sigma^2 \bar{X}_n}{2(M_{2,n} - \bar{X}_n^2)}$$

or using the (non-central) sample variance $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$,

$$\tilde{a}_n = \bar{X}_n + \frac{S_n^2}{\bar{X}_n}\gamma \quad \wedge \quad \tilde{b}_n = \frac{\sigma^2 \bar{X}_n}{2S_n^2}.$$

We have the following result about the consistency of the pseudo moments estimators.

Theorem 3 *If* $2ab > \sigma^2(\gamma+1)$*, the pseudo moments estimators* $\tilde{a}_n$ *and* $\tilde{b}_n$ *are a.s. consistent estimators for* a *and* b*.*

Proof Suppose that ξ is a random variable with the invariant gamma density $Gamma(\alpha, \beta)$, where $\alpha = \frac{2ab}{\sigma^2} - \gamma, \beta = \frac{2b}{\sigma^2}$. It is straightforward to prove the consistency of both estimators, since, using the ergodic theorem, we have that

$$\lim_{n\to\infty} \bar{X}_n = \mathbb{E}[\xi] = \frac{\alpha_0}{\beta_0}, \quad a.s.$$

and

$$\lim_{n\to\infty} S_n^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \mathbb{V}[\xi] = \frac{\alpha_0}{\beta_0^2}, \quad a.s.$$

Then,

$$\lim_{n\to\infty}\left(\bar{X}_n+\frac{S_n^2}{\bar{X}_n}\gamma\right)=\frac{\alpha_0}{\beta_0}+\frac{\gamma}{\beta_0}=a_0\quad a.s.$$

and

$$\lim_{n\to\infty}\frac{\sigma^2\bar{X}_n}{2S_n^2}=\frac{\beta_0\sigma^2}{2}=b_0\quad a.s.$$

proving the consistency of the estimators.

Remark 1 We have assumed that σ is known, if σ is unknown the problem of estimating σ can be solved using the quadratic variation of the process, and following [11] we get, a estimator for σ^2,

$$\hat{\sigma}_{1,n}^2=\frac{\sum_{i=1}^{n-1}(X_{i+1}-X_i)^2}{\sum_{i=1}^{n-1}X_i^{\gamma+1}\Delta_n},$$

or following [12]

$$\hat{\sigma}_{2,n}^2=\frac{1}{T}\sum_{i=1}^{n-1}\frac{(X_{i+1}-X_i)^2}{X_i^{\gamma+1}}.$$

4 Simulation and Data Analysis

In this section, we will compare through simulation the moments estimators with the approximate maximum likelihood estimators presented in [10] and the estimators based in discrete approximations of the log-likelihood function. We will suppose that the observations are equally spaced, that is, $t_{i+1}-t_i=\Delta, i=1,\ldots,n$.

The estimators for a and b based on the discretized continuous-time likelihood function, see [1] or [7], $\breve{a}_n$ and $\breve{b}_n$, are given by:

$$\breve{a}_n=\frac{\sum_{i=1}^{n-1}\frac{X_{i+1}-X_i}{X_i}\sum_{i=1}^{n-1}X_i^{\gamma+1}-(X_n-X_1)\sum_{i=1}^{n-1}X_i^{\gamma}}{\sum_{i=1}^{n-1}\frac{X_{i+1}-X_i}{X_i}\sum_{i=1}^{n-1}X_i^{\gamma}-(X_n-X_1)\sum_{i=1}^{n-1}X_i^{\gamma-1}}$$

and

$$\breve{b}_n=\frac{1}{\Delta}\frac{\sum_{i=1}^{n-1}\frac{X_{i+1}-X_i}{X_i}\sum_{i=1}^{n-1}X_i^{\gamma}-(X_n-X_1)\sum_{i=1}^{n-1}X_i^{\gamma-1}}{\sum_{i=1}^{n-1}X_i^{\gamma-1}\sum_{i=1}^{n-1}X_i^{\gamma+1}-\left(\sum_{i=1}^{n-1}X_i^{\gamma}\right)^2}.$$

For simulation purposes, we will perform the generation of the trajectories of the processes using the approximation strong Taylor scheme of order 1.5, see [8].

The iterative scheme used is the following:

$$
\begin{aligned}
Y_{i+1} = Y_i &+ b(a - Y_i)Y_i^{\gamma}\Delta + \sigma Y_i^{\frac{\gamma+1}{2}}\Delta B \\
&+ \frac{\sigma^2(\gamma+1)}{4}Y_i^{\gamma}((\Delta B)^2 - \Delta) + \sigma b(\gamma(a - Y_i) - Y_i)Y_i^{\frac{3\gamma-1}{2}}\Delta Z \\
&+ \frac{1}{2}\left(b^2(a - Y_i)(\gamma(a - Y_i) - Y_i) + \frac{1}{2}b\gamma\sigma^2(\gamma(a - Y_i) - a - Y_i)\right)Y_i^{2\gamma-1}\Delta^2 \\
&+ \left(\frac{\sigma b}{2}(a - Y_i) + \frac{\sigma^3(\gamma-1)}{8}\right)(\gamma+1)Y_i^{\frac{3\gamma-1}{2}}(\Delta B\Delta - \Delta Z) \\
&+ \frac{\sigma^3\gamma(\gamma+1)}{4}Y_i^{\frac{3\gamma-1}{2}}\left(\frac{1}{3}(\Delta B)^3 - \Delta\right)\Delta B,
\end{aligned}
$$

where $\Delta B = \sqrt{\Delta}U_1$, $\Delta Z = \frac{1}{2}\Delta^{3/2}(U_1 + U_2/\sqrt{3})$ and U_1 and U_2 are independent $N(0, 1)$ random variables.

We simulated 500 trajectories and for the estimation of the parameter a we present the results when $n = 500$ in each trajectory, for the parameter b we considered $n = 250$, 500, and 1000 observations in each trajectory. We estimated a and b using the pseudo moments estimators $\tilde{a}_n$ and $\tilde{b}_n$ and we compared them with the pseudo maximum likelihood estimators $\hat{a}_n$ and $\hat{b}_n$ and the estimators $\check{a}_n$ and $\check{b}_n$ obtained from the discretized likelihood function.

We considered $\sigma = 0.1$, we present Table 1 for $\gamma = 0$ and Table 2 for $\gamma = 1$ (for other values of γ we get very similar results), the true value for a is always 1 and for b we considered the values 0.1, 0.5, 1, and 2.

Table 1 Mean and S.D. (standard deviation) for the estimators of a and b when $\gamma = 0$

		$b = 0.1$		$b = 0.5$		$b = 1$		$b = 2$	
Num. obs.	Estimator	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
500	$\tilde{a}_n$	0.9997	0.0447	1.0001	0.0091	1.0001	0.0047	1.0001	0.0026
	$\hat{a}_n$	0.9997	0.0447	1.0001	0.0091	1.0001	0.0047	1.0001	0.0026
	$\check{a}_n$	0.9997	0.0455	1.0001	0.0092	1.0001	0.0047	1.0001	0.0026
250	$\tilde{b}_n$	0.1207	0.0347	0.5147	0.0734	1.0136	0.1151	2.0182	0.2033
500		0.1099	0.0228	0.5051	0.0512	1.0015	0.0819	1.9986	0.1469
1000		0.1037	0.0150	0.4981	0.0351	0.9927	0.0565	1.9851	0.1029
250	$\hat{b}_n$	0.1215	0.0338	0.5228	0.0724	1.0293	0.1137	2.0480	0.2009
500		0.1111	0.0223	0.5125	0.0504	1.0157	0.0809	2.0253	0.1452
1000		0.1050	0.0147	0.5051	0.0347	1.0060	0.0558	2.0099	0.1017
250	$\check{b}_n$	0.1125	0.0308	0.4048	0.0495	0.6400	0.0573	0.8705	0.0611
500		0.1045	0.0211	0.4007	0.0354	0.6375	0.0406	0.8683	0.0428
1000		0.0995	0.0138	0.3967	0.0249	0.6342	0.0284	0.8657	0.0300

Table 2 Mean and S.D. (standard deviation) for the estimators of a and b when $\gamma = 1$

		$b = 0.1$		$b = 0.5$		$b = 1$		$b = 2$	
Num. obs.	Estimator	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
500	$\tilde{a}_n$	0.9988	0.0462	1.0001	0.0093	1.0001	0.0048	1.0001	0.0026
	$\hat{a}_n$	0.9980	0.0461	0.9999	0.0093	1.0000	0.0048	1.0000	0.0026
	$\breve{a}_n$	0.9978	0.0469	1.0000	0.0093	1.0001	0.0048	1.0001	0.0027
250	$\tilde{b}_n$	0.1213	0.0350	0.5169	0.0722	1.0182	0.1143	2.0273	0.2038
500		0.1104	0.0227	0.5075	0.0505	1.0061	0.0816	2.0074	0.1474
1000		0.1041	0.0148	0.5007	0.0347	0.9976	0.0564	1.9943	0.1034
250	$\hat{b}_n$	0.1233	0.0349	0.5253	0.0717	1.0342	0.1131	2.0573	0.2013
500		0.1125	0.0230	0.5150	0.0500	1.0204	0.0807	2.0343	0.1457
1000		0.1059	0.0151	0.5079	0.0346	1.0110	0.0558	2.0192	0.1022
250	$\breve{b}_n$	0.1141	0.0322	0.4035	0.0496	0.6356	0.0572	0.8615	0.0609
500		0.1051	0.0217	0.3994	0.0357	0.6330	0.0410	0.8588	0.0430
1000		0.0997	0.0141	0.3955	0.0250	0.6296	0.0285	0.8561	0.0300

In all the outputs, we can see that the proposed estimators for a and b give good results, very close to the approximate maximum likelihood estimators and we can also see that the estimator for b, $\breve{b}_n$ based in the score function only produce good results when the true value of b is 0.1 (small).

5 Conclusion

In this paper we proposed moments estimators for some ergodic processes. The consistency proof of the proposed estimators and a simulation study to show the applicability of the estimators were provided. For future research, we have the open problem of proving the normality of the asymptotic distribution of the estimators.

Acknowledgements This work was partially supported by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the project UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

1. Bibby, B.M., Sørensen, M.: Martingale estimation functions for discretely observed diffusion processes. Bernoulli **1**(1/2), 17–39 (1995)
2. Bjork, T.: Arbitrage Theory in Continuous Time. Oxford University Press, New York (1998)
3. Cox, J., Ingersoll, J., Ross, S.: A theory of the term structure of interest rates. Econometrica **53**(2), 385–408 (1985)
4. Dixit, A.K., Pindyck, R.S.: Investment Under Uncertainty. Princeton University Press, New Jersey (1994)

5. Florens-Zmirou, D.: Approximate discrete-time schemes for statistics of diffusion processes. Statistics **20**(4), 547–557 (1989)
6. Genon-Catalot, V.: Maximum contrast estimation for diffusion processes from discrete observations. Statistics **21**(1), 99–116 (1990)
7. Kessler, M.: Estimation of an ergodic diffusion from discrete observations. Scand. J. Stat. **24**(2), 211–229 (1997)
8. Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1999)
9. Kutoyants, Y.A.: Statistical Inference for Ergodic Diffusion Processes. Springer, London (2004)
10. Mota, P., Esquível, M.L.: Estimation using the invariant density for some ergodic processes (2018, submitted)
11. Phillips, P.C.B., Yu, J.: A two-stage realized volatility approach to estimation of diffusion processes with discrete data. J. Econometrics **150**, 139–150 (2009)
12. Yoshida, N.: Estimation for diffusion processes from discrete observations. J. Multivar. Anal. **41**(2), 220–242 (1992)

Statistical Modelling of Counts with a Simple Integer-Valued Bilinear Process

Isabel Pereira and Nélia Silva

Abstract The aim of this work is the statistical modelling of counts assuming low values and exhibiting sudden and large bursts that occur randomly in time. It is well known that bilinear processes capture these kind of phenomena. In this work the integer-valued bilinear INBL(1,0,1,1) model is discussed and some properties are reviewed. Classical and Bayesian methodologies are considered and compared through simulation studies, namely to obtain estimates of model parameters and to calculate point and interval predictions. Finally, an empirical application to real epidemiological count data is also presented to attest for its practical applicability in data analysis.

1 Introduction

In the analysis of stationary integer-valued time series the class of INARMA models plays a central role. However, such models are unlikely to provide a sufficiently broad class capable of accurately capturing features often exhibited by data sets such as sudden burst of large values. For that purpose and using the concept of thinning operator, introduced by [12], conventional bilinear models can be adapted to the integer case leading to the class of integer-valued bilinear models. Doukhan et al. [3] proposed the first-order INBL(1,0,1,1) model

$$X_t = \alpha \circ X_{t-1} + \beta \circ (\epsilon_{t-1} X_{t-1}) + \epsilon_t. \tag{1}$$

I. Pereira (✉) · N. Silva
Departamento de Matemática and CIDMA, University of Aveiro, Aveiro, Portugal
e-mail: isabel.pereira@ua.pt; neliasilva@ua.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_25

where the thinning operators "$\alpha\circ$" and "$\beta\circ$"[1] are mutually independents, $\{\epsilon_t\}_{t\in Z}$ is a sequence of i.i.d. non-negative integer-valued random variables with finite mean and finite variance, independent of the operators. Doukhan et al. [3] derived conditions guaranteeing strictly and second-order stationarities of INBL(1,0,1,1) model. Drost et al. [4] also provided sufficient conditions for the existence of higher order moments of X_t, considering the superdiagonal INBL(p, q, m, n). One step towards the application of bilinear models to real data sets is the estimation of parameters. Considering Poisson thinning operators [3] have obtained moments estimators and derived their asymptotic distribution. In contrast, Bayesian analysis of INBL has not received much attention in the literature neither diagnostic analysis.

In this paper we consider the INBL(1,0,1,1) given in (1), with the following assumptions: the operators "$\alpha\circ$" and "$\beta\circ$" are mutually independents such as $\alpha\circ X_{t-1}|X_{t-1} \frown Bi(X_{t-1},\alpha)$, $\beta\circ(\epsilon_{t-1}X_{t-1})|X_{t-1},\epsilon_{t-1} \frown Bi(\epsilon_{t-1}X_{t-1},\beta)$ and $\{\epsilon_t\}_{t\in Z}$ is a sequence of i.i.d. Poisson random variables with mean λ, independent of the operators.

The class of stationary models defined in (1) is useful for representing time series that assume low values with high probability and exhibit sudden bursts of large values that occur randomly in time, hence can produce heavy-tailed data. As an illustration of this kind of data we present in Fig. 1 two time series of count data in epidemiology, originally studied by [3]. Data consist of the weekly number of E. coli infections and meningitis cases, both starting in January 1990 and corresponding to 143 observations for each series. Counts are typically small, skewed and both series contain a large quantity of zeros. Hereafter the weekly number of E.coli infections and weekly number of meningitis cases are denoted by the E.coli data and meningitis data, respectively.

In time series analysis we usually are interested in estimating the underlying model and in predictive capabilities of that model. Thus, the aim of this study is to establish a comparison between classical and Bayesian approaches in order to

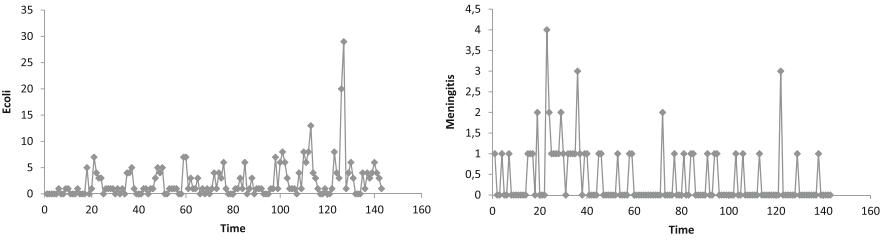


Fig. 1 Time series plots for E. coli data (*left*) and meningitis data (*right*)

[1] Steutel and van Harn operator "$\phi\circ$" is defined by $\phi\circ X=\sum_{i=1}^{X} Y_i$ where $\{Y_i\}$, $i=1,\ldots,X$, is a sequence of independent and identically distributed (i.i.d.) counting random variables with mean ϕ and X is a non-negative integer-valued random variable, independent of Y. If Y_i is a Bernoulli random variable, we have the binomial thinning operator.

conduct inference for model parameters and obtain predictions for future values. The rest of the paper is organized as follows. Classical and Bayesian methodologies are presented to obtain estimates of the model parameters in Sect. 2 and forecasting is addressed in Sect. 3. The performance of the above procedures is illustrated through a simulation study in Sect. 4. Section 5 provides applications of this model to real data sets. Finally, Sect. 6 contains some concluding remarks.

2 Parameters Estimation

One step forward the application of INBL models in practice is the estimation of their corresponding parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) = (\alpha, \beta, \lambda)$. The classical estimators studied are grouped according to two broad categories: regression-based and likelihood-based estimators. Furthermore, Bayesian estimation is also considered. In any estimation procedure, it is required to estimate the r.vs ϵ_t since they are not observable. From (1), the innovations ϵ_t can be recursively calculated through $\epsilon_t = X_t - \alpha \circ X_{t-1} + \beta \circ (\epsilon_{t-1} X_{t-1})$, $t = 1, \ldots, n.$, using an initial value for ϵ_1.

2.1 Conditional Least Squares Estimators

The CLS-estimators of $\boldsymbol{\theta}$ are obtained by minimizing

$$Q(\boldsymbol{\theta}) = \sum_{t=2}^{n} [X_t - E(X_t | X_{t-1}, \epsilon_{t-1})]^2 = \sum_{t=2}^{n} [X_t - \alpha X_{t-1} - \beta X_{t-1}\epsilon_{t-1} - \lambda]^2 ,$$

yielding to the following expressions for the parameters estimators

$$\begin{aligned}
\hat{\beta}_{\text{CLS}} &= \frac{S_{t;t-1} S_{t-1;(t-1,\epsilon-1)} - S_{t-1;t-1} S_{t;(t-1,\epsilon-1)}}{S^2_{t-1;(t-1,\epsilon-1)} - S_{t-1,\epsilon-1} S_{t-1;t-1}}, \\
\hat{\alpha}_{\text{CLS}} &= \frac{(n-1) S_{t;t-1} - \hat{\beta}_{\text{CLS}} S_{t-1;(t-1,\epsilon-1)}}{S_{t-1;t-1}}, \\
&\text{and} \\
\hat{\lambda}_{\text{CLS}} &= \frac{\sum_{t=2}^{n} X_t - \hat{\alpha}_{\text{CLS}} \sum_{t=2}^{n} X_{t-1} - \hat{\beta}_{\text{CLS}} \sum_{t=2}^{n} X_{t-1}\epsilon_{t-1}}{n-1},
\end{aligned}$$

with $i, j = 0, 1$,

$$\begin{aligned}
&\bar{X}_j = \tfrac{1}{n-1} \textstyle\sum_{t=2}^{n} X_{t-j} \\
&S_{t-i,t-j} = \textstyle\sum_{t=2}^{n} (X_{t-i} - \bar{X}_i)(X_{t-j} - \bar{X}_j) \\
&S_{t-i;(t-1,\epsilon-1)} = \textstyle\sum_{t=2}^{n} (X_{t-i} - \bar{X}_i)(X_{t-1}\epsilon_{t-1} - \bar{X}_1 \bar{\epsilon}_1), \quad \bar{\epsilon}_1 = \tfrac{1}{n-1} \textstyle\sum_{t=2}^{n} \epsilon_{t-1}, \\
&S_{t-1,\epsilon-1} = \textstyle\sum_{t=2}^{n} (X_{t-1}\epsilon_{t-1} - \bar{X}_1 \bar{\epsilon}_1)^2.
\end{aligned}$$

2.2 Conditional Maximum Likelihood Estimators

For fixed values of x_1 and ϵ_1, the conditional log-likelihood function for the INBL(1,0,1,1) model is given by

$$l(\boldsymbol{\theta}) := \ln L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1) = \sum_{t=2}^{n} \ln(p(x_t|x_{t-1}, \epsilon_{t-1})),$$

with $\mathbf{x}_n = (x_1, \ldots, x_n)$ and transition probabilities

$$p(x_t|x_{t-1}, \epsilon_{t-1}) = \sum_{k=0}^{m_t} \exp(-\lambda) \frac{\lambda^{x_t-k}}{(x_t-k)!} \times \sum_{j=M_t}^{m_{tk}} \binom{x_{t-1}}{j} \alpha^j (1-\alpha)^{x_{t-1}-j} \binom{x_{t-1}\epsilon_{t-1}}{k-j} \beta^{k-j} (1-\beta)^{x_{t-1}\epsilon_{t-1}-k+j},$$

$m_t = \min(x_t, x_{t-1} + x_{t-1}\epsilon_{t-1})$, $M_t = \max(0, k - x_{t-1}\epsilon_{t-1})$ and $m_{tk} = \min(k, x_{t-1})$, since the r.v. $X_t|x_{t-1}, \epsilon_{t-1}$ is the convolution between binomial distributions with parameters (X_{t-1}, α) and $(\epsilon_{t-1}X_{t-1}, \beta)$, respectively, and Poisson distribution with parameter λ. The CML-estimators are obtained by maximizing the conditional log-likelihood function. Due to the complexity of log-likelihood expression it is not possible to give explicit forms to the CML-estimators of α, β, and λ, thus it is necessary to use numerical procedures. The initial estimates required by such numerical procedures can be obtained by the method of least squares or using moment estimates given by [3] procedure.

2.3 Bayesian Approach

To implement the Bayesian version of the INBL(1,0,1,1) model we need to consider prior distributions for the parameters.Thus, for the parameters $0 < \alpha, \beta < 1$ we choose Beta priors with hyperparameters (a, b) and (c, d), respectively while for the positive parameter λ we choose a Gamma prior with hyperparameters (e, f). These priors are traditionally used for the PoINAR(1) by [11].

Given the particular sample $\mathbf{x}_n$, the updated information about $\boldsymbol{\theta}$ is expressed through Bayes theorem by the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$ given by

$$\pi(\boldsymbol{\theta}|\mathbf{x}_n) = \frac{L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1)\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1)\pi(\boldsymbol{\theta}),$$

with $\pi(\boldsymbol{\theta})$ representing the prior distribution.

Assuming independence assumptions on the parameters the posterior distribution is given by

$$\pi(\boldsymbol{\theta}|\mathbf{x}_n) \propto \lambda^{e-1}e^{-f\lambda}\alpha^{a-1}(1-\alpha)^{b-1}\beta^{c-1}(1-\beta)^{d-1}\left(\prod_{t=2}^{n}\sum_{k=0}^{m_t}\exp(-\lambda)\right.$$

$$\left.\frac{\lambda^{x_t-k}}{(x_t-k)!}\sum_{j=M_t}^{m_{tk}}\binom{x_{t-1}}{j}\alpha^j(1-\alpha)^{x_{t-1}-j}\binom{x_{t-1}\epsilon_{t-1}}{k-j}\beta^j(1-\beta)^{x_{t-1}\epsilon_{t-1}-k+j}\right),$$

with $\lambda > 0$, $0 < \alpha, \beta < 1$, $m_t = \min(x_t, x_{t-1} + x_{t-1}\epsilon_{t-1})$, $M_t = \max(0, k - x_{t-1}\epsilon_{t-1})$, and $m_{tk} = \min(k, x_{t-1})$.

Thus given the complexity of the posterior distribution, Markov Chain Monte Carlo (MCMC) techniques are required for sampling purposes. For the simulations we need the full conditional distributions for each parameter θ_i, denoted by $\pi(\theta_i|\boldsymbol{\theta}_{-i}, \mathbf{x}_n)$, which is the posterior distribution of θ_i conditional on all other parameters and the data $\mathbf{x}_n$. The full conditional distributions of α, β, and λ are, respectively:

$$\pi(\alpha|\beta, \lambda, \mathbf{x}_n) \propto \alpha^{a-1}(1-\alpha)^{b-1}L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1),$$

$$\pi(\beta|\alpha, \lambda, \mathbf{x}_n) \propto \beta^{c-1}(1-\beta)^{d-1}L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1)$$

and

$$\pi(\lambda|\alpha, \beta, \mathbf{x}_n) \propto \lambda^{e-1}e^{-f\lambda}L(\mathbf{x}_n; \boldsymbol{\theta}|x_1, \epsilon_1).$$

From the above expressions it is easy to conclude that the full conditional distributions will not be standard distributions and therefore a componentwise Metropolis- Hastings algorithm is used, particularly the Adaptive Rejection Metropolis Sampling (ARMS), as described in [7]. After having generated samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N)}$ sample central tendency measures are used to estimate the model parameters.

3 Prediction Future Observations

In this section we consider the problem of predicting the values of X_{n+h}, $h \in N$ for INBL (1,0,1,1) process based on the observed series up to time n. The usual way of producing forecasts is via the conditional predictive distribution and the most common procedure for obtaining predictions in time series models is to use conditional expectations, since we pretend to minimize the mean square error. Throughout this section we consider $\mathbf{B}_n = \{X_1, \ldots, X_n; \epsilon_1, \ldots, \epsilon_n\}$.

3.1 Classical Approach

3.1.1 Point Predictions

For $h \geq 1$ the h-step-ahead predictor can be obtained in a recursive way through

$$\hat{X}_{n+h} = \mathrm{E}(X_{n+h}|\mathbf{B}_n) = \alpha \hat{X}_{n+h-1} + \beta \mathrm{E}(\epsilon_{n+h-1} X_{n+h-1}|\mathbf{B}_n) + \lambda \tag{2}$$

with

$$\begin{aligned}\mathrm{E}(\epsilon_{n+h-1} X_{n+h-1}) &= \mathrm{E}[\epsilon_{n+h-1}(\alpha \circ X_{n+h-2} + \beta \circ (\epsilon_{n+h-2} X_{n+h-2}) + \epsilon_{n+h-1})|\mathbf{B}_n] \\ &= \alpha\lambda \hat{X}_{n+h-2} + \lambda\beta \mathrm{E}(\epsilon_{n+h-2} X_{n+h-2}) + \lambda^2 + \lambda.\end{aligned}$$

For the particular cases $h = 1$ and $h = 2$ we have, respectively,

$$\begin{aligned}\hat{X}_{n+1} &= \alpha X_n + \beta(X_n \epsilon_n) + \lambda, \\ \hat{X}_{n+2} &= \alpha^2 X_n + \alpha\beta(X_n \epsilon_n) + \alpha\lambda + \alpha\beta\lambda X_n + \beta^2 \lambda(X_n \epsilon_n) + \beta(\lambda^2 + \lambda) + \lambda.\end{aligned}$$

We can easily prove that

$$\lim_{h \longrightarrow +\infty} \hat{X}_{n+h} = \frac{\beta\lambda^2 + \beta\lambda + \lambda(1 - \lambda\beta)}{(1 - \lambda\beta)(1 - \alpha) - \alpha\beta\lambda}. \tag{3}$$

Since these predictors based on conditional expectation hardly produce integer-valued forecasts, we can alternatively use the median of h-step-ahead conditional distribution of $X_{n+h}|\mathbf{B}_n$, denoted by $\hat{M}_{n+h}$, to obtain coherent predictions of X_{n+h}, as suggested in [5].

3.1.2 Prediction Intervals for One-Step-Ahead Observation

The one-step-ahead prediction error

$$e_{n+1} = X_{n+1} - \hat{X}_{n+1} = X_{n+1} - \hat{\alpha} X_n - \hat{\beta}(X_n \epsilon_n) - \hat{\lambda}$$

is a discrete variable with probability function

$$\begin{aligned}&P(e_{n+1} = x_{n+1} - g) = P(X_{n+1} = x_{n+1}|\mathbf{B}_n) = f(x_{n+1}|\mathbf{B}_n; \theta) \\ &= \sum_{k=0}^{m_1} \exp(-\lambda) \frac{\lambda^{x-k}}{(x-k)!} \times \sum_{l=M_1}^{m_2} \binom{x_n}{l} \alpha^l (1-\alpha)^{x_n - l} \times \binom{x_n \epsilon_n}{k-l} \beta^{k-l} (1-\beta)^{x_n \epsilon_n - k + l}\end{aligned}$$

where $g = \hat{\alpha} X_n + \hat{\beta}(X_n \epsilon_n) + \hat{\lambda}$. Hence the γ level confidence interval for X_{n+1} is given by: $(\hat{X}_{n+1} + l_1, \hat{X}_{n+1} + l_2)$ where l_1 is the largest value of e_{n+1} such as $\mathrm{P}(e_{n+1} \leq l_1) \leq (1 - \gamma)/2$ and l_2 the smallest value of e_{n+1} such as $\mathrm{P}(e_{n+1} \leq l_2) \geq (1 + \gamma)/2$.

3.2 Bayesian Predictions

The Bayesian predictive probability function is based on the assumption that both the future observation X_{n+h} and $\boldsymbol{\theta}$ are unknown. The conditional distribution of X_{n+h} given $\mathbf{B}_n$ which can be viewed as containing all the accumulated information about the future, represents the h-step-ahead Bayesian posterior predictive distribution. It is defined by

$$\pi(x_{n+h}|\mathbf{B}_n) = \int_{\boldsymbol{\Theta}} f(x_{n+h}|\mathbf{B}_n;\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}_n)d\boldsymbol{\theta},$$

with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ being the vector of unknown parameters, $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$ the posterior density of $\boldsymbol{\theta}$, and $f(x_{n+h}|\mathbf{B}_n;\boldsymbol{\theta})$ the (classical) predictive distribution.

3.2.1 Point Predictions for the Future Observation

In the particular case of $h = 1$ the one-step-ahead Bayesian predictive distribution is given by

$$\pi(x_{n+1}|\mathbf{B}_n) == \int_\alpha \int_\beta \int_\lambda f(x_{n+1}|\mathbf{B}_n;\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}_n)d\alpha d\beta d\lambda,$$

with $m_1 = \min(x_{n+1}, x_n + \epsilon_n x_n)$, $m_2 = \min(x_n, k)$, $M_1 = \max(0, k - x_n\epsilon_n)$.

The Bayesian predictor of X_{n+1} can be obtained by any location measure of the predictive distribution. Its complexity does not allow work with it directly. However we can adapt to the integer case the Tanner composition method (as reported in [13]), to get an estimate of X_{n+h} using the sample mean or sample median of the generated values $(X_{n+h,1}, \ldots, X_{n+h,m})$. Similarly to the classical case we can use the recursive expression

$$\begin{aligned} E(X_{n+h}|\mathbf{B}_n) &= E[E(X_{n+h}|\boldsymbol{\theta}, \mathbf{B}_n)] = E\Big[E\Big(\alpha \circ X_n + \beta \circ (\epsilon_n X_n) + \epsilon_{n+1}\Big|\boldsymbol{\theta}, \mathbf{B}_n\Big)\Big] \\ &= \hat{\alpha}_B \hat{X}_{n+h-1} + \hat{\beta}_B E(\epsilon_{n+h-1} X_{n+h-1}|\mathbf{B}_n) + \hat{\lambda}_B \end{aligned} \tag{4}$$

where $\hat{\alpha}_B$, $\hat{\beta}_B$, and $\hat{\lambda}_B$ are the Bayesian estimates of the parameters. It is worth to mention that there is no need to do any plug-in as happened in classical approach.

3.2.2 HPD Predictive Intervals

We use an adaptive generalization of the method used to obtain Highest Posterior Density (HPD) intervals of model parameters, considering predictive distribution

instead of the posterior. Hence the 100 $\gamma\%$ HPD predictive interval for X_{n+1} is defined by $R(\gamma) = (X_L, X_U)$ if

$$P(X_L \leq X_{n+1} \leq X_U) = \sum_{x_{n+1}=X_L}^{X_U} \pi(x_{n+1}|\mathbf{B}_n) \geq K_\gamma,$$

with K_γ being the largest constant such as $P[X_{n+1} \in R(\gamma)] \geq \gamma$.

We can obtain an approximation to the HPD predictive interval for X_{n+1} using the algorithm developed by [1]. After computing the 100 $\gamma\%$ credible intervals

$$\hat{R}_i(\gamma) = \left(X_{n+1,i} \ , \ X_{n+1,i+[m\gamma]}\right), \ \ 1 \leq i \leq m - [m\gamma],$$

where $[m\gamma]$ is the integer part of $m\gamma$, the 100 $\gamma\%$ HPD interval, denoted by $\hat{R}(\gamma)$ is the one with the smallest amplitude.

4 Simulation Study

In this section we study the performance of the above classical and Bayesian procedures with count time series simulated by choosing various combinations of the parameters of INBL(1,0,1,1) model under stationarity conditions.

4.1 Inference

Through the simulation study we want to highlight the following issues: (a) how the results depend on the underlying bilinear parameter β; (b) what is the impact of sample size on the simulation results, and (c) what is the influence of the variance of the innovation process.

We simulated samples from INBL(1,0,1,1) model of length $n = 50$, 100, and 500 with 100 independent replicates.[2] In the absence of prior information we consider non-informative priors letting the hyperparameters equal to 0.0001. The MCMC algorithm was used with starting values based on the CLS-estimates and was run with 31,000 iterations in total, the 11,000 initial burn-in iterations were discarded and only the 20th value of the last iterations is kept to reduce the autocorrelation within the chain. Nevertheless, the stationarity of the chain and the convergence of the algorithm were duly analyzed with the usual diagnostic tests, respectively,

[2]The computation of Bayesian estimates is very demanding in terms of CPU time. Using an Inter Core i5 @ 1.8 GHz-4 GB RAM, the average computation time for producing the estimates of the parameters for samples with size $n = 100$ is approximately 3 days.

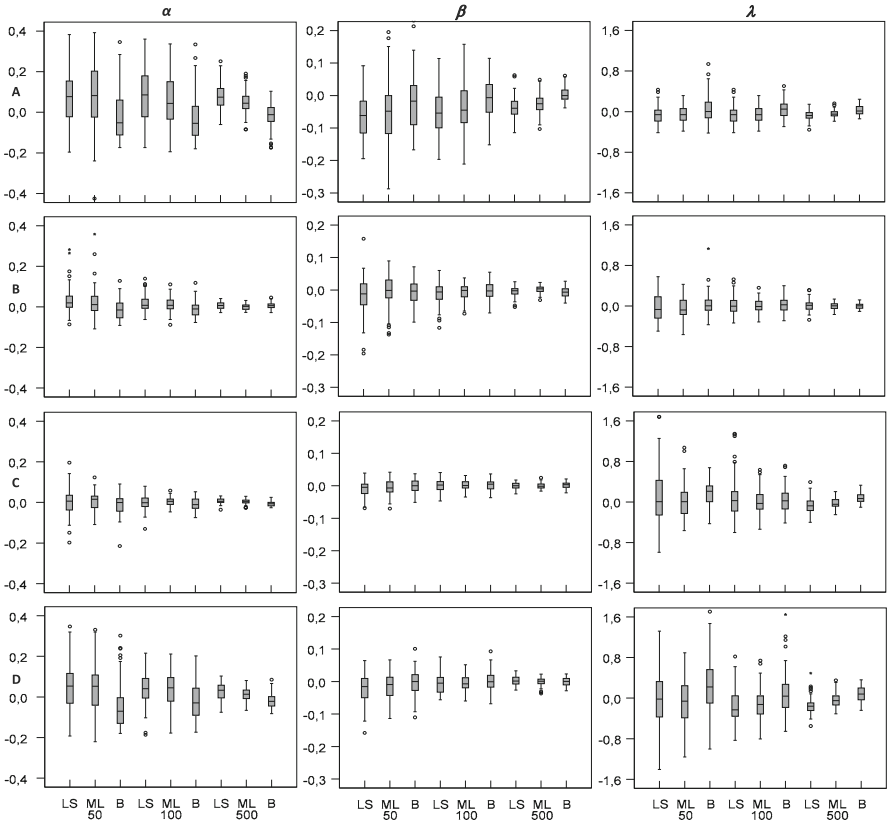


Fig. 2 Boxplots of the biases for $\boldsymbol{\theta} = (\alpha, \beta, \lambda)$ in models A:=(0.2,0.2,1), B:= (0.1,0.6,1), C:=(0.7,0.2,1) and D:=(0.2,0.2,2), with $n = 50, 100, 500$

[8] and [6] tests, which are available in package CODA. Figure 2 displays the boxplots of the biases of CLS-estimates, CML-estimates, and Bayesian estimates for $\boldsymbol{\theta}$, considering each model and the variation of sample size. Concerning the estimation of α a closer look at the figure reveals that classical estimators tend to overestimate the autoregressive parameter, in particular for the models A and D. On the other hand, β is underestimated by any methodology in models A, B, and D. Nevertheless considering all the parameters β is the one that is estimated more accurately. A comparison of the dispersion for the classical and Bayesian estimators shows the similarity for both small and large sample sizes. An important conclusion is that the value of the underlying bilinear parameter does not seem to interfere with the quality of the point estimates for this model. However the variance of the innovation process has large biases, which increases when the theoretical value of λ parameter rises, showing a significant degree of variability. As expected, both the bias and the skewness are also reduced when the sample size increases.

4.2 Prediction

To compare and analyze the different h-step-ahead predictors previously mentioned in Sect. 3 we simulated samples with sizes $n = 50, 100, 200$ from model (1). In order to obtain point or interval forecasts from classical approach the CML estimates were plugged in in (2) or in the predictive probability functions. To obtain Bayesian predictions, we used the expression (4) and Tanner algorithm [13]. It is worth to notice that the forecast performance depends on one hand, on the difference between x_n and x_{n+h}, $h \geq 1$, similarly to what happens in INAR(1) model as described by [11] and on the other hand, on the prediction errors e_n. This situation is illustrated as follows: the forecasts of $x_{168} = 8$ are $\hat{x}_{168,\mathrm{CML}} = 12.684$, $\hat{x}_{168,B} = 13.010$, $\hat{M}_{168}^{(\mathrm{CML})} = \hat{M}_{168}^{(B)} = 13$ are closer to $x_{167} = 13$, when $\alpha = 0.7$, $\beta = 0.2$, $\lambda = 1$.

In Fig. 3 the h-step-ahead predictions, considering two particular sets of parameters and $n = 100$, are plotted. These plots indicate that the obtained results for the predictions using the classical approach with CML-estimates and the Bayesian methodology are very similar. It must be emphasized that these predictions are closed to the limit values given by (3), corresponding to 5.33 for $\boldsymbol{\theta} = (0.1, 0.6, 1)$ and to 12 when $\boldsymbol{\theta} = (0.7, 0.2, 1)$. Figures 4 and 5 represent the amplitude means of the prediction intervals or the HPD predictive intervals for the future value and the frequencies of the simulated X_{n+1} belonging to the prediction interval, respectively. We observe that in general classical prediction intervals based on CML-estimates present smaller amplitude means than the Bayesian correspondents. Another

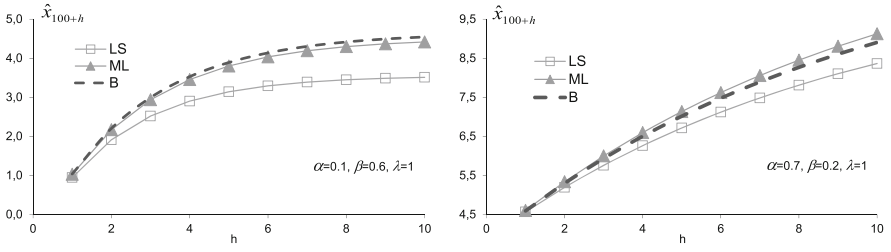


Fig. 3 h-step-ahead predictions for future observations

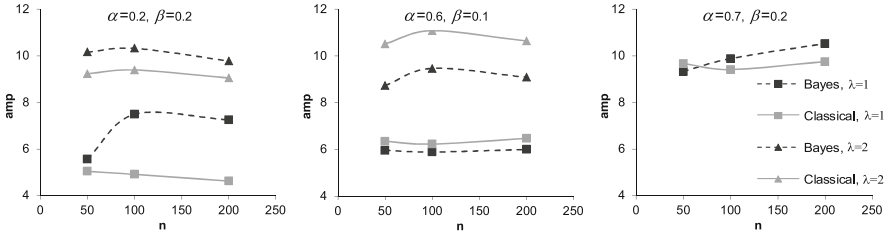


Fig. 4 Means of the prediction interval amplitudes of X_{n+1}

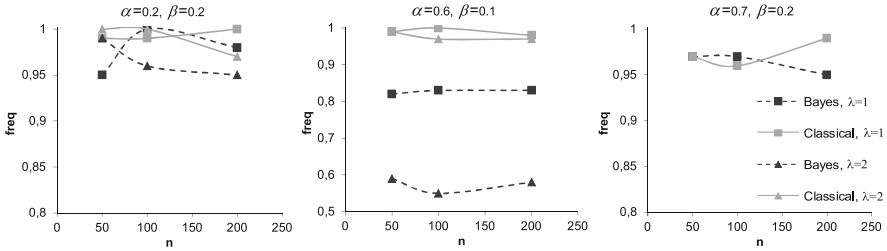


Fig. 5 Frequencies of X_{n+1} belonging to the prediction intervals

important feature exhibited is that the percentage of the simulated observation X_{n+1} belonging to the classical interval is greater than 96.

5 Application to Real Data

In this section, we illustrate the modelling procedure with the motivating examples presented in Fig. 1. It could be pointed out that both data sets are asymmetric with significant overdispersion in E.coli data, with empirical mean and variance being 2.3 and 13.03, respectively.

We should check the adequacy of the distributional assumptions of the model. For this purpose we use the nonrandomized version of PIT histogram, proposed by [2] (see [9], for further models evaluation based in its predictive performance). The graphical tools represented in Fig. 6 are the PIT histograms and the mean PIT charts applied to the data sets. From left to right, the PIT histograms are U-shaped and uniform indicating underdispersed and well-calibrated predictive distributions, respectively. These plots indicate that the probability structure addressed to the INBL(1,0,1,1) is misspecified in the E.coli data despite the Pearson residuals exhibit mean 0.0002, variance 0.9999 and no significant serial correlation. Results of the parameter estimates for the meningitis data are presented in Table 1. However the bilinear component β in the model seems to be very small, which may question its interest in the model. Finally, in order to evaluate and compare the different

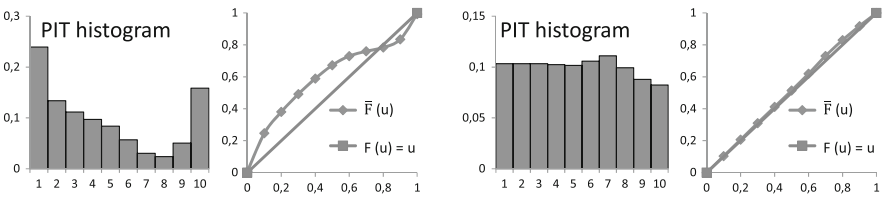


Fig. 6 PIT histograms and charts of mean PIT (denoted by $\bar{F}(u)$), applied to E.coli data (*left*) and meningitis data (*right*)

Table 1 Estimated model for the meningitis data

Data	$\hat{\alpha}_{\text{CLS}}$	$\hat{\beta}_{\text{CLS}}$	$\hat{\lambda}_{\text{CLS}}$	$\hat{\alpha}_{\text{CML}}$	$\hat{\beta}_{\text{CML}}$	$\hat{\lambda}_{\text{CML}}$	$\hat{\alpha}_B$	$\hat{\beta}_B$	$\hat{\lambda}_B$
Meningitis data	0.151	0.027	0.296	0.201	0.026	0.288	0.181	0.029	0.290

Table 2 h-step-ahead predictions using meningitis data

h	1	2	3
$\hat{X}^{(\text{CML})}_{140+h}$	0.296	0.357	0.369
$\hat{X}^{(B)}_{140+h}$	0.308	0.361	0.368
x_{140+h}	0	0	0

predictions methodologies, h-step-ahead forecasts ($h = 1, 2, 3$) are produced for the last 3 observations. From inspection of Table 2 it can be seen that the forecasts obtained by classical and Bayesian approaches are very similar (and close to the real values) which is not a surprising result since the correspondent parameter estimates are very closed. Regarding the predictions one step ahead of x_{141} and using the coherent predictions given by the medians $\hat{M}^{(\text{CML})}_{141}$ or $\hat{M}^{(B)}_{141}$ we obtain the value 1.

6 Concluding Remarks

In this work classical and Bayesian approaches to time series analysis and forecasting are applied to INBL (1,0,1,1) model. However much of the work for INBL processes remains to be done. We can point out some issues that are still open questions: invertibility conditions and the probabilistic structure of the process. This class of models, due to the cross term, can generate extreme observations and hence is suitable for modelling series of counts showing heavy tailed phenomena. However these features increase the difficulty in obtaining good predictions. Throughout this work we have seen that statistical modelling of INBL processes leads to likelihood functions based on convolutions. The difficulty of computing these functions exactly points towards the development of likelihood estimation by saddlepoint approximation, as suggested by [10], and the improvement of MCMC algorithms.

Acknowledgements This work was supported by Portuguese funds through the CIDMA—Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT—Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013.

References

1. Chen, M.-H., Shao, Q.-M.: Monte Carlo estimation of Bayesian credible and HPD intervals. J. Comput. Graph. Stat. **8**, 69–92 (1999)
2. Czado, C., Gneiting, T., Held, L.: Predictive model assessment for count data. Biometrics **65**, 1254–1261 (2009)

3. Doukhan, P., Latour, A., Oraichi, D.: A simple integer-valued bilinear times series models. Adv. Appl. Probab. **38**, 559–578 (2006)
4. Drost, F.C., van den Akker, R., Werker, B.J.: Note on integer-valued bilinear time series models. Stat. Probab. Lett. **78**, 992–996 (2008)
5. Freeland, R.K., McCabe, B.P.M.: Forecasting discrete valued low count time series. Int. J. Forecast. **20**, 427–434 (2004)
6. Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F. (eds.) Bayesian Statistics, vol. 4, pp. 169–193. Oxford University Press, Oxford (1992)
7. Gilks, W.R., Best, N.G., Tan, K.K.C.: Adaptive rejection metropolis sampling within Gibbs sampling. Appl. Stat. **44**, 455–472 (1995)
8. Heidelberger, P., Welch, P.D.: Simulation run length control in the presence of an initial transient. Oper. Res. **31**, 1109–1144 (1983)
9. Jung, R.C., Tremayne, A.R.: Useful models for time series of counts or simply wrong ones? AStA Adv. Stat. Anal. **95**, 59–91 (2011)
10. Pedeli, X., Davison, A.C., Fokianos, K.: Likelihood estimation for the INAR(p) model by saddlepoint approximation. J. Am. Stat. Assoc. **110**, 1229–1236 (2015)
11. Silva, N., Pereira, I., Silva, M.E.: Forecasting in INAR (1) model. REVSTAT - Stat. J. **7**, 119–134 (2009)
12. Steutel, F., van Harn, K.: Discrete analogues of self-decomposability and stability. Ann. Probab. **7**, 893–899 (1979)
13. Tanner, M.A.: Tools for Statistical Inference. Springer, New York (1996)

A Comparative Study of the Estimators for the Demand of Engineering Courses in Portugal

Raquel Oliveira, A. Manuela Gonçalves, and Rosa M. Vasconcelos

Abstract For the purpose of modeling the demand of Engineering Courses in Portugal we analyzed the possible regression models for panel count data models by establishing a comparison between the estimators obtained and then finding the most appropriate ones for our dataset. A precise quantification of the demand for each academic program is facilitated by the rules of access to higher education, in National Contest for Access and Admission to Higher Education, where candidates must list up to six preferences of institution and program. The data used in this paper covers the results of the national contest from 1997 to 2015 provided by the Portuguese Ministry of Education and Science. Multivariate methodologies were performed in order to allow a better understanding of the students' allocation behavior. The results seem to indicate that the negative binomial estimates fit better the dataset analyzed.

R. Oliveira (✉)
University of Minho, CMAT - Centre of Mathematics, Braga, Portugal

IPCA-EST, Vila Frescainha, Portugal
e-mail: rmoliveira@ipca.pt.

A. Manuela Gonçalves
University of Minho, CMAT - Centre of Mathematics, DMA - Department of Mathematics and Applications, Braga, Portugal
e-mail: nmneves@math.uminho.pt

R. M. Vasconcelos
University of Minho, 2C2T - Centre for Textile Science and Technology, DET - Department of Textile Engineering, Braga, Portugal
e-mail: rosa@det.uminho.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_26

1 Introduction

The idea of creating a European Higher Education Area (EHEA) was shared for the first time in the 800th anniversary of the University of Paris, Sorbonne Joint Declaration, in 1998 and was signed by the ministers of four countries: France, Germany, UK and Italy [7].

The decision to formalize this idea occurred 1 year later in Bologna, by 29–30 countries who expressed their willingness to commit to increase the competitiveness of EHEA through the 1999 Bologna Declaration [8], highlighting the need to further the independence and autonomy of all Institutions of Higher Education (IHE). These steps were followed by Ministerial Conferences in Prague in 2001, in Berlin in 2003, in Bergen in 2005, in London in 2007, in Leuven/Louvain-la-Neuve in 2009, in Budapest-Vienna in 2010, in Bucharest in 2012, and in Yerevan in 2015.

The aim of the Bologna Process (BP) was to "strengthen the competitiveness and attractiveness of European higher education and to foster student mobility and employability through the introduction of a system based on undergraduate and postgraduate studies with easily readable programs and degrees. Quality assurance has played an important role from the outset too" [13].

The Ministerial Conferences above mentioned provided more precise tools for implementing BP, such as modifying the structure of the undergraduate/postgraduate degrees, into a three-cycle system including the concept of qualifications frameworks, with an emphasis on learning outcomes, and introducing the concept of the social dimension of higher education and the recognition of qualifications as central to the European higher education policies. In Portugal, the first steps for BP implementation were given in 2005 by means of laws:

- nr 42/2005 which defines the regulatory instruments for the creation of the European Area of Higher Education. This law regulates the structure of the cycles of studies, the comparability of the degree structure, further to a comparable degree structure, the creation of a system of academic credits, whose accumulation and transferability across countries is guaranteed; this law also defines the mobility of students during and after their graduation;
- nr 49/2005, consisting of an amendment to the Law of the Education System Bases, including new areas and objectives of university education and polytechnic education.

Only in 2006, with decree nr 74/2006, changes were made to the existing law of the educational system bases that enable BP implementation, particularly the adoption of a new degree structure based on three cycles: the first cycle is bachelor degree (*licenciatura*—L), with a normal duration of 3 years, the second cycle is master (*mestrado*—M), with a normal duration of one-half or two years, and the third cycle is doctorate (*doutoramento*). The universities were also given the opportunity to offer a combined degree called integrated master (*mestrado integrado*—MI), with a duration of 5 or 6 years [1].

As part of the reorganization and rationalization of the European higher education system [10], BP implementation in Portugal was carried out by the Portuguese Ministry of Science Technology and Higher Education (MSTHE), the current Ministry of Education and Science (MES), which led to profound changes in the Portuguese higher education system.

The MSTHE determined that higher education institutions could restructure their study programs according to the Bologna principles beginning in 2006/2007 or in one of the two following years. Full implementation was achieved by 2009. This means that up to 2009 we had a variety of cases in IHE.

For the purpose of modeling the demand of Engineering Courses in Portugal, in this study we analyzed the possible regression models for panel count data models by establishing a comparison between the estimators obtained, and then finding the most appropriate ones for our dataset.

This study is organized in six sections: besides the "Introduction", the next section, "The Portuguese Higher Education System", presents the organization of higher education both in terms of the nature of the institutions and their tutelage; section "The Portuguese Higher Education System Access" explains the procedures put in place to access higher education and how this works for the public and private subsystems; section "Data and Descriptive Statistics" describes the data used in this study based in descriptive analysis; section "Statistical Analysis" expounds on the explanatory variables, specifies the models used to estimate the demand of Engineering Courses in Portugal, and presents the results obtained, exploring the significance of the work; finally, the last section presents the conclusions of the work and proposes suggestions for future research.

2 The Portuguese Higher Education System

Portugal has a binary higher education system, consisting of university and polytechnic education, each with distinct purposes that translate into specific curricular concepts [1].

University education, guided by a constant perspective of promoting research and knowledge creation, aims at ensuring a solid scientific and cultural preparation, by providing a technical training that qualifies for the exercise of professional and cultural activities and by promoting the development of design capabilities, innovation, and critical analysis.

Polytechnic education, guided by a constant perspective of applied research and development, aims at understanding and solving specific problems, at providing a solid cultural and technical level, and at developing the capacity for innovation, critical analysis, and its applications in the pursuit of professional activities.

University education is offered by public and private university institutions while polytechnic education is offered by public and private non-university institutions. Private higher education institutions must be subject to the previous recognition of the Ministry of Education and Science (The higher education system also comprises a concordatary institution) [1].

Both university and polytechnic institutions confer the degree of *licenciado* (bachelor). In polytechnic education, the cycle of studies leading to the degree of *licenciado* has a duration of 3 years of students' work and 180 credits. In certain cases, namely those covered by internal legislation or by European legislation, the cycle of studies can have up to 240 credits with a normal length of up to seven or eight curricular semesters of students' work. In university education, the cycle of studies that lead to the degree of *licenciado* has from 180 to 240 credits and a normal length of six to eight curricular semesters of students' work.

Both university and polytechnic institutions confer the degree of *mestre* (master). The cycle of studies leading to the degree of *mestre* has from 90 to 120 credits and a normal length of three to four curricular semesters of students' work or, in exceptional circumstances, 60 credits and a duration of two semesters, resulting from a stable and consolidated practice in that specific field at international level.

In university education, the degree of *mestre* may also be conferred after an integrated cycle of studies (integrated master), with 300–360 credits and a normal length of 10–12 curricular semesters of students' work, for cases in which access to the practice of a certain professional activity depends on that length of time established by legal European Union (EU) standards or resulting from a stable practice consolidated in the EU. In this cycle of studies, the degree of *licenciado* is conferred to those who have obtained 180 credits corresponding to the first six semesters of work. The degree of *doutor* (doctor) is conferred by universities and university institutes.

This study focuses on the publicly-funded higher education system that offers engineering study programs of bachelor or integrated master, since these programs include the majority of candidates; they are also representative in terms of supply of land area and their access is regulated by the Department of Higher Education (*Direcção Geral do Ensino Superior*—DGES).

3 The Portuguese Higher Education System Access

The MSTHE (MES), and more specifically the DGES, is in charge of the higher education sector and regulates access to the higher education system. Currently, access to higher education is conditioned by a system of *numerus clausus*, which defines the maximum number of students for each study program in both the public and private sectors. This number is defined by each institution, in fixed dates, and is subject to the approval of MES.

Numerus clausus works as a restriction on the supply side of the system, affecting the size and composition of the tertiary education sector [11]. Access to academic programs of bachelor or integrated master is done differently, whether in the public or in the private sector. Figure 1 illustrates the organization of access to higher education at this level.

Access to higher education for the public sector is done annually through a national contest based on the students' revealed preferences in their application. The national contest has two major phases: the first one takes place in July/August

Fig. 1 Organization of higher education access

and the second one in September and includes the vacancies that have not been filled in the first phase.

Each student ranks a maximum of six study program/institution pairs, from the most preferred (the first one) to the least preferred (the last one) alternative. The ensuing nationwide competition allocates the candidates based on their grade point average and the stated ranking of preferences. At each phase the applicant can only get a placement.

Students not allocated in the first phase, or allocated but not in the program/institution they want or those who had not applied in the first phase, may apply in the second phase.

4 Data and Descriptive Statistics

Due to differences in higher education access, most of the data available is related to the public sector disseminated by DGES according to the results of the different phases of the national contest. So, this study focuses on the publicly-funded higher education system that offers engineering study programs of bachelor or integrated master, since these programs include the majority of the candidates and they are also representative in terms of supply of land area and their access is regulated by DGES.

The data used is available online [6] and directly collected from the DGES archives. The data was collected for the period between 1997 and 2015, regarding the first phase, the most significant one, and the following variables are available:

- number of total applicants (representing the demand of pair institution/program);
- type of institution (University or Polytechnic);.
- academic program size (3 years program—Bachelor, 5 years program—Graduate, and 3+2 years program degree in two cycles—Graduate until 2006 and after 2006 the first cycle or integrated master);
- field of education and training courses (CNAEF) [5];
- number of vacancies available for pair institution/program;
- number of academic programs available for each institution;
- number of total allocated applicants in each program (total number of allocated students in pair institution/program, irrespective of their ranking);
- number of applicants by choice (1–6 preferences of pair institution/program);
- number of allocated applicants by choice (1–6 preferences of pair institution/program);

- grade point of the last place (GPLP);
- grade point average: applicants (GPAA), admission exams (GPAE), last year of secondary school, 12th year, 10th/11th year;
- number of male applicants;
- number of female applicants;
- number of men allocated;
- number of women allocated.

In Fig. 2 we present the number of IHE by period of time: Pre Bologna period (1997–2006) and Post Bologna period (2007–2015).

The number of IHE in both periods is similar: the minimum of IHE is 26 and the maximum is 30, which is not the case in its variation: it has an opposite variation.

Figures 3 and 4 represent, respectively, the total of applicants and the total of allocates for the national contest also by period.

Despite the higher number of applicants in the Post Bologna period (2008), generally the number of applicants has decreased over the years.

In relation to the allocated, the maximum and minimum values have been achieved in the Post Bologna period, and the most extreme variations occurred at the end of the two periods under observation.

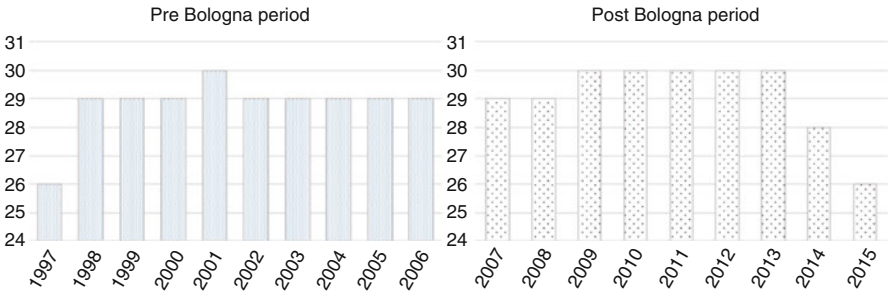


Fig. 2 Number of higher institutions

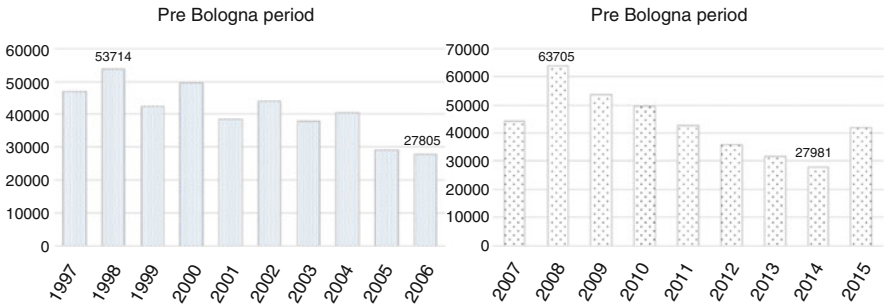


Fig. 3 Number of total applicants

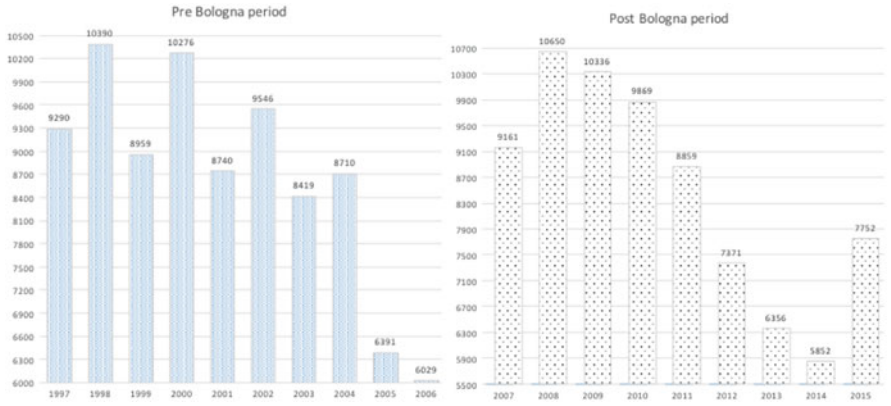


Fig. 4 Number of total allocated

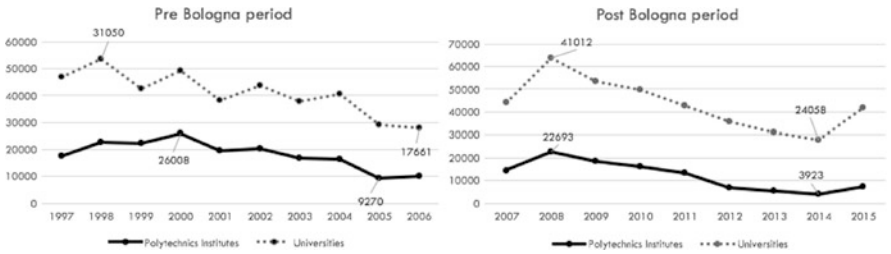


Fig. 5 Number of total applicants by IHE type

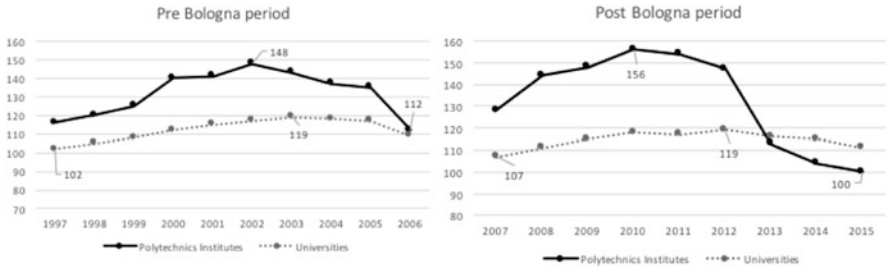


Fig. 6 Number of engineering programs by HEI type

Figures 5 and 6 illustrate, respectively, the number of total applicants and the number and engineering academic programs by IHE type.

To better understand the evolution of the degree (size of the program) of engineering academic programs offered over the years, we present the results in Fig. 7.

In 2006 coexisted all kinds of degrees. It was a "hybrid" year because IHE could choose to start implementing the BP curricula changes in 2006 or until 2008. However, we find that with regard to Engineering academic programs the restructuring of curricula changes was complete in 2007.

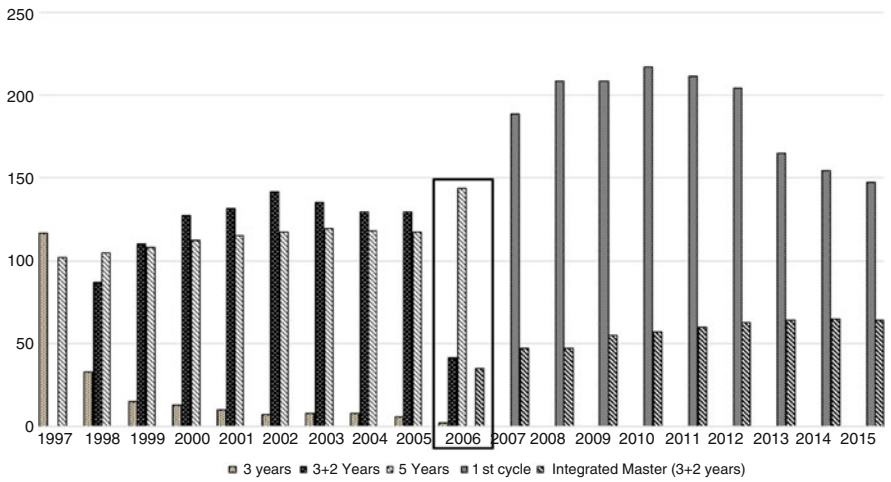


Fig. 7 Number of engineering programs by degree

Table 1 Summary of variables: number of male applicants, number of female applicants, number of men allocated, and number of women allocated

	Applicants		Allocated	
Years	Male	Female	Men	Women
1997	31,436	14,140	6287	2855
1998	36,907	17,111	7100	3187
1999	28,584	12,848	5954	2825
2000	32,409	15,534	6869	3404
2001	27,298	11,802	5953	2686
2002	30,811	12,983	6554	3009
2003	26,693	10,962	5903	2665
2004	29,436	10,955	6318	2393
2005	22,336	6692	4818	1573
2006	20,595	7001	4636	1553
2007	31,438	12,277	6538	2660
2008	44,962	18,391	7601	3049
2009	38,645	14,589	7427	2932
2010	35,712	13,654	7162	2716
2011	30,649	11,798	6477	2440
2012	25,454	10,116	5252	2068
2013	22,191	8991	4605	1732
2014	20,234	7383	4258	1554
2015	31,095	10,692	5655	2070

The summary of variables—number of male applicants, number of female applicants, number of men allocated and number of women allocated—is presented in Table 1.

5 Statistical Analysis

Statistical analysis was performed using the SPSS statistic software and STATA statistical data analysis software [3]. A Kolmogorov-Smirnov test was used to verify data normality (p-value < 0.001), for which the results indicated that non-parametric tests should be used for all comparisons.

In order to analyze if the number of applicants to Engineering programs (demand) depends on the type of institution or degree (size of program), some tests were performed, Tables 2 and 3 [9, 12].

The demand (number of applicants) of Engineering programs depended on the size of the programs for the two periods, but if we only consider the type of IHE then there is no dependency during the Pre Bologna period from 1999 to 2002.

5.1 Modeling Approach

The purpose of this study is to establish a comparison between the estimators for regression models by finding the most appropriate ones for our dataset, and so we describe the possible models to explain the number of applicants.

Table 2 Mann-Whitney and Kruskal-Wallis tests for the Pre Bologna period

		Mann-Whitney	Kruskal-Wallis
Pre Bologna		IHE type	Size of program
1997	Test statistics	−4.068	16.552
	p-value	<0.0001	<0.0001
1998	Test statistics	−3.245	20.686
	p-value	0.001	<0.0001
1999	Test statistics	−0.165	15.939
	p-value	**0.869**	<0.0001
2000	Test statistics	−0.346	11.503
	p-value	**0.729**	0.003
2001	Test statistics	−0.331	7.746
	p-value	**0.741**	0.021
2002	Test statistics	−1.397	8.095
	p-value	**0.162**	0.017
2003	Test statistics	−2.229	10.31
	p-value	0.026	0.006
2004	Test statistics	−3.19	16.367
	p-value	0.001	<0.0001
2005	Test statistics	−5.351	33.03
	p-value	<0.0001	<0.0001
2006	Test statistics	−3.257	45.001
	p-value	0.001	<0.0001

Bold: not statistically significant

Table 3 Mann-Whitney tests for the Post Bologna period

Post Bologna		Mann-Whitney IHE type	Mann-Whitney Size of program
2007	Test statistics	−6.438	−7.324
	p-value	<0.0001	<0.0001
2008	Test statistics	−7.287	−7.493
	p-value	<0.0001	<0.0001
2009	Test statistics	−7.99	−8.367
	p-value	<0.0001	<0.0001
2010	Test statistics	−8.761	−8.914
	p-value	<0.0001	<0.0001
2011	Test statistics	−9.143	−9.014
	p-value	<0.0001	<0.0001
2012	Test statistics	−10.352	−9.023
	p-value	<0.0001	<0.0001
2013	Test statistics	−8.897	−8.24
	p-value	<0.0001	<0.0001
2014	Test statistics	−8.697	−7.926
	p-value	<0.0001	<0.0001
2015	Test statistics	−8.841	−7.537
	p-value	<0.0001	<0.0001

Since the response variable is a nonnegative integer and since its distribution is skewed to the left, a count data type of model is appropriate [2]. As already mentioned the data for nineteen academic years is available, and so we have a panel structure with repeated observations on the same academic program and institution, which allows controlling for study program characteristics that are not observable but are assumed constant over time.

The starting point model for count data is the Poisson regression model with the exponential mean function [2]

$$\mu = \exp(\mathbf{x}^{'} \cdot \boldsymbol{\beta}). \tag{1}$$

In our data, descriptive statistics show that the dependent variable presents overdispersion, so the Negative Binomial regression model might be more appropriate for the data. Since we have repeated measures in individuals i over time t data for $i = 1, \ldots, n$ and $t = 1, \ldots, T$, and y_{it} are nonnegative integer-valued outcomes. So our data is in a panel structure. As established by [4], a major advantage of panel data is increased precision in estimation. This is the result of an increase in the number of observations owing to combining or pooling several time periods of data for each individual. However, for valid statistical inference one needs to control for likely correlation of regression model errors over time for a given individual. A second attraction of panel data is the possibility of consistent estimation of the fixed

effects model, which allows for unobserved individual heterogeneity that may be correlated with regressors.

Most disciplines in applied statistics treat any unobserved individual heterogeneity as being distributed independently of the regressors. Then the effects are called random effects, though a better term is purely random effects. Compared to fixed effects models, this stronger assumption has the advantage of permitting consistent estimation of all parameters, including coefficients of time invariant regressors. However, random effects and pooled estimators are inconsistent if the true model is one with fixed effects [4]. Therefore we performed the Hausman test with the null hypothesis: individual effects are uncorrelated with other regressors in the model that is the most appropriate for the data under study. We rejected the null hypothesis, meaning that we should use the fixed effects model.

Therefore, we estimated the data according to [12], following these regression models for panel count data:

1. Pooled Poisson and Negative Binomial regression models population-averaged (PPA and NBPA),

$$y_{it} = \alpha_i + \mathbf{x}_{it}^{'}\beta_i + \mu_{it}$$

where y_{it} is the scalar dependent variable, x_{it} is a $k \times 1$ vector of independent variables, and u_{it} is i.i.d. with mean 0 and variance σ_u^2.

2. Negative Binomial regression model with Fixed Effects (NBFE) and Random Effects (NBRE),

$$y_{it} = \mathbf{x}_{it}^{'}\beta_i + (\alpha_i + \varepsilon_{it})$$

   (a) Fixed effects:
   α_i is a random variable possibly correlated with $\mathbf{x}_{it}$.
   (b) Random effects:
   α_i is purely random (usually i.i.d. $N(0, \sigma_\alpha^2)$) uncorrelated with $\mathbf{x}_{it}$. Being

   (i) Poisson,

$$Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$$

   where y is the count for our dependent variable, $\mu = \exp(\mathbf{x}^{'} \cdot \boldsymbol{\beta})$.
   (ii) Negative Binomial,

$$Pr(Y = y|\mu, \alpha) = \frac{\Gamma(y + \alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}$$

where y is the count for our dependent variable, $\mu(\mathbf{x}) = \exp(\mathbf{x} \cdot \boldsymbol{\beta})$, $\alpha \geq 0$ is the overdispersion parameter, $\Gamma(.)$ is the gamma function, and $\mathbf{x}$ is a vector of regressors. This form assumes constant dispersion within groups, equal to $1 + \alpha\mu(\mathbf{x})$. The mean and variance of Y are defined as $\mu(\mathbf{x})$ and $(1 + \alpha\mu(\mathbf{x}))\mu(\mathbf{x})$, respectively.

We performed these models, the Poisson model was considered as reference, and the results are presented in Table 4.

Since we used the Poisson regression model as a starting point and since it cannot be applied because of overdispersion, we only analyzed the results from the negative binomial regression models. As one can be seen in Table 4, results can vary across models. For example, the significant predictors of the number of applicants for all the models are: IEH type, size of program, vacancies, number of allocated, grade point of the last place, number of applicants in 2nd, 3th, 4th, 5th, and 6th choices, number of allocated in 1st and 2th choices, number of female applicant, and finally the number of men and women allocated. These results indicate that the NBFE model most accurately explained the data, confirming the results obtained by the Hausman test [4].

When we performed in STATA a Negative Binomial regression model for panel count data it automatically computed a Wald test that evaluates the null hypothesis that the coefficients are equal to zero. In the presented analysis this test was statistically significant, p-value < 0.0001. Therefore, we could conclude that at least one coefficient differs significantly from zero.

6 Conclusions and Future Work

The aim of this paper was to understand the impact of the Bologna Process on the demand of Engineering courses in Portugal. Since the major changes caused by the Bologna Process were in the degrees attributed by the two types of Institution of Higher education (Universities or Polytechnics), first it was tested if the demand was influenced by these characteristics. To this end, we have performed the Mann-Whitney and Kruskal-Wallis tests, and found that there is a change in the behavior of the demand during the Pre Bologna (1997–2006) and Post Bologna (2007–2015) periods. Since 2007 the demand depends, without exception, on the type of degrees attributed (Graduate, Master, Doctor) or on the type of institution (Universities or Polytechnics). These results point to a behavioral change of the variable response under study (demand for Engineering courses in Portugal), showing that is effectively necessary to further deepen the study of the Bologna Process impact.

Furthermore, we have also estimated the demand during a 19 years period (1997–2015) by applying the models (Pooled Poisson and Negative Binomial regression models population-averaged, Negative Binomial regression model with fixed effects and random effects) deemed more suitable for the variables under study. The results

Table 4 Results of the modeling approach

	PPA	NBPA	NBRE	NBFE		PPA	NBPA	NBRE	NBFE
HEY type	0.0106	0.0765**	0.1641***	0.1365**	Allocated 1st choice	0.0029**	0.0070***	0.0030***	0.0028***
	(0.0379)	(0.0344)	(0.0326)	(0.0504)		(0.0009)	(0.0013)	(0.0006)	(0.0006)
CNAEF	0.0003	−0.0003	0.0001	0.0003	Allocated 2nd choice	−0.0013	−0.006***	−0.002**	−0.002**
	(0.0002)	(0.0002)	(.0002)	(0.0002)		(0.0008)	(0.0013)	(0.0008)	(0.0008)
Size program	−0.032**	−0.061***	−0.054***	−0.049***	Allocated 3th choice	−0.0045	−0.0012	−0.005**	−0.004**
	(0.0116)	(0.0123)	(0.0071)	(0.0072)		(0.0026)	(0.0014)	(0.0013)	(0.0013)
Vacancies	−0.007***	−0.002**	−0.005***	−0.006***	Allocated 4th choice	0.0029	0.0051	0.0028	0.0023
	(0.0009)	(0.0008)	(0.0005)	(0.0005)		(0.0037)	(0.0034)	(0.0022)	(0.0022)
Allocated	0.0081***	0.0057**	0.0095***	0.0096***	Allocated 5th choice	0.0007	−0.0003	−0.0007	−0.0003
	(0.0016)	(0.0018)	(0.0009)	(0.0009)		(0.0040)	(0.0031)	(0.0023)	(0.0023)
GPLP	−0.002**	−0.008***	−0.004***	−0.004***	Allocated 6th choice	−0.0026	−0.0013	0.0006	0.0013
	(0.0008)	(0.0009)	(0.0005)	(0.0005)		(0.0058)	(0,0038)	(0.0021)	(0.0021)
GPAA	−0.0026	−0.0001	−0.0032**	−0.0032**	Male applicant	0.0001	−0.0011	−0.0004**	−0.0002
	(0.0023)	(0.0020)	(0.0013)	(0.0014)		(0.0003)	(0.0007)	(0.0002)	(0.0002)
GPAE	−0.0015	0.0002	−0.0004	−0.0007	Female applicant	0.0008	0.0011**	0.0008***	0.0008***
	(0.0010)	(0.0009)	(0.0007)	(0.0007)		(0.0004)	(0.0005)	(0.0001)	(0.0001)

(continued)

Table 4 (continued)

	PPA	NBPA	NBRE	NBFE		PPA	NBPA	NBRE	NBFE
Applicants 1st choice	0.0000	−0.0011	−0.0004**	−0.0002	Men allocated	0.0013	0.0030**	0.0013 ***	0.0012***
	(0.0003)	(0.0007)	(0.0002)	(0.0002)		(0.0007)	(0.0009)	(0.0003)	(0.0003)
Applicants 2nd choice	0.0008	0.0011**	0.0008***	0.0008***	Women allocated	0.0007	0.0033***	0.0007**	0.0007**
	(0.0004)	(0.0005)	(0.0001)	(0.0001)		(0.0005)	(0.0007)	(0.0002)	(0.0003)
Applicants 3th choice	0.0013	0.0030**	0.0013 ***	0.0012***	Intercept	0.0060***	0.0096***	0.0059***	0.0057***
	(0.0007)	(0.0009)	(0.0003)	(0.0003)		(0.0010)	(0.0012)	(0.0005)	(0.0005)
Applicants 4th choice	0.0007	0.0033***	0.0007**	0.0007**	LL			−22444.1	**−19495.8**
	(0.0005)	(0.0007)	(0.0002)	(0.0003)					
Applicants 5th choice	0.0060***	0.0096***	0.0059***	0.0057***	Wald	2200.3***	3397.4***	14433.9***	13280.5***
	(0.0010)	(0.0012)	(0.0005)	(0.0005)					
Applicants 6th choice	0.0058***	0.0117***	0.0063***	0.0061***	DF	26	26	26	26
	(0.0008)	(0.0009)	(0.0005)	(0.0005)					

Standard errors in parentheses
Statistical significance: < 0.05; ** < 0.01; *** < 0.001
Bold: not statistically significant

seem to indicate that the best model for our data is the Negative Binomial regression model with fixed effects.

In future work, we intend to explore this model with more explanatory variables and to evaluate the model behavior during the Pre and Post Bologna periods in order to better assess the impact of the Bologna Process.

Acknowledgements A. Manuela Gonçalves and Raquel Oliveira were supported by the Research Centre of Mathematics of the University of Minho with the Portuguese Funds from the "FCT-Fundação para a Ciência e a Tecnologia," through the Project PEstOE/MAT/UI0013/2014. Rosa M. Vasconcelos was supported by the Foundation through "FCT - Fundação para a Ciência e Tecnologia," within the Project UID/MAT/00013/2013, by FEDER funds through the Competitivity Factors Operational Programme—COMPET and by national funds through FCT within the scope of the project POCI-01-0145-FEDER-007136.

References

1. Access to higher education in Portugal. http://www.dges.mctes.pt. Accessed 24 November 2016
2. Cameron, A.C., Trivedi, P.K.: Regression Analysis of Count Data. Cambridge University Press, Cambridge (1998)
3. Cameron, A.C., Trivedi, P.K.: Microeconometrics Using Stata. Stata Press, College Station (2009)
4. Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Stata Press, College Station (2010)
5. Classificação nacional das áreas de educação e formação. http://www.dados.gov.pt/PT/CatalogoDados/Dados.aspx?name=ClassNacionaldeareasdeeducacaoeformacao. Accessed 24 November 2016
6. Data of access to higher education in Portugal. http://www.dges.mctes.pt/DGES/pt/Estudantes. Accessed 24 November 2016
7. EHEA: a common vision. http://www.ehea.info/pid34248/history.html. Accessed 17 November 2016
8. European Ministers of Education. The Bologna Declaration (1999). http://www.dges.mctes.pt. Accessed 24 Nov 2016
9. Higgins, J.H.: Introduction to Modern Nonparametric Statistics. Thomson Toronto, Toronto (2004)
10. OECD, Organization for Economic Co-operation and Development. Reviews of national policies for education: tertiary education in Portugal. Examiner's Report. http://www.dges.mctes.pt/NR/rdonlyres/8B016D34-DAAB-4B50-ADBB-25AE105AEE88/2564/Backgoundreport.pdf. Accessed 17 November 2016 (2006)
11. Results of access to higher education in Portugal. http://www.dges.mctes.pt/DGES/pt/Estudantes/Acesso. Accessed 24 November 2016
12. Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioral Sciences, 2nd edn. McGraw-Hill, New York (1988)
13. The aim of Bologna process. http://www.europa.eu. Accessed 17 November 2016

Statistical Methods for Word Association in Text Mining

Anacleto Correia, M. Filomena Teodoro, and Victor Lobo

Abstract Text data has been growing dramatically in the last years, mainly due to the advance of web related technologies that enable people to produce an overwhelming amount of data. Many knowledge about the world is encoded in text data available through blogs, tweets, web pages, articles, and books.

This paper introduces some general techniques for text data mining, based on text retrieval models, that can be applicable to any text in any natural language. The techniques are targeted to problems requiring minimum or no human effort. These techniques, which can be used in many applications, allow the measurement of similarity of contexts, as well as the co-occurrence of terms in text data with different levels of granularity.

1 Introduction

The Web accelerated the textual revolution and made available a great amount of on-line information. Information and knowledge about almost any subject is encoded in text data available on-line, in articles or books. Text mining refers to the process of extracting high quality information from text data. The quality of the information derived is concerned with elicited patterns and trends. Text mining usually involves the process of structuring the input text (through parsing, and adding or removing linguistic features), deriving patterns within the structured data, and eventually analyzing and interpreting the output.

A. Correia (✉) · V. Lobo
CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, Almada, Portugal
e-mail: cortez.correia@marinha.pt; sousa.lobo@marinha.pt

M. F. Teodoro
CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, Almada, Portugal

CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, Lisbon, Portugal
e-mail: maria.alves.teodoro@marinha.pt

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8_27

Text mining, as an interdisciplinary field, benefits from contribution of several correlated disciplines, namely information retrieval, data mining, machine learning, probability, statistics, and computational linguistics. These various disciplines are combined together to build the text-mining process workflow. Through *information retrieval* (IR), documents that match submitted queries are collected. IR systems allow filtering the collection of documents to attain the most relevant of them to address a specific topic. IR is concerned on reducing the number of documents for text mining analysis in order that the processing of computationally-intensive algorithms can be sped up. On the other hand, *natural language processing* (NLP), a field of artificial intelligence, addresses the analysis of human language so that computers can understand natural languages in a similar way as humans do. Techniques from NLP includes *shallow parsers*, which identify the main grammatical elements in a sentence (e.g., noun phrases and verb phrases), as well as *deep parsers* for generation of sentences' grammatical structure. NLP contribution to text mining is on providing linguistic data (e.g., documents' annotations, part-of-speech tags, parsing results) on the information extraction phase.

Data Mining (DM) is the subprocess that allows the identification of patterns in large sets of data. The aim of DM is to uncover previously unknown, useful knowledge for decision making. When used in text mining, DM is applied to the facts generated by the information extraction phase. The results extracted by DM techniques can then be queried and visually represented. Using the *information extraction* (IE) subprocess is possible to automatically obtain structured data from unstructured natural language documents. Often this involves defining templates, which are used to guide the extraction process. The IE process relies itself on the data generated by the NLP process [11].

When looking at text data in any support, people may have expectations regarding its content, which can be: (1) to discover aspects about a specific natural language, its usage, as well as the patterns on it; (2) to mine knowledge from content of text data about the observed world, getting the essence of it or extracting information about relevant aspects of the world; (3) to mine knowledge about an observer, which means using text data to infer properties of a person; and (4) to make predictive analytics using text mining to infer real-world variables. When real-world variables are inferred they can also use intermediate results of other predictions. So, multiple types of knowledge can be mined from text in general [3, 8, 15].

As previously referred, information retrieval [9, 10] and text mining are very related domains for leveraging text data. However, whereas information retrieval aims fundamentally to turn raw text data into a smaller and relevant text data, for handling a specific problem or supporting a particular decision, text mining deals with processing text data to extract knowledge or synthesize information in order to be more easily processed by people. Text mining techniques are surveyed in several works [1, 2, 5, 6, 12–14]. In this paper we focus on statistical approach word associations used for extracting specific knowledge from documents [7].

So, Sect. 2 introduces the preliminaries of mining techniques from text data for word associations. Sections 3 and 4 describe word association as a form of analyzing the content of text in search of paradigmatic relation (context similarity) and the syntagmatic relation (co-occurrence of terms). In Sect. 5 we get some conclusions.

2 Word Associations

This section presents on mining associations of words from text data, following the presentation of the subject in [16]. In general there are two basic word relations: the *paradigmatic relation* and the *syntagmatic relation*. These two kind of relations are fundamental and they can be generalized to capture basic relations between units in arbitrary sequences. Also, they can be generalized to describe relations of any items in a language, such as words and even complex phrases.

The elements A and B are said to have paradigmatic relation if they can be *substituted* one for each other. This means that the two words are in the same semantic class, or syntactic class. In general, they can be replaced one by the other without affecting the understanding of the sentence, which means that the result would still be a valid sentence. In the case of a syntagmatical relation, on the other hand, the two words can be *combined* with each other. Therefore the elements A and B can be combined with each other in a sentence, since they are semantically related. However, in general, they cannot be replaced one by the other, since the sentence would become meaningless.

In another perspective, the relations can be seen as: (1) relations that occur in similar locations relative to the neighbors in the sequence (paradigmatic relation) or; (2) relations concerning co-occurrent elements that tend to show up in the same sequence (syntagmatical relation). These two basic relations of words are complementary.

For discovering paradigmatic relation, one can assume that words that have high context similarity also have paradigmatic relation. So, the context similarity of each word must be computed. To discover syntagmatic relation, one must search for words with high co-occurrences but relatively low individual occurrences. The justification is that those words tend to occur together. To compute the syntagmatic relation one must count how many times two words occur together in a context (a sentence, a paragraph, or even a document). Then a comparison should be made between the co-occurrences and their individual occurrences.

Both paradigmatic and syntagmatic relations are closely related since paradigmatically related words tend to have syntagmatic relation with the same word. They tend to be associated with the same word, which suggests that the discovery of the two relations can be done together. In the following sections some of the statistical methods used for discovering those kind of relations are introduced.

3 Paradigmatic Relations

The idea of discovering paradigmatic relations is to look at the context of each word and try to compute the similarity of those contexts. This can be done through two steps: 1—formally representing the context and; 2—defining a similarity function. The context contains in general lot of words usually regarded as bag of words.

The similarity function is, in general, a combination of similarities on different contexts [16].

Thinking in a bag of words is a useful mean for representation of vectors in a vector space model. The subjacent idea for this approach is to define each word in the vocabulary as one of the dimensions of the high dimensional space. Since there are N words in the vocabulary, then, there are also N dimensions in the space model. So, the context of a word, w_1, can be represented as a vector d_1, and a different word w_2, by another context d_2. The paradigmatic relation between the two words can then be measured computing the similarity of the two vectors. Therefore, by representing the context in the vector space model, the problem of paradigmatical relation discovery is converted into the problem of computing similarity of the vectors. For referring each vector we use the expression (1):

$$d_1 = (x_1, \ldots, x_N) \qquad \text{where each } x_i \text{ is given by} \qquad x_i = \frac{c(w_i, d_1)}{|d_1|}, \tag{1}$$

where $c(w_i, d_1)$ represents the total count of word w_1 in pseudo document d_1 and $|d_1|$ is the total amount of words in d_1.

Regarding the computation of similarity, there are several approaches developed for information retrieval that can be adapted to text mining. One of the approaches is to try to match the similarity of context based on the Expected Overlap of Words in Context (EOWC) method. The idea is to represent a context by a word vector where each word has a weight equal to the probability that a randomly picked word w_i, from the document vector, is the specific word w_i. In other words, x_i is defined as the normalized count of word w_i in the context, and this can be interpreted as the probability of randomly picking this word from the document d_j. Since these are normalized frequencies, the sum of $x_i's$ is one, which means the vector is in fact a probability of words distribution. According to this method each context is represented by a vector that specifies the probability of each word in the context. Consequently, the similarity is defined (2) as the dot product of the two vectors:

$$\text{Sim}(d_1, d_2) = d_1 \cdot d_2 = x_1\, y_1 + \cdots + x_n\, y_n = \sum_{i=1}^{N} x_i\, y_i. \tag{2}$$

With this similarity function one can compute the probability of two randomly words from the two contexts being identical with EOWC approach measuring the overlap of words in the contexts.

The EOWC method, however, has two problems, namely: 1—it favors matching frequent terms over matching distinct terms; 2—it treats every word equally, meaning that even a common word would contribute equally as others more relevant to the current content.

Retrieval heuristics, used in the text retrieval domain, can be used to solve these problems. To address the first problem, a sublinear transformation named Term Frequency (TF) is used, instead of the raw frequency count of the terms, to represent

the context. In the TF transformation, denoted by TF(w, d), the raw count of a word is converted into a weight that reflects the belief about the importance of the word. An implementation of the transformation in (3), called a BM25 transformation, was used in information retrieval to solve the same problem of overemphasizing a frequent word. This transformation

$$\mathrm{TF}(w, d) = \frac{(k+1)\,x}{x+k}, \tag{3}$$

where $k \in [0, +\infty[$ is a parameter and x is the raw count of a word, has an upper bound of $k + 1$, which puts a constraint on high frequency.

To solve the second problem, one must penalize popular terms and put more weight on rare terms. The heuristic (4) used in text retrieval is called Inverse Document Frequency (IDF) term weighting. Document frequency means the count of the total number of documents that contain a particular word. The IDF measure is defined as a logarithm function of the document frequency

$$\mathrm{IDF}(W) = \log \frac{(M+1)}{k}, \tag{4}$$

where k is the document frequency and M is the total number of documents in the collection. The IDF function gives a higher value for lower k, which means that it rewards a rare term. It reaches the maximum value on $\log(M + 1)$, for a very rare term that occurs just once in the context. The lowest value of IDF, close to zero, is when k reaches its maximum of M.

TF and IDF heuristics are used to improve the similarity function for paradigmatic relation mining. The document vector is defined as containing elements representing normalized BM25 values. The new weight reflects now the frequency of occurrence of the word in the context. In the document vector (1), each x_i is now given by (5):

$$x_i = \frac{\mathrm{BM25}(w_i, d_1)}{\sum_{j=1}^{N} \mathrm{BM25}(w_j, d_1)}. \tag{5}$$

The weight of each word is normalized by the sum of the weights of all the words. This ensures that all the x_i's will sum to 1 in the vector that represents now the word distribution.

The formula in (6) allows the definition of the document vector, giving to high frequency terms lower weight. This helps to control the influence of the high frequency terms. So, in (5) the weight computed for each word x_i in document d_1 is

$$\mathrm{BM25}(w_i, d_1) = \frac{(k+1)\,c(w_i, d_1)}{c(w_i, d_1) + k\left(1 - b + \frac{b^* |d_1|}{\mathrm{avdl}}\right)}, \tag{6}$$

with $k \in [0, +\infty[$ and $b \in [0, 1]$. $c(w_i, d_1)$ is the counting of word w_i in document d_1, and is introduced to achieve the sublinear normalization. The parameter k is generally a positive number that controls the upper bound and the extent of the linear transformation. The parameter b controls the length of the normalization. The normalization formula has also the average document length avdl, which is computed by taking the average of the lengths of all the context documents. This average is constant for any given collection of documents and only affects the context document length $|d_1|$ and the parameter b.

The similarity function in (2) becomes the one in (7), when the IDF function is included, weighting the importance of each specific word w_i and a common word worthing less than a rare word

$$\text{Sim}(d_1, d_2) = \sum_{i=1}^{N} \text{IDF}(w_i)\, x_i\, y_i. \tag{7}$$

With this modification, the new function similarity function based on BM addresses the two mentioned problems regarding EOWC method.

Summarizing, when a document vector is used to represent the context, it turns out that some words will have higher weights, and other lower weights. This allows to use the weights to discover the words that are strongly associated with the context. Applying IDF weighting allows to re-weigh the words in order that highly similar word pairs can be treated as having paradigmatic relations, which means these words share similar contexts,

$$\text{IDF-weighted}\, d_1 = (x_1 * \text{IDF}(w_1), \ldots, x_n * \text{IDF}(w_n)). \tag{8}$$

The term vector presented in previous expression to represent the context some terms would have higher weights, while others have lower weights. Depending on how weights are assigned to these terms, the expression in (8) might be used to discover the words that are strongly associated with a candidate of word in the context.

The context presented in expression (8) can also be used to discover syntagmatic relations. The claim is that if a term is highly scored in the document vector, then it is strongly related with other terms highly weighted. To understand such conclusion one may bear in mind that when IDF is applied to frequent terms they are re-weighted. This means that common words are penalized so the highest weighted terms in the final document vector will not be the common terms because they have lower IDFs. Instead, highest weighted terms will be the ones that are frequent in the context but not frequent in the overall collection. Those are the words more relevant in the context. For this reason, the highly weighted terms of the weighted vector (8) can also be considered as candidates for syntagmatic relations.

This conclusion regarding syntagmatic relation discovery is a complementary result raised up from the study of paradigmatic relations. In the following section specific methods concerning syntagmatic relation are presented.

4 Syntagmatic Relations

As previously mentioned in Sect. 2, syntagmatic relations hold between words that have correlated co-occurrences. The concept of entropy, introduced in information theory [4], is used for discovering syntagmatic relations.

For the purpose of text mining, the entropy function is treated as a function defined on a random variable X_W. So, the problem can be formally defined as predicting the value of a binary random variable X_W, with W denoting a word. Each random variable is associated only with one word. The degree of randomness of the stochastic variable X_W indicates the difficulty on predicting the word W in a segment of words. When the value of the variable $X_W = 1$, it means the word is present. When $X_W = 0$, it means the word is absent. The randomness of the variable X_W is quantified by measuring the entropy. So, the entropy expression in (9) is a way of quantitatively measuring whether a word is hard or easy to predict in a segment. A higher entropy (9) is expected for words hard to predict

$$H(X_w) = \sum_{v \in \{0,1\}} -p(X_W = v) \log_2 p(X_W = v) \tag{9}$$

$$= -p(X_W = 0) \log_2 p(X_W = 0) - p(X_W = 1) \log_2 p(X_W = 1). \tag{10}$$

When a different kind of scenario is addressed, and a prior information about the text segment is known, the concept of conditional entropy is used. Supposing the presence of word W_2, this means having knowledge regarding another random variable X_{W_2}, which allows the use of conditional probability. As a consequence, using the conditional probabilities in the entropy function, we'll get the conditional entropy (11):

$$H(X_{W_1}|X_{W_2}) = \sum_{u \in \{0,1\}} [p(X_{W_2} = u) \sum_{v \in \{0,1\}} [-p(X_{W_1} = v|X_{W_2} = u) \log_2 p(X_{W_1} = v|X_{W_2} = u)]]. \tag{11}$$

Conditional entropy helps to capture syntagmatic relation, because it gives a way to measure directly the association of two words. This is because it measures the extent in which a word can be predicted, given the knowledge about the presence or absence of another word.

In general, for any discrete random variables X_{W_1} and X_{W_2}, the following relation is verified:

$$H(X_{W_1}) \geq H(X_{W_1}|X_{W_2}).$$

$H(X_{W_1}|X_{W_2})$ reaches its minimum 0 when X_{W_1}=X_{W_2}, and its maximum $H(X_{W_1})$ when there is no relation between X_{W_1} and X_{W_2}. The algorithm for mining

syntagmatic relations using conditional entropy has the following steps: For each word W_1:

- enumerate all other words W_2 and compute for each one $H(X_{W_1}|X_{W_2})$.
- sort all the candidate in ascending order of $H(X_{W_1}|X_{W_2})$.
- take the top-ranked candidate words, below a certain a threshold, as words that have potential syntagmatic relations with W_1.

However, this algorithm does not help mining the strongest k syntagmatical relations, irrespective of the words, for a complete word segment. To achieve it, comparability between conditional entropies of different words is required, such as in $H(X_{W_1}|X_{W_2})$ and $H(X_{W_1}|X_{W_3})$. However, $H(X_{W_1}|X_{W_3})$ and $H(X_{W_3}|X_{W_2})$ are not comparable because they have different outer bounds. That is why comparability of conditional entropy among different pairs of words is needed, in order to discover the k syntagmatical relations. This can be achieved through the concept of *mutual information*, $I(X_{W_1}; X_{W_2})$, which measures the entropy reduction of X_{W_1} obtained from knowing X_{W_2}, or, conversely, the entropy reduction of X_{W_2} obtained from knowing X_{W_1},

$$I(X_{W_1}; X_{W_2}) = H(X_{W_1}) - H(X_{W_1}|X_{W_2}) = H(X_{W_2}) - H(X_{W_2}|X_{W_1}). \tag{12}$$

The properties of mutual information function are summarized by relations (13)–(15):

- Non-negativity:

$$I(X_{W_1}; X_{W_2}) \geq 0; \tag{13}$$

- Symmetry:

$$I(X_{W_1}; X_{W_2}) = I(X_{W_2}; X_{W_1}); \tag{14}$$

- Independence:

$$I(X_{W_1}; X_{W_2}) = 0 \text{ iff } X_{W_1} \text{and } X_{W_2} \text{ are independent.} \tag{15}$$

The ranking obtained through mutual entropy is the same as got with the conditional entropy of X_{W_1} given X_{W_2}. However, the mutual information function is more general since it can also be used to compare different pairs of X_{W_1} and X_{W_2}. This makes mutual information more useful, from a practical point of view.

To compute the mutual information a method from information theory [4] is used, which allows to mathematically rewrite the mutual information into the

Kullback–Leibler divergence (KL-divergence). This method measures the divergence between two joint distributions. The numerator in (16)

$$I(X_{W_1}; X_{W_2}) = \sum_{u \in 0,1} \sum_{v \in 0,1} p(X_{W_1} = u, X_{W_2} = v) \log_2 \frac{p(X_{W_1} = u, X_{W_2} = v)}{p(X_{W_1} = u) p(X_{W_2} = v)}. \tag{16}$$

gives the observed joint distribution of the two random variables. On the other hand, the denominator can be interpreted as the expected joint distribution of the two random variables, if they are independent. The larger this ratio (or divergence), the higher the mutual information.

The intuition for using mutual information for syntagmatical relation mining is that words that are strongly associated have a high mutual information, whereas words that are not related have lower mutual information.

5 Conclusions

Word association is a form of analyzing the content of text data in search of relations between terms. In paradigmatic relations, a particular kind of word associations, the aim is to compute similarity of candidate words context documents, after collecting these context through a bag of words. The highly similar word pairs can then be treated as having paradigmatic relations, i.e. those words share similar contexts. From the several different approaches to implement the notion of paradigmatic relation, some related with text retrieval models were introduced, in order to help designing similarity functions to compute the paradigmatic relations. Specifically the BM25 and IDF weighting were used to discover paradigmatic relation.

For syntagmatic relations, the general idea is counting how many times two words occur together in a context.The co-occurrences of words must be compared with their individual occurrences. The assumption is that words with high co-occurrences but relatively low individual occurrences have syntagmatic relations. Conditional entropy and mutual information are the two approaches introduced for discovering syntagmatic relations.

The two mentioned relations are in fact closely related, since paradigmatic related words tend to have syntagmatic relation with the same word. This allows both relations being jointly searched.

Acknowledgements This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Portuguese Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2013.

References

1. Berry, M.W.: Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, New York (2004)
2. Berry, M.W., Castellanos, M.: Survey of Text Mining II: Clustering, Classification, and Retrieval. Springer, New York (2008)
3. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New Jersey (2012)
5. Hotho, A., Nrnberger, A., Paa, G.: A brief survey of text mining. LDV Forum - GLDV J. Comput. Linguist. Lang. Technol. **20**(1), 19–62 (2005)
6. Inzalkar, S., Sharma, J.: A survey on text mining-techniques and application. Int. J. Res. Sci. Eng. Techno-Xtreme **16**, 488–495 (2015)
7. Jiang, S., Zhai, C.: Random walks on adjacency graphs for mining lexical relations from big text data. In: Proceedings of IEEE International Conference on Big Data 2014. https://doi.org/10.1109/BigData.2014.7004272 (2014)
8. Lu, Y., Mei, Q., Zhai, C.: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf. Retr. **14**(2), 178–203 (2011)
9. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
10. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, Cambridge (2008)
11. Miner, G.: Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic, New York (2012)
12. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
13. Patel, M.R., Sharma, M.G.: A survey on text mining techniques. Int. J. Eng. Comput. Sci. **3**(5), 5621–5625 (2014)
14. Tated, R.R., Ghonge, M.M.: A survey on text mining-techniques and application. Int. J. Res. Advent Technol. **ICATEST2015**, 380–385 (2015)
15. Zhai, C., Massung, S.: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. Morgan & Claypool, Williston (2016)
16. Zhai, C.: Text mining and analytics. Available from https://www.coursera.org/learn/text-mining. Cited 25 May 2016 (2016)

Index

T. A. Oliveira et al. (eds.), *Recent Studies on Risk Analysis and Statistical Modeling*, Contributions to Statistics,
https://doi.org/10.1007/978-3-319-76605-8