Springer Proceedings in Mathematics & Statistics

Christos P. Kitsos
Teresa A. Oliveira
Alexandros Rigas
Sneh Gulati *Editors*

Theory and Practice of Risk Assessment

ICRA 5, Tomar, Portugal, 2013

Springer

Springer Proceedings in Mathematics & Statistics

Volume 136

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including operation research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

More information about this series at http://www.springer.com/series/10533

Christos P. Kitsos · Teresa A. Oliveira
Alexandros Rigas · Sneh Gulati
Editors

Theory and Practice of Risk Assessment

ICRA 5, Tomar, Portugal, 2013

Springer

Editors

Christos P. Kitsos
Technological Educational Institute of Athens
Athens
Greece

Teresa A. Oliveira
Universidade Aberta
Palácio Ceia, Lisbon
Portugal

Alexandros Rigas
Democritus University
Thrace
Germany

Sneh Gulati
Florida International University
Miami, FL
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-18028-1 ISBN 978-3-319-18029-8 (eBook)
DOI 10.1007/978-3-319-18029-8

Library of Congress Control Number: 2015937949

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Preface

Everything humans venture to do has some degree of risk involved in it. Since the degree of risk they face is random, it is only natural that we statisticians feel the need to jump in and try to analyze it. The plethora of papers in this field being published in the field of risk analysis eventually led to the launching of a conference series where statisticians could present their findings.

The International Committee on Risk Analysis initially launched the series as a conference on cancer risk assessment in Athens, Greece on August 22, 2003. The second and the third conferences held in 2007 and 2009, respectively, still retained cancer risk assessment as the main focus; however, after the third conference it was decided to broaden the theme of the conference. This led to the new acronym of the ICRA (**I**nternational **C**onference on **R**isk **A**nalysis) series, so the 4th meeting took place in 2011 at Limassol, Cyprus. Proceedings of these meetings are available.

The fifth conference on risk analysis, ICRA5 was held at Tomar, Portugal, from May 30 to June 1, 2013. The papers presented at this conference covered a number of topics on risk analysis with applications in both biological and industrial fields. This book forms a proceedings volume and includes most of the papers presented at the conference. The book itself is divided into two main parts based on the subject matter covered:

Part I is devoted to *Risk Methods for Bioinformatics* while
Part II focuses on *Risk Methods for Management and Industry.*

The papers in Part I mainly cover topics from Life Sciences and Environmetrics and are divided into subsections based on the primary focus of the papers included. The first subsection in The Bioinformatics section deals with the original theme of the conference: "cancer research" and consists of three chapters covering various aspects of cancer risk analysis. The second subsection consists of two chapters which consider the applications of first time hitting models while the third subsection considers papers on general quantification of risk for diseases. Finally, the last subsection in this first part considers risk analysis as it pertains to the environment.

The second part of the book, “Risk Analysis in Management and Industry,” is also divided into four subsections on the basis of the subject matter covered. The first chapter in this part though deals with sampling strategies and stands on its own. Section 2 considers papers on Industrial Quality Control and consists of two chapters. Next in Section 3, the focus is on Extreme Value Theory with the papers looking at various ways of quantifying extreme quantiles in natural catastrophic events. The last subsection in this part is devoted to general papers in reliability and survival analysis and consists of six chapters.

The total of 30 papers presented in this volume cover a diverse range of topics on risk analysis and we hope our readers find it useful for their research.

Contents

Part II Risk Methods for Management and Industry

Introduction

"Nothing Ventured, Nothing Gained," a well-known proverb implying that to attain something, one has to be willing to take risks. Thus, it is only natural that no professional venture or field can be devoid of risk. As an example, in medical studies, one is concerned with the risk of patient death, the risk of a lack of a cure, the risk of side effects from medications, etc. Engineers are concerned with the risk of structural and mechanical failures; manufacturers are concerned with the risk of producing defective products and so on. Mitigating risk and analyzing it then are integral components of any area, however, risk analysis per se, is a field specific to applied statistics. In an attempt to recognize the role that statistics plays in risk analysis, the International Committee on Risk Analysis decided to launch a conference series to serve as a forum for researchers in this area to get together and discuss their methodologies.

The main focus of the first conference was cancer research, and so the series was actually launched as the International Conference on Cancer Risk Assessment (ICCRA) in Athens, Greece on August 22, 2003. The second conference was held on the island of Santorini, Greece during May 25–27, 2007 (still as ICCRA) while the third and final ICCRA was held at Porto Heli, Greece during July 16–18, 2009. Thereafter it was decided to broaden the theme of the conference and the first ICRA (**I**nternational **C**onference on **R**isk **A**nalysis) labeled ICRA4 took place at Limassol, Cyprus, during May 26–29, 2011. From its inception, support for the series was provided by its main sponsor, the ISI Committee on Risk Analysis. One of the main aims of the committee was to improve and expand the role of statistics in risk analysis. While the committee retained human health, welfare, and survival as its main focus, it decided to more actively pursue risk analysis in other fields such as the environment, ecology, engineering, etc. Toward that end it was formally decided that the committee would engage more actively in conferences with a broader coverage of risk analysis, including identification and quantification of risk (http://www.isi-web.org/sections/44-com/com/126-ra).

The first realization of this goal was manifested in ICRA5 held at Tomar, Portugal, from May 30 to June 1, 2013. The conference drew together a number of scientists working on various aspects of risk analysis with applications in both biological and industrial fields. This book forms a proceedings volume and includes most of the papers presented at the conference.

While the past ICRA conferences have led to a number of publications, none have been as broad in subject coverage as this book. Examples of statistical analysis using real data are found throughout this book, which we hope will spark interest in the related theoretical results. Despite the fact that research in risk analysis has been burgeoning, the new methodologies have not had the deepest possible penetration among the practitioners of the field. We believe that this is because the relevant articles and papers are scattered in too many journals with different foci. We hope to remedy the situation with the publication of this volume, totally devoted to methods on Risk Analysis.

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

Part I is devoted to *Risk Methods for Bioinformatics* while
Part II focuses on *Risk Methods for Management and Industry.*

The papers in Part I mainly cover topics from Life Sciences and Environmetrics and are divided into subsections based on the primary focus of the papers included. We now briefly describe some of the topics covered in this section:

The first subsection in The Bioinformatics section deals with the original theme of the conference: "cancer research." "Generalized Information Criteria for the Best Logit Model" considers the use of Entropy measures to quantify relative risk and applies it to compute the relative risk of breast cancer for women based on their risk factors which include (but are not limited to): age, use of oral contraceptives, hormone replacement therapy. "Fractal Case Study for Mammary Cancer: Analysis of Interobserver Variability" uses a Fractal Case Study to classify different types of malignancies in cancer tissues. "On Analytical Methods for Cancer Research" uses Statistical Dynamic Shape Analysis to quantify cancer risk. The authors argue that temporal shaping in medicine has to consider the medical relevance for certain time points in the measurement and landmarks to describe the object at these time points. Their analysis is shown to have a distinct advantage in oncology compared to traditional approaches.

The second subsection on Bioinformatics looks at the applications of First Time Hitting Models. In remission studies, one often encounters long-term survivors, or a "cured fraction" of units, which will never experience the event of interest. As a result, the empirical survival function for such studies never tends to zero. First hitting time (FHT) models can be used to account for such phenomena in lifetime models, and an example relevant to treatment of drug users is presented in "Modelling Times Between Events with a Cured Fraction Using a First Hitting Time Regression Model with Individual Random Effects". First time hitting models can also be used to estimate the number of disease-free years lost to occupational exposure and "Acceleration, Due to Occupational Exposure, of Time to Onset of a Disease" uses such a model to estimate the "expected number of disease free years lost due to exposure to asbestos."

The next subsection focuses on general risk quantification for diseases. Relationships between relative risk, odds ratios, and their respective confidence intervals are discussed in "Transformations of Confidence Intervals for Risk Measures?". "Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification" presents an overview of the application of compound tests to classify individuals into two groups based on the presence or absence of a disease. In the same spirit, "A Maximum Likelihood Estimator for the Prevalence Rate Using Pooled Sample Tests" presents maximum likelihood methods to determine the prevalence rate of a disease. "On Intra-Individual Variations in Hair Minerals in relation to Epidemiological Risk Assessment of Atopic Dermatitis" discusses risk analysis in a Cohort Study of 842 mother-infant pairs for Atopic Dermatitis in Japan. The chapter looks at the association between hair minerals at one month and the onset of atopic dermatitis (AD) at ten months after birth with the aim of identifying infants with a high risk of getting the disease. "Assessing Risk Factors for Periodontitis Using Multivariable Regression Analysis" presents a deterministic mathematical model to evaluate risk factors for periodontitis (using data from Portugal) and the authors conclude that periodontitis is significantly associated with High Density Lipoproteins (HDL). "COPD: on Evaluating the Risk for Functional Decline" considers patients with Chronic Obstructive Pulmonary Disease (COPD) and uses a longitudinal study to measure their risk of becoming dependent on others for day-to-day activities. The goal is early intervention and assistance in order to reduce their dependence on others. "Microarray Experiments on Risk Analysis Using R" presents several designs to conduct microarray analysis using R, a technique that is being increasingly used to identify individuals at risk of getting a certain disease as well to identify the relevant risk factors. Finally, the last chapter in this subsection, "Risk Assessment of Complex Evolving Systems Involving Multiple Inputs" looks at complex nuerophysical systems with multiple inputs to evaluate whether some of the inputs inhibit the occurrence of new events.

The final subsection in this section deals with risk analysis for environmental sciences. While most of us want to live in a world completely free of pollution, realistically speaking that is an impossible dream. Hence, environmental policies focus on achieving "optimal pollution levels" where the marginal damage cost is equal to the marginal abetment cost. The authors "Monitoring Environmental Risk by a Methodology Based on Control Charts" argue that it is more efficient to focus instead on maximizing the net benefit (difference between abatement costs and damage costs.) They present different methods to evaluate the benefit area, which allows the comparison of different environmental policies. In the last chapter of this section, "Risk Problems Identifying Optimal Pollution Level", the authors propose a method for monitoring environmental risk through the use of control charts when the contaminant concentration follows a Birnbaum-Saunders distribution.

Next we turn to Part II: Risk Analysis in Management and Industry. The success of any industry is heavily dependent on its ability to deliver a product that is consistent and of high quality. Thus risk of failures, breakdowns, losses, inferior quality, etc., must all be identified and mitigated. Some methodologies to do just that are presented in this next section. As in the previous part, this section is also divided into subsections, although the first chapter in this section is in its own subsection since it deals with sampling strategies.

"Finite Populations Sampling Strategies and Costs Control" presents a brief and compact review of Sampling Techniques. As pointed out by the authors, sophisticated statistical techniques are useless when they use bad data. Thus the authors present a quick overview of sampling strategies to show how to deal with cost control in nonideal circumstances (where the practitioner is unable to sample randomly without replacement).

Thereafter, the section moves on to papers with applications to real-life problems. The first subsection is on Quality Control. Process Control Charts have revolutionized the concept of monitoring quality. However, standard quality control charts are predicated on the assumption of normality, which is often not the case with real data. "Industrial Production of Gypsum: Quality Control Charts" deals with the use of Box-Cox transformations to normalize data in order to construct appropriate control charts. An application to the production of gypsum (marketed only if it meets required specifications) is presented. The benefit of a tolerance region, rather than a confidence region for problems in Industry and Management, is explained and discussed in "Risk Analysis with Reference Class Forecasting Adopting Tolerance Regions".

The previous subsection provides a natural bridge to the next subsection on "Extreme Value Theory." In order to estimate guarantee values and tolerance limits, most manufacturers are concerned with the estimation of extreme quantiles, especially in the context of heavy-tailed distributions. Heavy-tailed distributions are a norm in financial and insurance data. Extreme quantiles in these settings are often called Value at Risk at level q (Var_q) or Probable Maximum Loss (PML). No risk analysis strategy in the business world is complete without an evaluation of Var_q and a number of papers in this section deal with the estimation of the same. Research methodologies presented deal with the enumeration of stable extreme value laws along with a characterization of their domains in "Randomly Stopped ***k***th Order Statistics", the use of a new class of skew-normal distributions to model heavy-tailed distributions in "The Role of Asymmetric Families of Distributions in Eliminating Risk", estimation of extreme rainfall levels using parametric and semiparametric methods in "Parametric and Semi-Parametric Approaches to Extreme Rainfall Modelling", the use of Pareto Probability Weighted Moments (PPWM) and semi-parametric methods to estimate extreme quantiles in "A Log Probability Weighted Moment Estimator of Extreme Quantiles" and "A Mean-of-order-*p* Class of Value-at-Risk Estimators" respectively. The last chapter in this subsection, "Adaptive Choice and Resampling Techniques in Extremal Index Estimation" presents the use of resampling techniques to estimate the extreme value index, an important parameter in the estimation of PML.

The remaining papers fall under the general heading of "Applications in Reliability and Survival Analysis" and are summarized below:

In a number of experiments in industrial quality control and reliability, observed data often consist of record-breaking values where only successive maxima or minima are recorded. Such data also routinely arise in fields like Climatology, Geosciences, and athletics. The authors in "Some Estimation Techniques in Reliability and Survival Analysis Based on Record-Breaking Data", present a review of the results on statistical inference from records that can be used in Reliability and Survival Analysis, including inferential results for heavy-tailed distributions. Commercial credit management is a matter of great importance for most small and medium enterprises (SMEs), since it represents a significant portion of their assets. Commercial lending involves assuming some credit risk due to exposure to default. Thus, the Management of Trade Credit and payment delays are strongly related to the liquidation and bankruptcy of these enterprises. The relationship between Trade Credit Management and the level of risk in SMEs is extensively discussed in "Risk Scoring Models for Trade Credit in Small and Medium Enterprises". The concept of signature is a powerful tool in the analysis of reliability systems and networks. However, most papers on this topic have dealt with a system of i.i.d components. "Signatures of Systems with Non-exchangeable Lifetimes: Some Implications in the Analysis of Financial Risk" considers the expansion of this concept to the non-exchangeable case, which allows applications to systems in different fields, such as Economics, Financial Risk, Environmental Sciences, etc. One of the key factors in quality control and risk quantification is the identification of outliers. In autoregressive time series one encounters two basic types of outliers: additive outliers (AO), affecting only a particular observation, and innovative outliers (IO), which act as an addition to the noise at a point in the entire series. Tests to detect the two types of outliers are presented in "Detecting an IO/AO Outlier in a Set of Time Series". Response Surface Methodology (RSM) is becoming more and more important as a risk assessment tool in this ever-changing world where big data and several dependent variables are the norm. Thus in "Response Surface Methodology: A Review of Applications to Risk Assessment" presents a review of the various aspects on the use of RSM as a risk assessment tool in the environmental, financial and public health fields. The final chapter in our book, "FF-type Multivariate Models in an Enforced Regression Paradigm" considers the use of "Enforced Regression Theory" to describe the relationship between a dependent variable and several independent variables.

In conclusion, the 30 papers included in this volume are diverse in nature, some applied, some theoretical, with a number of them providing the essential bridge between the two. In addition, several review papers included here fulfill the mission of the committee to put forth publications with papers that review various methodologies in risk assessment.

Given its scope and the straightforward nature of the presentation, we believe that this book will help a new generation of statisticians and practitioners to solve complex problems in risk analysis. Therefore, this book can easily serve as a textbook for a special topics course in risk analysis.

All of the papers collected here 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 their submission of papers, but also for their expeditious revisions and for incorporating the reviewer's suggestions.

Thanks also go out to the ISI Committee on Risk Analysis for sponsoring the conference. We would also like to express our sincere gratitude to all the people who worked on various committees, served as chairpersons, reviewed papers, and of course presented their own work at ICRA5. Without them ICRA5 would have never seen the light of day.

Last but not least, the editors would also like to express their heartfelt thanks and gratitude to SPRINGER for their help and support with this volume, especially, Dr. Eva Hiripi and Udhayakumar Panneerselvam without whose valuable assistance we could never have realized this manuscript.

Christos P. Kitsos
Chair, ISI Committee on Risk Analysis
Technological Educational Institute of Athens

Teresa A. Oliveira
Secretary, ISI Committee on Risk Analysis
Universidade Aberta and Center of Statistics and
Applications of University of Lisbon

Alexandros Rigas
Democritus University of Thrace

Sneh Gulati
Florida International University, Miami, FL

Part I
Risk Methods for Bioinformatics

Generalized Information Criteria for the Best Logit Model

Christos P. Kitsos and Thomas L. Toulias

Abstract In this paper the γ–order Generalized Fisher's entropy type Information measure (γ–GFI) is adopted as a criterion for the selection of the best Logit model. Thus the appropriate Relative Risk model can be evaluated through an algorithm. The case of the entropy power is also discussed as such a criterion. Analysis of a real breast cancer data set is conducted to demonstrate the proposed algorithm, while algorithm's realizations, through MATLAB scripts, are cited in Appendix.

Keywords Fisher's entropy measure · Logit model · Relative Risk · Breast Cancer

1 Introduction

The two main lines of thought are adopted as far as the Fisher's information measure is concerned: The parametric approach and the entropy power [2]. In this paper we shall use the generalized form of the Fisher's entropy type information measure, as developed in Sect. 2, as well as a generalized form of the usual normal distribution. In Sect. 3 the binary response model is related to the developed theory of Sect. 2. An algorithm is proposed to choose the best binary response model for evaluating the Relative Risk. As a binary response case that demonstrates the algorithm the breast cancer problem is studied, see [16] among others.

The pioneering work of Jaynes [10] on the maximum entropy principle in Statistical Thermodynamics led to the adoption of this principle to other fields of interest. In the following, a compact form of various parametric and non–parametric information measures is discussed.

C.P. Kitsos (✉) · T.L. Toulias
Technological Educational Institute of Athens, Ag. Spyridonos
and Palikaridi Str., 12210 Egaleo, Athens, Greece
e-mail: xkitsos@teiath.gr

T.L. Toulias
e-mail: t.toulias@teiath.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_1

Let X be a random variable with probability density $f(x;\ \theta)$, with $\theta \in \Theta$ being the parameter vector from the parameter space $\Theta \subseteq \mathbb{R}^p$. Let

$$U(\theta) := \tfrac{\partial}{\partial\theta} \log f(x;\ \theta), \quad \theta \in \Theta \subseteq \mathbb{R}^p,$$

be the parametric score function. Then, the parametric information measure $\mathrm{I}(\theta)$ can be defined as

$$\mathrm{I}(\theta) := g\left(\mathrm{E}[h(U(\theta))]\right), \quad \theta \in \Theta \subseteq \mathbb{R}^p,$$

where g and h being defined as real–valued functions, and $\mathrm{E}[\cdot]$ denoting the expected value operator with respect to the parameter θ. For the univariate case, if $g :=$ id. the Fisher's information measure $\mathrm{I}_F(\theta)$ is defined when $h(U) := U^2$, and the Vajda's information measure $\mathrm{I}_V(\theta)$ when $h(U) := |U|^\lambda$, $\lambda \geq 1$. When $g(A) =: A^k$, the Mathai's information measure $\mathrm{I}_M(\theta)$ is obtained when $h(U) := |U|^\lambda$ and $k = 1/\lambda$, $\lambda \geq 1$, while the Boeke's information measure $\mathrm{I}_B(\theta)$ is defined with $h(U) := |U|^{\lambda/(\lambda-1)}$ and $k = \lambda - 1$, $1 \neq \lambda > 0$. That is,

$$\mathrm{I}(\theta) = \begin{cases} \mathrm{I}_F(\theta), & g := \text{id.}, \quad h(U) := U^2, \\ \mathrm{I}_V(\theta), & g := \text{id.}, \quad h(U) := U|^\lambda,\ \lambda \geq 1, \\ \mathrm{I}_M(\theta), & g(A) := A^k,\ h(U) := |U|^\lambda,\ k = 1/\lambda,\ \lambda \geq 1, \\ \mathrm{I}_B(\theta), & g(A) := A^k,\ h(U) := |U|^{\frac{\lambda}{\lambda-1}},\ k = \lambda - 1,\ \lambda \in \mathbb{R}_+ \setminus 1. \end{cases} \tag{1}$$

Some of the merits of Fisher's information measure, $\mathrm{I}_F(\theta)$, it remains invariant under orthogonal transformation, provides the well known Cramer–Rao lower bound, and plays an important role in optimal experimental design theory, see [8, 12, 23]. Therefore, $\mathrm{I}(\theta)$ as in (1), defines the $\mathfrak{F}_1$ family of information measures.

There are two main problems in applications concerning Fisher's $\mathrm{I}_F(\theta)$ measure: is the measure singular or ill-conditioned? (see also [23]).

The elements of the non–parametric family $\mathfrak{F}_2$ of information measures are defined, through two given distributions $f_i = \frac{dP_i}{d\mu}$, $P_i \ll \mu$, $i = 1, 2$ with μ a σ–finite measure, that is

$$\mathfrak{F}_2 := \left\{ \mathrm{I}(f_1, f_2) : \quad \mathrm{I}(f_1, f_2) := g\left(\int h(f_1, f_2) d\mu \right) \right\}. \tag{2}$$

Some known information measures (i.m.), or divergences, such as the Kullback–Leibler i.m. $\mathrm{I}_{KL}(f_1, f_2)$, the Vajda's i.m. $\mathrm{I}_V(f_1, f_2)$, the Kagan i.m. $\mathrm{I}_K(f_1, f_2)$, the Csiszar i.m. $\mathrm{I}_C(f_1, f_2)$, the Matusita i.m. $\mathrm{I}_M(f_1, f_2)$, as well as the Rényi's divergence $\mathrm{I}_R(f_1, f_2)$, are defined as follows

$$\mathrm{I}(f_1, f_2) = \begin{cases} \mathrm{I}_{KL}(f_1, f_2), \; g(A) := A, & h(f_1, f_2) := f_1 \log(f_1/f_2), \\ \mathrm{I}_V(f_1, f_2), \quad g(A) := A, & h(f_1, f_2) := f_1 \left|1 - (f_2/f_1)\right|^{\lambda}, \; \lambda \geq 1, \\ \mathrm{I}_K(f_1, f_2), \quad g(A) := A, & h(f_1, f_2) := f_1 \left|1 - (f_2/f_1)\right|^{2}, \\ \mathrm{I}_C(f_1, f_2), \quad g(A) := A, & h(f_1, f_2) := f_2 \phi(f_1/f_2), \; \phi \text{ convex}, \\ \mathrm{I}_M(f_1, f_2), \quad g(A) := \sqrt{A}, & h(f_1, f_2) := (\sqrt{f_1} - \sqrt{f_2})^2, \\ \mathrm{I}_R(f_1, f_2), \quad g(A) := \frac{\log A}{1-\lambda}, & h(f_1, f_2) := f_1^{\lambda} f_2^{1-\lambda}, \; 1 \neq \lambda > 0. \end{cases} \tag{3}$$

We can obtain the corresponding parametric information measures from the non-parametric ones, through the following general scheme, [6, 22]:

$$\mathrm{I}(\theta) = \lim_{\Delta\theta \to 0} \inf_{\Delta\theta} \left\{ \frac{1}{\Delta\theta^2} \mathrm{I}\left(f(x;\theta), f(x;\theta + \Delta\theta)\right) \right\}. \tag{4}$$

Then, for the univariate case and under certain regularity conditions [2], it can be proved that the parametric K–L information measure $\mathrm{I}_{KL}(\theta)$, i.e. (4) with $\mathrm{I}(\cdot, \cdot)$ being the K–L measure $\mathrm{I}_{KL}(\cdot, \cdot)$ as in (3), is the half of the Fisher's $\mathrm{I}_F(\theta)$ as in in (1), and $2/\lambda$ of Reyni's $\mathrm{I}_R(\theta)$, or

$$\mathrm{I}_{KL}(\theta) = \tfrac{1}{2}\mathrm{I}_F(\theta) = \tfrac{2}{\lambda}\mathrm{I}_R(\theta).$$

The two afore mentioned families of information measures, $\mathfrak{F}_1$ and $\mathfrak{F}_2$ are considered for the parametric case. Some of these measures attract special interest in ecological studies, see [1]. In the next Section we present the entropy type information measures.

2 Entropy Type Information Measures

Now as far as the entropy type information measures are concerned, notice that the well known Fisher's entropy type information measure $\mathrm{J}(X)$ of a p–variate random variable is given by

$$\mathrm{J}(X) = \int\limits_{\mathbb{R}^p} [\nabla f(x)][\nabla \log f(x)] dx = \int\limits_{\mathbb{R}^p} f(x) \|\nabla \log f(x)\|^2 dx. \tag{5}$$

Recall that the Shannon entropy H of a r.v. X is defined as, [2],

$$\mathrm{H}(X) := -\int\limits_{\mathbb{R}^p} f(x) \log f(x) dx, \tag{6}$$

while the corresponding entropy power is given by

$$\mathrm{N}(X) = \nu e^{\frac{2}{p}\mathrm{H}(X)}, \tag{7}$$

with $\nu = (2\pi e)^{-1}$, see [2] for details.

Introducing an extra parameter, say δ, Kitsos and Tavoularis in [15] defined the Generalized (entropy type) Fisher's Information measure (δ–GFI), J_δ as follows:

$$\mathrm{J}_\delta(X) := \int\limits_{\mathbb{R}^p} f(x)\|\nabla \log f(x)\|^\delta dx. \tag{8}$$

For parameter value $\delta = 2$ we get the known Fisher's information, i.e. $\mathrm{J}_2(X) = \mathrm{J}(X)$. The extension of the entropy power, the Generalized Entropy Power (δ–GEP) is defined for $\delta \in \mathbb{R} \setminus [0, 1]$, as

$$\mathrm{N}_\delta(X) := \nu_\delta e^{\frac{\delta}{p}\mathrm{H}(X)}, \tag{9}$$

where

$$\nu_\delta := \left(\tfrac{\delta-1}{\delta e}\right)^{\delta-1} \pi^{-\delta/2} \left[\frac{\Gamma(\frac{p}{2}+1)}{\Gamma(p\frac{\delta-1}{\delta}+1)}\right]^{\delta/p}, \quad \delta \in \mathbb{R} \setminus [0, 1], \tag{10}$$

see [15]. Trivially, when $\delta = 2$, (9) is reduced to the known entropy power $\mathrm{N}(X)$, i.e. $\mathrm{N}_2(X) = \mathrm{N}(X)$ as $\nu_2 = \nu$. Moreover, it can be shown [15] that

$$\mathrm{J}_\delta(X)\mathrm{N}_\delta(X) \geq p, \tag{11}$$

with p being the number of the involved parameters, i.e. $\Theta \subseteq \mathrm{R}^p$. Therefore, $\mathrm{J}_\delta(X) \approx p/\mathrm{N}_\delta(X)$.

The above extensions give rise to a generalization of the multivariate normal distribution. This new distribution plays the same role as the classical normal distribution for the Fisher's entropy type information, and we shall call it the γ–order Generalized Normal Distribution (γ–GND).

Recall the p-dimensional random variable X_γ is said to follow the γ–GND, denoted by $X_\gamma \sim \mathcal{N}_\gamma^p(\mu, \varSigma)$, with mean vector $\mu \in \mathbb{R}^{p\times 1}$ and positive definite scale matrix $\varSigma \in \mathbb{R}^{p\times p}$, when the density function, f_{X_γ}, is of the form, see [15, 17],

$$f_{X_\gamma}(x;\ \mu, \varSigma) := C_\gamma^p(\Sigma) \exp\left\{-\tfrac{\gamma-1}{\gamma} Q(x)^{\frac{\gamma}{2(\gamma-1)}}\right\}, \quad x \in \mathbb{R}^{p\times 1}, \tag{12}$$

with quadratic form $Q(x) := (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$ and the normality factor $C_\gamma^p(\Sigma)$ defined as

$$C_\gamma^p(\Sigma) := \pi^{-p/2} \frac{\Gamma(\frac{p}{2}+1)}{\Gamma\left(p\frac{\gamma-1}{\gamma}+1\right)} (\tfrac{\gamma-1}{\gamma})^{p\frac{\gamma-1}{\gamma}} \, |\det \varSigma|^{-1/2}. \tag{13}$$

Notice that the 2–GND coincides with the usual multivariate (elliptically contoured) Normal distribution, i.e. $\mathscr{N}_2^p(\mu, \Sigma) = \mathscr{N}^p(\mu, \Sigma)$, while the 1–GND and the ($\pm\infty$)–GND reduced in limit to the multivariate (elliptically contoured) Laplace and Uniform distributions respectively, i.e. $\mathscr{N}_1(\mu, \Sigma) = \mathscr{U}^p(\mu, \Sigma)$ and $\mathscr{N}_{\pm\infty}^p(\mu, \Sigma) = \mathscr{L}^p(\mu, \Sigma)$. Moreover, for dimensions $p = 1, 2$ the 0–GND is reduced to the degenerate Dirac distribution, i.e. $\mathscr{N}_0^p(\mu, \Sigma) = \mathscr{D}^p(\mu)$. See [21] for details.

Proposition 1 *The Shannon entropy of a random variable $X_\gamma \sim \mathscr{N}_\gamma^p(\mu, \Sigma)$, is of the form*

$$\mathrm{H}(X_\gamma) = p\tfrac{\gamma-1}{\gamma} - \log C_\gamma^p(\Sigma). \tag{14}$$

Proof Consider the p.d.f. f_{X_γ} as in (12). From the definition (6) we have that the Shannon entropy of X is

$$\mathrm{H}(X_\gamma) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\Sigma)\tfrac{\gamma-1}{\gamma} \int\limits_{\mathbb{R}^p} Q(x)^{\frac{\gamma}{2(\gamma-1)}} \exp\left\{-\tfrac{\gamma-1}{\gamma} Q(x)^{\frac{\gamma}{2(\gamma-1)}}\right\} dx.$$

Applying the linear transformation $z = (x-\mu)^{\mathrm{T}}\Sigma^{-1/2}$ with $dx = d(x-\mu) = \sqrt{|\det\Sigma|}dz$, the $\mathrm{H}(X_\gamma)$ above is reduced to

$$\mathrm{H}(X_\gamma) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)\tfrac{\gamma-1}{\gamma} \int\limits_{\mathbb{R}^p} \|z\|^{\frac{\gamma}{\gamma-1}} \exp\left\{-\tfrac{\gamma-1}{\gamma} \|z\|^{\frac{\gamma}{\gamma-1}}\right\} dz,$$

where $\mathbb{I}_p$ denotes the $p \times p$ identity matrix. Switching to hyperspherical coordinates, we get

$$\mathrm{H}(X_\gamma) = -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)\tfrac{\gamma-1}{\gamma}\omega_{p-1} \int\limits_{\mathbb{R}_+} \rho^{\frac{\gamma}{\gamma-1}} \exp\left\{-\tfrac{\gamma-1}{\gamma} \rho^{\frac{\gamma}{\gamma-1}}\right\} \rho^{p-1} d\rho,$$

where $\omega_{p-1} := 2\pi^{p/2}/\Gamma\left(\frac{p}{2}\right)$ is the volume of the $(p-1)$–sphere. Applying the variable change $du := d(\frac{\gamma-1}{\gamma}\rho^{\gamma/(\gamma-1)}) = \rho^{1/(\gamma-1)}d\rho$ we obtain successively

$$\begin{aligned}\mathrm{H}(X_\gamma) &= -\log C_\gamma^p(\Sigma) + C_\gamma^p(\mathbb{I}_p)\omega_{p-1} \int\limits_{\mathbb{R}_+} ue^{-u}\rho^{\frac{(p-1)(\gamma-1)-1}{\gamma-1}} du \\ &= -\log C_\gamma^p(\Sigma) + p\tfrac{\gamma-1}{\gamma}\Gamma(p\tfrac{\gamma-1}{\gamma})C_\gamma^p(\mathbb{I}_p)\omega_{p-1}.\end{aligned}$$

Finally, by substitution of the volume ω_{p-1} and the normalizing factors $C_\gamma^p(\Sigma)$ and $C_\gamma^p(\mathbb{I}_p)$ and as in (13), relation (14) is obtained.

Example 1 Substituting (14) into (9), the δ–GEP is then

$$\mathrm{N}_\delta(X_\gamma) = \left(\tfrac{\delta-1}{e\delta}\right)^{\delta-1} \left(\tfrac{e\gamma}{\gamma-1}\right)^{\delta\frac{\gamma-1}{\gamma}} \left[\frac{\Gamma\left(p\dfrac{\gamma-1}{\gamma}+1\right)}{\Gamma\left(p\dfrac{\delta-1}{\delta}+1\right)} \right]^{\delta/p} |\det\Sigma|^{\frac{\delta}{2p}}. \tag{15}$$

Moreover, the generalized Fisher's entropy type information measure $\mathrm{J}_\delta(X_\gamma)$ with X_γ spherically contoured, i.e. $X_\gamma \sim \mathscr{N}_\gamma(\mu, \sigma^2\mathbb{I}_p)$, is given by the formula, [20],

$$\mathrm{J}_\delta(X_\gamma) = \left(\tfrac{\gamma}{\gamma-1}\right)^{\frac{\delta}{\gamma}} \frac{\Gamma\left(\frac{\delta+p(\gamma-1)}{\gamma}\right)}{\sigma^\delta\Gamma\left(p\frac{\gamma-1}{\gamma}\right)}. \tag{16}$$

Example 2 For the usual entropy power of the γ–GND, i.e. for the second–GEP of the r.v. $X_\gamma \sim \mathscr{N}_\gamma(\mu, \Sigma)$, we have that

$$\mathrm{N}(X_\gamma) = \tfrac{1}{2e}\left(\tfrac{e\gamma}{\gamma-1}\right)^{2\frac{\gamma-1}{\gamma}} \left[\frac{\Gamma\left(p\frac{\gamma-1}{\gamma}+1\right)}{\Gamma\left(\frac{p}{2}+1\right)} \right]^{2/p} |\det\Sigma|^{1/p}.$$

Note that for the limiting cases of X_1 (1–GND) and $X_{\pm\infty}$ ($\pm\infty$–GND) we obtain the usual entropy power for the multivariate (and elliptically contoured) Uniform and Laplace distributions respectively, i.e.

$$\mathrm{N}(X_1) = \lim_{\gamma\to1^+}\mathrm{N}(X_\gamma) = \frac{|\det\Sigma|^{1/p}}{2e\Gamma^{2/p}\left(\frac{p}{2}+1\right)},$$

$$\mathrm{N}(X_{\pm\infty}) = \lim_{\frac{\gamma-1}{\gamma}\to1^+}\mathrm{N}(X_\gamma) = 2^{\frac{2-p}{p}} e \left[\frac{(p-1)!\sqrt{|\det\Sigma|}}{\Gamma\left(\frac{p}{2}\right)} \right]^{2/p}.$$

Finally, for the r.v. X_2 we obtain the usual entropy power for the multivariate Normal, i.e. $\mathrm{N}(X_2) = \sqrt[p]{|\det\Sigma|}$ (Fig. 1).

Table 1 provides an evaluation of $\mathrm{N}_\delta(X_\gamma)$ with $X_\gamma \sim \mathscr{N}_\gamma^1(0,1)$ for certain γ and $\delta \geq 1$ values.

Example 3 We make the following observations here. When $\delta = \gamma$, from (15), we have that

$$\mathrm{N}_\gamma(X_\gamma) = |\det\Sigma|^{\frac{\gamma}{2p}}. \tag{17}$$

Thus, $\mathrm{N}_0(X_0) = 1$, i.e. the 0–GEP of the Dirac distributed, $X_0 \sim \mathscr{D}(0)$ is 1 while $\mathrm{N}_\delta(X_0) = +\infty$ for every defined $\delta \in \mathbb{R}\setminus[0,1]$ as it is derived through (15).

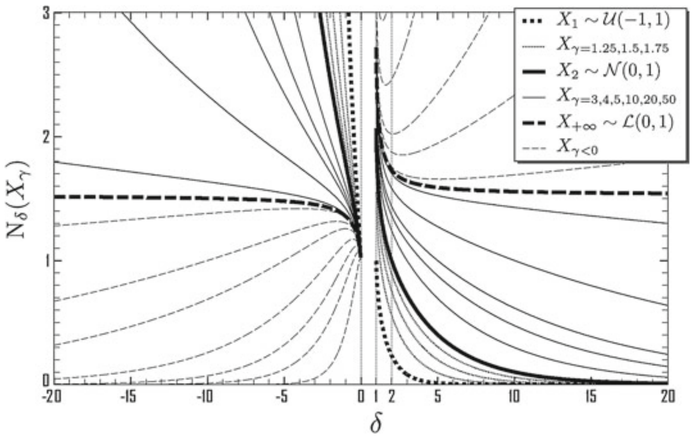


Fig. 1 Graphs of $\mathrm{N}_\delta(X_\gamma)$ along δ for various γ values, where $X_\gamma \sim \mathcal{N}_\gamma(0, \sigma^2)$ with $\sigma = 0.8, 1, 1.5$

Table 1 Evaluation of $\mathrm{N}_\delta(X_\gamma)$ with $X_\gamma \sim \mathcal{N}_\gamma^1(0, 1)$ for various γ and $\delta \geq 1$ parameters

γ/δ	1	3/2	2	3	5	10	50	$+\infty$
-50	2.7412	1.8834	1.7598	1.684	1.6566	1.6912	2.3295	$+\infty$
-10	2.8309	1.9766	1.8769	1.8549	1.9461	2.3340	11.660	$+\infty$
-5	2.9393	2.0912	2.0233	2.0761	2.3482	3.3981	76.283	$+\infty$
-2	3.2430	2.4235	2.463	2.7884	3.8392	9.0834	10411.	$+\infty$
-1	3.6945	2.9469	3.1967	4.1230	7.3677	33.452	7.05e+6	$+\infty$
$-1/10$	8.3767	10.0610	16.434	48.057	441.49	1.2e+5	$\approx +\infty$	$+\infty$
1	1.0	0.4150	0.2342	0.0818	0.0107	7.06e–5	2.95e–22	0.0
3/2	1.7974	1.0	0.7566	0.4748	0.2008	0.0248	1.59e–9	0.0
2	2.0664	1.2327	1.0	0.7214	0.4032	0.1002	1.74e–6	0.0
3	2.3040	1.4513	1.2433	1.0	0.6950	0.2977	0.0004	0.0
5	2.4779	1.6187	1.438	1.2440	1.0	0.6163	0.0149	0.0
10	2.6009	1.7406	1.5842	1.4384	1.2739	1.0	0.1684	0.0
50	2.6952	1.8362	1.7012	1.6007	1.5223	1.4281	1.0	0.0
$\pm\infty$	2.7183	1.8598	1.7305	1.6422	1.5886	1.5552	1.5317	1.0

Moreover, $\mathrm{N}_1(X_1) = |\det\Sigma|^{1/(2p)}$ while $\mathrm{N}_1(X_{\pm\infty}) = e|\det\Sigma|^{1/(2p)}$. For the Laplace distributed $X_{\pm\infty}$, (17) implies

$$\mathrm{N}_{\pm\infty}(X_{\pm\infty}) = \begin{cases} 0, & |\det\Sigma| < 1, \\ 1, & |\det\Sigma| = 1, \\ +\infty, & |\det\Sigma| > 1. \end{cases}$$

See Appendix 1 for the minimum/maximum analysis of the generalized entropy power N_δ of a γ–GND random variable.

3 Generalized Fisher's Information and Relative Risk

Consider a subject with attributes given by the input vector $X = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$. In risk analysis, the focus is on the parameter $p(x)$, i.e. the probability that this subject has a certain characteristic C, given that the input vector takes on the real vector value x, i.e. $X = x$, and measures the odds ratio or the Relative Risk (RR):

$$RR = \frac{p(x)}{1-p(x)} \quad \text{with} \quad \log\frac{p(x)}{1-p(x)} = x^{\mathrm{T}}\beta, \tag{18}$$

where β being an appropriate vector of regression parameters, see also [11, 13].

Due to the Central Limit Theorem, the involved Binomial distribution $\mathscr{B}(n, P)$, $P = p(x)$, corresponding to the binary response model under investigation, approximated by the Normal distribution, i.e.

$$\mathscr{B}(n, P) \approx \mathscr{N}\left(nP, nP(1-P)\right) = \mathscr{N}_2^1\left(nP, nP(1-P)\right).$$

For the normally distributed $X \sim \mathscr{N}\left(\mu, \sigma^2\right) := \mathscr{N}\left(nP, nP(1-P)\right)$, the Shannon entropy is

$$\mathrm{H}(X) = \tfrac{1}{2} + \log\sqrt{2\pi nP(1-P)},$$

while the entropy power is

$$\mathrm{N}(X) = \tfrac{1}{2\pi e}e^{2\mathrm{H}(X)} = \tfrac{1}{2\pi e}e^{1+\log\{2\pi nP(1-P)\}} = nP(1-P){:=}\sigma^2.$$

The generalized entropy power $\mathrm{N}_\delta(X)$ introduced in (9), for this case is

$$\mathrm{N}_\delta(X) = (\tfrac{\pi}{2})^{\frac{\delta}{2}}e^{1-\frac{\delta}{2}}(\tfrac{\delta-1}{\delta})^{\delta-1}\Gamma^{-\delta}(\tfrac{\delta-1}{\delta}+1)[nP(1-P)]^{\delta}. \tag{19}$$

Given the above discussion we propose the following algorithm for the examination of the optimum variable entering the logit model, based on the methodology of [11, 18], while for the maximum entropy see [22]. That is, for each k–risk variable model the maximum δ-GFI models (with respect to the parameter δ) is chosen, and among them, we obtain the one with the minimum γ-GFI value. The steps of the proposed algorithm for a bioassay, [7] are presented as follows:

3.1 Algorithm

Step 1. For the k risk variables X_j, $j = 1, 2, \dots, k$ involved to the bioassay, the average value $\bar{P}_j$ is given by

$$\bar{P}_j = \tfrac{1}{n} \sum_{i=1}^{n} P_{ji},$$

where $P_{ji} = p_i(x_{ji}) = \mathrm{Logit}^{-1}(x_j^{\mathrm{T}} \beta)$, $i = 1, 2, \dots, n$. Following the Central Limit Theorem, we have $X_j \sim \mathscr{N}(\mu_j, \sigma_j^2)$, $\mu_j = n\bar{P}_j$ and $\sigma_j^2 = n\bar{P}_j(1 - \bar{P}_j)$, $j = 1, 2, \dots, k$.

Step 2. Choose the parameter γ, say γ_0, which provides the minimum $\sigma^2_{j;\gamma}$ value i.e.

$$\min_{\gamma}\{\sigma^2_{j;\gamma}\} = \sigma^2_{j;\gamma_0}, \tag{20}$$

where $\sigma^2_{j;\gamma}$ is the scale parameter of the extended γ–GND risk variables $X_{j;\gamma} \sim \mathscr{N}_\gamma(\mu_j, \sigma_j^2)$ whose variance is σ_j^2 from Step 1.

Step 3. For all the $X_{j;\gamma}$ models calculate the δ–GFI $\mathrm{J}_\delta(X_{j;\gamma_0})$ values and choose the parameter δ, say δ_0, which maximizes the above i.e.

$$\max_{\delta}\{\mathrm{J}_\delta(X_{j;\gamma_0})\} = \mathrm{J}_{\delta_0}(X_{j;\gamma_0}). \tag{21}$$

Step 4. Finally, we choose the model obtained from the optimum input variable X_{opt} as

$$\mathrm{J}_{\delta_0}(X_{opt}) = \min_{j}\{\mathrm{J}_{\delta_0}(X_{j;\gamma_0})\} = \min_{j} \max_{\delta}\{\mathrm{J}_\delta(X_{j;\gamma_0})\}. \tag{22}$$

We apply the above algorithm to the example discussed below.

3.2 Application

A number of breast cancer risk factors have been established in research studies [3, 4, 14], such as the late age at first childbirth, early age of menarche, use of oral contraceptives or hormone replacement therapy, etc. In his section, we extensively discuss the results in [11] following the breast cancer analysis in [16] where the collected data for 98 breast cancer patients and 125 healthy controls through the logit model are considered. As input variables were considered: the `Age` of the woman, the years of `Menarche` and `Menopause`, as well as the frequencies of estrogen biosynthesis `CYP17`, and the inactivation `COMT`, see [9]. Variable `CYP17` attracts particular interest in bibliography, see [5] and for applications see [16, 19].

Table 2 Coefficients of the logit model for the one input variable model

Variables	β_0	β_1
`Age`	−3.45	0.55
`COMT`	−0.195	−0.03
`CYP17`	−0.419	0.095
`Menarche`	0.08902	−0.026
`Menopause`	−3.55	0.067

Table 2 provides the coefficients β_0 and β_1 of the one variable logit model for the denoted input variables, see also [11] for details.

Step 1. Let $X_j \sim \mathscr{N}\left(\bar{P}_j, p\bar{P}_j(1-\bar{P}_j)\right)$, $j = 1, 2, \ldots, 5$ correspond to the variables `Age`, `COMT`, `CYP17`, `Menarche` and `Menopause` respectively. Also let m_j be the minimum ($\delta > 1$)–GEP value for every X_j, i.e. $m_j := \min_{\delta>1}\{\mathrm{N}_\delta(X_j)\}$ obtained for parameter $\delta = \delta_j^{min}$, i.e. $\mathrm{N}_{\delta_j^{min}}(X_j) = m_j$. Also let M_j be the maximum ($\delta < 0$)–GEP value for every X_j, i.e. $M_j := \max_{\delta<0}\{\mathrm{N}_\delta(X_j)\}$ obtained for parameter $\delta = \delta_j^{max}$, i.e. $\mathrm{N}_{\delta_j^{max}}(X_j) = M_j$, $j = 1, 2, \ldots, 6$. Table 3 provides the corresponding numerical evaluations, extending results in [11], see Appendix 3.

Step 2. We consider the extended γ–GND r.v. $X_{j;\gamma} \sim \mathscr{N}_\gamma(n\bar{P}_j, \sigma_{j;\gamma}^2)$, $j = 1, 2, \ldots, 5$, such that $\mathrm{Var}(X_{j;\gamma}) = \sigma_j^2 = n\bar{P}_j(1-\bar{P}_j)$ for all $j = 1, 2, \ldots, 5$. For $\gamma = 2$ the classic Normal distribution is obtained, and hence $X_{j;2} = X_j$ and $\sigma_{j;2} = \sigma_j$, $j = 1, 2, \ldots, 5$. Recall the variance of the γ–GND $X_{j;\gamma}$ after [21],

$$\sigma_j^2 = \mathrm{Var}(X_{j;\gamma}) = (\tfrac{\gamma}{\gamma-1})^{2\frac{\gamma-1}{\gamma}} \frac{\Gamma(3\frac{\gamma-1}{\gamma})}{\Gamma(\frac{\gamma-1}{\gamma})}\sigma_{j;\gamma}^2, \quad j = 1, 2, \ldots, 5. \tag{23}$$

Thus, (15) can be written, with $\delta = 2$, as

$$\mathrm{N}(X_{j;\gamma}) = \mathrm{N}_2(X_{j,\gamma}) = \tfrac{2}{\pi} e^{2\frac{\gamma-1}{\gamma}-1} (\tfrac{\gamma-1}{\gamma})^2 \frac{\Gamma^3(\frac{\gamma-1}{\gamma})}{\Gamma(3\frac{\gamma-1}{\gamma})} n\bar{P}_j(1-\bar{P}_j), \quad j = 1, 2, \ldots, 5.$$

Table 3 Evaluation of μ_j, σ_j^2 as well as m_j, δ_j^{min} and M_j, δ_j^{max} for the jth input variable (Normal approximation)

j	X_j	μ_j	σ_j^2	m_j	δ_j^{min}	M_j	δ_j^{max}
1.	`Age`	222.999	0.0006	0.	$+\infty$	$+\infty$	$-\infty$
2.	`COMT`	98.0708	54.9413	14.725	1.0425	1.005	–0.0101
3.	`CYP17`	97.9825	54.9306	14.723	1.0424	1.005	–0.0101
4.	`Menarche`	97.9969	54.9324	14.724	1.0424	1.005	–0.0101
5.	`Menopause`	96.2594	54.7084	14.692	1.0425	1.005	–0.0101

Table 4 Evaluation of the parameter $\sigma^2_{j,\gamma}$ of $X_{j;\gamma} \sim \mathscr{N}_\gamma(\mu_j, \sigma^2_{j;\gamma})$ such that $\mathrm{Var}(X_{j;\gamma}) = \sigma^2_j = n\bar{P}_j(1 - \bar{P}_j)$

γ	COMT	CYP17	Menarche	Menopause
	$X_{2;\gamma}$	$X_{3;\gamma}$	$X_{4;\gamma}$	$X_{5;\gamma}$
1	164.82	167.792	164.79	164.12
1.5	70.759	70.7451	70.747	70.459
2	54.941	54.9506	54.932	54.708
5	36.033	36.0260	36.027	35.880
50	28.220	28.2150	28.216	28.100
$+\infty$	**27.471**	**25.558**	**27.466**	**27.354**

Table 4 presents the $\sigma^2_{j;\gamma}$ parameter values obtained from (23), for various $\gamma \geq 1$ values and $j = 2, 3, 4, 5$, see Appendix 3. The variable X_1 corresponding to the risk factor Age is omitted as the σ^2_1 values of the input variable X_1 (Age) are too small. From Table 4 it is clear that the shape parameter's limiting value $\gamma_0 = 50$ ("close" to infinity), is the one which provides minimum $\sigma^2_{j;\gamma}$ values for $\gamma > 1$, i.e.

$$\min_{\gamma>1}\{\sigma^2_{j;\gamma}\} = \sigma^2_{j;+\infty} \approx \sigma^2_{j;50}, \quad j = 2, \ldots, 5.$$

Notice that when $\gamma < 0$ we derive $\gamma_0 = 0$, which is beyond this study.

Therefore, the minimum $\sigma^2_{j;\gamma>1}$, $j = 2, 3, \ldots, 5$, values (bold values in Table 4) correspond to the Laplace distribution. For the form of the risk variable COMT $= X_2$, see Fig. 2, where $X_{2;\text{Laplace}} := X_{2;+\infty}$.

Step 3. In order to choose the parameter δ which provides maximum values for $\mathrm{J}_\delta(X_{j;+\infty})$, we refer to Table 5 for the appropriate evaluations of $\mathrm{J}_\delta(X_{j;\gamma_0})$, where $\gamma_0 = 50 \approx +\infty$ due to Step 3, see Appendix 3. Thus,

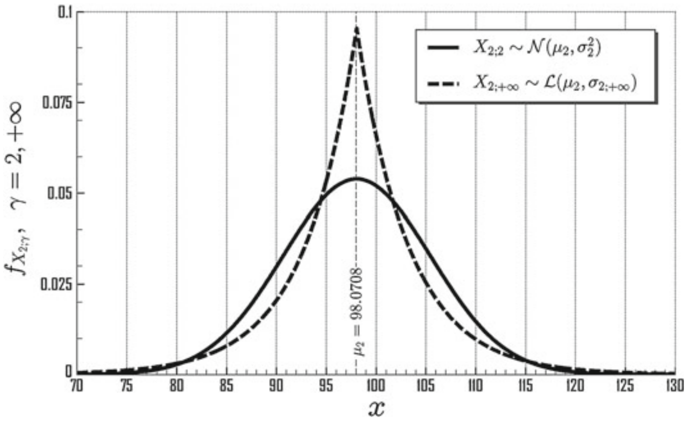


Fig. 2 Graphs of the p.d.f. of COMT $= X_2 \sim \mathscr{N}(\mu_2, \sigma^2_2)$ and $X_{2;\text{Laplace}} \sim \mathscr{L}(\mu_2, \sigma_{2;\text{Laplace}})$

Table 5 Evaluation of $\mathrm{J}_\delta(X_{j;\gamma_0})$ where γ_0 provides the minimum $\sigma^2_{j,\gamma}$ value

γ	COMT	CYP17	Menarche	Menopause
	$X_{2;\gamma}$	$X_{3;\gamma}$	$X_{4;\gamma}$	$X_{5;\gamma}$
1	**0.19079**	0.19081	0.19081	0.19120
1.5	0.08333	0.08335	0.08335	0.08360
2	0.03640	0.03641	0.03640	0.03656
5	0.00025	0.00025	0.00025	0.00025*
10	6.39e–8	9.17e–8	6.4e–8	6.53e–8*

*For $\gamma \geq 5$ the values of $\mathrm{J}_\delta(X_{j;\gamma})$ are close to zero

$$\max_{\delta\geq 1}\{\mathrm{J}_\delta(X_{j;+\infty})\} = \max_{\delta\geq 1}\{\sigma^{-\delta}_{j;+\infty}\} = \max_{\delta\geq 1}\{2^{\delta/2}\sigma_j^{-\delta}\} = \sqrt{2}/\sigma_j = \mathrm{J}_1(X_{j;+\infty}), \tag{24}$$

because of (16) and (23) with $\gamma \to +\infty$. It is clear from (24) that, for all $j = 2, 3, 4, 5$, the parameter value $\delta_0 = 1$ (which is a limiting value of $\mathrm{J}_{\delta>1}$) provides maximum for $\mathrm{J}_\delta(X_{j;+\infty})$.

Step 4. We choose the logit model which provides the optimum input variable X_{opt} as

$$\mathrm{J}_1(X_{opt}) = \min_{j=2,3,4,5}\{\mathrm{J}_1(X_{j;Laplace})\}.$$

Hence, from Table 5, we choose as $X_{opt} = X_{2;Laplace}$, i.e. $X_{opt} = X_{2;+\infty}$, and therefore the optimum variable to participate to the model is the risk variable COMT corresponding to the minimum value

$$\mathrm{J}_1(\texttt{COMT}) = 0.19079.$$

Thus COMT appears to be the appropriate selected risk variable according to the above min–max δ–GFI criteria. Moreover, this is done through the Laplace distribution and not through the Normal.

As far as the general minimum/maximum behaviour of $\mathrm{N}_\delta(X_{j;\gamma})$ see Appendix 2.

4 Discussion

The method we proposed appears to provide, as appropriate, a set of variables which differ from the classical one. This is due to the different criteria we apply. We believe that the γ–GFI as a criterion offers a “safe” alternative for the researcher to choose the appropriate Logit model.

The logit model with all the variables is presented in Table 4 [16], while the final model is in Table 5. The specific model provided by our method is chosen through the min-max entropy (amongst the worst we choose the best). Although it seems computationally tedious, it can be handled using packages such as MATLAB, see Appendix 3 for example.

In this paper we worked with the generalized Fisher's entropy type information measure J_δ. Similar algorithms can be adopted utilizing the generalized entropy power N_δ instead of GFI.

In the following Appendices 1 and 2 we provide the appropriate theoretical background concerning the min–max behavior of the generalized entropy power as an alternative criterion to the generalized Fisher's information measure.

Acknowledgments The authors would like to thank the referees who improved the language as well as the content of this paper, with their valuable comments.

Appendix 1

A study on the min–max behavior of the generalized entropy power $\mathrm{N}_\delta(X_\gamma)$ applied on a γ-order normally distributed random variable is given below.

Let $X_\gamma \in \mathcal{N}_\gamma^p(\mu, \Sigma)$. For $\delta > 1$, numerically can be verified (see Fig. 1) that,

$$\max_{\gamma>1} \mathrm{N}_\delta(X_\gamma) = \min_{\gamma<0} \mathrm{N}_\delta(X_\gamma) = \mathrm{N}_\delta(X_{+\infty}),$$

$$\min_{\delta\geq 1}\max_{\gamma>1} \mathrm{N}_\delta(X_\gamma) = \mathrm{N}_{+\infty}(X_{+\infty}) = 1, \quad \text{when} \quad |\det\Sigma| = 1,$$

$$\max_{\delta\geq 1}\min_{\gamma<0} \mathrm{N}_\delta(X_\gamma) = \mathrm{N}_1(X_{+\infty}) = e|\det\Sigma|^{\frac{1}{2p}}.$$

The dual case hold for $\delta < 0$ (max / min reversion). Moreover, for $|\det\Sigma| \leq 1$, see also Fig. 3, the following holds:

$$\max_{\delta\geq 1} \mathrm{N}_\delta(X_\gamma) = \mathrm{N}_1(X_\gamma), \quad \gamma > 1,$$

$$\min_{\gamma\geq 1}\max_{\delta\geq 1} \mathrm{N}_\delta(X_\gamma) = \mathrm{N}_1(X_1) = |\det\Sigma|^{\frac{1}{2p}},$$

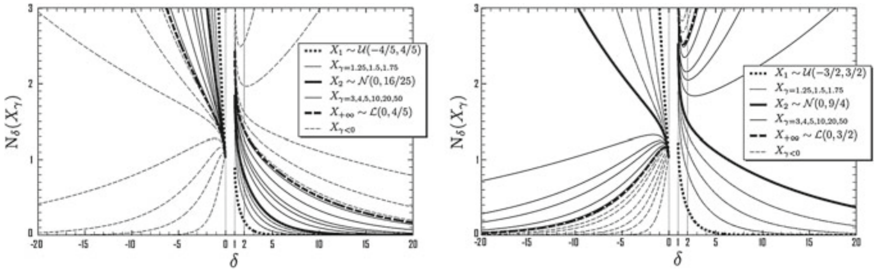


Fig. 3 Graphs of $\mathrm{N}_\delta(X_\gamma)$ along δ for various γ values, where $X_\gamma \sim \mathcal{N}_\gamma(0, \sigma^2)$ with $\sigma = 0.8, 1, 1.5$

For $|\det \Sigma| \geq 1$, Fig. 3, we have

$$\max_{\gamma>1}\left\{\min_{\delta\geq 0}\mathrm{N}_\delta(X_\gamma) \neq 0\right\} = \min_{\delta>1}\mathrm{N}_\delta(X_{+\infty}).$$

Appendix 2

A study on the min–max behavior of $\mathrm{N}_\delta(X_{j;\gamma})$ and $\mathrm{J}_\delta(X_{j;\gamma})$ applied on the γ-order normally distributed risk variables $X_{j,\gamma}$, $j = 1, 2, \ldots, 5$ is given below.

Figure 4, verifies that

$$\max_{\gamma>1}\left\{\mathrm{N}_\delta(X_{j;\gamma})\right\} = \mathrm{N}_\delta(X_{j;2}) = \mathrm{N}_\delta(X_j), \quad \delta > 1,$$

i.e. the $\max_{\gamma>1}\mathrm{N}_\delta(X_{j;\gamma})$ corresponds, for $\delta > 1$, to the usual Normal distribution, and therefore,

$$\min_{\delta>1}\max_{\gamma>0}\left\{\mathrm{N}_\delta(X_{j;\gamma})\right\} = m_j,$$

with m_j as in Table 3. Moreover, the $\max_{\gamma>1}\mathrm{N}_\delta(X_{j;\gamma})$ for $\delta < 0$ corresponds to the usual Laplace distribution, i.e.

$$\max_{\gamma>1}\left\{\mathrm{N}_\delta(X_{j;\gamma})\right\} = \mathrm{N}_\delta(X_{j;-\infty}), \quad \delta < 0.$$

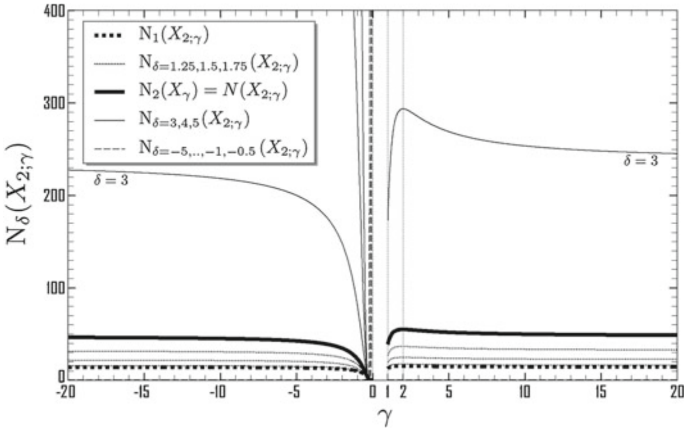


Fig. 4 Graphs of $\mathrm{N}_\delta(X_{2;\gamma})$ along γ for various δ values, where $X_{2;\gamma} \sim \mathcal{N}_\gamma(\mu_2, \sigma^2_{2;\gamma})$

Table 6 Evaluations of m_j^* and m_j^{**}, $j = 2, 3, 4, 5$

j	X_j	m_j^*	m_j^{**}
2.	`COMT`	14.247	0.072732
3.	`CYP17`	14.247	0.072746
4.	`Menarche`	14.246	0.073042
5.	`Menopause`	14.217	0.071991

while for $\delta > 1$, through (23) with $\frac{\gamma-1}{\gamma} \to 1$, is given by

$$\max_{\gamma<0} \left\{\mathrm{N}_\delta(X_{j;\gamma})\right\} = \mathrm{N}_\delta(X_{j;-\infty}) = e(\tfrac{\delta-1}{\delta})^{\delta-1}\Gamma^{-\delta}\left(\tfrac{\delta-1}{\delta}+1\right)\sigma_j/\sqrt{2}, \quad \delta > 1, \quad \text{i.e.}$$

$$m_j^* := \min_{\delta>1}\max_{\gamma<0}\left\{\mathrm{N}_\delta(X_{j;\gamma})\right\} = \mathrm{N}_1(X_{j;-\infty}) = \tfrac{1}{2}e\sqrt{2}\sigma_j,$$

with m_j^* as in Table 6 (compared with Table 3).

For the δ–GFI, we have that $\mathrm{J}_\delta(X_\gamma)$ is a monotone function of $\delta \geq \gamma(1.4628 - 1) + 1$ for all $X_\gamma \sim \mathcal{N}_{\gamma\geq1}(\mu, 1)$, see proof of Proposition 3.1 in [20]. Thus $\mathrm{J}_\delta(X_\gamma)$ is an increasing function of $\delta \geq 2$. Therefore,

$$\min_{\delta\geq2} \mathrm{J}_\delta(X_\gamma) = \mathrm{J}_2(X_\gamma) = \mathrm{J}(X_\gamma), \quad \gamma \geq 1.$$

The above relation holds for $X_\gamma \sim \mathcal{N}_{\gamma\geq1}(\mu, \sigma^2)$, due to (16), assuming $\sigma \geq 1$, and hence

$$\min_{\delta\geq2} \mathrm{J}_\delta(X_{j;\gamma}) = (\tfrac{\gamma-1}{\gamma})^{2/\gamma}\frac{\Gamma(\frac{3\gamma-1}{\gamma})}{\Gamma(\frac{\gamma-1}{\gamma})}\sigma_{j;\gamma}^{-2}, \quad \gamma \geq 1.$$

Through (23) we have

$$\min_{\delta\geq2} \mathrm{J}_\delta(X_{j;\gamma}) = (\tfrac{\gamma-1}{\gamma})^{2\frac{2-\gamma}{\gamma}}\frac{\Gamma(3\frac{\gamma-1}{\gamma})\Gamma(\frac{3\gamma-1}{\gamma})}{\Gamma^2(\frac{\gamma-1}{\gamma})}\sigma_j^{-2}, \quad \gamma \geq 1.$$

Numerically we can derive that (see Table 6)

$$m_j^{**} := \max_{\gamma\geq1}\min_{\delta\geq2} \mathrm{J}(X_{j;\gamma}) = \mathrm{J}_\delta(X_{j;+\infty}).$$

Appendix 3

Appendix 3.1. *MATLAB script for the evaluations of Table* 3. The 5th column of the `Data` matrix correspond to the data for `Age`, `COMT`, `CYP17`, `Menarche` and `Menopause` risk variables, while the `b0` and `b1` arrays contain the coefficients from Table 2.

```
Data = [Age COMT CYP17 Menarche Menopause]; [n,k] = size(Data);
b0 = [-3.45 -0.195 -0.419  0.08902 -3.55];
b1 = [ 0.55 -0.03  -0.095 -0.026    0.067];
Variable = {'Age','COMT','CYP17','Menarche','Menopause'};
Title = @(i,m,v) ...
   [Variable{i} ':' char(' '*ones(1,15-length(Variable{i}))) ...
   num2str(m) '  ' num2str(v)];
disp(' '), disp('                mu_j     Var_j')

Mu = zeros(1,k); Var = Mu;
for i = 1:k
    e = exp(b0(i)+b1(i)*Data(:,i)); p = e./(1+e);
    mp = mean(p); m = n*mp; v = m*(1-mp);
    disp(Title(i,m,v)); Mu(i) = m; Var(i) = v;
end

N = @(a,d,p,s,e) ( ((d-1)./(d*e)).^(d-1) ).*...
                 ( (e./a).^(d.*a) ).*( s.^(d/(2*p)) ).*
                 (( gamma(1+p*a)./gamma(1+p*((d-1)./d)) ).^(d/p));

A = @(g) (10^-8)*(g == 1)+((g-1)./g).*(g > 1 | g < 0)+isinf(g);
aU = A(1); aN = A(2); aL = 1;

D = -5:0.0001:-0.0001; s2 = Var;
disp('=== M_j ========================');
for k = 1:length(s2);   Y = N(aN,D,1,s2(k),exp(1)); Y(isinf(Y)) = 0;
    [maxN i] = max(Y); disp([maxN D(i)])
end;

D = 1.0001:0.0001:5;
disp('=== m_j ========================');
for k = 1:length(s2)
    [minN i] = min(N(aN,D,1,s2(k),exp(1))); disp([minN D(i)])
end;
```

Appendix 3.2. *MATLAB script for the evaluations of Table* 4. Array s2 and variable k are defined in the previous script.

```
Var = @(a,p,s2) (a.^(-2*a) ).*( gamma((p+2)*a)./gamma(p*a) )*s2;

A = @(g) (10^-12)*(g == 1)+((g-1)./g).*(g > 1 | g < 0)+isinf(g);
G = union([-50 -10 -5 -2 -1 -0.1],[1 3/2 2 3 5 10 50 inf]);
ng = length(G); AG = A(G); AG(isnan(AG)) = 1;

Lsep = 72; sepTitle = @(m,Title,sepstr,n)...
  disp([char(sepstr*ones(1,m)) ' ' Title ' ' ...
  char(sepstr*ones(1,n-m-length(Title)-2))]);
sep = @(sepstr,n) disp(char(sepstr*ones(1,n)));

format compact; format short; M = zeros(ng,k);
sep(' ',Lsep); sepTitle(3,'s2_j;gamma','=',Lsep);
disp([NaN V]); sep('-',Lsep);
i = 1; for v = V;   M(:,i) = v./Var(AG,1,1)'; i = i+1; end;
disp([G' M]); sep('=',Lsep); format loose;
```

Appendix 3.3. *MATLAB script for the evaluations of Table 5*. Array V and MATLAB function Var are defined in the previous script.

```
J = @(a,d,p,s2) a.^(d*(1-a)).*...
              (gamma(p*a+d*(1-a))./gamma(p*a))./(s2.^(d/(2*p)));

D = union([-inf -10 -5 -2 -1 -0.1 -0.01],[1 3/2 2 3 5 10 inf]);
nd = length(D); aL = 1;

format compact; format short g; M = zeros(nd,k); Lsep = 92;
sep(' ',Lsep); sepTitle(3,'Jd(X_j;gamma0=Inf)','=',Lsep);
disp([NaN V]); sep('-',Lsep); i = 2;
for v = V(2:k); M(:,i) = J(aL,D,1,v./Var(aL,1,1))'; i = i+1; end;
M(:,1) = J(A(-0.01),D,1,v./Var(aL,1,1))';   %%% Jd(X_j;gamma0=Inf)
disp([D' M]); sep('=',Lsep); format loose;
```

References

1. Burnham, K.P., Anderson, D.R.: Kullback-Leibler information as a basis for strong inference in ecological studies. Wild Life Res. **28**, 111–119 (2001)
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken (2006)
3. Dippon, J., Fritz, P., Kohler, M.: A statistical approach to case based reasoning, with application to breast cancer data. Comp. Stat. Data Anal. **40**, 579–602 (2002)
4. Edler, L., Kitsos, C.P.: Recent Advances in Qualitative Methods in Cancer and Human Health Risk Assessment. Wiley, Chichester (2005)
5. Feigelson, S.H., McKean-Cowdlin, R., Henderson, E.B.: Concerning the CYP17 MspA1 polymorphism and breast cancer risk: a meta-analysis. Mutagenesis **17**, 445–446 (2002)
6. Ferentinos, K., Papaioannou, T.: New parametric measures of information. Inf. Control **51**, 193–208 (1981)
7. Finney, D.J.: Statistical Method in Biological Assay. Hafner, New York (1952)
8. Ford, I., Kitsos, C.P., Titterington, D.M.: Recent advances in nonlinear experimental design. Technometrics **31**, 49–60 (1989)
9. Huang, C.S., Chern, H.D., Chang, K.J., Cheng, C.W., Hsu, S.M., Shen, C.Y.: Breast risk cancer associated with genotype polymorphism of the estrogen-metabolizing CYP17, CYP1A1 and COMT—a multigenic study on cancer susceptibility. Cancer Res. **59**, 4870–4875 (1999)
10. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620–630 (1957)
11. Kitsos, C.P., Toulias, T.L.: A min-max entropy algorithm for relative risk problems. Submitted
12. Kitsos, C.P., Toulias, T.L.: On the information matrix of the truncated cosinor model. Br. J. Math. Comput. Sci. **3**(3) (2013)
13. Kitsos, C.P.: On the logit methods for Ca problems. In: Statistical Methods for Biomedical and Technical Systems, pp. 335–340. Limassol, Cyprus (2006)
14. Kitsos, C.P.: The Ca risk assessment as a bioassay. In: 55th Session of the International Statistical Institute. Sydney (2005)
15. Kitsos, C.P., Tavoularis, N.K.: Logarithmic Sobolev inequalities for information measures. IEEE Trans. Inf. Theory **55**(6), 2554–2561 (2009)
16. Kitsos, C.P.: Estimating the relative risk for breast cancer. Biom. Lett. **47**(2), 133–146 (2010)
17. Kitsos, C.P., Toulias, T.L.: New information measures for the generalized normal distribution. Information **1**, 13–27 (2010)
18. Kitsos, C.P.: Invariant canonical form for the multiple logistic regression. Math. Eng. Sci. Aerosp. (MESA) **38**(1), 267–275 (2011)

19. Kitsos, C.P.: Cancer Bioassays: A Statistical Approach. LAMBERT Academic Publishing, Saarbrucken (2012)
20. Kitsos, C.P., Toulias, T.L.: Bounds for the generalized entropy type information measure. J. Commun. Comput. **9**(1), 56–64 (2012)
21. Kitsos, C.P., Toulias, T.L., Trandafir, C.P.: On the multivariate γ-ordered normal distribution. Far East J. Theor. Stat. **38**(1), 49–73 (2012)
22. Soofi, E.S.: Principal information approaches. J. Am. Stat. Assoc. **95**(452), 4352–4374 (2007)
23. Zarikas, V., Gikas, V., Kitsos, C.P.: Evaluation of the optimal design "cosinor model" for enhancing the potential of robotic theodolite kinematic observations. Measurement **43**(10), 1416–1424 (2010)

Fractal Case Study for Mammary Cancer: Analysis of Interobserver Variability

Philipp Hermann, Sarah Piza, Sandra Ruderstorfer, Sabine Spreitzer and Milan Stehlík

Abstract This paper discusses some features of the distribution of box-counting fractal dimension measured on a real data set from mammary cancer and masthopathy patients. During the study we found several reasons why mammary cancer and its following distribution cannot be easily represented by single box-counting dimension. The main problem is that without a histopathological examination of the tumor a simple algorithm based only on single box-counting dimension is difficult to be constructed. We have tried to understand the distribution underlying the real data, especially its departures from normality. Both normal and gamma distributions are related to the Tweedy distributions, which are given by multi-fractal dimension spectra present in histopathological images. Without having a histological examination of the data multifractality is unavoidable as can be seen from several analysis in this paper. We have seen a fair differentiation between cancer and masthopathy. Finally we studied the depths of the data based on the information divergence. Some practical conclusions are also given.

Keywords Depth · Discrimination · Mammary cancer · Multifractality · Skewness

P. Hermann · S. Piza · S. Ruderstorfer · S. Spreitzer · M. Stehlík (✉)
Department of Applied Statistics, Johannes-Kepler-University Linz,
Altenbergerstraße 69, 4040 Linz a. D., Austria
e-mail: Milan.Stehlik@jku.at, mlnstehlik@gmail.com

P. Hermann
e-mail: philipp.hermann@jku.at

S. Piza
e-mail: sarah.piza@aon.at

S. Ruderstorfer
e-mail: sandra.ruderstorfer@gmail.com

S. Spreitzer
e-mail: sabine.spreitzer@gmx.net

M. Stehlík
Departamento de Matemática, Universidad Técnica Federico Santa María,
Casilla 110-V, Valparaíso, Chile

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_2

1 Introduction

Breast cancer is one of the most common cancers. The chance of curing cancer primarily relies on its early diagnosis and the selection of the treatment depends on its malignancy kind. Therefore it is critical to distinguish cancerous tissues from healthy ones and identify its malignancy kind. There is a common understanding in the cancer research, that many of the typical cancer tissues have some fractal geometry features (see [1, 6, 9, 16]). It is clear that not all of cancer tissues will follow rigid fractal geometry (e.g. Wilms tumors, see [12]). However, mammary cancer is understood to have some fractal background. There is a hope that stochastic and deterministic models of cancer growth can help to better differentiate between cancer and various forms of masthopathy. Several procedures have been developed to make this discrimination only based on fractal dimension. However, this is oversimplifying the real situation. In this paper we show some complexities of dependencies within a dataset containing both mammary cancer and masthopathy histological images. For the other approaches to estimate the relative risk for breast cancer and related issues see [2, 7, 8].

The paper is organized as follows. In the next section we introduce the data. In Sect. 2.1 we study the deviations of the subsamples from the normality. In the Sect. 2.2 we study the skewness and kurtosis individually, together with heterogeneity of the data. Afterwards in Sect. 2.3 we introduce the depth based on both the heuristic algorithms and parametric assumption of the underlying gamma distribution. Section 3 contains a deeper analysis of the variance. Within this section in Sect. 3.1 a simple discrimination between masthopathy and mammary cancer, based on the box-counting dimension is done and in Sect. 3.2 the different groups are tested for normality.

2 Case Study

The cancer data contains the observation-number, the box-counting dimensions measured on the data (see [16]), the characteristics (mamca or masto) and the percentages. In total there are 391 observations. The data has been taken from [10] and are collected from histological examination, with 512 × 512 pixel image resolutions. Histologic exams usually look at physical examples under a miroscope and assign a tumor grade. Fractal analysis looks at images of breast tissue specimens and provides a numeric description of tumor growth patterns as a number between 1 and 2. This number, the fractal dimension, is an objective and reproducible measure of the complexity of the tissue architecture of the biopsy specimen.

The minimum of the sample dimensions is 1.1039 and the maximum is 1.8715. The value of the range is therefore 0.7676. Moreover the comparison of the mean (1.587391) and the median (1.5972) shows that the data is not strongly biased due to outliers. Firstly, the analysis of data was conducted and a graphical representation of

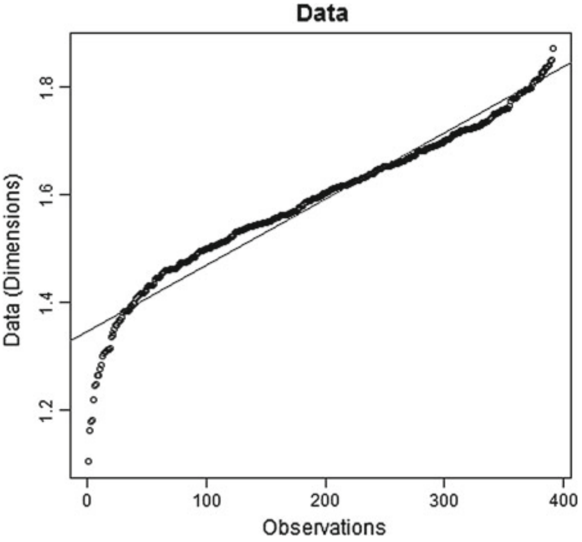


Fig. 1 Scatterplot of data with fitted regression-line

the ordered dimensions follows in Fig. 1. The skewness has a value of −0.47. This figure indicates that the data is left-skewed. The kurtosis of 3.01 means that the data is peaked.

Obviously it is noticeable that the data points are placed closely at the regression line. Especially at the lower quantile of the data, many points differ strongly compared to the regression line of the dimensions. For this reason the data was analyzed to detect if there are possible candidates for outliers by comparing the values with upper (b_u) and lower (b_l) borders. Those were calculated with the following formulas, where IQR is the interquantile range.

$$b_l = q_{0.25} - (1.5 \cdot IQR(dimensions)) \tag{1}$$

$$b_u = q_{0.75} + (1.5 \cdot IQR(dimensions)) \tag{2}$$

The calculations with the Eqs. (1) and (2) result in four detected outliers, which were deleted from the dataset. The next step was to investigate if the data can be considered as normally distributed. By creating a histogram in the program EViews, a test for normality is calculated additionally. In this case a p-value of 0.000729 and a Jarque-Bera-value of 14.44795 were computed. For this reason it cannot be assumed that the data is normally distributed.

A Shapiro-Wilk-Test for normality was approached in the statistical program R [12]. The Shapiro-Wilk-Test has a good property in context of robust testing for normality [15]. This p-Value of 0.0001104 is approximately in the range of the test-statistic of the Jarque-Bera-Test. These two tests allow rejecting the null hypothesis, hence the data is not normally distributed. The p-value improves with the modified data to 0.001153, but normality can still not be assumed.

It needs to be investigated which distribution fits to the data, because it cannot be assumed that the data is normally distributed. For this reason a two-sided Kolmogorov-Smirnov-Test for a gamma distribution with a shape of 125.379 and a scale parameter of 78.98 was computed (see [16]). For an alternative statistical approach see [17]. The p-value of this calculation (0.2161) allows to assume that the data is gamma distributed. The same test was done with the modified data and still yields a p-value ($p = 0.142$), which also allows to maintain the null hypothesis of a gamma distribution.

Moreover we decided to test whether the mean of the data fits to the expected mean with given shape and scale parameters of the gamma distribution. In this case the expected value is $E(X) = \frac{\lambda}{\beta}$, where λ is the shape and β the scale parameter. This calculation yields 1.587477, which is almost equal to the mean of the data (1.587391). This naive check is another indication that the gamma distribution fits to the dimensions of the slices.

As previously given in [16], the data can be assumed gamma distributed. Nevertheless the data is investigated also for Weibull distribution. In order to test this, the Kolmogorov-Smirnov-Test for a Weibull distribution was used. Stehlík et al. [16] suggests using the values 13.68 and 1.648 for the shape and the scale parameters. Analogously to the test for gamma distribution, the original and the modified data were tested for Weibull distribution.Thereby a p-value of 0.3568 for the complete data was calculated and the test for the modified data delivered an even better p-value of 0.9576, which says that the data can be assumed as Weibull distributed. These two significantly different p-values are another hint, that the four neglected values are outliers.

2.1 Jarque-Bera-Statistic

The following formula was used to implement an exact Jarque-Bera-Test for normality in *R* [12]. The calculation with the following formula (3) results in a value of 14.44795.

$$jb = \frac{n}{6} * \left(\frac{\hat{\mu_3}}{\hat{\mu_2}^{\frac{3}{2}}} \right)^2 + \frac{n}{24} * \left(\frac{\hat{\mu_4}}{\hat{\mu_2}^2} - 3 \right)^2, \tag{3}$$

where $\hat{\mu_i}$ stands for the central moments of the data. The second central moment ($\hat{\mu_2}$) is the same as the variance. The third moment equals to the skewness ($\hat{\mu_3}$) and the fourth moment ($\hat{\mu_4}$) describes the kurtosis. By inserting the moments of the data into the formula the same figure as in EViews is received.

2.1.1 Algorithm of the Simulation

1. Step: Simulate 391 random variables ~N(0,1)

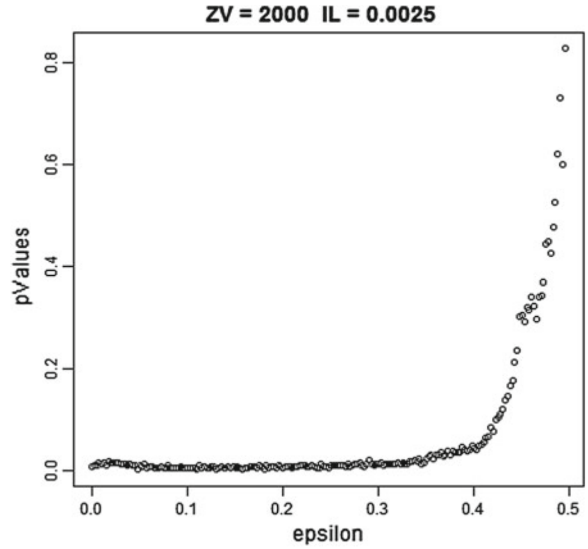
2. Step: Calculate a Jarque-Bera Statistic with formula (3) and compare with the "true value" 14.44795. Add to a temporary variable 1, if the new value is bigger than the old one.
3. Step: Repeat steps 1. and 2., 10,000 times and divide counter by 10,000.

The result of this algorithm was a p-value of 0.0053. Analogously to the previous procedure the implemented test was now approached with the data without outliers. This test delivered a new p-value 0.0216. Compared to the previous one, it has significantly improved, but normality still needs to be rejected.

2.1.2 Epsilon-Belt

The aim of the simulation is to transform to a normal distribution by truncating the data at a special ϵ-point (this is a specific form of deleting outliers), which is prespecified by a pathologic expert. To test the normality, simulations were conducted. First the data was cut on both sides with the same ϵ-value. Afterwards the data was cut just on one side, because of the left-skewed distribution of the data. To show the effect of the ϵ-belts, the values of ϵ were plotted against the corresponding p-values. In this case the value of ϵ is equal to the upper or lower quantile. Precisely if $p = 0.1$, this means, that the upper and lower 10 % of the data were cut. Figure 2 shows these calculated p-values. It needs to be mentioned, that e.g. $ZV = 2000$ means that 2000 random variables were calculated and $IL = 0.0025$ says that the differences between the epsilon values are 0.0025.

Fig. 2 JB-Test with data cut on both sides



2.1.3 Heterogeneity Between the Different Groups

Another very interesting aspect concerning tests for normality is the difference between the two groups. The p-values computed by the Shapiro-Wilk-Test differ between these ones significantly. First this test was done with the observations where masthopathy was detected. The calculated p-value smaller than 0.0001 shows that the data cannot be assumed as normal distributed. Following to that the same test was approached with those observations where cancer was observed. The p-value with this set of data is 0.04519. Although the null hypothesis cannot be rejected, the significantly improved p-value shows that there is a difference in the distribution between these two groups. Due to the knowledge described above, the previously named Jarque-Bera-Test was accomplished with these two groups. These outputs are compared to each other on Fig. 3.

Figure 3 demonstrates the significant difference between the p-values concerning masthopathy and cancer. Thereby the data was cut only on the left side. Another difference to the previous simulations is that in this case the epsilon stands for the minimum dimension of the data for which the p-values were calculated. This test was only computed for observation which have higher dimension than given ϵ. In average the p-values for cancer are 10 times higher compared to the ones for the masthopathy group. We constructed this by visual discrimination and further research should be conducted to get the levels of significance for the difference.

2.2 *Skewness and Kurtosis*

In this section we investigate skewness and kurtosis of the empirical distribution of estimated box-counting dimensions. The aim is to modify the data to improve the values of skewness and kurtosis to be more adequate for normality. Without the

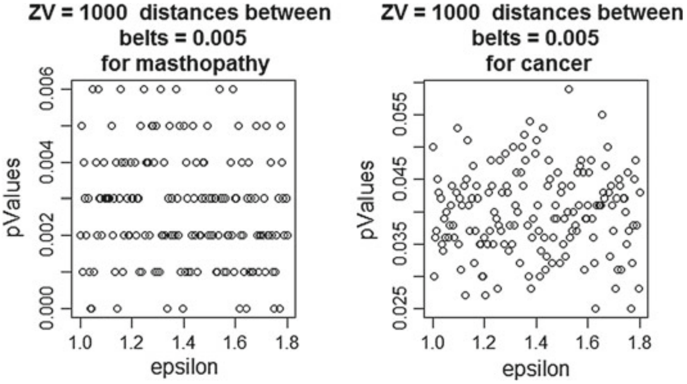


Fig. 3 Difference between masthopathy and cancer

modification of the data skewness and kurtosis have the following values: −0.001336; 0.001211. The value of the skewness is approximately zero and concerning only this value, the data could be assumed to be normally distributed. Nevertheless the value for kurtosis is also around zero and definitely too small for normality. The following modification of the data is done similarly to the previous approach. For this reason data outside of the fixed quantile is cut and only the remained data is investigated. Figure 4 shows the calculated values for skewness on the left and for kurtosis on the right side.

The next step is to cut off the data which is bigger than a fixed value of ϵ. These limits can be read on the ordinate of the plot. The split of the output is similar to the previous figure.

As visible on Fig. 5 the data still cannot be assumed to be normally distributed. Although the value of skewness converges roughly to zero, kurtosis is still too low. The desired value of 3 is still unreachable. We can conclude that there is more structural departure from normality than the one studied here given by outliers.

2.3 Depth-Plot

With the following two depth functions (see [18]) it is possible to visualize the 3-dimensional median. The x-coordinate represents the data (dimensions) and the y-coordinate the characteristic feature of cancer. On the left side the depth plot was accomplished with the method of Tukey and the plot of the right side shows Liu's version.

For the plot on the right side it seems that the mean is nearer to the healthy tissue (0), than to the others. This can be seen, because the peak at the x-value of 0 is higher compared to the depth of the data at a x-value of 1 (Fig. 6).

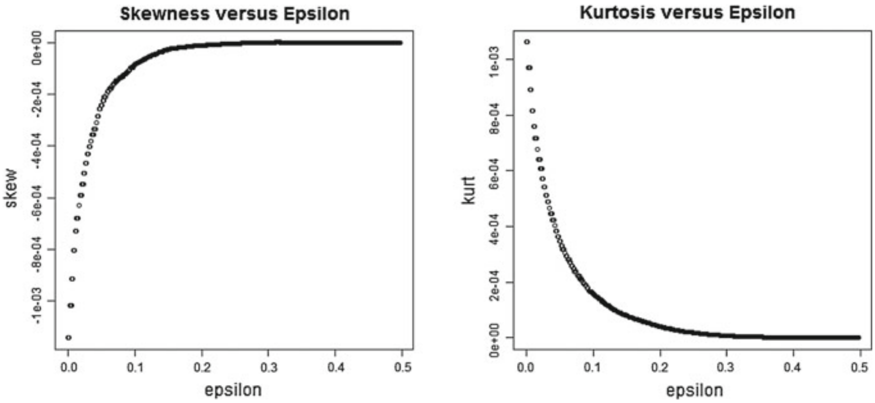


Fig. 4 Changes in skewness and kurtosis by modifying the data

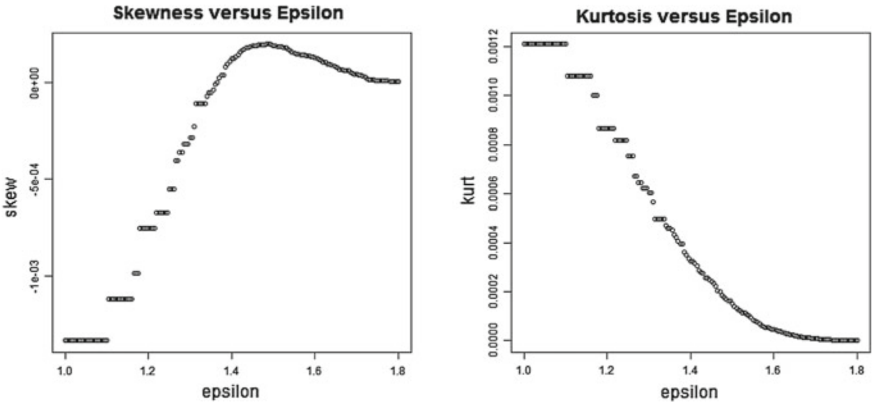


Fig. 5 Changes in skewness and kurtosis by modifying the data

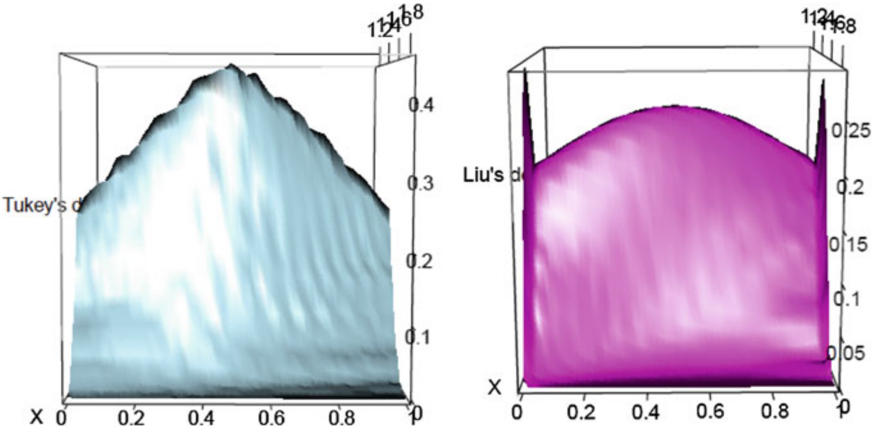


Fig. 6 Depth plot of Tukey (*left*) and Liu (*right*)

The left hand side of Fig. 7 shows a depth contour plot. Here it is nicely visible where the mean of the data is placed. Unfortunately it is not possible to plot the whole data, so we took 200 points from the middle of the data. At the x-coordinate the value is climbing up to 1.58 and at the y-coordinate to 0.55. So here with the reduced data no difference between the characteristic features of cancer and masthopathy is visible. The right plot of Fig. 7 shows the confidence intervals for the parameters of β, which are used for the gamma distribution. The calculation has been accomplished by the divergence from data vector to canonical parameter according to Pázman (see [11, 14]):

$$I_N(y,\gamma) = I(\gamma_y,\gamma) = -\sum_{i=1}^{N}(v_i - v_i \cdot ln(v_i)) + \sum_{i=1}^{N}(y_i\gamma_i - v_i \cdot ln(y_i\gamma_i)) \quad (4)$$

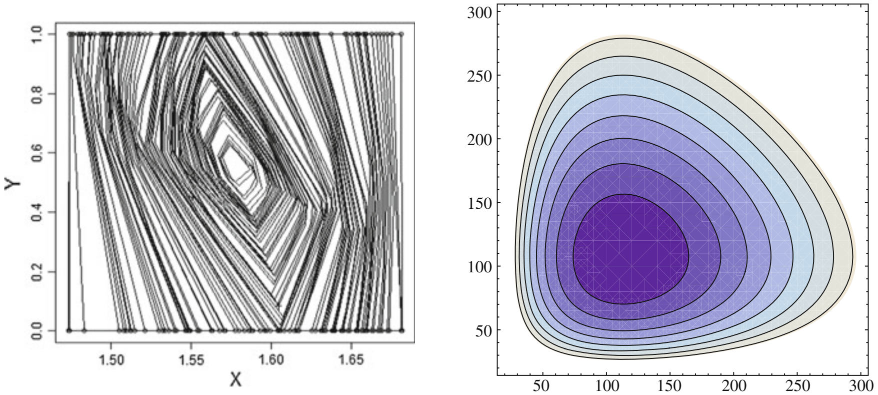


Fig. 7 Contour Plot (*left*) and Confidence Intervals for the parameter β (*right*)

In this case the parameters have been chosen as follows: $\beta = \gamma = 70$. β is the scale parameter and ν is the shape parameter of the gamma distribution, which has previously been fixed with approximately 125.

3 Graphical Analysis for Variance

Our next step will be to investigate the variances of the groups masthopathy and mammary cancer. By using the Eqs. (1) and (2) outliers for both of the groups at the lower end can be detected. The dimensions of the group masthopathy are on average higher compared to the dimensions of mammary cancer tissue. This can be easily proven by comparing the minimum (1.24; 1.10) and the maximum (1.87; 1.82) of the groups masthopathy (first) and mammary cancer (second).

3.1 Simple Discrimination Between Masthopathy and Mammary Cancer Based on the Box-Counting Dimension

If we will follow the simple concept that higher dimension is more risky, the issue is that we will arrive with this dataset to some sort of contradiction. The problem is that the median of the box-counting dimension is 1.5972. When we make a simple clustering based on ordering the box-counting dimension and decide to tell that more risky tissue have a box-counting dimension bigger than the median and non-risky tissue is below, then we only classified 135 of mamca and 60 of mastho below. Therefore it needs to be mentioned that 199 observations contain the characteristic masthopathy and 192 observations mammary cancer. Even using the arithmetic mean of 1.587391 decreases the number of classified cancer tissues to 128 for mamca and 56 for masto. Based on this simple example we can conclude that we need a more sophisticated procedure based on the box-counting dimension to discriminate between the two groups and we should take more detailed characteristics of the tissue into account. In extremal case there is no possibility to develop automatic clustering based on box-counting dimension, which can avoid histological expert examination.

If we take a look on the left plot of Fig. 8 we see that to force usage only of one box-counting dimension establishes inverse problems, which are ill posed. Loosely saying we need a continuous dimension spectrum, e.g. multi-fractal dimension spectra. It has already been used in breast cancer discrimination by George and Kamal (see [5]). A multifractal system is a generalization of a fractal system in which a single exponent (the fractal dimension) is not enough to describe its dynamics; instead, a continuous spectrum of exponents (the so-called singularity spectrum) is needed. Several fractal objects have been recognized (see [3, 4]). This relates also to Tweedie exponential dispersion models, which as a special case contains both normal and gamma distributions. This is further justification for these two simple distributional families: in the case of our empirical data we have found a strong deviation from normality and therefore we used gamma distribution.

The right plot of Fig. 8 gives a good overview of the distribution of the observations of the different groups. It is visible that mammary cancer tissue has on average higher dimensions (red) compared to masthopathological tissue. These conclusions were already recognizable due to the comparison of the mean.

We can observe that the difference between the green and red curve is well visible and differentiates the two groups. To use the fact, that the sum of the squares of standard normal distributed and independent random variables are Chi-Square distributed, it is essential to prove first this conditions. For this reason it is necessary to show that the differences between the curves are standard normal distributed. Therefore the Shapiro-Wilk-test was used. Because of the small p-value it cannot be assumed that the differences are standard normal distributed. We have to show, that the squared differences are Chi-Square distributed with one degree of freedom. In case that this occurs we are allowed to use the Chi-Square-test, which is realized below.

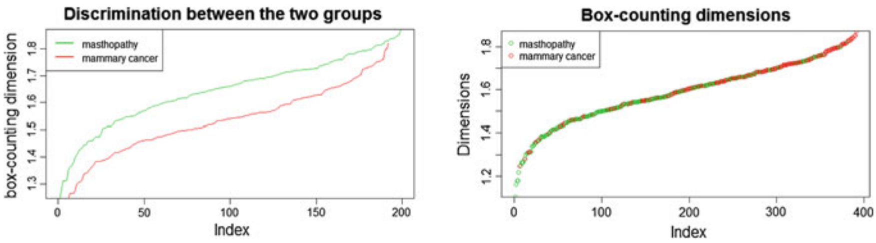


Fig. 8 Plot of the dimensions discriminated between masthopathy and mammary cancer

Table 1 Simulation of p-values with given shape and scale parameter

Shape	Scale	p-value
0.45491680	$\frac{1}{0.45729929}$	8.354e-05
0.48	$\frac{1}{0.4573}$	3.7e-12
0.44	$\frac{1}{0.4673}$	1.65e-11
0.44	$\frac{1}{0.4673}$	5.55e-10
0.42	$\frac{1}{0.48}$	2.22e-16
0.43	$\frac{1}{0.47}$	1.403e-10
0.425	$\frac{1}{0.48}$	1e-9
0.42	$\frac{1}{0.48}$	0.004855
0.425	$\frac{1}{0.48}$	0.04488
0.43	$\frac{1}{0.48}$	0.009656
0.42	$\frac{1}{0.485}$	0.1272
0.415	$\frac{1}{0.485}$	0.0996

Computing the sum of the squared differences delivers a value of 5.38. A Chi-Square-test was accomplished to test whether the two groups are equal or different. The p-value of the distribution function of the Chi-Squared distribution with 199 degrees of freedom is approximately one. This value is another proof that the two groups are different. Furthermore we made a standardization of the previously calculated differences. Therefore we reduced the differences with its mean and divided those values by the standard deviation. With these values the null hypothesis "The two groups are different" is tested. The sum of the standardized squared differences is 198. The distribution function at this value and 199 degrees of freedom is 0.49331. Due to this p-value it can be once more recognized that the two groups are different.

χ_1^2 is not fitting completely well, because the p-value is very small. However, we believe that the distribution can be gamma with the real valued shape parameter. We have produced the plots and also checked the p-values with different shape and scale parameters of the Gamma distribution. By reducing the shape parameter and in contrast to that increasing the rate, which reduces the shape parameter, a convergence to higher p-values is obviously (Table 1).

We can see that we will find a rather good fit for specific values of the parameters. These values were tested by the version of Kolmogorov-Smirnov in R (See [12]). Thereby the shape parameter of 0.415 and a scale parameter of 1/0.485 delivered an accurate p-value of 0.09956. The test with a shape parameter of 0.42 delivered an even better p-value of 0.1272. Therefore it can be assumed that the standardized difference is Gamma distributed with a shape parameter which lies in between the

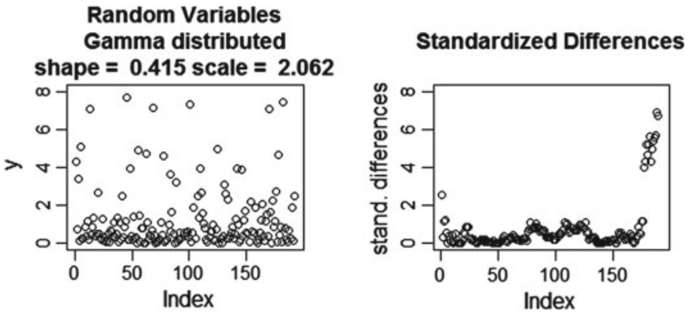


Fig. 9 Comparison of standardized differences with random variables of a gamma distribution

range [0.415, 0.42] and the scale parameter is about 2.06 ($\frac{1}{0.485}$). However it can be caused by random generators implemented in R-Software (see [12]), thus producing more entropy in the samples. Moreover it is quite useful to compare the standardized differences with generated random variables of a gamma distribution, with a shape parameter of 0.415 and a scale parameter of 2.062. This comparison can be seen in the Fig. 9.

3.2 Testing for Normality of the Groups

Due to the fact that within this paper normality plays an important role it needs to be investigated, whether one of these groups is normally distributed or not. Therefore two Shapiro-Wilk-Tests were computed to see if the groups can be assumed to be normal distributed. The p-value of masthopathy was smaller than 0.0001 and signalized, that the data can not be seen as normal distributed. In contrast to that the p-value of mammary cancer tissue (0.04519) was ten times higher. However this value is still too small (for a significance level of 95 %) to state that the box-counting dimension of mammary cancer tissue is normal distributed.

The QQ-Plots in Fig. 10 are another indication that masthopathy is not normal distributed, but normal distribution of the dimensions of mammary cancer cannot be rejected. Indeed the lower quantile differs significantly from the comparative line.

As previously seen it cannot be safely rejected, that the box-counting dimension of mammary cancer tissue is normally distributed. The p-value is very close to the rejecting area, so that we still can assume that this data is normal distributed. It would be useful to calculate the possible candidates for outliers (with formulas (1) and (2)), cut them off the data and compute another p-value. The same formula as on page 4 was used to compute the values of these outliers. This p-value has improved significantly up to 0.5716 and therefore it can be assumed that the data is normally distributed. Another manifestation for the normality of this group are the

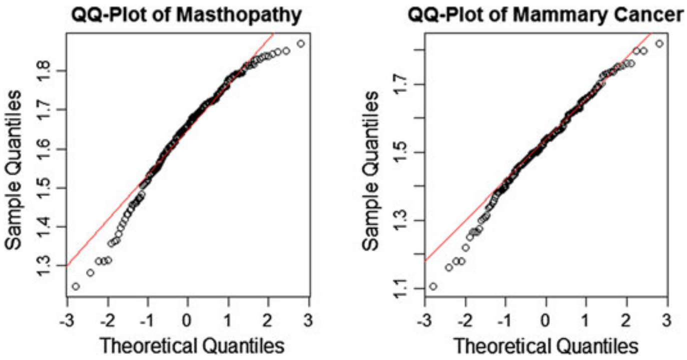


Fig. 10 QQ-Plot of the groups masthopathy (*left*) and mammary cancer (*right*)

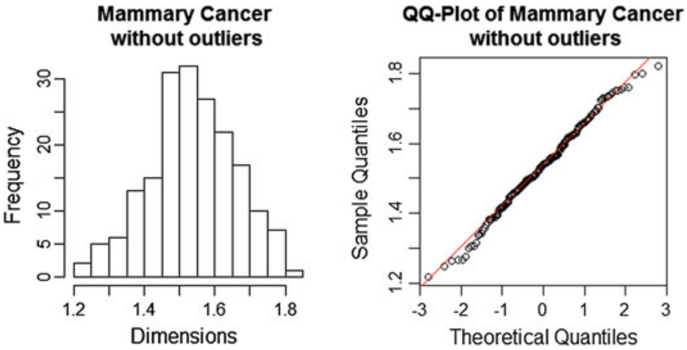


Fig. 11 Histogram and QQ-Plot of the data without outliers (n = 188)

histogram and QQ-Plot of the modified data in Fig. 11. Both of the plots suggest that the modified box counting dimension is normal distributed.

For this reason the next step will be to create a linear regression model. Owing to the fact that only the box-counting dimensions of mammary cancer tissue can be seen as normal distributed, a regression model was calculated just for this group. Thereby a significant intercept term and moreover a significant explanatory variable were computed. The p-value for both of the parameters are smaller than 2e-16 and the values are 1.34 (intercept) and 1.276e-03 (variable X, which contains the characteristics masthopathy or mammary cancer). The very high value of R^2 (0.9116) is another indication that this model fits well to the data. The following Fig. 12 shows the plot of the regression model.

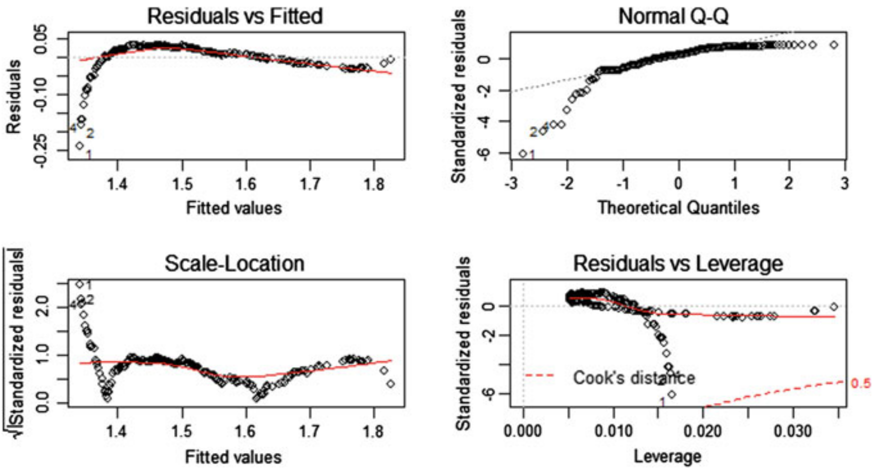


Fig. 12 Plot of the regression model

4 Discussion

In this paper we empirically model the distribution of the box-counting dimension from histological images of mammary tissues. We have discussed departures from normality, depth, heterogeneity and some complexities of distributions modeling entirely the fractal dimension. Our suggestion can be made to the practitioners to study such datasets in a deeper way to understand the depths and distribution deviation from normal or gamma samples. Several open problems remain e.g. relation between mathematical model of fractal dimension and tissue growth, parametric distribution of the box-counting dimension in mammary cancer. Such problems will be the valuable directions for future research.

Acknowledgments The research received partial support from the WTZ Project No. IN 11/2011 “Thermal modelling of cancer”. The corresponding author acknowledges Proyecto Interno 2015, REGUL. MAT 12.15.33, Modelaciòn del crecimientode tejidos con aplicaciones a la Investigaciòn del càncer. First author thanks to the support of *ANR project Desire FWF I 833-N18*. We also thank the editor and reviewers, whose insightful comments helped us to sharpen the paper considerably.

References

1. Baish, J.W., Jain, R.K.: Fractals and cancer. Perspect. Cancer Res. **60**, 3683–3688 (2000)
2. Breslow, N.E., Day, N.E.: Statistical Methods in Cancer Research, Volume 1: The Analysis of Case Controls Studies. IARC, Lyon (1980)
3. Chakravarthi, S., Choo, Z.W., Nagaraja, H.S.: Susceptibility to renal candidiasis due to immunosuppression induced by breast cancer cell lines. Sci. World J. **5**(1), 5–10 (2010)
4. Enby, E.: A breast cancer tumor consisted of a spore-sac fungus (Ascomycotina). 3rd Millennium Health Care Sci. **18**(1), 8–10 (2013)
5. George L.E., Kamal H.S.: Breast cancer diagnosis using multi-fractal dimension spectra. In: 2007 IEEE International Conference on Signal Processing and Communications (ICSPC'07) (2007)
6. Hermann, P., Mrkvička, T., Mattfeldt, T., Minárová, M., Helisová, K., Nicolis, O., Wartner, F., and Stehlík, M.: Fractal and stochastic geometry inference for breast cancer: a case study with random fractal models and Quermass-interaction process. Statistics in Medicine (2015) doi:10.1002/sim.6497
7. Kitsos, C.P.: Cancer Bioassays: A Statistical Approach, p. 110. LAMBERT Academic Publisher, Saarbrucken. ISBN 978-3-659-29451-8 (2012)
8. Kitsos, C.P.: Estimating the relative risk for the breast cancer. Biom. Lett. **47**(2), 133–146 (2010)
9. Mandelbrot, B.: The Fractal Geometry of Nature. W.H. Freeman and Co., New York (1982)
10. Mrkvička, T., Mattfeldt, T.: Testing histological images of mammary tissues on compatibility with the Boolean model of random sets. Image Anal. Stereol. **30**, 11–18 (2011)
11. Pázman, A.: Nonlinear statistical Models (chapters 9.1 and 9.2). Kluwer Academic Publication, Dordrecht (1993)
12. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, 2010. http://www.R-project.org [12 June 2013]
13. Stehlík, M., Giebel, S.M., Prostakova, J., Schenk, J.P.: Statistical inference on fractals for cancer risk assessment. Pakistan J. Statist. **30**(4), 439–454 (2014)
14. Stehlík, M.: Distributions of exact tests in the exponential family. Metrika **57**(2), 145–164 (2003)
15. Stehlík, M., Fabián, Z., Střelec, L.: Small sample robust testing for normality against Pareto tails. Commun. Stat.—Simul. Comput. **41**(7), 1167–1194 (2012b)
16. Stehlík, M., Wartner, F., Minárova, M.: Fractal analysis for cancer research: case study and simulation of fractals. PLISKA—Studia Mathematica Bulgarica **22**, 195–206 (2013)
17. Wosniok, W., Kitsos, C., Watanabe, K.: Statistical issues in the application of multistage and biologically based models. In: Prospective on Biologically Based Cancer Risk Assessment, pp. 243–273. Plenum Publication (NATO Pilot Study Publication by Cogliano, Luebeck, Zapponi (eds.)) (1998)
18. Zuo, Y., Serfling, R.: Nonparametric notions of multivariate "scatter measure" and "more scattered" based on statistical depth functions. J. Multivar. Anal. **75**(1), 62–78 (2000)

On Analytical Methods for Cancer Research

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

Abstract The use of image recognition and classification of objects according to images is becoming extremely popular, especially in the field of medicine. A mathematical procedure allows us, not only to evaluate the amount of data per se, but also ensures that each image is processed similarly. Our study has two focal points: The first one is the automated data entry and the second one is the evaluation in a manageable way. We propose the use of mathematical procedures to support the applicants in their evaluation of magnetic resonance images (MRI) of renal tumours. Therapy of renal tumours in childhood based on therapy optimizing SIOP (Society of Pediatric Oncology and Hematology)-study protocols in Europe. The most frequent tumour is the nephroblastoma (over 80 %). Other tumour entities in the retroperitoneum are clear cell sarcoma, renal cell carcinoma and extrarenal tumours, especially neuroblastoma. Radiological diagnosis is produced with the help of cross sectional imaging methods (computer tomography CT or Magnetic Resonance Images MRI). Our research is the first mathematical approach on MRI of retroperitoneal tumours for transversal images (of 40 patients). We use MRI in 3 planes and evaluate their potential to differentiate other types of tumours. We determine the key points or three dimensional landmarks of retroperitoneal tumours in childhood by using the edges

S. Giebel
University of Luxembourg, Luxembourg, Luxembourg
e-mail: Stefan.Giebel@gmx.de

P. Hermann · M. Stehlík (✉)
Department of Applied Statistics, Johannes-Kepler-University Linz,
Altenbergerstraße 69, 4040 Linz, Austria
e-mail: Milan.Stehlik@jku.at; mlnstehlik@gmail.com

P. Hermann
e-mail: philipp.hermann@jku.at

M. Stehlík
Departamento de Matemática, Universidad Técnica Federico Santa María,
Casilla 110-V, Valparaíso, Chile

J.-P. Schenk
Department of Diagnostic and Interventional Radiology,
University Hospital Heidelberg, Heidelberg, Germany
e-mail: Jens-peter_Schenk@med.uni-heidelberg.de

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_3

of the platonic body (C60) and test the difference between the groups (nephroblastoma versus non-nephroblastoma). The size is not eliminated like in former studies. All objects are comparable. For other important references see [1, 4, 9].

Keywords Diagnostics · Discrimination · Landmarks · Renal tumours · Wilms tumours

1 Introduction

The aim in our study is to differentiate tumours, not to detect them like in [6]. There are different kinds of renal tumours: Nephroblastoma (Wilms' tumour) [10] is the typical tumour of the kidneys appearing in childhood, as it is shown in Fig. 1. Therapy is organised in therapy-optimizing studies of the Society of Pediatric Oncology and Hematology (SIOP). Indication of preoperative chemotherapy is based on radiological findings. The preferred radiological methods are sonography and MRI. Both methods avoid radiation exposure, which is of great importance in childhood. Preoperative chemotherapy is performed without prior biopsy [8].

Information of the images of magnetic resonance tomography, especially the renal origin of a tumour and the mass effect with displacement of other organs, is needed for diagnosis. Besides nephroblastomas other tumours of the retroperitoneum exist, which are difficult to differentiate [7]. Renal tumours in childhood are classified into three groups of malignancy (I, II, III). Typical Wilms tumours mostly belong to stage II. In group II different subtypes of nephroblastoma tissue exist [3].

In our sample of tumours, four different types of retroperitoneal tumours are represented: nephroblastoma, neuroblastoma, clear cell carcinoma, and renal cell

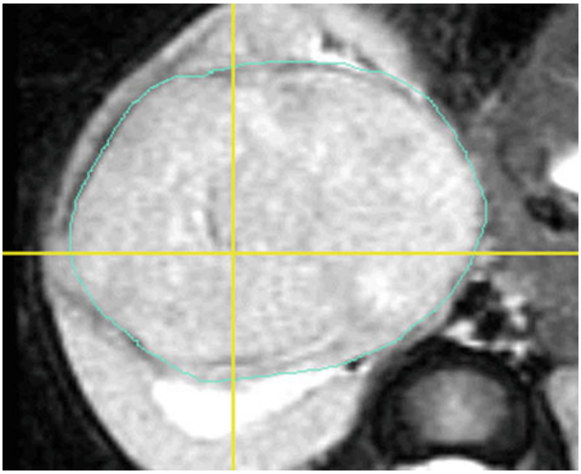


Fig. 1 Transversal image of a renal tumour

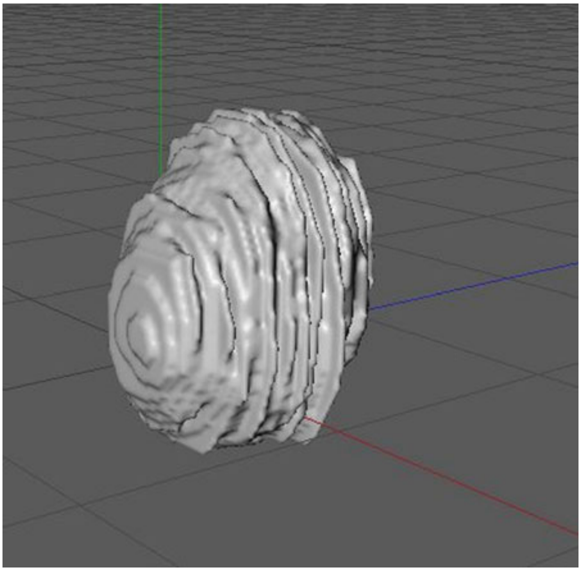


Fig. 2 3D Image of a renal tumour

carcinoma. Renal cell carcinomas are very rare in childhood. They represent the typical tumours of adult patients. They do not have high sensitivity for chemotherapy. Clear cell sarcomas are very rare in childhood and are characterised by high malignancy. Neuroblastomas are the typical tumours of the sympathetic nervous system and suprarenal glands. Infiltration of the kidney is possible.

The tumour grows with encasement of vessels. Because of the high importance of radiological diagnosis for therapy, it is of great interest to find markers for a good differentiation of tumours. MRI produces 2D-images. From the two dimensional data a three dimensional object has to be computed as it is shown in Fig. 2. Our aim is to find and to develop mathematical methods to support diagnosis.

2 Data Analysis

Landmarks are points to describe the object. If there are theoretical concepts of a group of objects like in anatomy, landmarks can be selected easily. Without theoretical concepts we need a procedure for finding landmarks. In the following study we get $3D$ landmarks by constructing a three dimensional object of the tumor. Then we take as landmarks the cut-points between the surface of the tumor and the vector of the edge of the platonic body C60.

Within the following analysis the data was investigated for normal-distribution as well as for differences between the group means of the x-, y- and z-coordinates of the Landmarks. Before starting to compare these values it is necessary to give a short descriptive overview of the three variables. The first necessary step was to separate

Table 1 Descriptive overview of the x-, y- and z-coordinate

Coordinate	Wilms	Minimum	Median	Mean	Maximum
x	No	−44.23	4.51	10.66	89.28
x	Yes	−90.47	−11.56	−8.18	103.22
y	No	−469.98	0.86	−1.15	49.09
y	Yes	−51.11	0.54	0.81	55.46
z	No	−102.04	−10.20	−14.82	49.14
z	Yes	−97.06	0.74	−1.33	85.05

the data according to whether Wilms tumour was diagnosed or not. The sample size consists of 40 patients, whereby 30 patients were diagnosed with Wilms, 7 with Non-Wilms and the cancer characteristic of the rest (3 patients) is not known at the moment of observation. We want to emphasize that the patients differ in their malignancy kind of cancer and for the sake of simplification we name patients diagnosed with other malignancy kind as "Non-Wilms group" hereafter. Moreover for every observation 60 landmarks were measured. A count of these points shows that between 186 and 6638 points were measured to get the exact location of the 60 necessary landmarks. This exact location was calculated by geometric methods. On an average 1666 measurements were needed for the desired result. It has to be mentioned that every one of these landmarks consists of a x-, y- and a z-coordinate. The data was differentiated on the basis of these three coordinates. This enables to compare the values of each of the coordinates between the two groups Wilms and Non-Wilms. Table 1 shows the minimum, maximum, median and mean of the three coordinates.

The first two columns of Table 1 show the coordinate and whether Wilms or Non-Wilms group was investigated. The next four columns present the values of the minimum, median, mean and the maximum. It is obvious, that the differences within the groups of the x- and z-coordinates are higher than at the y-coordinate. So the differences between the mean or the median are bigger than approximately 10 but are not exceeding around19 (see mean of the x-coordinates). Another outstanding aspect is that the the distance between the minima of the x-coordinates for the two groups is quite big, as is the case for the maxima of the z-coordinates. The distance between the minima is about 45 and the range between the maxima is approximately 36. The differences in the minimum and maximum of the y-coordinate are approximately 18 and 6. However computing differences of the median and the mean of this coordinate do not deliver values, such that a significant difference between the groups could be assumed (0.34 and 1.96).

Differences in the Group Means

The following histograms and QQ-Plots compare the x-, y- and z-coordinates of the landmarks for the two groups. The histograms are on the left side of the plot, whereby the histogram of the group without Wilms tumour is the left one, while that for the group with cancer is the right one. Figure 3 shows the comparison between the two groups for the x-coordinate. The differences can already be seen, especially in the

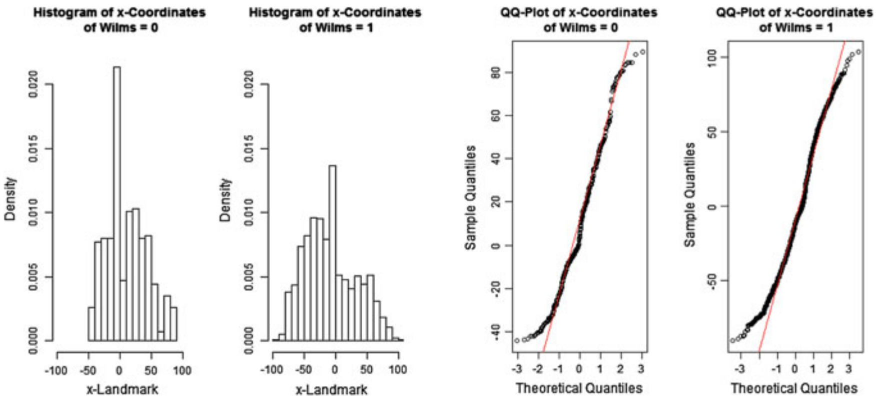


Fig. 3 Histogram and Quantile-Quantile-plot of the x-coordinates

negative region of the x-coordinate of the histogram on the left side. None of the landmarks where Wilms tumour was not diagnosed has a x-value which is smaller than −50. Therefore a Wilcoxon-Test was conducted to see whether this difference is random or statistically significant. In this case the calculated p-value is smaller than 0.0001, which means that the difference between these two groups is statistically significant. It needs to be mentioned that all the Wilcoxon-tests of this paper are based on a 95 % level of significance. This evidence supports the earlier assumptions regarding the differences, which can be seen in Table 1.

The same procedure was conducted for the y-(Fig. 4) and the z-(Fig. 5) coordinates and their outputs are shown in the following figures. The computed p-value of the Wilcoxon-Test was for the second coordinate approximately 0.2493 and for the last one again smaller than 0.0001. This means that the differences between the means of the y-coordinate are not statistically significant. Indeed, by having a closer look at the histograms of Fig. 4, no big differences between these two groups can be detected. For this reason it can be said that the means of the two groups of the y-coordinate are approximately equal.

In contrast to the last comparison the differences concerning the z-coordinate between the two observed groups are statistically significant at a confidence level of 95 %. This can be seen quite easily in the histogram of Fig. 5, since the z-coordinates of all of the all the observations from the "Non-Wilms group" were less than 50. Certainly some of the x-coordinates of the Wilms-tumour group are higher than 50 and the frequency in the range between −100 and −50 is lower in the cancer group. Generally it can be said, that the z-coordinates of the patients, who are not affected with cancer are on an average smaller. This assumption can be made, because it is obvious that a higher part of the density of the z-coordinate of the "Non-Wilms group" is on the left side compared to the cancer group. The calculated values for skewness of the cancer and "Non-Wilms group" are 0.19 and −0.80. Hence the values of kurtosis (3.86 and 4.16) are another hint for the different distribution of the groups.

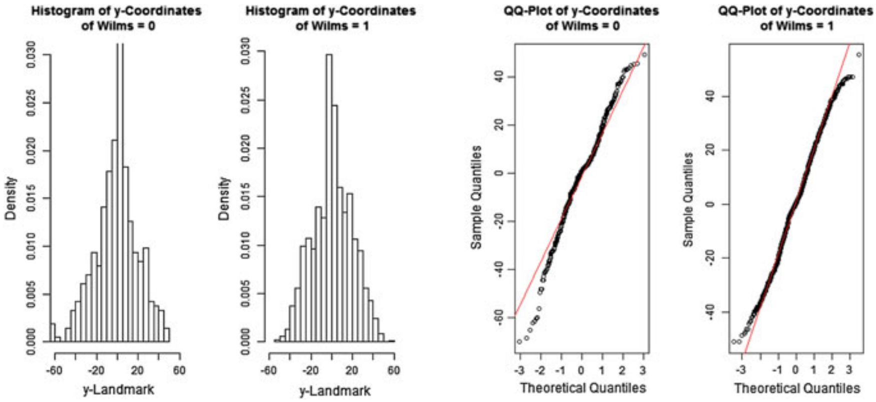


Fig. 4 Histogram and Quantile-Quantile-plot of the y-coordinates

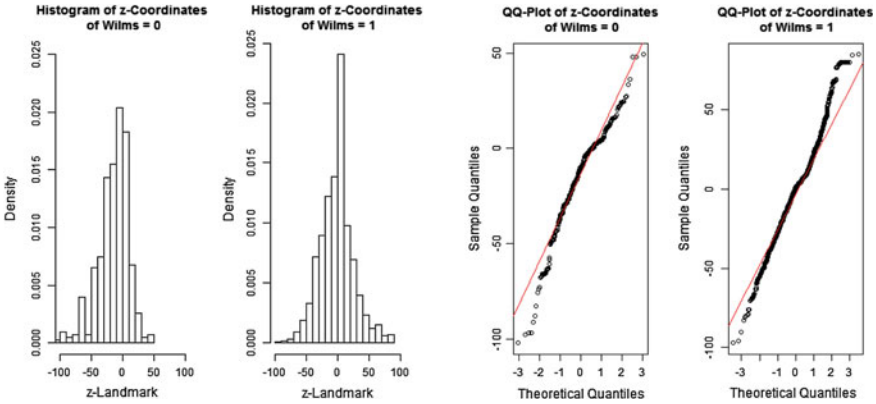


Fig. 5 Histogram and Quantile-Quantile-plot of the z-coordinates

Testing for Normal-Distribution

The next part is to test the data concerning normality. For further tests and investigations a normal distribution of the data is, as known, desirable. In this case study QQ-Plots and Shapiro-Wilk-Tests were accomplished to have a mathematical as well as a visual check. The plots on the right side of the previous Figs. 3, 4 and 5 show the QQ-Plots, where on the right side of the QQ-Plots the observations with Wilms tumour are placed. At first view in Fig. 3 the visual test for normality seems to be quite useful, nevertheless the upper and lower end of the plot shows that it cannot be said that the x-coordinates of both groups are normally distributed. The p-values of the Shapiro-Wilk-Tests are smaller than 0.0001 for both of the groups, which confirm the assumption that none of the groups is normally distributed.

The two graphics on the right side of Fig. 4 show the Quantile-Quantile plots for the y-coordinates. Due to greater deviations in the lower quantiles and additionally a p-value smaller than 0.0001 display that the values of the second coordinate are also not normally distributed. Although the QQ-Plot for the group where cancer was diagnosed does not look that bad, the p-value smaller than 0.0001 clearly shows that this group is not normally distributed neither, because of the greater deviations at the upper and lower quantiles.

Even prior to calculating the p-values for the z-coordinates it can be seen from the QQ-Plots of Fig. 5 that none of the groups will be normally distributed, because the deviations between the empirical and the theoretical quantiles are too big. The very small p-values (both smaller than 0.0001) show that our first impression was correct.

3 Conclusions

Even in the case of transversal images Wilms- and Non-Wilms tumours can be differentiated. In contrast to the approach of Giebel [2] three dimensional landmarks were used. The sample of 37 cancer patients equates to a third of the total population with renal tumours per year. Hence our results give a new approach for a support in diagnosis. The validity in oncology is questionable like for prostata [5]. Even if the sample is bigger, the question of validity has to be solved by a bootstrap procedure. The crucial points are all unknown patients in the future.

Furthermore we are using our data analysis procedure in a field in medicine, where decisions according to images are necessary to get a suitable therapy. Our procedure could be a first step in the development of a tool to help the diagnostics of images for all patients. Especially, in the light of the sample with renal tumours, our results give a new approach for a support in diagnosis.

Further studies with an extension on $4D$ are a future option, especially to study changes in the shape of the tumor during chemotherapy. Another important study object should be the shape analysis in organs with fast movements, e.g. shape of heart ventricles or aorta during cardiac cycle.

Acknowledgments The research received partial support from the WTZ Project No. IN 11/2011 “Thermal modelling of cancer”. We also thank the editor and reviewers, whose insightful comments helped us to sharpen the paper considerably.

References

1. Furtwängler, R., Schenk, J.P., Reinhard, H., et al.: Nephroblastom- Wilms-Tumor. Onkologie **11**(1) (2005)
2. Giebel, S.: Statistical analysis of the shape of renal tumors in childhood. Diploma thesis, University Kassel (2007)
3. Graf, N., Semler, O., Reinhard, H.: Die Prognose des Wilms-Tumors im Verlauf der SIOP-Studien. Der Urologe, Ausgabe A **43**(4), 421–428 (2004)

4. 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)
5. McCarthy, M.: PSA screening said to reduce prostate cancer deaths, or does it? Lancet **351**, 1563 (1998)
6. Ratto, C., Sofo, L., Ippoliti, M., Merico, M., Doglietto, G.B., Crucitti, F.: Prognostic factors in colorectal cancer. Literature review for clinical application. Dis. Colon Rectum **41**, 1033–1049 (1998)
7. 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)
8. Schenk, J.P., Schrader, C., Zieger, B., Furtwängler, R., Leuschner, I., Ley, S., Graf, N., Tröger J.: Reference radiology in nephroblastoma: accuracy and relevance for preoperative chemotherapy. Rofo **178**(1), 38–45 (2006)
9. Schenk, J.P., Waag, K.L., Graf, N., Wunsch, R., Jourdan, C., Behnisch, W., Tröger J., Günther, P.: 3D-visualization by MRI for surgical planning of Wilms tumors. Rofo **176**(10), 1447–1452 (2004)
10. Wilms, M.: Die Mischgeschwülste der Niere, pp. 1–90. Verlag von Arthur Georgi, Leipzig (1889)

Modelling Times Between Events with a Cured Fraction Using a First Hitting Time Regression Model with Individual Random Effects

S. Malefaki, P. Economou and C. Caroni

Abstract The empirical survival function of time-to-event data very often appears not to tend to zero. Thus there are long-term survivors, or a “cured fraction” of units which will apparently never experience the event of interest. This feature of the data can be incorporated into lifetime models in various ways, for example, by using mixture distributions to construct a more complex model. Alternatively, first hitting time (FHT) models can be used. One of the most attractive properties of a FHT model for lifetimes based on a latent Wiener process is that long-term survivors appear naturally—corresponding to failure of the process to reach the absorbing boundary—without the need to introduce special components to describe the phenomenon. FHT models have been extended recently in order to incorporate individual random effects into their drift and starting level parameters and also to be applicable in situations with recurrent events on the same unit with possible right censoring of the last stage. These models are extended here to allow censoring to occur at every intermediate stage. Issues of model selection are also considered. Finally, the proposed FHT regression model is fitted to a dataset consisting of the times of repeated applications for treatment made by drug users.

Keywords Recurrent events · Wiener process · Long-term survivors

S. Malefaki (✉)
Department of Mechanical Engineering and Aeronautics, University of Patras,
26504 Rion-Patras, Greece
e-mail: smalefaki@upatras.gr

P. Economou
Department of Civil Engineering, University of Patras, 26504 Rion-Patras, Greece
e-mail: peconom@upatras.gr

C. Caroni
Department of Mathematics, School of Applied Mathematical and Physical Sciences,
National Technical University of Athens, 157 80 Athens, Greece
e-mail: ccar@math.ntua.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_4

1 Introduction

In survival and reliability studies it is commonly observed that some experimental units in the sample have not undergone the event of interest by the time that data collection ceases. These data are recorded as right-censored observations. Although the usual interpretation is that the time until the event for these units is larger than the time available for the study, an alternative is that some or all of them are long-term survivors, or a "cured fraction" of units which will never experience the event of interest no matter for how long the study continues.

Many different approaches to modelling a cured fraction can be found in the literature. One is a mixture model in which it is assumed that a proportion of units will never experience the event of interest (failure, death etc.) because they were never actually at risk. This idea seems to go back to [6, 7]; also see, for example, [18]. Other approaches include Yakovlev et al.'s model [24] and more recently a model that proposes a different underlying mechanism as the source of the long-term survivors [8]. Another alternative interpretation of the nature of the cured fraction can be obtained through a first hitting time (FHT) or threshold regression model. It is among the most attractive properties of the FHT model for lifetimes based on a latent Wiener process that long-term survivors appear naturally—corresponding to failure of the process to reach the absorbing boundary—without the need to introduce special components into the model in order to describe the phenomenon. Reviews of FHT models and some of their applications can be found in [1, 13, 15], for example; see also [14, 16, 23] and references therein. A brief discussion of the potential use of the FHT models as cured fraction models can be found in [16]. Several extensions and modifications of the FHT model were suggested by [4] in order to improve the fit of the adopted model.

One emphasis of this paper is consequently on how different cured fraction models represent the nature of the long-term survivors and how these arise in the context of FHT models, which is an area still under development. Our modelling is carried out within the framework of a recent extension of FHT regression models to recurrent events [10]. Our other emphasis is on the further extension of this model here in Sect. 2 in order to allow censoring to occur in every intermediate stage and not only at the last stage. In the same section an MCMC algorithm for simulating observations from the posterior distribution is presented briefly. Model selection criteria are also discussed. In Sect. 3 the nature of the long-term survivors under this FHT regression model is discussed in detail. A case study, in the form of the application of the model to recurrent events data on drug users, is presented in Sect. 4. The paper concludes with a short discussion.

2 Model Definition and Estimation

2.1 The Model

Time-to-event data take the simple form of the time (on an appropriate scale—real time or operating time) from an origin (a machine commences operation, a patient undergoes surgery) until a well-defined event occurs (such as the breakdown of the machine or the death of the patient). These simple data may, however, arise from a complex underlying process. The breakdown of the machine is a result of the increasing wear and tear on its components; the death of the patient follows from the deterioration of his or her state of health. The chief characteristic of FHT models for the time until an event occurs is that it recognizes this structure, by postulating that the observable event occurs when an underlying stochastic process reaches a certain boundary or threshold for the first time. This stochastic process could itself be observable—for example, the size of a crack in a machine component. More often, however, it will be regarded as a latent process. For example, the patient's condition might be measured by many indicators, but in the FHT context will just be thought of as an unobservable "state of health" whose value changes stochastically in time. If this latent construct is presumed to take non-negative values, then the death of the patient is supposed to happen when the state of health falls to zero.

In this context, the underlying continuous time stochastic process is denoted by $\{X(t),\ t \in \mathscr{T},\ X(t) \in \mathscr{X}\}$, where $\mathscr{X}$ is its continuous state space. The distribution of lifetimes is given by the first passage time from the initial state at time zero, $X(0) = x_0$, to the boundary or threshold, $\mathscr{B} \subset \mathscr{X}$. A useful and common choice for the parent stochastic process is a Wiener process (e.g. [14]). In this case, the observed lifetimes follow an inverse Gaussian distribution with parameter values that are functions of the initial state x_0, variance σ^2 and drift μ of the Wiener process [16].

This model can easily be extended to a sequence of latent processes in order to describe recurrent events in independent individuals. For individual i ($1 \leq i \leq n$; the sample size), n_i stages are observed, demarcated by the time points: $0 = t_{0i} < t_{1i} < t_{2i} < \cdots < t_{n_i i}$. These are realizations of first passage times in the sequence of processes

$$\{X_m(t),\ t \in [T_{m-1}, +\infty),\ x \in \mathscr{X}\},\ m = 1, 2, \ldots,$$

(suppressing the index i for brevity) where $\mathscr{X}$ is the common state space of the processes, $0 = T_0 < T_1 < T_2 < \cdots$, and $X_m(T_{m-1}) = x_m$ is the initial value of the m^{th} stage of the process. Stage m ends at time T_m, with duration $S_m = T_m - T_{m-1}$ since the previous event, whereupon the process restarts (a machine is repaired, a chronically ill patient receives treatment). The starting value and the parameters of the underlying Wiener process will be allowed to vary between stages. The threshold could also be allowed to vary, but in our models it will always be assumed to be zero. In many applications, only the last stage can be censored but this is not an essential restriction. In the proposed model, we shall allow censoring to occur in every intermediate stage.

Heterogeneity among individuals in an FHT regression model is usually introduced by allowing the initial state x_m and the drift μ_m to depend on the values of the available individual-level covariates. In addition, in the context of recurrent events, we will also introduce dependence on two process-related covariates: the number of stages $m - 1$ previously completed by the individual and the total time elapsed t_{m-1} until the beginning of the current stage. However, the implication that two different individuals with the same covariates will have the same initial state and drift, although possibly realistic in some reliability studies, is not reasonable in human studies. The extra heterogeneity that is usually observed between individuals and cannot be explained by the available covariates can be introduced into the model through individual random effects. These random effects are expressed at each stage for the drift as normal random variables with an additive effect on this characteristic of the process. For the initial states of the processes, we assume that they are generated by a normal distribution truncated to the left at zero (the threshold value) with variance τ^{-1} and a location parameter that depends on the extended vector of the covariates (the initial covariates, the number of previous stages and the total time so far t_{m-1}).

More specifically, for the initial state x_{mi} and drift μ_{mi} of the ith individual's mth stage, the following regression structures are assumed:

$$x_{mi} \sim N_{+}\left(\tilde{\mathbf{u}}'_{mi}\boldsymbol{\alpha}^{*}, \tau^{-1}\right) \tag{1}$$

$$\mu_{mi} = \sum\nolimits_{k=1}^{m} b_{ki} + \tilde{\mathbf{v}}'_{mi}\boldsymbol{\beta}^{*} \tag{2}$$

where

$$\begin{aligned}\tilde{\mathbf{u}}'_{mi}\boldsymbol{\alpha}^{*} &= \mathbf{u}'_{i}\boldsymbol{\alpha} + \alpha_{ns}(m-1) + \alpha_{s}t_{(m-1)i}\\ \tilde{\mathbf{v}}'_{mi}\boldsymbol{\beta}^{*} &= \mathbf{v}'_{i}\boldsymbol{\beta} + \beta_{ns}(m-1) + \beta_{s}t_{(m-1)i}\quad .\end{aligned}$$

The vectors $\tilde{\mathbf{u}}_{mi}$ and $\tilde{\mathbf{v}}_{mi}$ denote the covariates associated with the starting level and drift, respectively, of this stage of that individual, including the covariates $\mathbf{u}_i$ and $\mathbf{v}_i$ (not necessarily disjoint) of individual characteristics that remain unchanged through the stages. More generally, their values could change between stages, so that $\mathbf{u}_i$ would become $\mathbf{u}_{mi}$, with a similar change in $\mathbf{v}_i$. The vectors $\boldsymbol{\alpha}^{*'} = (\boldsymbol{\alpha}', \alpha_{ns}, \alpha_s)$ and $\boldsymbol{\beta}^{*'} = (\boldsymbol{\beta}', \beta_{ns}, \beta_s)$ are the regression coefficients associated with the starting level and drift, respectively. Finally, the $b_{ki} \sim N\left(0, \lambda^{-1}\right)$, $k = 1, \ldots, n_i$, are the individual's random effects.

2.2 The Likelihood

When the parent stochastic process is a Wiener process, it is absorbed at the boundary with probability one if the drift parameter $\mu_{mi} \leq 0$. In this case, the random variable $S_{mi} = T_{mi} - T_{m-1,i}$ which expresses the time elapsed between two successive events, i.e. the time required for the mth stage of the process in the ith individual to reach the threshold level for the first time, follows an inverse Gaussian distribution with mean $-\frac{x_{mi}}{\mu_{mi}}$ and scale parameter $\left(\frac{x_{mi}}{\sigma}\right)^2$ [9]. The scale of the unobserved underlying process is arbitrary, hence the model is over-parameterized and therefore one parameter must be fixed. Because we are modelling the dependence of the drift on the available covariates, the parameter that can be fixed is the variance of the Wiener process, which is set equal to one: $\sigma^2 = 1$. With this modification the pdf is

$$f(s_{mi}|x_{mi}, \mu_{mi}) = \frac{x_{mi}}{\sqrt{2\pi s_{mi}^3}} \exp\left(-\frac{(\mu_{mi}s_{mi} + x_{mi})^2}{2s_{mi}}\right), \qquad s_{mi} > 0 \tag{3}$$

and the survival function is

$$\bar{F}(s_{mi}|x_{mi}, \mu_{mi}) = \Phi\left[\frac{\mu_{mi}s_{mi} + x_{mi}}{\sqrt{s_{mi}}}\right] - \exp\left(-2x_{mi}\mu_{mi}\right)\Phi\left[\frac{\mu_{mi}s_{mi} - x_{mi}}{\sqrt{s_{mi}}}\right], \tag{4}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. If the drift parameter $\mu_{mi} > 0$, then absorption is not certain; the probability of reaching the boundary is $\exp(-2x_{mi}\mu_{mi})$. The inverse Gaussian distribution with this property has been characterized as a "defective" inverse Gaussian distribution [25]. The distribution given by (3) and (4) continues to apply, however, conditionally on reaching the boundary.

Under the assumed regression model every stage is described by a Wiener process which given its starting point and drift is independent of other stages. The event time for any uncensored stage follows an inverse Gaussian distribution and therefore its contribution to the likelihood function is

$$\frac{x_{mi}}{\sqrt{2\pi s_{mi}^3}} \exp\left(-\frac{(\mu_{mi}s_{mi} + x_{mi})^2}{2s_{mi}}\right) \tag{5}$$

where s_{mi} denotes the inter–event time of the mth stage of the ith individual, given by $s_{mi} = t_{mi} - t_{m-1,i}$ with $t_{0i} = 0$. The event time for a censored stage is not observed and therefore its contribution to the likelihood is the survival function (4), assuming uninformative censoring as usual. Alternatively, if we could observe for the censored mth stage of the ith individual the level ℓ_{mi} of the stochastic process

$\{X_{mi}(t),\ t \in [T_{m-1,i}, +\infty)\}$, or equivalently of the $\{X_{mi}(s),\ s \in [0, +\infty)\}$ at time s_{mi}, then the contribution to the likelihood would be given by the pdf of the stochastic process [9]:

$$p(\ell_{mi}, s_{mi}|x_{mi}; \mu_{mi}) = \frac{1}{\sqrt{2\pi s_{mi}}} \exp\left(-\frac{(\ell_{mi} - x_{mi} - \mu_{mi}s_{mi})^2}{2s_{mi}}\right) \times \left[1 - \exp\left(-\frac{2\ell_{mi}x_{n_i i}}{s_{n_i i}}\right)\right]. \tag{6}$$

It is feasible to use this function in the likelihood if we treat the unobserved level ℓ_{mi} for a censored stage as an unknown parameter and then sample it from its full conditional posterior distribution in the MCMC algorithm that will be described below.

To sum up, the likelihood for the regression model is given by

$$\prod_{i=1}^{n}\prod_{m=1}^{n_i} \left(\frac{x_{mi}}{\sqrt{2\pi s_{mi}^3}} \exp\left(-\frac{(\mu_{mi}s_{mi} + x_{mi})^2}{2s_{mi}}\right)\right)^{d_{mi}} \times \left(\frac{1}{\sqrt{2\pi s_{mi}}} \exp\left(-\frac{(\ell_{mi} - x_{mi} - \mu_{mi}s_{mi})^2}{2s_{mi}}\right)\left[1 - \exp\left(-\frac{2\ell_{mi}x_{mi}}{s_{mi}}\right)\right]\right)^{1-d_{mi}}$$

where d_{mi} is the usual censoring indicator with $d_{mi} = 1$ for an uncensored stage and $d_{mi} = 0$ for a censored one.

Before we present the MCMC algorithm, it is also necessary to assign priors to the parameters of the model. Conjugate priors are assigned to the regression parameters $\boldsymbol{\beta}^{*'} = (\boldsymbol{\beta}', \beta_{ns}, \beta_s)$ and to λ, the inverse of the variance of b_{mi}:

$$\begin{aligned} \boldsymbol{\beta}^* &\sim N_q(\boldsymbol{\beta}^*_{prior}, \boldsymbol{\Sigma}_{\beta^*}) \\ \lambda &\sim Gamma(\gamma_1, \gamma_2) \end{aligned}$$

where q is the dimension of $\boldsymbol{\beta}^*$ (including the constant term). The choice of $\boldsymbol{\Sigma}_{\beta^*}$ reflects the degree of confidence in the point estimate $\boldsymbol{\beta}^*_{prior}$, which has possibly been obtained from previous analyses. If it is not considered to be a good estimate, then a reasonable choice under the assumption of no multicollinearity could be $\boldsymbol{\Sigma}_{\beta^*} = 10^4\boldsymbol{I}_q$, where $\boldsymbol{I}_q$ is the identity matrix of size q.

A suitable prior for λ is the Gamma distribution with mean γ_1/γ_2 and variance γ_1/γ_2^2. The choices for the hyperparameters γ_1 and γ_2 could reflect information from previous studies. Otherwise, a diffuse prior such as $Gamma(2, 1/2)$ can be used.

The prior distribution assigned for the regression parameters $\boldsymbol{\alpha}^*$ of the initial states x_{mi} is a p dimensional normal distribution:

$$\boldsymbol{\alpha}^* \sim N_p(\boldsymbol{\alpha}^*_{prior}, \boldsymbol{\Sigma}_{\alpha^*})$$

where p is the dimension of $\boldsymbol{\alpha}^*$, and $\boldsymbol{\alpha}^*_{prior}$ and $\boldsymbol{\Sigma}_{\alpha^*}$ are chosen in a similar way to $\boldsymbol{\beta}^*_{prior}$ and $\boldsymbol{\Sigma}_{\beta^*}$, respectively.

It is unlikely that information would be available concerning the unobserved levels ℓ_{mi} of the latent stochastic process of a censored stage, so we assign an improper, non-informative flat prior to ℓ_{mi}, given $d_i = 0$:

$$\pi(\ell_{mi}|d_{mi} = 0, \cdot) \propto 1(\ell_{mi} > 0) \tag{7}$$

where $1(\cdot)$ stands for the indicator function.

A gamma prior could be assigned to the parameter τ. However, we follow the recommendation of Pennell et al. [19] to take a fixed moderately small value in order to avoid identifiability problems. Their investigation found that results were not sensitive to the choice of τ. Our own findings from a sensitivity analysis reached the same conclusion and thus are omitted.

Combining all the above, the posterior distribution for the regression model allowing right censoring of any stage is given by

$$\begin{aligned}
L \propto & \prod_{i=1}^{n} \left\{ \prod_{m=1}^{n_i} \left(\frac{x_{mi}}{\sqrt{2\pi s_{mi}^3}} \exp\left(-\frac{(\mu_{mi} s_{mi} + x_{mi})^2}{2 s_{mi}} \right) \right)^{d_{mi}} \times \right. \\
& \left(\frac{1}{\sqrt{2\pi s_{mi}}} \exp\left(-\frac{(\ell_{mi} - x_{mi} - \mu_{mi} s_{mi})^2}{2 s_{mi}} \right) \left[1 - \exp\left(-\frac{2\ell_{mi} x_{mi}}{s_{mi}} \right) \right] \right)^{1-d_{mi}} \times \\
& \left. \sqrt{\lambda} \exp\left(-\frac{1}{2} \lambda b_{mi}^2 \right) \cdot \frac{\exp\left(-\frac{\tau (x_{mi} - \tilde{\mathbf{u}}'_{mi} \boldsymbol{\alpha}^*)^2}{2} \right)}{\left(1 - \Phi\left(-\tau^{\frac{1}{2}} \tilde{\mathbf{u}}'_{mi} \boldsymbol{\alpha}^* \right) \right)} \right\} \times \\
& \lambda^{\gamma_1 - 1} \exp(-\gamma_2 \lambda) \cdot \exp\left(-\frac{1}{2} (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^*_{prior})' \boldsymbol{\Sigma}_{\alpha^*}^{-1} (\boldsymbol{\alpha}^* - \boldsymbol{\alpha}^*_{prior}) \right) \times \\
& \exp\left(-\frac{1}{2} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^*_{prior})' \boldsymbol{\Sigma}_{\beta^*}^{-1} (\boldsymbol{\beta}^* - \boldsymbol{\beta}^*_{prior}) \right) \cdot \prod_{i=1}^{n} \prod_{\substack{m=1 \\ d_{mi}=0}}^{n_i} 1(x_i > 0)
\end{aligned}$$

where

$$\tilde{\mathbf{u}}'_{mi} \boldsymbol{\alpha}^* = \mathbf{u}'_{1i} \boldsymbol{\alpha} + \alpha_{ns}(m-1) + \alpha_s t_{(m-1)i}$$

and

$$\mu_{mi} = \sum_{k=1}^{m} b_{ki} + \mathbf{v}'_{1i} \boldsymbol{\beta} + \beta_{ns}(m-1) + \beta_s t_{(m-1)i} \,.$$

In the following subsection we describe briefly an MCMC algorithm for sampling from the posterior distribution (see also [10, 19]).

2.3 The MCMC

A hybrid Gibbs Sampler is proposed for simulation from the posterior distribution. More specifically, an accept–reject algorithm is used in order to simulate the unobserved level ℓ_{mi} when the stage m is censored. The shape of the target distribution, which is the full conditional distribution of ℓ_{mi} given the remaining parameters of the model, depends on the value of the quantity $\mu_{mi}s_{mi} + x_{mi}$, thus two different proposals are used. A similar approach is used for the starting level x_{mi} when the stage m is censored, because its full conditional distribution has the same form as that of the unobserved level ℓ_{mi}. When the stage m is not censored, an accept–reject algorithm is also used but with proposal distribution the truncated normal distribution with mean equal to the mode of its full conditional distribution. The regression coefficients of the starting levels are simulated using Griffiths' method [11] for sampling from truncated multivariate normal distributions. The remaining model parameters are simulated using a regular Gibbs step, using their full conditional distributions since they have a known form. For more details of the algorithm see [10] and also [19].

2.4 Model Selection Criteria

Model selection is a key issue in data analysis and there is an extensive literature on this topic: see, for example, [17, 20] for reviews. We suggest that two criteria, namely the Deviance Information Criterion (DIC) [21] and the Information Criterion (IC) [3], are suitable measures for selecting from among a set of candidate models in the case of FHT models.

The DIC is defined as

$$DIC = -2E_{\theta|y}(\log f(y|\theta)) + P_D,$$

where

$$P_D = 2\log f(y|\hat{\theta}) - 2E_{\theta|y}(\log f(y|\theta))$$

and $\hat{\theta}$ is the posterior mean of the model parameters. Although this criterion is easy to compute, it suffers from the disadvantage that it tends to select overfitted models. In order to overcome this problem, Ando [2] proposed the Bayesian Predictive Information Criterion (BPIC). Its form does not allow easy calculations and thus it is not readily applicable, especially in models as complex as the ones proposed in this paper. Subsequently, Ando [3] proposed the IC which is claimed to incorporate the advantages of both the DIC and BPIC. It is easy to compute in any model without overfitting. The IC is a modification of the DIC, given by

$$IC = -2E_{\theta|y}(\log f(y|\theta)) + 2P_D.$$

The IC usually prefers a simpler model than the one selected by DIC since its penalty term is twice that of the DIC. For both DIC and IC, a model with a smaller value of the criterion is preferred over models with larger values.

3 The Nature of the Cured Fraction

Before moving on to an illustrative application of the proposed model, it is worth comparing and contrasting the way in which the cured fraction appears in different models. As already mentioned, an FHT model for lifetimes based on a latent Wiener process incorporates the presence of long-term survivors in a natural way, without the need to introduce any special component into the model. We discuss here this feature in more detail in order to understand the nature of the cured fraction in this approach in comparison to the classical cured fraction mixture model and the model proposed recently by [8]. In the cured fraction mixture model [7] the survival function is given by

$$S_c(t) = \pi_c + (1 - \pi_c)S_{T_c}(t)\,,$$

where π_c is the proportion of units that are never at risk of the event and $S_{T_c}(t)$ is the survival function of the random variable T_c which represents the time to the event of interest in the population of units that are susceptible to the event. Under this approach it is assumed firstly that the proportion π_c is constant over time and secondly that these cured units were actually "cured" from the beginning and remain so throughout the study.

Under an alternative hidden competing risk model for censored observations proposed in [8], all the units start out as susceptible to the event of interest but may move into a non-susceptible group if another event intervenes. This results in a continuously increasing proportion of "cured" units. Both these models divide, at any given time, the population into two clearly defined groups, one that is susceptible to the event of interest and one that is not. This assumption may not always be realistic and in some cases the following situation may seem more reasonable. We assume that some members of the population will eventually experience the event of interest with probability one (these members belong to the susceptible group) but for the rest of the population, this probability is less than one but not zero. This means that the population is not separated into susceptible and non-susceptible groups, but rather into a susceptible group and a potentially susceptible group.

A characteristic example, which is actually the application of the current work, is the time that passes until a drug user applies for treatment. It could be supposed that the decision to approach a treatment service depends on the user's mental and physical state. This unobservable state is not fixed but varies over time and the random walk of the Wiener process may be a reasonable representation of it. The particular individual characteristics of the drug user may indicate that he or she has a value of the drift parameter that will eventually lead to an application for treatment. On the other hand, for the majority of the drug users this is not the case and it is not certain that the drug user will ever seek treatment again.

This situation arises naturally in the Wiener-based FHT model because of the non-zero probability $\exp(-2x_m\mu_m)$ that the mth stage of the process fails to reach the threshold when $\mu_m > 0$, which means that long-term survivors appear automatically without the need to introduce special components into the model in order to describe the phenomenon. This was an important feature of the application in [25]; it is considered in detail, along with various extensions, by [4].

This property provides an alternative to the usual explanation for the long-term survivors that appear in the sample as right-censored observations. A right-censored observation in this FHT model does not necessarily correspond to an individual who has not yet experienced the event of interest, nor to an individual who will never experience the event of interest because he or she was never actually at risk, as in a mixture model. Instead, it may arise from an individual for whom failure was a possibility, but not an ultimate certainty. The interpretation in the case of drug users' applications for treatment is that, if the parameter $\mu > 0$ then depending on the course of the underlying stochastic process that describes his or her psychological and physical state, the user might reach the point of requesting treatment but possibly will never do so. This is more realistic than supposing that the drug user started this stage in such a state that guaranteed that he or she would never seek treatment.

Under the proposed model, both of the parameters x_{mi} and μ_{mi} depend on the values of the covariates, the number of previous stages and the total time t under observation until the beginning of the current stage, through the relationships (1) and (2). The expected value of the drift is

$$E(\mu_{mi}) = \mathbf{v}_i'\boldsymbol{\beta} + \beta_{ns}(m-1) + \beta_s t_{(m-1)i} \,.$$

The expected value of the starting level is given by the mean of the corresponding truncated normal distribution which can be expressed as

$$E(x_{mi}) = \mathbf{u}_i'\boldsymbol{\alpha} + \alpha_{ns}(m-1) + \alpha_s t_{(m-1)i} + \frac{h(\zeta_{mi})}{\sqrt{\tau}}$$

where

$$\zeta_{mi} = -\tau^{\frac{1}{2}}\left(\mathbf{u}_i'\boldsymbol{\alpha} + \alpha_{ns}(m-1) + \alpha_s t_{(m-1)i}\right)$$

and $h(\cdot)$ is the hazard function of the standard normal distribution [5]. The expected time to absorption at the threshold (hence, the expected duration of this stage) is

$$E(S_{mi}) = x_{mi}/|\mu_{mi}|$$

where $S_{mi} = T_{mi} - T_{m-1,i}$. If $\mu_{mi} > 0$ this expectation still holds conditioning on the event's occurrence but, as stated earlier, in this case there is a probability $1 - \exp(-2x_{mi}\mu_{mi})$ that the Wiener process will never reach the threshold on this stage. It is apparent that larger values of x_{mi} make it less likely that this mth stage will end in an event, and increase the expected time to event conditioning on its occurrence. Larger values of $\mu_{mi} > 0$ also make it less likely that an event will occur, but decrease the expected time to its occurrence when it does occur.

Further properties of the FHT regression model based on an underlying Wiener process, focusing on comparison with the widely used semi-parametric Cox regression model, are given in [23].

4 Case Study

4.1 The Data

Health care services generally have to deal with the same client repeatedly. The times of the client's successive contacts with the service form a sequence of recurrent events. In the illustrative example that we will be analyzing here, the clients are approaching services in order to seek treatment for their drug use problems. Each approach is added to a central database. The time origin $t_0 = 0$ for the ith individual is the time of his or her first approach and entry into the database. The recurrent events are further approaches, if any. The purpose of the present analysis is to identify which, if any, individual characteristics affect the time until recontacting services. It is restricted to 1553 individuals who reported that they were primarily using some substance other than heroin. Only 188 of them (12.1 %) made a further contact with services (and therefore experienced an "event") and 36 of these (19.2 %) reappeared again. Altogether, 1791 stages were recorded, of which 238 ended in an event. Only final stages were censored. As the duration of the study was 10 years, this small number of events may indicate the presence of a "cured fraction" in the population. The Kaplan-Meier estimates [12] of the survival function separately for data from the first and the second stages are presented in Fig. 1. From these plots it is clear that the survival functions do not tend to zero, which indicates the presence of a proportion of drug users who will never apply again for treatment ("cured fraction"). This proportion seems to be smaller among those who have already been recorded in the database, who apparently tend to reapply earlier (more rapidly decreasing survival function).

Covariates recorded at the time of the initial entry into the database were: place of residence, whether or not the client had received any treatment before the database started operation in 2001 (reported by 89.6 %), gender (81.9 % male), age (mean 29.9 years, SD 9.4) and the year of initial entry (coded as years after 2001). As place of residence was recorded in three categories (Athens/Piraeus, Thessaloniki, Other), it was entered into the analysis as two indicator variables, one denoting Athens/Piraeus (73.7 % of the sample) and the other denoting Thessaloniki (15.6 %).

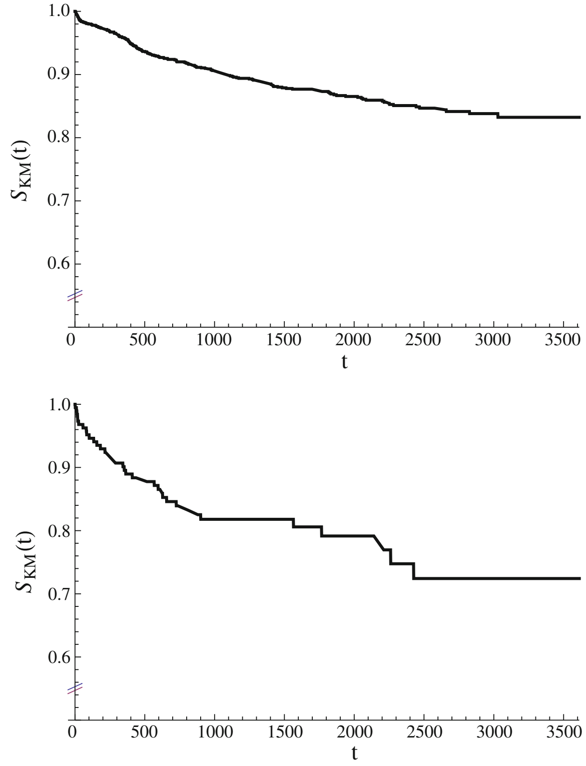


Fig. 1 Kaplan-Meier estimates of the survival functions for the first (*upper plot*) and the second stages (*lower plot*)

4.2 Results

Although in some applications it might be possible to say that a certain covariate logically cannot affect one or the other of the drift and starting level parameters of the process [22], that was not the case here. Consequently all six covariates listed above (five variables but with place of residence represented by two dummy variables) were associated with both the parameters. Thus the vectors $\mathbf{u}_i'$ and $\mathbf{v}_i'$ were identical. We also included in both vectors the number of previous stages (applications for treatment) and the total time t_{m-1} under observation until the beginning of the present stage. Consequently, including the constant terms in the linear predictors, fitting the model requires the estimation of 19 parameters (9 for the drift, 9 for the starting level and λ, the inverse of the variance of b_{mi}) The regression structures are as follows:

$$\mu_{mi} = \sum_{k=1}^{m} b_{ki} + \beta_0 + \beta_1 \text{Athens}_i + \beta_2 \text{ Thess.}_i + \beta_3 \text{Previous}_i + \beta_4 \text{ Male}_i \\ + \beta_5 \text{ Age}_i + \beta_6 \text{Year}_i + \beta_{ns}(m-1) + \beta_s t_{(m-1)i}.$$

and

$$x_{mi} \sim N_{+}\left(\tilde{\mathbf{u}}'_{mi}\boldsymbol{\alpha}^{*}, \tau^{-1}\right)$$

where

$$\begin{aligned}\tilde{\mathbf{u}}'_{mi}\boldsymbol{\alpha}^{*} &= \alpha_0 + \alpha_1 \text{ Athens}_i + \alpha_2 \text{ Thess.}_i + \alpha_3 \text{ Previous}_i + \alpha_4 \text{ Male}_i \\ &\quad + \alpha_5 \text{ Age}_i + \alpha_6 \text{ Year}_i + \alpha_{ns}(m-1) + \alpha_s t_{(m-1)i}\end{aligned}$$

We assigned flat priors to the regression parameters with

$$\begin{aligned}\alpha^{*}_{prior} &= (1, 0, 0, 0, 0, 0, 0, 0, 0), \quad \boldsymbol{\Sigma}_{\alpha^{*}} = 10^4 \boldsymbol{I}_9 \\ \beta^{*}_{prior} &= (1, 0, 0, 0, 0, 0, 0, 0, 0), \quad \boldsymbol{\Sigma}_{\beta^{*}} = 10^4 \boldsymbol{I}_9 \\ \lambda_{prior} &\sim Gamma(2, 1/2)\end{aligned}$$

where the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ correspond to the covariates in the order listed in Tables 1 and 2. These tables, and Table 3 for the parameter λ, present descriptive statistics for the posterior distribution of the parameters obtained from running the MCMC algorithm for a total of 110,000 iterations with the first 30,000 iterations discarded as burn-in. Figures 2, 3 and 4 show the trace (with the ergodic mean superimposed) and a smooth kernel estimate of the density of the posterior distribution for a representative selection of parameters, with the mean indicated.

The MCMC estimates lead to some clear conclusions regarding the latent process that describes the psychological and physical status of the drug users. From the descriptive statistics for the posterior distributions of the parameters of the model we can describe the behavior of the drug users regarding the starting point and the drift of their latent psychological and physical status until their next application for treatment.

The starting level of the latent processes does not appear to be affected by the variables Stages $(m - 1)$ and Time (t_{m-1}) that describe the previous stages of the process, because the posterior distributions of their coefficients are concentrated on zero (Table 1). Having had a treatment before 2001 also does not have a clear effect. In other words, the drug user's starting point in each stage seems to be independent of his or her treatment history. The remaining covariates all appear to have effects, and all with negative coefficients as shown by the high values of $\mathrm{P}(\alpha_i < 0|\mathrm{Data})$ for these covariates (Table 1). This means that a lower initial psychological and physical status is expected among residents of the major cities (Athens/Piraeus and Thessaloniki), males, older users and those who entered the database earlier. For example, it is clear that the older the drug user's age, the lower his or her starting point. All of these attributes tend to lead to a shorter time before the next application for treatment.

Concerning the drift, the positive signs of the posterior means of nearly all the coefficients (Table 2) indicate that $\mu_m > 0$ for the vast majority of the individuals, if not all. This is consistent with the presence of a "cured fraction" in the population

Table 1 Descriptive statistics for the posterior distributions of the regression parameters $\boldsymbol{\alpha}$ (initial state) of the model fitted to the data on drug users

	Mean	Mode	$q_{0.025}$	Q1	Median	Q3	$q_{0.975}$	P($\alpha_i < 0$|Data)
Constant	30.148	30.930	23.793	27.736	30.129	32.384	37.623	<0.0001
Athens	−6.598	−5.812	−11.790	−7.895	−6.347	−5.079	−2.510	>0.9999
Thess.	−10.168	−9.710	−15.331	−11.518	−9.944	−8.674	−5.852	>0.9999
Previous	−2.180	−1.625	−5.190	−3.213	−2.086	−1.140	0.556	0.933
Male	−4.023	−3.435	−7.445	−5.142	−3.874	−2.895	−0.894	0.994
Age	−0.103	−0.105	−0.187	−0.131	−0.103	−0.074	−0.019	0.991
Year	−0.607	−0.623	−0.909	−0.720	−0.612	−0.503	−0.260	>0.9999
Stages	−0.476	−0.408	−2.404	−1.176	−0.501	0.211	1.588	0.683
Time	0.0014	0.0008	−0.0018	0.0002	0.0014	0.0027	0.0050	0.215

Previous = treatment before 2001; Stages = number of previous stages; Time = total time before this stage

Table 2 Descriptive statistics for the posterior distributions of the regression parameters $\boldsymbol{\beta}$ (drift) of the model fitted to the data on drug users

	Mean	Mode	$q_{0.025}$	Q1	Median	Q3	$q_{0.975}$	$P(\beta_i < 0 \mid \text{Data})$
Constant	0.049	0.051	−0.020	0.024	0.048	0.073	0.124	0.087
Athens	0.039	0.039	−0.0040	0.025	0.039	0.053	0.080	0.036
Thess.	0.006	0.005	−0.044	−0.009	0.006	0.021	0.050	0.392
Previous	0.036	0.038	−0.0002	0.024	0.036	0.048	0.072	0.026
Male	−0.014	−0.016	−0.046	−0.024	−0.014	−0.003	0.017	0.803
Age	0.0013	0.0013	-2.2×10^{-5}	8.0×10^{-4}	0.0013	0.0017	0.0027	0.027
Year	0.019	0.019	0.014	0.017	0.019	0.021	0.026	<0.0001
Stages	0.141	0.142	0.099	0.126	0.141	0.155	0.185	<0.0001
Time	1.4×10^{-5}	1.1×10^{-5}	-3.3×10^{-5}	-2.7×10^{-6}	1.3×10^{-5}	3.0×10^{-5}	6.4×10^{-5}	0.288

Table 3 Descriptive statistics for the posterior distributions of the λ of the model fitted to the data on drug users

	Mean	Mode	$q_{0.025}$	Q1	Median	Q3	$q_{0.975}$
λ	62.273	60.600	35.769	56.781	61.840	67.589	100.133

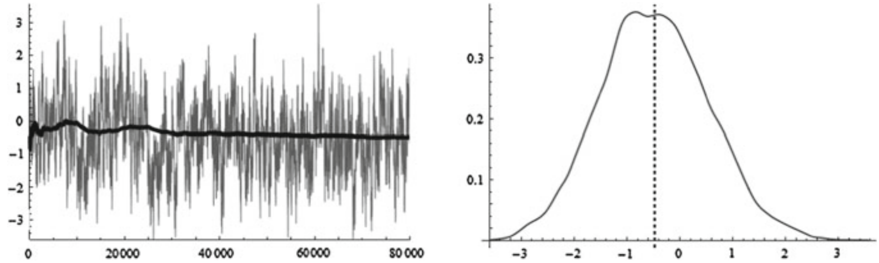


Fig. 2 The trace (*left plot*) with the ergodic mean superimposed (*heavy black line*) and a smooth kernel estimate of the density of the posterior distribution (*right plot*, with mean indicated by *dashed line*) of α_7 (coefficient of $m - 1$) in the model fitted to the data on drug users

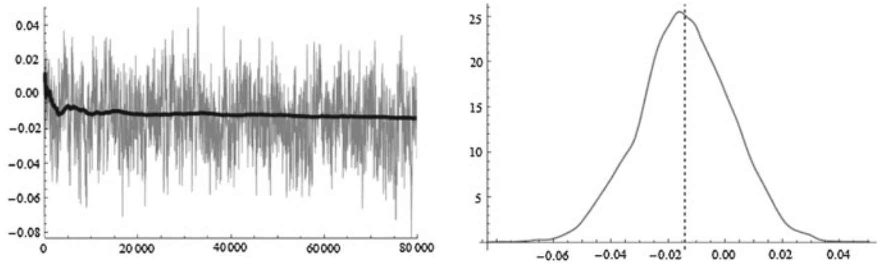


Fig. 3 The trace (*left plot*) with the ergodic mean superimposed (*heavy black line*) and a smooth kernel estimate of the density of the posterior distribution (*right plot*, with mean indicated by *dashed line*) of β_4 (coefficient of age) in the model fitted to the data on drug users

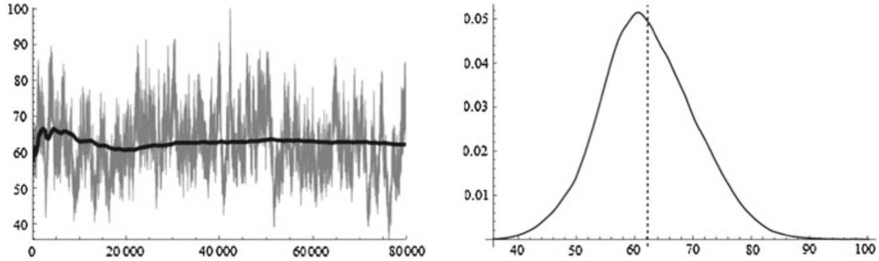


Fig. 4 The trace (*left plot*) with the ergodic mean superimposed (*heavy black line*) and a smooth kernel estimate of the density of the posterior distribution (*right plot*, with mean indicated by *dashed line*) of λ (inverse of variance of random effects b_{mi}) in the model fitted to the data on drug users

as observed in Fig. 1. As with the starting level, the variable t_{m-1} does not seem to have a significant effect. The same holds for the variables Thessaloniki and Gender, because the posterior distributions of the corresponding parameters are concentrated on zero. In contrast to the starting level, the treatment history does play a role in the drift, in the shape of both having been treated before 2001 and especially the variable Stages ($m - 1$; this reflects the difference between stages seen in Fig. 1). Because the posterior distributions of their coefficients are concentrated on the positive axis, more previous involvement with treatment increases the probability of no event, that is, of not returning to treatment. This is logical—otherwise, treatment would appear to be ineffective. It is clear that the random effects b_{mi} in the drift have little or no impact because their variance is essentially zero (Table 3). Their possible omission from the model will be investigated in the following subsection.

4.3 Model Selection

The results in the previous subsection clearly indicate that the covariate Time (t_{m-1}) is not associated with either the starting level or the drift of the latent process. Furthermore, the random effects b_{mi} seem to play no role in the drift. These findings suggest that we should compare the following four models to find the best one among them: the full model including all 19 parameters (Model 1, fitted in the previous Sect. 4.2), the model without the individual random effects for the drift (Model 2), the model without the covariate t_{m-1} (Model 3) and the model from which both t_{m-1} and the individual random effects for the drift are omitted (Model 4).

The comparison between these models was based on the criteria presented in Sect. 2.3. In order to specify the models without the individual random effects for the drift, all b_{mi}'s were set equal to zero and the prior distribution of λ was deleted from the likelihood. Priors for the remaining parameters took the same form as those in the full Model 1, but centred at the posterior means presented in Tables 1 and 2. All models were fitted by running the MCMC algorithm taking initial values at random from these priors. This procedure should provide good starting values. Thus we ran the algorithm for each model for 11,000 iterations, discarding the first 1,000 iterations as burn-in.

Based on Table 4, the preferred model is clearly Model 4 which omits the covariate t_{m-1} and the individual random effects for the drift, because this is the one with lowest values of the criteria. The omission of the random effects b_{mi} means that the

Table 4 Model selection: values of the DIC and IC criteria for the four models under consideration

	Model 1	Model 2	Model 3	Model 4
DIC	390131.26	310584.94	351011.34	276316.44
IC	388157.37	311379.61	352351.85	277156.06

Table 5 Descriptive statistics for the posterior distributions of the regression parameters α (initial state) in the preferred Model 4

	Mean	Mode	$q_{0.025}$	Q1	Median	Q3	$q_{0.975}$	P($\alpha_i < 0$|Data)
Constant	30.126	28.919	23.959	28.173	29.581	31.662	40.618	<0.0001
Athens	−5.446	−5.562	−10.725	−6.350	−5.405	−4.323	−1.410	>0.9999
Thess.	−11.005	−10.954	−15.871	−11.849	−10.970	−10.106	−6.540	>0.9999
Previous	−0.897	−0.514	−5.425	−1.700	−0.808	0.041	2.666	0.740
Male	−3.469	−3.121	−6.900	−4.400	−3.305	−2.560	0.282	0.9979
Age	−0.082	−0.068	−0.190	−0.110	−0.079	−0.055	0.016	0.9932
Year	−0.487	−0.545	−0.901	−0.585	−0.507	−0.392	0.004	0.9998
Stages	−0.479	−0.373	−2.234	−0.883	−0.485	−0.109	1.165	0.801

Table 6 Descriptive statistics for the posterior distributions of the regression parameters β (drift) in the preferred Model 4

	Mean	Mode	$q_{0.025}$	Q1	Median	Q3	$q_{0.975}$	$P(\beta_i < 0 \mid \text{Data})$
Constant	−0.00032	−0.00022	−0.03686	−0.00639	−0.00040	0.00579	0.03396	0.518
Athens	0.01473	0.01568	−0.00425	0.01129	0.01487	0.01832	0.03569	0.0025
Thess.	0.01087	0.01060	−0.00918	0.00719	0.01097	0.01458	0.02902	0.025
Previous	0.00745	0.00728	−0.00877	0.00467	0.00744	0.01028	0.02238	0.038
Male	−0.00015	0.00027	−0.01557	−0.00295	−0.00006	0.00275	0.01295	0.506
Age	0.00045	0.00045	−0.00030	0.00033	0.00045	0.00057	0.00104	0.0075
Year	0.00784	0.00778	0.00475	0.00728	0.00785	0.00843	0.01102	<0.0001
Stages	−0.00140	−0.00143	−0.01076	−0.00323	−0.00147	0.00040	0.00915	0.704

heterogeneity in drift between individuals is fully accounted for by the measured covariates. Furthermore, the length of the previous stages does not affect the process in any way.

Tables 5 and 6 present descriptive statistics for the posterior distributions of the regression parameters of the preferred Model 4. Compared with the results of the full model in Tables 1 and 2 the only significant changes to the posterior distributions of the parameters concern the ranges of the distributions of the coefficients for the covariates Previous (α_3) and Year (α_6) for the starting value and for the covariates Thess. (β_2) and Stages (β_7) for the drift of the process. The posterior distributions of the parameters α_6 and β_2 now seem to be concentrated on zero, indicating that these covariates may not be really associated with both the starting level and the drift of the latent process.

5 Conclusions

In this paper the extension of FHT regression model to recurrent events proposed by Economou et al. [10] is further extended to allow censoring to occur at every intermediate stage and is applied to the study of the psychological and physical health of one category of drug users in Greece. This approach offers a flexible and easy to interpret model describing the health status of these people. For instance, a poorer initial health status is expected for males, residents of the large cities, older users and those who entered the database earlier. The drift parameter of the stochastic model of their psychological and physical health turns out to be positive for the majority of the drug users. This implies that for the majority of them there is a non-zero probability that they will never apply again for a new treatment (the event of interest). This is an especially attractive property of models of this type because, in this way, they incorporate naturally a cured fraction (long-term survivors). These long-term survivors arise because the corresponding underlying stochastic process failed to reach the boundary, which in the present application can be interpreted as implying that the drug user's state never deteriorated sufficiently to prompt a request for treatment. To sum up, the proposed model offers an interesting interpretation of the nature of the observed data that can be applied in many other reliability and survival studies. Nonetheless, further work is needed on various aspects of these models, including statistical inference, convergence diagnostics, and further evaluation of the model selection procedures used in this work or alternatives to them. Furthermore, introducing additional extensions, as we have done in allowing censoring to occur in every intermediate stage, will allow even wider application of the model in statistical data analysis.

References

1. Aalen, O.O., Gjessing, H.K.: Understanding the shape of the hazard rate: a process point of view. Stat. Sci. **16**, 1–22 (2001)
2. Ando, T.: Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika **94**, 443–458 (2007)
3. Ando, T.: Predictive Bayesian model selection. Am. J. Math. Manag. Sci. **31**, 13–38 (2011)
4. Balka, J., Desmond, A.F., McNicholas, P.D.: Review and implementation of cure models based on first hitting times for Wiener processes. Lifetime Data Anal. **15**, 147–176 (2009)
5. Barr, D.R., Sherrill, E.T.: Mean and variance of truncated normal distributions. Am. Stat. **53**, 357–361 (1999)
6. Berkson, J., Gage, R.P.: Survival curves for cancer patients following treatment. J. Am. Stat. Assoc. **47**, 501–515 (1952)
7. Boag, J.W.: Maximum likelihood estimates of the proportion of patients cured by cancer therapy. J. R. Stat. Soc., Ser. B **11**, 15–34 (1949)
8. Caroni, C., Economou, P.: A hidden competing risk model for censored observations. Braz. J. Probab. Stat. **28**, 333–352 (2014)
9. Cox, D.R., Miller, H.D.: The Theory of Stochastic Processes. Chapman and Hall, New York (1965)
10. Economou, P., Malefaki, S., Caroni, C.: A threshold regression model with random effects for recurrent events. Meth. Comput. Appl. Prob. (2015)
11. Griffiths, W.: A Gibbs' sampler for the parameters of a truncated multivariate normal distribution. Working Paper Series, 856. Department of Economics, The University of Melbourne, Melbourne (2002)
12. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**, 457–481 (1958)
13. Lawless, J.F.: Statistical Models and Methods for Lifetime Data, 2nd edn. Wiley, New York (2003)
14. Lee, M.-L.T., Garshick, E., Whitmore, G.A., Laden, F., Hart, J.: Assessing lung cancer risk to rail workers using a first hitting time regression model. Environmetrics **15**, 501–512 (2004)
15. Lee, M.-L.T., Whitmore, G.A.: First Hitting Time Models for Lifetime Data. In: Balakrishnan, N., Rao, C.R. (eds.) Handbook of Statistics, vol. 23, pp. 537–543. Elsevier, Amsterdam (2004)
16. Lee, M.-L.T., Whitmore, G.A.: Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat. Sci. **21**, 501–513 (2006)
17. Linhart, H., Zucchini, W.: Model Selection. Wiley, New York (1986)
18. Maller, R.A., Zhou, X.: Survival Analysis with Long-term Survivors. Wiley, Chichester (1992)
19. Pennell, M.L., Whitmore, G.A., Lee, M.-L.T.: Bayesian random-effects threshold regression with application to survival data with nonproportional hazards. Biostatistics **11**, 111–126 (2010)
20. Raftery, A.E.: Bayesian model selection in social research (with discussion by Andrew Gelman, Donald B. Rubin and Robert M. Hauser). In: Marsden, P.V. (ed.) Sociological Methodology, pp. 111–196. Blackwell, Oxford (1995)
21. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit (with discussion). J R Stat. Soc., Ser. B **64**, 583–639 (2002)
22. Stogiannis, D., Caroni, C.: Issues in fitting inverse Gaussian first hitting time regression models for lifetime data. Commun. Stat. - Simul. Comput. **42**, 1948–1960 (2013)
23. Stogiannis, D., Caroni, C., Anagnostopoulos, C.E., Toumpoulis, I.K.: Comparing first hitting time and proportional hazards regression models. J. Appl. Stat. **38**, 1483–1492 (2011)
24. Yakovlev, A.Y., Tsodikov, A.D., Bass, L.: A stochastic model of hormesis. Math. Biosci. **116**, 197–219 (1993)
25. Whitmore, G.A.: An inverse Gaussian model for labour turnover. J R Stat. Soc., Ser. A **142**, 468–478 (1979)

Acceleration, Due to Occupational Exposure, of Time to Onset of a Disease

A. Chambaz , D. Choudat and C. Huber-Carol

Abstract Occupational exposure to pollution may accelerate or even induce the onset of diseases. A pecuniary compensation, to be paid by the state or the company, is then due to the exposed workers. The computation of the amount of this compensation might be based on the so-called “expected number of years of disease-free life” lost by the workers due to their occupational exposure. In order to estimate this number of years, we propose a method based on the threshold regression model also known as first hitting time (FHT) model. This model was initially developed for cohort studies. As our motivating example is a case-control study conducted in France to evaluate the link between lung cancer occurrence and occupational exposure to asbestos, we define a FHT model, adapt it to case-control data, and finally derive, for each worker in the study, the estimated expected number of years of disease-free life lost due to their occupational exposure to asbestos.

Keywords Case-control data · Cross-validation · First-hitting time model · Occupational exposure

1 Introduction

Quality of life is a central concern in medicine. Thus, the challenge of defining and estimating the expected number of years of life free of some disease lost due to an occupational exposure frequently arises for the sake of characterizing

A. Chambaz (✉)
Modal’X, Université Paris Ouest Nanterre, 92001 Nanterre, France
e-mail: achambaz@u-paris10.fr

D. Choudat
Département de médecine du travail, Assistance Publique-Hôpitaux de Paris,
Université Paris Descartes, Sorbonne Paris Cité, 75014 Paris, France
e-mail: dominique.choudat@parisdescartes.fr

C. Huber-Carol
MAP5, Université Paris Descartes, Sorbonne Paris Cité, 75270 Paris Cedex 06, France
e-mail: catherine.huber@parisdescartes.fr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_5

the amount of a pecuniary compensation due to a worker, on a case-by-case basis. Our motivating example is a French case-control study on the occurrence of lung cancer for workers exposed to asbestos [8]. As our objective is not to evaluate the risk of developing the disease, the logistic model, which is typically used for case-control studies, would not be the proper choice here. Several other models, though, can be used in order to solve our problem. One of the simplest is the Cox model involving the occupational exposure as a covariate together with other risk factors that could also induce lung cancer, like family history of cancer and tobacco consumption. But we chose to adapt the threshold regression model initially developed for cohort studies [6] to the case control study. The FHT model was initially developed for cohort studies [6], so this means that we have to adapt the standard FHT model to case-control studies. This model allows us to deal with the occupational exposure as an accelerator of the time leading possibly to a quicker onset of the disease, while the other covariates are divided into two classes, depending on how they act on the time to onset: the genetic ones and the ones pertaining to lifestyle like tobacco consumption. The expected number of years lost due to professional exposure to asbestos is then derived from the model by replacing for each exposed subject his time to onset by the decelerated time they would have had when exposure is removed and all other factors in the model remain the same [1].

2 Motivation of the Choice of FHT Model

2.1 Preliminary Studies

We start with a preliminary non parametric study of the data. The Kaplan-Meier estimators of the survival functions of subsets of the data based on high or low occupational exposures may show a possible influence of the amount of occupational exposures and other factors.

Then, we could consider a Cox model that puts all covariates $Z = (Z_1, \dots, Z_k)$ including exposure covariates as well as personal biological and behavioral covariates, on the same level. The model reads

$$\lambda(t|z) = \lambda_0(t) \times \exp(<\theta, z>), \tag{1}$$

where $\lambda(t|z)$ is the incidence rate at time t of a subject whose covariates Z is equal to z, $\lambda_0(t)$ is a baseline incidence rate, and θ is a k-dimensional real parameter. Denoting $\Lambda_0(t) = \int_0^t \lambda_0(u)du$, the resulting survival function for a subject with covariates Z equal to z is then

$$S(t|Z = z) = \exp(\Lambda_0(t) \times \exp(<\theta, z>).$$

But, considering that, actually, the covariates are not all of the same kind, we choose a FHT model that enables us to separate the covariates into three different kinds based on their actions on the health status of the patient.

2.2 FHT Model

When estimating the influence of an occupational exposure on the onset of a disease, three different types of covariates are distinguished by considering how they act on (or account for), the decrease of the latent "amount of health". Specifically, the three types of covariates are as follows:

- the initial covariates which act on the initial amount of health of the patient, including genetic factors, gender and past family disease history;
- the lifestyle and biological covariates which act on (or account for) the "decrease" of the initial amount of health. They may include, for example, cholesterol level and tobacco consumption;
- the occupational exposure under study which may accelerate the time to onset of the considered disease.

The time T to occurrence of the disease is modeled by a stochastic process $X(t)$ which represents the amount of health of the subject at time t: the disease occurs when this amount of health hits the boundary 0 for the first time (hence the expression FHT). Let B be a Brownian motion. For any real numbers $h > 0$ and $\mu \leq 0$, the process X(t) is defined as:

$$X(t) = h + \mu t + B(t), \tag{2}$$

where h plays the role of an initial amount of health relative to the disease, and μ a rate of decay of the amount of health. The value of h depends on the initial covariates while the value of μ depends also on the personal lifestyle and biological covariates. Then

$$T(h, \mu) = \inf\{t \geq 0 : X(t) \leq 0\}, \tag{3}$$

the first time the drifted Brownian motion $X(t)$ hits 0. The distribution of $T(h, \mu)$ is known as the inverse Gaussian distribution with parameter (h, μ). It is characterized by its cumulative distribution function (cdf)

$$F(t|h, \mu) = 1 + e^{-2h\mu}\Phi\left((\mu t - h)t^{-1/2}\right) - \Phi\left((\mu t + h)t^{-1/2}\right), \tag{4}$$

where Φ is the standard normal cdf.

As $\mu \leq 0$, the drifted Brownian motion $X(t)$ will almost surely reach the boundary (i.e. $T(h, \mu) < \infty$). Therefore $T(h, \mu)$ is also characterized by its density

$$f(t|h, \mu) = \frac{h}{(2\pi t^3)^{1/2}} \exp\left(-\frac{(h - |\mu|t)^2}{2t}\right). \tag{5}$$

$T(h, \mu)$ has mean $h/|\mu|$ whenever $\mu < 0$.

The effect of occupational exposure is taken into account through an acceleration function R that is nondecreasing and continuous on $I\!R^+$ such that $R(t) \geq t$ for all t. The acceleration function R depends on the occupational exposure and, given R, we define

$$T(h, \mu, R) = \inf\{t \geq 0 : h + \mu R(t) + B(R(t)) \leq 0\}, \tag{6}$$

the first time the drifted Brownian motion $(X(B(R(t)))$ hits 0 along the modified time scale derived from R, so that the cdf of $T(h, \mu, R)$ at t is $F(R(t)|h, \mu)$, and its density at t is $R'(t) f(R(t)|h, \mu)$ as long as R is differentiable.

Conditional on $[T \geq x - 1]$, the survival function and density of T at $t \geq x - 1$ are respectively:

$$G(t|h, \mu, R) = \frac{1 - F(R(t)|h, \mu)}{1 - F(R(x-1)|h, \mu)}, \tag{7}$$

$$g(t|h, \mu, R) = \frac{R'(t) f(R(t)|h, \mu)}{1 - F(R(x-1)|h, \mu)}. \tag{8}$$

3 The Data Set

3.1 Description of the Data Set

The matched case-control study took place between 1999 and 2002 at four Parisian hospitals and consisted of $n = 1761$ patients, among which 860 were cases and 901 were controls. The non-occupational information on each patient comprised of six covariates, the hospital, $W_0 \in \{1, 2, 3, 4\}$, the gender $W_1 \in \{0, 1\}$ (0 for men, 1 for women), the occurrence of lung cancer in close family, $W_2 \in \{0, 1\}$, (1 for occurrence and 0 for no occurrence), the tobacco consumption $W_3 \in \{0, 1, 2, 3\}$ respectively for pack-year $\in \{0, [1\ ; 25], [26\ ; 45], > 45\}$, the age at interview $X(\tau)$ where τ is calendar time, the age at incidence of lung cancer T, with convention $T = \infty$ if no lung cancer occurred yet. The indicator of a case, equal to 1 for cases and 0 for controls, is thus

$$Y = 1\{T \leq X\}.$$

Matching was done based on hospital, gender and age at recruitment ±2.5 years. In the sequel, we denote

$$V = (W_0, W_1, X)$$

the matching variable.

The other items observed on the patients deal with informations on occupational exposure up to the time of interview. The occupational history up to age X is measured, for each patient, on each of their successive jobs they held, by its duration together with three indicators of the exposure to asbestos: its probability, frequency and intensity of exposure, each with 3 levels $(1, 2, 3)$. A probability index equal to $1, 2$ or 3 corresponds respectively to a passive exposure, a possible direct exposure or a very likely or certain direct exposure. A frequency index equal to $1, 2$ or 3 corresponds respectively to exposures occurring less than once a month, more than once a month and during less than half of the monthly working hours or during more than half of the monthly working hours. An intensity index equal to $1, 2$ or 3 corresponds respectively to a concentration of asbestos fibers less than 0.1 f/mL, between 0.1 and 1 f/mL and more than 1 f/mL. Adding a category $0 = (0, 0, 0)$ for no exposure at all, the set $\mathscr{E}$ of categories of exposure has $27 + 1 = 28$ elements $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$.

Amongst the 8432 jobs held by the participants, 7009 were without any significant exposure. But although this leaves 1423 jobs featuring a significant exposure, it can be seen in Table 1, which contains many 0's, that several profiles in $\mathscr{E}$ are not represented, which gives evidence of an over-parametrization.

Let $a_i(t)$ be the exposure of subject i at time t, $\widetilde{a_i}(t) = a_i|_0^t$ be the exposure from time 0 to time t of subject i and $\widetilde{a}$ his history of exposure along his lifetime up to the occurrence of cancer if he is a case or to the time of interview if he is a control. The function $\widetilde{a}$ is piecewise constant.

For example, let subject i be a patient who started their first job at the age of 20 and experienced an occupational exposure of $\varepsilon = (2, 1, 3)$ for the next 15 years. Then they started their second job at the age of 35 and for the next 10 years had an occupational exposure of $\varepsilon = (2, 2, 3)$. They are diagnosed with lung cancer at age

Table 1 Number of jobs for each possible "probability/frequency/intensity" description

Exposure	Count	Exposure	Count	Exposure	Count
111	213	211	53	311	138
112	167	212	6	312	105
113	3	213	6	313	24
121	150	221	5	321	136
122	46	222	3	322	189
123	3	223	3	323	22
131	0	231	2	331	1
132	0	232	0	332	3
133	0	233	0	333	0

$T_i = 45$ years, at which point they become a case. Then, for $0 < t < 20$, $a_i(t) = 0$, for $20 \leq t < 35$, $a_i(t) = (2, 1, 3)$, and for $t \geq 35$, $a_i(t) = (2, 2, 3)$; moreover $\widetilde{a}_i(30) = a_i|_0^{30}$, and $\widetilde{a}_i = a_i|_0^{45}$.

3.2 Specific Problems Due to the Data Set

Several problems arise due to the way the data set was collected:

First of all, the data set contained information pertaining to occupational exposures, like silica and aromatic hydrocarbons. However, the preliminary non parametric analysis using Kaplan-Meier estimates for sub-samples of the data set revealed that these two factors did not have much influence on the age at onset. Moreover, very few people were exposed to silica and/or aromatic hydrocarbons and in very small quantities. Thus we decided to restrict attention to asbestos.

Second, the actual matching pattern is not available. What is known is only on which covariates the pairing was done.instead of giving up on the matching, we choose to artificially determine a random matching pattern, based on the same covariates, and also make sure that our results are preserved when using several different valid patterns.

Third, the FHT model, initially developed for cohort data, has to be adapted to a case-control survey via a weighing of the log-likelihood.

Fourth and finally, the way the exposure to asbestos is defined is complex and leads to an over-parametrization of the current model. The exposure is defined, for each job of each worker, by four quantities: its duration, the probability, frequency, and intensity of exposure of the job and its duration. Each of the three first quantities has three levels (1, 2, 3) so that, including 0 for no exposure at all, this leads to 28 exposure levels. This great number of parameters has to be reduced in a sensible way that we explain in Sect. 4.2.

4 Data Analysis

4.1 Preliminary Studies

We first apply the non parametric Kaplan-Meier method to estimate the survival of four sub-samples having low/high tobacco consumption and low/high asbestos exposure. The cutting points are the respective medians. We obtain the corresponding survival functions for the time to onset of the lung cancer (see Fig. 1). We see from this figure that there is probably an impact on lung cancer occurrence of both tobacco consumption and asbestos exposure. Applying the same process to the other exposures present in the data set, like silica and aromatic hydrocarbons, gave no

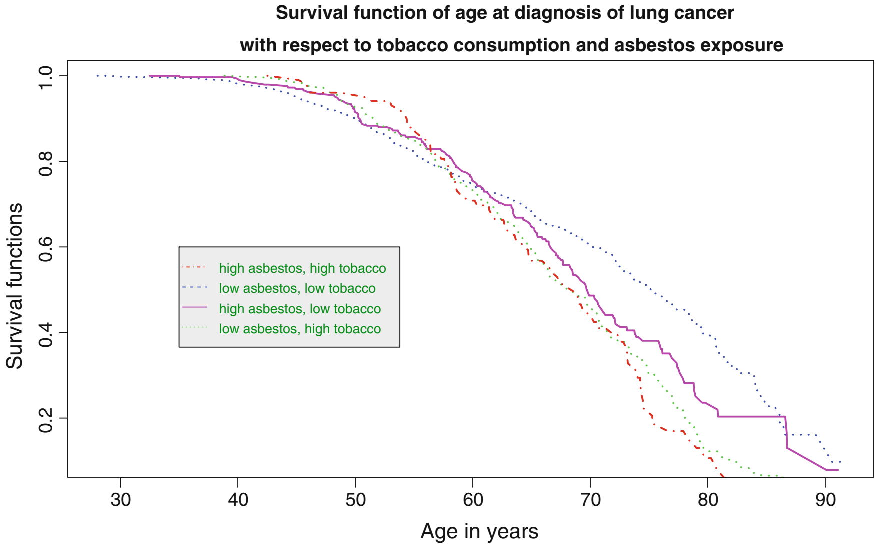


Fig. 1 Survival functions of age at diagnosis of lung cancer for high or low tobacco consumption and exposure to asbestos

evidence. This may be due to the fact that there are very few jobs featuring a significant exposure to silica or aromatic hydrocarbons.

A naive application of an FHT model to those data would consist in defining $\log(h)$ as a linear function of gender (W_1) and past family history of lung cancer (W_3), μ as a linear function of W_1, W_3 and also tobacco consumption (W_2) and the acceleration $R(t)$ as $\sum_{j=1}^{J} m_j \times a_j(t)$ for a patient having experienced J jobs, where m_j is the acceleration parameter attached to the category ε_j. The dimension of this naive model equals $3 + 4 + 28 = 35$, and the underlying assumption of linearity of $\log(h)$ and μ seems quite restrictive. In contrast, taking into account the excessive number of 0s in Table 1, we reduce the number of parameters m_j and let $\log(h)$ and depend in a most flexible way than the linear one on the observations W_1, W_2 and W_3. We build a more general though slightly more economic FHT model of dimension 27, and consider it as a maximal model containing simpler models among which we select, based on our data, a better model described in Sect. 4.4.

4.2 Acceleration Due to Occupational Exposure

Initially, for each job, the acceleration is a function of three variables, each of them having three values. We replace it by the product of three functions of one variable, $M_1(\varepsilon_1)$, $M_2(\varepsilon_2)$, $M_3(\varepsilon_3)$. Each of these three functions, M_1 for probability, M_2 for frequency and M_3 for intensity is assumed to be non negative, non decreasing and having 1 as maximum value. In this view, define

$$\mathscr{M} = \Big\{(M_0, (M_k(l))_{k,l \leq 3} \in I\!R^+ \times (I\!R^+)^{2\times 3}, \\ 0 \leq M_k(1) \leq M_k(2) \leq M_k(3) = 1, k = 1, 2, 3\Big\}. \quad (9)$$

Then the rate yielded by description $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathscr{E} \setminus \{0\}$ is expressed as

$$M(\varepsilon) = 1 + M_0 \times M_1(\varepsilon_1) \times M_2(\varepsilon_2) \times M_3(\varepsilon_3)$$

with convention $M(0) = 1$. Note that $M(0) = 1 \leq M(\varepsilon) \leq M(3, 3, 3) = 1 + M_0$. Exposure $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ can then be written as a fraction M_ε of the maximal acceleration, where

$$M_\varepsilon = \frac{M(\varepsilon) - 1}{M_0} = M_1(\varepsilon_1) \times M_2(\varepsilon_2) \times M_3(\varepsilon_3).$$

This parametrization is identifiable and reduces the number of parameters needed to associate every category of exposure with an acceleration rate, from 28 to 7.
Set $M \in \mathscr{M}$ and a generic longitudinal description $\widetilde{a}$ be as presented in Sect. 3. For convenience, we consider a continuous approximation to the piecewise constant function $t \mapsto M(\varepsilon(t))$, which we denote as $r(M, \widetilde{a})$ (see [1] for details). Then every pair $(M, \widetilde{a})$ thus gives rise to the nondecreasing and differentiable acceleration function

$$R(M, \widetilde{a})(t) = \int_0^t r(M, \widetilde{a})(s)ds \leq t. \quad (10)$$

4.3 The Case-Control Weighed Log-likelihood

Let us recapitulate the parametrization of our FHT model:

$$\begin{array}{lll} \log(h) & = \alpha(W_1, W_2) & \in I\!R^4, \\ \log(-\mu) & = \beta(W_1, W_2, W_3) & \in I\!R^{16}, \\ R & = R(M, \widetilde{a})(X) \;\;, & M \in \mathscr{M}, \\ \theta & = (\alpha, \beta, M) & \in \Theta. \end{array}$$

Recall that V and Y are respectively the matching variable and the case indicator, and define $Z = \min(T, X)$. We rely on weights whose characterization requires the prior knowledge of the joint probability of (V, Y) which implies the knowledge of the conditional probabilities

$$\begin{array}{rl} q_v^*(y) & := P(Y = y|V = v), \\ q_y(v) & := P(V = v|Y = y)). \end{array}$$

Recall the definitions of G and g from (7) and (8), and let case i be matched by J_i controls. Then the weighed log-likelihood is

$$\mathrm{loglik}(\theta) = \sum_{i=1}^{n} \left\{ q_1(V_i) \log g(Z_i|\theta) + q_0(V_i) \frac{1}{J_i} \sum_{j=1}^{J_i} \log G(Z_i|\theta) \right\}.$$

Asymptotic properties of the resulting estimators are derived in [1].

4.4 Model Selection by Cross Validation

The maximal model Θ gives rise to a collection of sub-models Θ_k obtained by adding constraints on the maximal parameter $\theta = (\alpha, \beta, M) \in \Theta$. We define a large collection $\{\Theta_k : k \in \mathscr{K}\}$ of sub-models of interest. Then we let the data select a better sub-model $\Theta_{\hat{k}}$ based on a multi-fold likelihood cross validation criterion. The sample is divided into ten sub-samples of equal size. The initialization goes like this: in turn, we exclude one of the ten sub-samples and use the nine others to estimate the parameter of the maximal model, then we compute the likelihood of the tenth sub-sample under the estimated value of the parameter. The average likelihood, L_0, is finally computed. We then consider all one-step sub-models obtained by excluding W_1 or W_2 from the parametrization of h and μ, or by putting additional constraints on M. We compute their cross-validated scores, say L_1, in the same manner as L_0 was computed. If one sub-model at least satisfies $L_1 - L_0 > g$ for some pre-specified $g > 0$, then the model yielding the largest increase is selected. The above process is repeated with the best sub-model in place of the maximal model, until no gain is observed or if the gain is no larger than $c \times g$ for some pre-specified c (Tables 2, 3 and 4).

The final model features that there are no constraints either on "frequency" or on "intensity", but on "probability", and the conclusion is that when $\varepsilon_1 = 1$, there is no effect of exposure and that there is no difference between $\varepsilon_1 = 2$ and $\varepsilon_1 = 3$.

4.5 Model Estimation

First, we fit the best model by maximum likelihood on the whole data set. Then, we derive confidence intervals through percentile bootstrap [2] with Bonferroni correction [3–5, 9, 10]: B = 1000, on 95 % of the sample repeatedly re-sampled from the 860 cases together with the corresponding controls.

Table 2 Confidence intervals for the initial health h as a function of gender W_1 (0 for men)

W_1	h	h_{min}	h_{max}
0	23.82	23.42	24.13
1	25.09	24.86	25.40

Table 3 Confidence intervals for drift $= -100\,\mu$ as a function of gender, W_1 and tobacco, W_3

W_3	$W_1 = 0$			$W_1 = 1$		
0	0.69	0.08	1.46	0.02	0.01	0.03
1	7.70	6.91	8.28	6.63	5.73	7.68
2	13.89	13.25	14.46	10.55	9.63	11.80
3	17.67	17.11	18.38	14.79	13.65	17.77

Table 4 Confidence intervals for acceleration parameters

$M_0 = 1.19$	CI = [0.34 2.00]	
$M_1(1) = 0$		
$M_1(2) = 0.97$	CI = [0.96 0.99]	$M_1(3) = 1$
$M_2(1) = M_2(2)$		
$M_2(2) = 0.93$	CI = [0.90 0.98]	$M_2(3) = 1$
$M_3(1) = 0.02$	CI = [0.00 0.09]	
$M_3(2) = 0.09$	CI = [0.00 0.27]	$M_3(3) = 1$

5 Conclusion

This method allows us to derive, for each patient, the expected number of years free of a disease due to occupational exposure in a simple way: once the model is estimated, the expected number of life free of lung cancer lost due to asbestos exposure of any patient i may be computed by decelerating his time T_i to onset of lung cancer by its estimated acceleration. The difference between the decelerated time and the observed time T_i is an estimation of the expected number of years free of lung cancer due to asbestos exposure. Denoting $\widehat{R_i}(t)$ the estimated acceleration function for subject i, the estimated expected number of years free of disease lost by him, denoted L_i, due to his occupational exposure is

$$L_i = \widehat{R_i}^{-1}(T_i) - T_i.$$

Examples of such values are given in Table 5.

Table 5 Six examples of expected number of years free of lung cancer lost due to occupational asbestos exposure

Sex	Age	Asbestos	Family	Tobacco	Years lost
0	65	228	0	1	3.1
0	57	125	0	1	2.5
0	60	25	0	1	2.7
1	41	36.0	0	1	1.6
0	66	24.0	1	1	3.0
1	61	78.0	0	0	3.4

Although Fig. 1 supports our choice to neglect a possible interaction between tobacco consumption and occupational exposure to asbestos, future research would profitably consist in enriching the maximal model to account for this possible interaction and letting the data decide whether the added complexity is worth keeping or not.

References

1. Chambaz, A., Choudat, D., Huber-Carol, C., Pairon, J.-C., van der Laan, M.J.: Analysis of the effect of occupational exposure to asbestos based on threshold regression modeling of case-control data. Biostatistics **15**(2), 327–340 (2014)
2. Hall, P.: The Bootstrap and Edgeworth Expansion. Springer Series in Statistics. Springer, New York (1992)
3. Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika **75**, 800–803 (1988)
4. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
5. Hommel, G.: A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika **75**, 383–386 (1988)
6. Lee, M.-L.T., Whitmore, G.A.: Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Stat. Sci. **21**(4), 501–513 (2006)
7. Lee, M.-L.T., Whitmore, G.A.: Proportional hazards and threshold regression: their theoretical and practical connections. Lifetime Data Anal. **16**(2), 196–214 (2010)
8. Pairon J.-C., Legal-Regis B., Ameille J., Brechot J.-M., Lebeau B., Valeyre D., Monnet I., Matrat M., and Chammings B., Housset S.: Occupational lung cancer: a multicentric case-control study in Paris area. European Respiratory Society, 19th Annual Congress. Vienna (2009)
9. Shaffer, J.P.: Multiple hypothesis testing. Annu. Rev. Psychol. **46**, 561–576 (1995)
10. Wright, S.P.: Adjusted P-values for simultaneous inference. Biometrics **48**, 1005–1013 (1992)

Transformations of Confidence Intervals for Risk Measures?

Karl-Ernst Biebler and Bernd Jäger

Abstract The relative risk and the odds ratio are discussed briefly in connection with their point estimations and confidence estimations. At an example is proved that the numeric conversion of the confidence limits of the odds ratio does not yield a confidence interval for the relative risk necessarily.

Keywords Relative risk · Odds ratio · Confidence estimation · Transformation of confidence bounds

1 Data Structures and Risk Measures

The simplest data structure at risk assessments is the 2×2 table resulting from two independent samples of binomially distributed random variables $X \sim B(n_1, p_1)$ and $Y \sim B(n_2, p_2)$, respectively. The sampling results are arranged as seen in Table 1.

Usual risk measures are the risk difference $RD = p_1 - p_2$, the relative risk $RR = p_1/p_2$, and the odds ratio $OR = (p_1/(1 - p_1))/(p_2/(1 - p_2))$. Their domains are $[-1, 1]$, $[0, \infty]$ and $[0, \infty]$, respectively. The given three risk measures are non-linear functions of each other. The different interpretations of these risk measures are not discussed here.One takes into account that at the comparison of two 2×2 tables with identical RD, the RR and OR as functions of p_1 and p_2 vary.

K.-E. Biebler (✉) · B. Jäger
Institute of Biometry and Medical Informatics,
Ernst-Moritz-Arndt-University of Greifswald, Greifswald, Germany
e-mail: kebiebler@web.de

B. Jäger
e-mail: bjaeger@biometrie.unigreifswald.de

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_6

Table 1 Arrangement of the data of two independent samples from binomially distributed random variables in a 2 × 2 table

Sample	1	2	Sum
Sample size	a	b	m_1
	c	d	m_2
	n_1	n_2	N

2 Asymptotic Confidence Intervals for *RR* and *OR*

Different estimation methods were developed and investigated for the calculation of confidence limits of RR and of OR. Expansions of the topic submitted to regard on different studies designs, stratification and adjustment. There is an extensive literature on this field. One may read more in monographs (e.g. [1, 6]) or in the help menus of software packages (e.g. SAS®).

Background of the most applied confidence estimation methods for the here observed sampling is the normal approximation of a binomial distribution. The confidence level is named with ε as usual. An asymptotic $(1-\varepsilon)$-confidence interval for the parameter p of a binomially distributed random variable X from a sample of size N with result $X = k$ is then

$$(\hat{p}_\ell, \hat{p}_u) = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}\ , \tag{1}$$

$\hat{p} = k/n$ the point estimate of the parameter and $z_{1-\alpha/2}$ the respective standard normal quantile for $\alpha = \varepsilon$.

So-called exact confidence limits $(\hat{p}_{\ell,ex}, \hat{p}_{u,ex})$ one obtains as follows: The lower $(\hat{p}_{\ell,ex})$ and the upper $(\hat{p}_{u,ex})$ bound of the confidence region is calculated respectively regarding $\varepsilon/2$ to given ε. $\hat{p}_{\ell,ex}$ is the maximal p with $\varepsilon/2 = \sum_{i=0}^{k} P(X=i)$ and $\hat{p}_{u,ex}$ is the minimal p with $\varepsilon/2 = \sum_{i=k}^{N} P(X=i)$, $P(X=i)$ the respective binomial probabilities. These tail probabilities are available via Beta functions also for large N in several software packages, e.g. MATHEMATCA® and SAS®.

A third method uses the logit transformation $\theta = f(p) = log[p/(1-p)]$ and the asymptotic normality of the estimator $\hat{\theta} = f(\hat{p}) = log[\hat{p}/(1-\hat{p})]$. One yields an asymptotic $(1-\varepsilon)$-confidence interval for the parameter p as

$$(\hat{p}_\ell, \hat{p}_u) = \left[\frac{e^{\hat{\theta}_\ell}}{1+e^{\hat{\theta}_\ell}}, \frac{e^{\hat{\theta}_u}}{1+e^{\hat{\theta}_u}}\right] \tag{2}$$

from $(\hat{\theta}_\ell, \hat{\theta}_u) = \hat{\theta} \pm z_{1-\alpha/2}\sqrt{\hat{\theta}(1-\hat{\theta})/N}$ and using the logistic function

$$p = f^{-1}(p) = \frac{e^{f(p)}}{1+e^{f(p)}} = \frac{e^{\theta}}{1+e^{\theta}}. \tag{3}$$

It is well-founded in the monotony of the logistic function that this way one gets a $(1-\varepsilon)$-confidence interval.

An asymptotic confidence estimation of RR due to the last mentioned method reads as follows (see e.g. [6]). Let be $\psi = log(RR)$ and $\widehat{RR} = an_2/bn_1$ the point estimate of RR. The variance of $log(RR)$ can be consistently estimated for $p_1 \neq p_2$ by $\hat{\sigma}_1^2 = \hat{V}[log(\widehat{RR})] = (1/a - 1/n_1 + 1/b - 1/n_2)$.

Then $(log(\widehat{RR}) - log(RR))/\hat{\sigma}_1$ is asymptotically standard normal distributed. $(\hat{\psi}_\ell, \hat{\psi}_u) = log(\widehat{RR}) \pm z_{1-\alpha/2}\hat{\sigma}_1$ is an approximate $(1-\varepsilon)$-confidence interval for ψ which results in the desired asymptotic $(1-\varepsilon)$-confidence interval for the relative risk RR,

$$(\widehat{RR}_\ell, \widehat{RR}_u) = (exp(\hat{\psi}_\ell), exp(\hat{\psi}_u)). \tag{4}$$

One uses $\hat{\sigma}_0^2 = (N/m_1 - 1)(N/n_1n_2)$ instead of $\hat{\sigma}_1^2$ in case $p_1 = p_2$.

An asymptotic $(1-\varepsilon)$-confidence interval for the odds ratio OR one obtains analogously as

$$(\widehat{OR}_\ell, \widehat{OR}_u) = (exp(\hat{\delta}_\ell), exp(\hat{\delta}_u)) \tag{5}$$

with $\delta = log(OR)$, $\widehat{OR} = ad/bc$ the point estimate of OR, $(\hat{\delta}_\ell, \hat{\delta}_u) = log(\widehat{OR}) \pm z_{1-\alpha/2}\hat{\sigma}_1$ and the variance estimator $\hat{\sigma}^2 = \hat{V}[log(\widehat{OR})] = (1/a + 1/b + 1/c + 1/d)$. This method dates back to 1955 (see [9]) and is explained in detail in several textbooks, e.g. [6].

3 Transformations of Confidence Intervals: An Instructive Example

Odds ratio and relative risk are very often used in the research to characterize different influence of an exposure or a treatment on the status of human beings. These risk measures have different meanings and this is sometimes misunderstood. Therefore one would like to express the more suitable relative risk as a function of $\widehat{OR}$. The authors of [10] convert $\widehat{OR}$ in $\widehat{RR}$ by the formula

$$\widehat{RR} = T(\widehat{OR}) = \widehat{OR} \Big/ \left[1 - \frac{c}{m_2} + \frac{c}{m_2} \cdot \widehat{OR}\right]. \tag{6}$$

They calculate a $(1-\varepsilon)$-confidence interval for the relative risk by this way converting the bounds of the $(1-\varepsilon)$-confidence interval for the odds ratio. This transformation was already proposed by Holland (see [5]). The bias of that method is explained in [4]. McNutt et al. [7] stated again that the use of this formula will deliver biased estimates. They report that the use of this transformation of $\widehat{OR}$ into $\widehat{RR}$ and also of the confidence bounds has gained increasing popularity in the field of medical research. Nevertheless, it was again explained in detail and applied without any quantitative characterization of the bias to an example in a methodological overview article of [8]. This is astonishing because the paper [7] is among the references in the last mentioned article.

The coverage probability of the interval $I_{RR} = (T(\widehat{OR_\ell}), T(\widehat{OR_u}))$ for the relative risk, which is obtained from the transformed bounds of the $(1 - \varepsilon)$-confidence interval $(\widehat{OR_\ell}, \widehat{OR_u})$ for the odds ratio, is investigated now. It will be characterized by means of simulations of each 10,000 runs with regard to the parameter OR. 10,000 tables each representing two independent binomial samples were randomly generated. The interval $(\widehat{OR_\ell}, \widehat{OR_u})$ and from that the interval $I_{RR} = (T(\widehat{OR_\ell}), T(\widehat{OR_u}))$ were calculated for each of the tables. Predefined were $n_1 = 400$, $n_2 = 500$, $p_1 = 1/3$ and $OR = 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2, 3, 4, 5, 6, 7, 8.5, 10, 12.5, 15$. The number K counts the cases $\widehat{RR} \notin I_{RR}$ in 10,000 runs. Provided that $K > 500$ is valid, I_{RR} is not a 0.95-confidence interval for the relative risk RR.

Figure 1 illustrates the estimated coverage probability of I_{RR} depending on OR. Obviously, I_{RR} is far from being a 0.95-confidence interval for the relative risk RR in general.

From the same tables, both the 0.95-confidence interval for the odds ratio according to formula (5) and the 0.95-confidence interval for the relative risk according to formula (4) were calculated. The numbers K of non-coverings sway around the level of 500 (see Fig. 2).

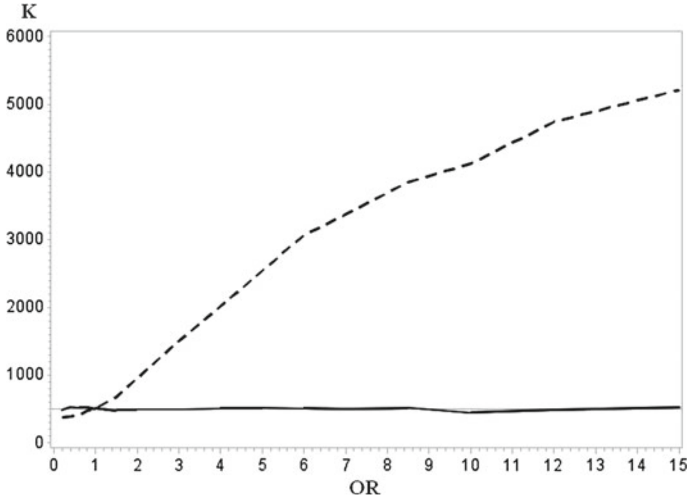


Fig. 1 Numbers K of non-coverings of the approximate 0.95—confidence interval for the odds ratio (*black line*) and of the interval $I_{RR} = (T(\widehat{OR_\ell}), T(\widehat{OR_u}))$ (*dashed line*) as functions of the odds ratio in each 10,000 runs. For the simulation parameters see text

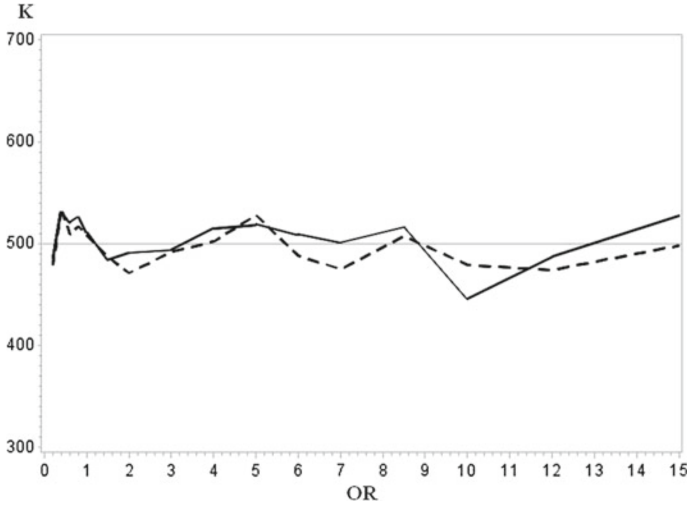


Fig. 2 Numbers K of non-coverings of the approximate 0.95-confidence interval for the odds ratio (*black line*) and of the approximate 0.95—confidence interval for the relative risk (*dashed line*) as functions of the odds ratio in each 10,000 runs. For the simulation parameters see text

4 Conclusion

Obviously, the transformations $T(\widehat{OR_\ell})$ and $T(\widehat{OR_u})$ of the bounds of the confidence interval for the odds ratio OR does not yield any confidence interval for the relative risk RR for OR different from 1. However, it is completely unproblematic to calculate both an approximate confidence interval for the odds ratio and an approximate confidence interval for the relative risk from the data of a cohort study.

There are also confidence estimation methods dealing with more complicated study designs. A model-based confidence interval estimation method mainly basing on variance estimations by means of resampling methods can be found in [3], for example.

References

1. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley-Interscience, Hoboken (2003)
2. Gart, J.J.: Approximate confidence limits for the relative risk. J. R. Stat. Soc. Ser. B **24**, 454–463 (1962)
3. Greenland, S.: Model-based estimation of relative risks and other epidemiologic measures in studies of common outcomes and in case-control studies. Am. J. Epidemiol. **160**, 301–305 (2004)
4. Greenland, S., Holland, P.W.: Estimating standardized risk differences from odds ratios. Biometrics **47**, 319–322 (1991)

5. Holland, P.W.: A note on the covariance of the Mantel-Haenszel log-odds-ratio estimator and the sample marginal rates. Biometrics **45**, 1009–1016 (1989)
6. Lachin, J.M.: Biostatistical Methods: The Assessment of Relative Risks. Wiley, New York (2000)
7. McNutt, L.A., Wu, C., Xue, X., et al.: Estimating the relative risk in cohort studies and clinical trials of common out-comes. Am. J. Epidemiol. **157**, 940–943 (2003)
8. Schmidt, C.O., Kohlmann, T.: When to use the odds ratio or the relative risk? Int. J. Public Health **53**, 165–167 (2008)
9. Woolf, B.: On estimating the relation between blood group and disease. Ann. Hum. Genet. **19**, 251–253 (1955)
10. Zhang, J., Yu, K.F.: What's a relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. JAMA **280**, 1690–1691 (1998)

Discrete Compound Tests and Dorfman's Methodology in the Presence of Misclassification

Rui Santos, João Paulo Martins and Miguel Felgueiras

Abstract Compound tests can be used to save resources for classification or estimation purposes in clinical trials and quality control. Nevertheless, the methodologies that are usually applied are restricted to qualitative group tests. Moreover, when quantitative compound tests are applied the problem is to ascertain whether the amount of some substance of any individual in the group is greater or lower than a prefixed threshold. An overview of the applications of the discrete compound tests highlights the advantages (to save resources) and disadvantages (higher probability of misclassification), and suggests criteria to assess the suitability of applying Dorfman's methodology.

Keywords Compound tests · Misclassification · Simulation · Quality measure

1 Introduction

Let p be the prevalence rate of some infection in a population with N individuals and the Bernoulli trials $X_i, i = 1, \ldots, N$, denote the presence ($X_i = 1$) and absence ($X_i = 0$) of the infection in the ith member of the population. Moreover, the random variables (r.v.s) $X_i \sim \text{Ber}(p)$ will be considered mutually independent due to random

R. Santos (✉) · J.P. Martins
School of Technology and Management, CEAUL — Center of Statistics and Its Applications,
Polytechnic Institute of Leiria, Campus 2, Morro do Lena – Alto do Vieiro, Apartado
4163, 2411–901 Leiria, Portugal
e-mail: rui.santos@ipleiria.pt

J.P. Martins
e-mail: jpmartins@ipleiria.pt

M. Felgueiras
School of Technology and Management, CEAUL — Center of Statistics and Its Applications,
CIIC— Computer Science and Communications Research Centre of Polytechnic Institute
of Leiria, Polytechnic Institute of Leiria, Campus 2, Morro do Lena – Alto do Vieiro, Apartado
4163, 2411–901 Leiria, Portugal
e-mail: mfelg@ipleiria.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_7

sampling. Thus, the number $\mathrm{I}^{[n]}$ of infected members in a group of size n is a r.v. with binomial distribution, i.e. $\mathrm{I}^{[n]} \sim \mathrm{Bin}(n, p)$. In addition, suppose that the clinical trial for identification of the infected is carried out by counting the number of a certain type of bacteria in a milliliter of blood. If this number is greater than a given threshold t the individual is classified as infected, otherwise it is classified as not infected (the opposite inequalities could also be applied with the appropriate adaptations). Furthermore, let the number of bacteria in one ml of blood be given by the r.v. Y_i. If $X_i = 0$ (uninfected individual), then $Y_i = Y_i^- \sim \mathbf{D}_0(\boldsymbol{\theta_0})$ where $\mathbf{D}_0$ denotes some count distribution with support $\mathrm{S}_0 \subseteq \mathbb{N}_0$ and parameter vector $\boldsymbol{\theta_0}$. If $X_i = 1$ (infected individual), then $Y_i = Y_i^+ \sim \mathbf{D}_1(\boldsymbol{\theta_1})$ where $\mathbf{D}_1$ is some count distribution with support $\mathrm{S}_1 \subseteq \mathbb{N}_0$ and parameter vector $\boldsymbol{\theta_1}$. In fact, the r.v.s Y^- and Y^+ can have the same support (at least partly, i.e. $\mathrm{S}_0 \cap \mathrm{S}_1 = \mathrm{S} \neq \emptyset$). Furthermore, any classification methodology has a nonzero probability to return an erroneous classification for those individuals in S, thereby leading to the presence of misclassification in the individual test (if $\mathrm{S} = \emptyset$ then there would be no problem of misclassification). Nevertheless, Y_i^- and Y_i^+ must have some relationship to ensure that the probability of misclassification is small, for instance Y_i^+ shall stochastically dominate Y_i^- (i.e. $F_{Y^+}(x) \leq F_{Y^-}(x), \forall x \in \mathbb{R}$), otherwise the test would not have much information concerned with the infection. In some applications $\mathbf{D}_0$ and $\mathbf{D}_1$ denote the same distribution with a change of location and a possible alteration of scale. The usual applied distributions are the Poisson, the negative binomial, and the binomial distributions. The previously illustrated context will be used throughout this work in the description of the proposed methodologies. However, the applicability of these methods is not restricted to blood analysis, such as the screening of infectious diseases like HIV (see [25, 26]), since it can easily be adjusted to be applied to any fluid that can be mixed, e.g. in industrial quality control cf. [3, 9].

An overview of the applications of the discrete compound tests will be given in this section. We will describe the goals (estimation and classification) and implementation of the compound tests as well as the usual main measures applied to evaluate the accuracy of the test results. Hereafter, a new evaluation measure of the application of compound tests in the presence of misclassification is proposed in order to identify the situations in which the compound test can be applied without an excessive probability of misclassification (Sect. 2). Section 3 provides a detailed description of two different methodologies to perform the compound tests and, therefore, to set up its cut off points. In Sect. 4 the performance of the two proposed methodologies is investigated via simulation. Finally, the main conclusions are outlined in Sect. 5.

1.1 Group/Compound Tests

In an individual test we are performing the statistical hypothesis test formalized by $\mathbf{H_0} : X_i = 0$ versus $\mathbf{H_1} : X_i = 1$. If the cut off point is the threshold t then the significance level is $\alpha = \mathrm{P}(Y_i > t |\, X_i = 0) = \mathrm{P}\left(Y_i^- > t\right)$ and the power of the test

$1-\beta = \mathrm{P}\left(Y_i > t \mid X_i = 1\right) = \mathrm{P}\left(Y_i^+ > t\right)$. To perform a compound test, one ml of blood is taken from each of the n members of the group and then mixed. Thus, we get n ml of pooled blood where the number of bacteria is given by $B_n = \sum_{i=1}^n Y_i$. After being turned into a totally homogeneous fluid, one ml of this fluid is withdrawn for testing. The number B_1 of bacteria in this ml of fluid can be computed using hierarchical models, more precisely by the application of a binomial filter, taking into account that each of the B_n bacteria in the n ml of blood has probability n^{-1} of being drawn to the chosen mixed ml. Therefore, $B_1 \sim \mathrm{Bin}\left(B_n, \frac{1}{n}\right)$, cf. [21]. This ml of pooled blood will be used to perform the compound test. The main idea of this test is to identify if there are any infected member in the group. Thus, if the test results negative then all member in the group are uninfected. Otherwise, at least one member of the group must be infected. Therefore, the statistical hypothesis test to perform is $\mathbf{H_0} : \sum_{i=1}^n X_i = 0$ versus $\mathbf{H_1} : \sum_{i=1}^n X_i \geq 1$. In fact, it aims to identify if any of the individuals in the group ($i = 1, \ldots, n$) can be classified as infected, i.e. if $\max\left(Y_1, \ldots, Y_n\right) > t$, using the only available information B_1—the number of bacteria in the ml of pooled blood. The significance level of the compound test is $\alpha = \mathrm{P}\left(B_i > t^* \mid \sum_{i=1}^n X_i = 0\right)$ and the power of the test $1-\beta = \mathrm{P}\left(B_i > t^* \mid \sum_{i=1}^n X_i \geq 1\right)$, where t^* denotes the applied cut off point of the compound test.

1.2 Classification and Prevalence Rate Estimation

The use of compound test has two main goals: classification (categorization of each individual as infected or not infected) and the estimation of the prevalence rate p. Dorfman's methodology was first used in the second World War in order to identify all American soldiers infected with syphilis, cf. [5]. In this classification methodology the whole population is divided into groups with n individuals and a compound test is performed within each group. As outlined in the previous subsection, if the result is negative then all members are classified as not infected. Otherwise at least one member should be classified as infected, and individual tests are performed in order to identify which elements are indeed infected. The main goal is to compute the optimal dimension n (as a function of the prevalence rate p) in order to minimize the expected number of tests required to identify all infected individuals. Whereas in individual tests N tests are required to classify all N individuals in the population, in Dorfman's methodology the expected number of tests is $\mathrm{E}[T_n] = N\left(\frac{1}{n} + 1 - (1-p)^n\right)$. Hence, the relative cost (expected number of test for the classification of each individual) is $\mathrm{RC} = \frac{n+1}{n} - (1-p)^n$, $n \geq 2$, which can be used to establish the optimal size n for each prevalence rate p. If $p \leq 0.3$ then $\mathrm{RC} \leq 1$ and, therefore, is better to use the Dorfman's methodology than individual tests. The optimal size n for some prevalence rates p can be found in [5, 21]. For $p \leq 0.12$ a linear approximation can be used and the optimal group size is approximately $n \approx \frac{1}{\sqrt{p}} + 0.5$ with good results, cf. [6]. In what follows we assume that the cost of mixing samples is negligible [16] and therefore the number of required tests is the only relevant cost to be taken into account in the classification procedure.

The first few classification methodologies considered the absence of misclassification and were restricted to qualitative group tests (identification of the presence or absence of some substance in the compound fluid). Subsequently, new algorithms have been proposed in order to minimize the number of tests required for the correct classification of all individuals in the population, mainly applying halving nested procedures or hierarchical algorithms (generalizations of the Dorfman's methodology in which positive groups are repeatedly divided into smaller non-overlapping subgroups until all members have been individually tested, cf. [6, 8, 12, 15, 23, 24]), square array testing (with the use of overlapping pools, cf. [13, 19, 28]), and multidimensional array algorithms (for an extension to higher dimensional arrays, cf. [1, 20]).

In addition, the use of compound tests can also be useful in the estimation of the prevalence rate p, cf. [22]. Under certain conditions, these estimators have better performance than the estimators based on individual tests, allowing the reduction of the number of performed tests and simultaneously to achieve more accurate estimates with respect to the bias, efficiency as well as robustness, cf. [4, 7, 10, 14, 17, 22] among others. Some packages with applications of several compound testing estimators, such as *binGroup* for the R software [2], are available.

Nevertheless, in both cases (classification and estimation) the use of compound tests should only be performed for low prevalence rates. In this work the compound tests will be used for classification purposes applying the Dorfman's methodology.

1.3 Misclassification Evaluation

The usual applied measure to evaluate the misclassification problem are the specificity, the sensitivity, the positive predictive value, and the negative predictive value. These measures can be defined for the individual test, for the compound test, and for the application of a specific classification methodology.

1.3.1 Misclassification in Individual Tests

Performing individual tests, the individual specificity is the probability of getting a negative result (X_i^-) from a not infected individual, i.e. $\varphi_e = \mathrm{P}\left(X_i^- \mid X_i = 0\right)$, and the individual sensitivity is the probability of getting a positive result (X_i^+) from an infected individual, i.e. $\varphi_s = \mathrm{P}\left(X_i^+ \mid X_i = 1\right)$. The positive predictive value is the probability of having an infected sample in a positive individual test, i.e. $\mathrm{PPV} = \mathrm{P}\left(X_i = 1 \mid X_i^+\right)$, and the negative predictive value is the probability of having an uninfected sample in a negative individual test, i.e. $\mathrm{NPV} = \mathrm{P}\left(X_i = 0 \mid X_i^-\right)$.

1.3.2 Misclassification in Compound Tests

Let us now consider compound tests performed to groups of size n, and let $X^{[+,n]}$ and $X^{[-,n]}$ denote respectively a positive and a negative compound result. Hence, the compound specificity is given by $\varphi_e^{[n]} = \mathrm{P}(X^{[-,n]}|\ \mathrm{I}^{[n]} = 0)$, and the compound sensitivity by $\varphi_s^{[n]} = \mathrm{P}(X^{[+,n]}|\ \mathrm{I}^{[n]} \geq 1)$. Nevertheless, $\varphi_s^{[n]}$ depends on the number of infected members in the group due to the dilution and consequent rarefaction of the number of bacteria. Thus, setting $\varphi_s^{[j,n]} = \mathrm{P}(X^{[+,n]}|\ \mathrm{I}^{[n]} = j)$, the rarefaction factor can be added in the $\varphi_s^{[n]}$ computation, by doing $\varphi_s^{[n]} = \sum_{j=1}^{n} \varphi_s^{[j,n]}\ \mathrm{P}(\mathrm{I}^{[n]} = j|\ \mathrm{I}^{[n]} \geq 1)$, cf. [21]. Moreover, $\varphi_s^{[n]} \approx \varphi_s^{[1,n]}$ for low prevalence rates (see [21]). Similarly, the compound positive predictive value is $\mathrm{PPV}^{[n]} = \mathrm{P}(\sum_{i=1}^{n} X_i \geq 1|\ X^{[+,n]})$ and the compound negative predictive value $\mathrm{NPV}^{[n]} = \mathrm{P}(\sum_{i=1}^{n} X_i = 0|\ X^{[-,n]})$. Generally, the compound sensitivity decreases as the group size increases, due to dilution. In the literature there are different procedures to model the dilution factor, cf. [11, 21, 27, 29]. The selection of the most suitable procedure for group testing depends on the sensitivity, on the specificity, and on the cost involved [16].

1.3.3 Misclassification in Classification Methodologies

The misclassification measures previously defined can be generalized in order to measure the misclassification in some classification methodology $\mathscr{M}$. Thus, the same definitions are applied as in the individual tests, but the probabilities are computed taking into consideration the application of the methodology under investigation. Hence, the $\mathscr{M}$ methodology specificity is the probability of an uninfected individual being classified as uninfected by the application of methodology $\mathscr{M}$. The Dorfman's methodology specificity is given by (cf. [21])

$$\varphi_{e_n} = \mathrm{P}\left(X_i^-|\ X_i = 0\right) = \sum\nolimits_{i=0}^{n-1} \mathrm{P}\left(X_1^-|\ X_1 = 0, \mathrm{I}^{[n-1]} = i\right) \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right), \tag{1}$$

and, analogously, the Dorfman's methodology sensitivity is given by

$$\begin{aligned} \varphi_{s_n} &= \mathrm{P}\left(X_1^+|\ X_1 = 1\right) = \sum_{i=0}^{n-1} \mathrm{P}\left(X_1^+|\ X_1 = 1, \mathrm{I}^{[n-1]} = i\right) \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right) \\ &= \varphi_s \sum\nolimits_{i=0}^{n-1} \varphi_s^{[i+1,n]} \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right). \end{aligned} \tag{2}$$

Dorfman's positive predictive value $\mathrm{PPV}_{\mathscr{D}_n}$ and Dorfman's negative predictive value $\mathrm{NPV}_{\mathscr{D}_n}$ follow using Bayes' inversion formula.

2 A Proposal to Measure the Quality of the Individual Test

In carrying out individual tests, if the threshold t increases then the specificity increases and the sensitivity decreases. Hence, if we improve the specificity (sensitivity) then the sensitivity (specificity) gets worse, as it happens with the probabilities of error type I and type II in a statistical hypothesis testing. Hence, appropriate tuning of the threshold is the key to achieve balance of specificity and sensitivity. The threshold t can be set to an equilibrium value t^e in order to equalize the sensitivity to the specificity (or, if not possible, to minimize their difference). The value of these probabilities (sensitivity and specificity), denoted by ϕ, defines a measure of the quality of the individual test performance.

Definition 1 In the individual test, the probability ϕ which verifies $\varphi_s = \varphi_e = \phi$ for some threshold t^e is the quality measure of the individual test performance. If this value does not exist (e.g. in the use of count distributions), the distance between φ_s and φ_e shall be minimized and $\phi = \frac{\varphi_s+\varphi_e}{2}$.

A high value of ϕ (in the neighborhood of the unit, $\phi \approx 1$) implies that the hypothesis test has a low probability of misclassification. Otherwise, if ϕ is much lower than 1, it implies a high probability of misclassification.

In fact, some information is lost when bloods are mixed. Hence, the compound tests should not be applied in cases in which ϕ is quite low, because the mixture will still further increase the probability of misclassification. In consequence, compound tests should be applied only if the individual tests have a good performance. Therefore, in Sect. 4 this quality measure of the individual test performance will be used in simulations in order to assess if it can be also used to measure the adequacy of the application of compound tests.

3 Methodologies for the Compound Tests

In this section two different methodologies to perform the compound tests are described. As in statistical hypothesis tests, it is impossible to improve the two probabilities of misclassification and therefore only one can actually be controlled. Thus, the goal of each of the applied methodologies is to control either sensitivity or specificity. The more usual first methodology $\mathbf{M_1}$ controls the compound specificity, while the second methodology $\mathbf{M_2}$ fixes the compound sensitivity.

3.1 The Usual Methodology—$\mathbf{M_1}$

The usually applied hypothesis test (using methodology $\mathbf{M_1}$) is, cf. [21],

$$\mathbf{H_0}: \sum\nolimits_{i=1}^{n} X_i = 0 \ \left(\mathrm{I}^{[n]} = 0\right) \ \text{ versus } \ \mathbf{H_1}: \sum\nolimits_{i=1}^{n} X_i \geq 1 \ \left(\mathrm{I}^{[n]} \geq 1\right).$$

Hence, the test size is given by $\alpha = \mathrm{P}(X^{[+,n]} | \sum_{i=1}^{n} X_i = 0) = 1 - \varphi_e^{[n]}$ and, therefore, the compound specificity is fixed. Thus, the specificity is controlled by setting the value of α, but it neglects the sensitivity, i.e. the occurrence of false negatives. Moreover, by (1) φ_{e_n} is equal to (see [21])

$$\mathrm{P}\left(\mathrm{I}^{[n-1]} = 0\right)\left[\varphi_e^{[n]} + \left(1 - \varphi_e^{[n]}\right)\varphi_e\right] + \sum\nolimits_{i=1}^{n-1} \mathrm{P}\left(\mathrm{I}^{[n-1]} = i\right)\left[\varphi_s^{[i,n]}\varphi_e + \left(1 - \varphi_s^{[i,n]}\right)\right]$$

and, therefore, $\varphi_{e_n} \geq \varphi_e^{[n]}$. Hence, the Dorfman's methodology specificity verifies $\varphi_{e_n} \geq 1 - \alpha$ and the size of test sets a lower limit for φ_{e_n}.

3.2 Alternative Methodologies—$\mathbf{M_2}$ *and* $\mathbf{M_2^\star}$

In most applications, it is essential to control the occurrence of false negative results (i.e. to set the sensitivity). With this goal we propose an alternative methodology $\mathbf{M_2}$, implemented by an hypothesis test

$$\mathbf{H_0}: \sum\nolimits_{i=1}^{n} X_i \geq 1 \ \left(\mathrm{I}^{[n]} \geq 1\right) \ \text{ versus } \ \mathbf{H_1}: \sum\nolimits_{i=1}^{n} X_i = 0 \ \left(\mathrm{I}^{[n]} = 0\right).$$

The test size α is given by $\alpha = \mathrm{P}(X^{[-,n]} | \sum_{i=1}^{n} X_i \geq 1) = 1 - \varphi_s^{[n]}$. Hence, the compound sensitivity is fixed and, therefore, the probability of false negative results is controlled. An obvious drawback of $\mathbf{M_2}$ when compared to $\mathbf{M_1}$ is the complexity of the cut off point computation due to the different scenarios in $\mathbf{H_0}$. In practice, a simplified methodology $\mathbf{M_2^\star}$ can be implemented in order to easily compute the cut off point, performing the following hypothesis test:

$$\mathbf{H_0}: \sum\nolimits_{i=1}^{n} X_i = 1 \ \left(\mathrm{I}^{[n]} = 1\right) \ \text{ versus } \ \mathbf{H_1}: \sum\nolimits_{i=1}^{n} X_i = 0 \ \left(\mathrm{I}^{[n]} = 0\right).$$

The results of applying this simplified $\mathbf{M_2^\star}$ are quite similar to $\mathbf{M_2}$ because the probability of getting more than one infected individual in the group is quite low (an obvious requirement for the sensible use of compound tests). On the other hand, in this simplified $\mathbf{M_2^\star}$ the significance level is given by $\alpha = \mathrm{P}(X^{[-,n]} | \sum_{i=1}^{n} X_i = 1) = 1 - \varphi_s^{[1,n]}$, and therefore α will set $\varphi_s^{[1,n]}$. In addition, as having just one infected individual in the group corresponds to the worst case scenario, then $\varphi_s^{[1,n]} \leq \varphi_s^{[2,n]} \leq \cdots \leq \varphi_s^{[n,n]}$, and consequently $\varphi_s^{[1,n]} \leq \varphi_s^{[n]}$. Thus, the α value will set up a lower limit for the compound sensitivity $\varphi_s^{[n]}$. Nevertheless, by (2) the Dorfman's sensitivity is lower than the compound sensitivity $\varphi_{s_n} \leq \varphi_s^{[n]}$. Hence, the Dorfman's sensitivity can be lower or higher than $1 - \alpha$. $\mathbf{M}_2^{\star}$ had already been proposed in [18] but without any examination, which will be carried out in the simulations performed in Sect. 4.

4 Simulation

The main goal of these simulations is to investigate the use of $\mathbf{M}_2^{\star}$, as well as the ensuing quality measure of the individual test performance ϕ.

4.1 Simulation Settings

All simulations were performed in software R using 10^6 groups in each simulation and applying different prevalence rates p, significance levels α, group dimensions n and the quality measure of the individual test performance ϕ. The case $n = 1$ corresponds to the restricted use of individual tests. For an infected individual the distribution $\mathbf{D}_1$ of Y_i^+ was defined through a change of location $Y_i^+ = \mu' + Y_i^-$ in which μ' is computed in order to verify a specific value for ϕ. The investigated measures were the Dorfman's sensitivity φ_{s_n}, specificity φ_{e_n}, positive $\mathrm{PPV}_{\mathscr{D}_n}$ and negative $\mathrm{NPV}_{\mathscr{D}_n}$ predictive values, and the relative cost RC.

4.2 Results and Discussion

Table 1 compares results when applying $\mathbf{M}_1$ and $\mathbf{M}_2^{\star}$, using a significance level of 5 %, a prevalence rate of 1 % and $Y_i^- \sim \text{Poisson}(100)$. In addition, several group dimensions have been considered, although for $p = 0.01$ the more efficient size in Dorfman's methodology (without misclassification) is 11 individuals in each group ($n^* = 11$). Nevertheless, the efficient size can correspond to a case with high probability of misclassification, a possibility which should be avoided.

The simulation results clearly show that the two methodologies fulfill their goals: whereas the significance level in $\mathbf{M}_1$ controls φ_{e_n}, implying $\varphi_{e_n} \geq 1 - \alpha$ and in most cases $\varphi_{e_n} \approx 0.99$, in $\mathbf{M}_2^{\star}$ it controls φ_{s_n}, despite of φ_{s_n} having a higher variability in the case of $\phi = 0.95$. Besides, the observed φ_{s_n} converges quickly to zero whenever n increases in $\mathbf{M}_1$, but φ_{e_n} still exhibits good results in $\mathbf{M}_2^{\star}$ even when n increases. The $\mathrm{NPV}_{\mathscr{D}_n}$ values are quite reasonable, but $\mathrm{PPV}_{\mathscr{D}_n}$ do not perform as well. These results are a consequence of working with low prevalence rates, and as such, there are many uninfected groups and few infected ones in the 10^6 simulated. In $\mathbf{M}_1$ a very

Table 1 Simulations with $\alpha = \mathbf{0.05}$, $p = 0.01$ $(n^* = 11)$, $Y_i^- \sim$ Poisson(100), and 10^6 groups

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2^\star}$				
	φ_{s_n}	φ_{e_n}	$\mathrm{PPV}_{\mathscr{D}_n}$	$\mathrm{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\mathrm{PPV}_{\mathscr{D}_n}$	$\mathrm{NPV}_{\mathscr{D}_n}$	RC
	$\phi = \mathbf{0.95}$									
$n = 1$	94.11	**95.71**	18.02	99.94	100	**94.11**	95.71	18.02	99.94	100
$n = 2$	71.93	**98.29**	29.65	99.71	55.73	**91.33**	96.33	20.00	99.91	**74.10**
$n = 3$	57.65	**98.66**	30.25	99.57	39.57	**91.29**	96.19	19.51	99.91	74.42
$n = 5$	44.39	**98.90**	29.02	99.44	27.15	**90.44**	96.16	19.20	99.90	77.86
$n = 7$	35.05	**99.10**	28.36	99.34	21.05	**90.12**	96.13	19.07	99.90	80.20
$n = 10$	28.51	**99.23**	27.15	99.28	16.77	**90.29**	96.07	18.81	99.90	84.16
$n = 20$	22.44	**99.30**	24.58	99.22	12.99	**90.37**	96.01	18.63	99.90	88.66
$n = 30$	18.36	**99.40**	23.56	99.18	**11.02**	**90.70**	95.97	18.48	99.90	90.94
$n = 50$	17.89	**99.37**	22.22	99.17	11.31	**91.16**	95.91	18.36	99.91	93.40
$n = 100$	16.81	**99.36**	21.06	99.16	11.78	**91.48**	95.88	18.32	99.91	95.12
	$\phi = \mathbf{0.99}$									
$n = 1$	99.88	**95.71**	19.26	100	100	**94.26**	99.88	88.99	99.94	100
$n = 2$	95.33	**98.13**	34.15	99.95	56.82	**94.66**	98.29	35.94	99.94	56.17
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	**48.79**
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 7$	56.15	**99.06**	37.67	99.55	22.45	**95.65**	96.61	22.20	99.95	63.57
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
$n = 30$	30.16	**99.25**	28.79	99.30	13.77	**96.19**	96.09	19.86	99.96	86.46
$n = 50$	26.65	**99.26**	26.69	99.26	**13.53**	**96.44**	96.01	19.63	99.96	90.23
$n = 100$	25.69	**99.19**	24.22	99.25	15.37	**97.32**	95.90	19.35	99.97	94.45
	$\phi = \mathbf{0.999}$									
$n = 1$	100	**95.70**	19.37	100	100	**94.00**	100	99.83	99.94	100
$n = 2$	99.61	**98.28**	36.98	100	56.26	**94.87**	99.68	75.00	99.95	52.29
$n = 3$	97.17	**98.57**	40.71	99.97	41.21	**94.48**	99.02	49.35	99.94	38.88
$n = 5$	86.32	**98.92**	44.81	99.86	28.83	**94.82**	98.07	33.33	99.95	**36.82**
$n = 7$	75.01	**99.03**	43.98	99.75	23.75	**95.00**	97.53	28.00	99.95	43.23
$n = 10$	62.65	**99.11**	41.62	99.62	20.00	**95.14**	97.11	24.96	99.95	52.43
$n = 20$	46.52	**99.14**	35.56	99.46	16.69	**95.94**	96.50	21.75	99.96	72.12
$n = 30$	39.71	**99.17**	32.53	99.39	**15.66**	**96.33**	96.28	20.68	99.96	80.31
$n = 50$	36.82	**99.09**	29.06	99.36	17.02	**96.57**	96.12	20.09	99.96	86.68
$n = 100$	35.09	**98.98**	25.79	99.34	19.80	**97.92**	95.91	19.49	99.98	94.05

Table 2 Simulations with $\phi = \mathbf{0.99}$ and $Y_i^- \sim \text{Poisson}(100)$

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2^\star}$				
	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC
	$\alpha = \mathbf{0.10},$ and $p = 0.01$									
$n = 3$	92.38	**96.40**	20.61	99.92	45.58	**89.85**	97.06	23.59	99.89	42.96
$n = 5$	79.19	**97.10**	21.69	99.78	32.93	**89.68**	95.50	16.82	99.89	42.61
$n = 10$	59.02	**97.62**	20.01	99.58	23.66	**90.38**	93.94	13.07	99.90	55.28
$n = 20$	46.83	**97.72**	17.17	99.45	20.76	**91.07**	93.04	11.66	99.90	68.31
	$\alpha = \mathbf{0.05},$ and $p = 0.01$									
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
	$\alpha = \mathbf{0.01},$ and $p = 0.01$									
$n = 3$	62.76	**99.83**	78.50	99.62	36.10	**98.14**	99.19	55.26	99.98	66.37
$n = 5$	40.97	**99.88**	77.23	99.41	22.82	**98.05**	99.17	54.38	99.98	76.30
$n = 10$	23.16	**99.91**	72.96	99.23	12.81	**98.21**	99.15	53.88	99.98	91.66
$n = 20$	16.84	**99.92**	68.24	99.17	08.48	**98.28**	99.14	53.57	99.98	96.78
	$p = \mathbf{0.05}$ and $\alpha = 0.05$									
$n = 3$	85.83	**98.44**	74.45	99.25	49.44	**95.06**	97.44	66.22	99.73	58.08
$n = 5$	73.82	**98.43**	71.18	98.62	40.11	**96.26**	96.68	60.33	99.80	67.42
$n = 10$	60.22	**98.42**	66.60	97.92	33.71	**96.48**	96.28	57.60	99.81	80.82
$n = 20$	57.95	**98.20**	62.91	97.79	36.11	**98.01**	95.95	56.00	99.89	93.50
	$p = \mathbf{0.01}$ and $\alpha = 0.05$									
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
$n = 20$	34.16	**99.23**	31.01	99.33	14.55	**95.96**	96.19	20.27	99.96	82.07
	$p = \mathbf{0.001},$ and $\alpha = 0.05$									
$n = 3$	83.49	**98.72**	6.19	99.98	38.16	**94.28**	97.60	3.82	99.99	46.63
$n = 5$	66.27	**99.04**	6.38	99.97	25.10	**94.79**	96.95	2.97	99.99	52.07
$n = 10$	44.04	**99.31**	6.02	99.94	15.30	**94.59**	96.54	2.67	99.99	65.76
$n = 20$	27.18	**99.48**	4.93	99.93	10.34	**95.00**	96.28	2.49	99.99	77.77

Table 3 Simulations with $p = 0.01, \phi = \mathbf{0.99}, \alpha = \mathbf{0.05}$

	Methodology $\mathbf{M_1}$					Methodology $\mathbf{M_2^\star}$				
	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC	φ_{s_n}	φ_{e_n}	$\text{PPV}_{\mathscr{D}_n}$	$\text{NPV}_{\mathscr{D}_n}$	RC
	$Y_i^- \sim$ Poisson(10)									
$n = 3$	84.52	**98.49**	36.11	99.84	40.33	**94.70**	97.34	26.42	99.95	47.77
$n = 5$	64.21	**98.87**	36.42	99.64	27.17	**93.80**	96.77	22.67	99.94	49.12
$n = 10$	43.99	**99.08**	32.70	99.43	17.88	**94.70**	96.08	19.65	99.94	65.49
	$Y_i^- \sim$ Negative Binomial$(1, \frac{1}{11})$									
$n = 3$	75.90	**98.19**	29.84	99.75	40.45	**94.48**	97.25	25.88	99.94	45.11
$n = 5$	53.83	**98.56**	27.38	99.53	27.40	**94.89**	96.52	21.52	99.95	45.15
$n = 10$	37.03	**98.90**	25.36	99.36	17.74	**95.34**	96.00	19.41	99.95	59.94
	$Y_i^- \sim$ Binomial$(20, \frac{1}{2})$									
$n = 3$	82.34	**99.23**	51.83	99.82	40.27	**92.44**	98.73	42.33	99.92	47.95
$n = 5$	63.05	**99.45**	53.58	99.62	27.26	**93.20**	98.46	37.99	99.93	54.00
$n = 10$	40.25	**99.60**	50.55	99.40	17.14	**94.79**	98.21	34.83	99.95	71.92
	$Y_i^- \sim$ Poisson(100)									
$n = 3$	84.78	**98.67**	39.22	99.84	40.30	**94.89**	97.58	28.45	99.95	48.79
$n = 5$	69.06	**98.93**	38.89	99.66	27.85	**95.01**	96.92	23.83	99.95	54.86
$n = 10$	48.45	**99.09**	34.89	99.48	19.22	**95.18**	96.48	21.41	99.95	69.21
	$Y_i^- \sim$ Negative Binomial$(1, \frac{1}{101})$									
$n = 3$	74.54	**98.20**	29.34	99.74	40.21	**94.66**	97.16	25.05	99.94	45.21
$n = 5$	53.63	**98.54**	27.15	99.53	27.37	**94.96**	96.43	21.23	99.95	44.99
$n = 10$	37.34	**98.85**	24.63	99.37	17.89	**95.07**	95.84	18.72	99.95	59.51
	$Y_i^- \sim$ Binomial$(200, \frac{1}{2})$									
$n = 3$	82.85	**98.79**	40.87	99.82	40.33	**94.44**	97.66	28.98	99.94	51.03
$n = 5$	64.83	**99.05**	40.85	99.64	27.61	**94.68**	97.15	25.12	99.94	57.03
$n = 10$	46.03	**99.20**	36.74	99.46	18.83	**95.13**	96.78	22.95	99.95	71.48
	$Y_i^- \sim$ Poisson(1000)									
$n = 3$	84.84	**98.51**	36.65	99.84	40.48	**94.93**	97.15	25.27	99.95	50.40
$n = 5$	68.91	**98.75**	35.59	99.69	28.36	**95.15**	96.45	21.19	99.95	56.28
$n = 10$	46.51	**99.04**	32.92	99.46	18.61	**95.01**	95.97	19.26	99.95	69.48
	$Y_i^- \sim$ Negative Binomial$(1, \frac{1}{1001})$									
$n = 3$	77.40	**98.12**	29.58	99.77	40.54	**95.03**	97.14	25.26	99.95	45.23
$n = 5$	53.26	**98.56**	27.37	99.52	27.22	**95.46**	96.36	21.01	99.95	45.65
$n = 10$	37.13	**98.85**	24.61	99.36	17.79	**95.68**	95.78	18.61	99.95	60.45
	$Y_i^- \sim$ Binomial$(2000, \frac{1}{2})$									
$n = 3$	84.87	**98.56**	37.31	99.85	40.52	**94.66**	97.31	26.24	99.94	49.95
$n = 5$	68.09	**98.85**	37.50	99.67	28.04	**94.53**	96.71	22.56	99.94	54.33
$n = 10$	48.80	**99.02**	33.56	99.48	19.39	**94.91**	96.19	20.15	99.95	69.12

low RC can be attained, but with high probability of misclassification, while in $\mathbf{M_2^\star}$ the efficiency is not so good, but both φ_{s_n} and φ_{s_n} have good performance.

Table 2 investigates different prevalence rates $p \in \{0.05, 0.01, 0.001\}$ and multiple significance levels $\alpha \in \{0.1, 0.05, 0.01\}$ with $\phi = 0.99$. The results are as expected, i.e. the φ_{e_n} (φ_{s_n}) decreases and the φ_{s_n} (φ_{e_n}) increases in methodology $\mathbf{M_1}$ ($\mathbf{M_2^\star}$) when the significance level increases. Moreover, the use of different prevalence rates (all used rates are low, because compound test should not be used otherwise) does not seem to have a major impact either on the sensitivity or in the specificity.

Different distributions (Poisson, negative binomial, and binomial) and different parameters values (maintaining the same expected value in the distinct distributions) were analyzed in Table 3 in order to evaluate the impact in the misclassification. Moreover, the same quality value for the individual test performance ($\phi = 0.99$) has been applied in all investigated cases. Hence, it seems that the shape of the applied distribution is not important to assess the problem of misclassification, but exclusively the value of ϕ.

5 Final Remarks

The usual $\mathbf{M_1}$ methodology can be applied to control the specificity and minimize the cost, as for screening cases, while the methodology $\mathbf{M_2^\star}$ can be applied in order to control the sensitivity, namely in epidemic situations. The optimum group size n depends on the purpose of the investigator, since it can be greater if the main goal is to save resources (and accepting having a higher probability of misclassification) or lower if the main goal is to control the problem of misclassification (cases in which we should use a smaller dimension n to ensure a low probability of misclassification). Furthermore, the compound tests should be applied only if the individual test has good performance (low probability of misclassification). Hence, the proposed quality measure of the individual test performance ϕ can be used to identify those situations. Moreover, the distribution $\mathbf{D}_0$ and the prevalence rate p do not seem to have a major impact on Dorfman's methodology misclassification problem if the same ϕ value is provided. A high ϕ value ensures a high φ_{s_n} and φ_{e_n} in the new $\mathbf{M_2^\star}$ methodology, although the RC remains quite high. The same high ϕ value only guarantees a high φ_{e_n} in the usual $\mathbf{M_1}$ methodology and, therefore, it must be used with caution. Nevertheless, in the $\mathbf{M_1}$ methodology, the RC rapidly decreases with n.

Acknowledgments The authors thank the referees for their very useful comments. Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal — FCT under the projects PEst-OE/MAT/UI0006/2011 and PEst-OE/MAT/UI0006/2014.

References

1. Berger, T., Mandell, J.W., Subrahmanya, P.: Maximally efficient two-stage screening. Biometrics **56**, 833–840 (2000)
2. Bilder, C.R., Zang, B., Schaarschmidt, F., Tebbs, J.M.: binGroup: a package for group testing. R J. **2**, 56–60 (2010)
3. Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P.: Annotated bibliography of composite sampling, part A: 1936–92. Environ. Ecol. Stat. **3**, 1–50 (1996)
4. Chen, C., Swallow, W.: Sensitivity analysis of variable-sized group testing and its related continuous models. Biom. J. **37**, 173–181 (1995)
5. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Stat. **14**, 436–440 (1943)
6. Finucan, H.M.: The blood testing problem. Appl. Stat.—J. R. St. C **13**, 43–50 (1964)
7. Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R.: Probability estimation with sample compositing techniques. J. Off. Stat. **5**, 365–374 (1989)
8. Gastwirth, J.L.: The efficiency of pooling in the detection of rare mutations. Am. J. Hum. Genet. **67**, 1036–1039 (2000)
9. Hughes-Oliver, J.M.: Pooling experiments for blood screening and drug discovery. Screening—Methods for Experimentation in Industry, Drug Discovery, and Genetics, pp. 48–68. Springer, New York (2006)
10. Hung, M., Swallow, W.: Robustness of group testing in the estimation of proportions. Biometrics **55**, 231–237 (1999)
11. Hwang, F.K.: Group testing with a dilution effect. Biometrika **63**, 671–673 (1976)
12. Johnson, N.L., Kotz, S., Wu, X.: Inspection Errors for Attributes in Quality Control. Chapman and Hall, New York (1991)
13. Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, C.: Comparison of group testing algorithms for case identification in the presence of testing errors. Biometrics **63**, 1152–1163 (2007)
14. Lancaster, V.A., Keller-McNulty, S.: A review of composite sampling methods. J. Am. Stat. Assoc. **93**, 1216–1230 (1998)
15. Litvak, E., Tu, X.M., Pagano, M.: Screening for the presence of a disease by pooling sera samples. J. Am. Stat. Assoc. **89**, 424–434 (1994)
16. Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, T.C.: Cost analysis in choosing group size when group testing for potato virus Y in the presence of classification errors. Ann. Appl. Biol. **159**, 491–502 (2011)
17. Loyer, M.W.: Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. **37**, 57–59 (1983)
18. Martins, J.P., Santos, R., Sousa, R.: Testing the maximum by the mean in quantitative group tests. In: Pacheco et al. (eds.), New Advances in Statistical Modeling and Applications, pp. 55–63. Springer, Berlin (2014)
19. Phatarfod, R.M., Sudbury, A.: The use of a square array scheme in blood testing. Stat. Med. **13**, 2337–2343 (1994)
20. Roederer, M., Koup, R.A.: Optimized determination of T cell epitope responses. J. Immunol. Methods **274**, 221–228 (2003)
21. Santos, R., Pestana, D., Martins, J.P.: Extensions of Dorfman's theory. In: Oliveira, P.E., et al. (eds.) Studies in Theoretical and Applied Statistics, Recent Developments in Modeling and Applications in Statistics, pp. 179–189. Springer, Berlin (2013)
22. Sobel, M., Elashoff, R.: Group testing with a new goal, estimation. Biometrika **62**, 181–193 (1975)
23. Sobel, M., Groll, P.A.: Group testing to eliminate efficiently all defectives in a binomial sample. Bell Syst. Tech. J. **38**, 1179–1252 (1959)
24. Sterret, A.: On the detection of defective members of large populations. Ann. Math. Stat. **28**, 1033–1036 (1957)

25. Tu, X.M., Litvak, E., Pagano, M.: Studies of AIDS and HIV surveillance, screening tests: can we get more by doing less? Stat. Med. **13**, 1905–1919 (1994)
26. Tu, X.M., Litvak, E., Pagano, M.: On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening. Biometrika **82**(2), 287–297 (1995)
27. Wein, L.M., Zenios, S.A.: Pooled testing for HIV screening: capturing the dilution effect. Oper. Res. **44**, 543–569 (1996)
28. Woodbury, C.P., Fitzloff, J.F., Vincent, S.S.: Sample multiplexing for greater throughput in HPLC and related methods. Anal. Chem. **67**, 885–890 (1995)
29. Zenios, S., Wein, L.: Pooled testing for HIV prevalence estimation exploiting the dilution effect. Stat. Med. **17**, 1447–1467 (1998)

A Maximum Likelihood Estimator for the Prevalence Rate Using Pooled Sample Tests

João Paulo Martins, Rui Santos and Miguel Felgueiras

Abstract Since Dorfman's seminal work, research on methodologies involving pooled sample tests has increased significantly. Moreover, the use of pooled samples refers not only to the classification problem (identifying all the infected individuals in a population), but also refers to the problem of estimating the prevalence rate p, as Sobel and Elashoff stated. The use of compound tests is not restricted to hierarchical algorithms where the most common example is Dorfman's two-stage procedure. Matrix schemes such as the square array algorithm or multidimensional matrices schemes in certain cases outperform Dorfman's procedure. Maximum likelihood estimates are quite difficult to compute when a procedure does not classify all individuals. This paper presents two innovative methods to compute maximum likelihood estimates in both type of procedures.

Keywords Compound tests · Maximum likelihood estimator · Prevalence rate

1 Introduction

The use of group testing procedures to screen for a binary characteristic is usually said to have started from Dorfman's (cf. [3]) seminal work. His purposed procedure proved to be less expensive than applying only individual tests in the detection of the World War II soldiers infected with syphilis. The new strategy was to gather groups

J.P. Martins (✉) · R. Santos · M. Felgueiras
School of Technology and Management, Polytechnic Institute of Leiria,
Apartado 4163, 2411-901 Leiria, Portugal
e-mail: jpmartins@ipleiria.pt

J.P. Martins · R. Santos · M. Felgueiras
CEAUL – Center of Statistics and Applications of the University of Lisbon,
University of Lisbon, Lisbon, Portugal
R. Santos
e-mail: rui.santos@ipleiria.pt

M. Felgueiras
e-mail: mfelg@ipleiria.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_8

of n individuals into pools and then perform a pooled sample test. A negative result of the pooled mixture indicates that all of them are free of the disease. A positive result indicates that at least one of the n individuals has the disease, but we do not know who or how many. In this case, performing individual tests is advised to identify the infected individuals in the sample. The main issue is to determine the optimal batch size which minimizes the expected number of tests as it is a good measure of monetary costs, since the cost of mixing samples is usually negligible (cf. [11]).

Pooled samples may be used in two types of problems: a classification problem or an estimation problem. Identifying all the subjects that are infected or have a high level of blood sugar are examples of classification problems. In both examples, we determine, for each individual, if fall in the class of interest. Estimating the prevalence rate of a disease or of a gene in some population are examples of estimation problems. In this case, the performance of individual tests is only optional, since the goal is no longer to identify the infected individuals (cf. [2]). The use of only pooled samples has also the advantage of anonymity of the infected members, given that they are not identified. Furthermore, the estimators obtained by applying compound tests have, under certain conditions, better performance than the traditional estimators based on individual tests, cf. [5, 12, 15]. The bias, the efficiency and the robustness of these estimators has been reviewed in several works, such as those from [2, 6, 10]. Bilder et al. [1] proposes the use of the package *binGroup* for the *R* software, which includes applications of several compound testing estimators. Thus, the estimators based on group testing not only allow one to obtain monetary gains (by decreasing the number of performed tests), but also allow us to achieve more accurate estimates, compared to those obtained on the basis of individual tests.

Group testing application can be done in several ways (cf. [8]). The main reason for having different procedures is related to the misclassification problem, as an individual can be wrongly classified. The sensitivity and the specificity of the test (see Definition 1) may be used for measuring the accuracy of the test results. In particular, the sensitivity of a test generally decreases as the pooled sample size increases. The choice for a particular group testing procedure depends on the number of samples available and the sensitivity, the specificity and the monetary costs of the process (cf. [11]). For an overview for this problem, known as the dilution problem, see [7, 14, 17, 18].

The outline of this work is as follows. Section 2 introduces the binomial model assumption and discusses some considerations about the prevalence rate maximum likelihood (ML) estimator when pooled samples are used. Section 3 describes the two main types of group testing procedures and is the core of this work as new ways of computing ML estimates are provided. For the hierarchical algorithms dealt in Sect. 3.1, we propose a method for the classification of an individual for estimation purposes that does not need any individual tests. This allows the application of the traditional ML estimators even when in the last stage of the algorithm no individual tests are performed. The array-based group testing procedures are presented in Sect. 3.2. In these kind of procedures, the computation of the ML estimates is very difficult to perform. Hence, in order to easily compute reasonable estimates an iterative method is proposed. In Sect. 4 some final remarks are discussed.

2 The Binomial Model

Let p denote the probability that an individual is infected, n be the pool sample size and t the number of performed tests. The total number of individuals is $N = n \times t$. Let us also assume that the individuals status (infected/not infected) within a pooled sample are independent. The probability of having an infected pooled sample is $\pi_n = 1 - (1 - p)^n$. Hence the total number of infected samples is described by a binomial random variable $I \frown \text{Bin}\,(t, \pi_n)$. The ML estimator of π_n is

$$\widehat{\pi_n} = \frac{I}{t}. \tag{1}$$

As p is given by a simple transformation of π_n, it is straightforward to prove, applying the proprieties of the ML estimators, that the ML estimator of p is

$$\widehat{p} = 1 - \left(1 - \frac{I}{t}\right)^{1/n}. \tag{2}$$

For $n = 1$, $\hat{p} = 1 - \left(1 - \frac{I}{t}\right) = \frac{I}{t}$ is an unbiased estimator of p. For $n > 1$, the estimator is positively biased. Expressions for the expected value and variance of the estimator may be found in [6].

As screening errors may occur, the above is, in practice, unrealistic. Let us consider the problem of estimating the prevalence rate of some disease and let $X_i = 1$ denote an infected individual and $X_i = 0$ denote a non-infected individual. In addition, X_i^+ will stand for a positive test result and X_i^- stand for a negative test result. In order to assess the sources of error two measures will be considered.

Definition 1 Consider an individual X_i who is tested individually. The probability $\varphi_s = \text{P}\left(X_i^+ | X_i = 1\right)$ is called the **test sensitivity** and $\varphi_e = P\left(X_i^- | X_i = 0\right)$ is called the **test specificity**.

When a pooled sample test is performed the probability of having a positive result from an infected sample may decrease. As the amount of substance per unit of volume is less or equal to the amount of substance found in a unit of volume collected from an infected individual, it may be difficult to screen the infected pool sample as positive. However, the probability of getting a negative outcome on a non-infected sample is equal to φ_e as there are no dilution problems. Thus, [14] defines the concepts of specificity and sensitivity of some specific methodology of classification or estimation $\mathscr{M}$ (these concepts are closely related to the pooling sensitivity and pooling specificity concepts defined in [8]). These measures assess the quality of an outcome provided by some methodology $\mathscr{M}$.

Definition 2 The **methodology sensitivity** or the procedure sensitivity is the probability of an infected individual being correctly identified by the methodology $\mathscr{M}$, that is, $\varphi_s^{\mathscr{M}} = P_{\mathscr{M}}\left(X_i^+ | X_i = 1\right)$. The **methodology specificity** or the procedure specificity stands for the probability of a non-infected individual being correctly classified by the methodology $\mathscr{M}$, that is, $\varphi_e^{\mathscr{M}} = P_{\mathscr{M}}\left(X_i^- | X_i = 0\right)$.

For an individual testing procedure the sensitivity (specificity) methodology is equal to the test sensitivity (specificity). For instance, in Dorfman's procedure, assuming there is no dilution effect, the probability of an infected individual being screened as positive is

$$\varphi_s^{\mathscr{M}} = \varphi_s^2, \tag{3}$$

as it is required that both pooled and subsequent individual test outcomes to be positive. Note that the methodology sensitivity is less than the test sensitivity, i.e.,

$$\varphi_s^{\mathscr{M}} \leq \varphi_s, \tag{4}$$

for $n > 1$.

For computing the probability of a non-infected individual being correctly classified it is necessary to account for three possible situations:

- the pooled sample is not infected and the pooled test outcome is negative;
- the pooled sample is not infected but the pooled test outcome is positive and the subsequent individual test outcome is negative;
- the pooled sample is infected and the pooled test outcome is positive but in the subsequent individual test the subject is correctly classified as non-infected.

Hence,

$$\varphi_e^{\mathscr{M}} = \varphi_e\,(1-p)^{(n-1)} + (1-\varphi_e)\varphi_e\,(1-p)^{(n-1)} + \varphi_s\varphi_e\left(1-(1-p)^{(n-1)}\right). \tag{5}$$

The exponent $n-1$ in Eq. (5) is due to the fact that we are computing a conditional probability. These probabilities allows us to compute the real bias of

$$\widehat{p} = \frac{T}{N} \tag{6}$$

where T stands for the number of specimens classified as positive when Dorfman's procedure is applied. T is a binomial random variable described by $T \frown \text{Bin}\,(N, p^*)$. It depends on the methodology specificity $\varphi_e^{\mathscr{M}}$ and on the methodology sensitivity $\varphi_s^{\mathscr{M}}$, therefore, $p^* = \psi\,(\varphi_s, \varphi_e, p)$. Santos et al. [14] computed the value of p^* by

$$\begin{aligned} p^* &= P\left(X_i^+ | D\right) P\,(D) + P\left(X_i^+ | \overline{D}\right) P\left(\overline{D}\right) \\ &= \varphi_s^{\mathscr{M}} p + \left(1-\varphi_e^{\mathscr{M}}\right)(1-p) \\ &= 1-\varphi_e^{\mathscr{M}} + \left(\varphi_s^{\mathscr{M}} + \varphi_e^{\mathscr{M}} - 1\right) p. \end{aligned} \tag{7}$$

where D stands for an infected individual and $\overline{D}$ stands for a non-infected individual.

Hence, the estimator is, in general, biased. The bias is equal to

$$\text{Bias}\left(\widehat{p}\right) = p^* - p \tag{8}$$

and the estimator variance is

$$\text{Var}\left(\widehat{p}\right) = \frac{p^*\left(1 - p^*\right)}{N}. \tag{9}$$

The mean square error (MSE) of the estimator is, by definition,

$$\text{MSE}\left(\widehat{p}\right) = \left[\text{Bias}\left(\widehat{p}\right)\right]^2 + \text{Var}\left(\widehat{p}\right). \tag{10}$$

Note that, for instance, if $\varphi_e^{\mathscr{M}} = \varphi_s^{\mathscr{M}} = p = 0.5$ the estimator is actually unbiased. The mean square error is a possible measure for assessing the quality of the estimates in each procedure. This measure may be used to combine different prevalence rate estimates. Chen et al. [2] uses a logistic regression where the parameters are computed iteratively but the quality of each estimate is measured by just using the pooled sample size. Martins et al. [13] provides an iterative meta-analysis-based procedure that uses the mean square error as weights for achieving a single estimate. The content in Sect. 3.2 enhances the work of [13] as it provides a computational method for estimating the prevalence rate from an array-based group testing algorithm and, even more important for the meta-analysis technique, it provides a way to estimate the MSE of the estimator.

3 ML Estimators in Several Group Testing Procedures

On a pooled sample-based procedure there are two goals: minimizing the sources of error and providing a less expensive method than individual testing for achieving the investigation goal. To assess the savings of some procedure $\mathscr{M}$, the **relative cost** will be used as a measure of the **methodology efficiency**, RC ($\mathscr{M}$), that is, the expected number of tests per specimen since the cost of mixing samples is usually negligible. When only individual tests are performed the methodology efficiency is equal to one. In general, the methodology efficiency is high for low prevalence rates as the pooled samples sizes tend to increase when p decreases. For instance, in the traditional Dorfman procedure the maximum efficiency for a prevalence rate equal to 0.1, 0.01 and to 0.001 is obtained by using a pooled sample size equal to 4, 11 and 32, respectively (cf. [3]).

The most commonly used pooled sample procedures can be binned in the following two groups:

- **Hierarchical algorithms**—a pooled sample is tested and if the test outcome is positive it is divided into smaller nonoverlapping groups until eventually all individuals have been tested;

Table 1 Correct and wrong decisions at the sth stage

		Pooled sample at the sth stage	
		Infected	Not infected
$X_i = 0$	Test result +	✓	×
	Test result −	✓	✓
$X_i = 1$	Test result +	✓	Not possible
	Test result −	×	Not possible

- **Array-based group testing algorithms**—in its simplest two-stage version (square array), a sample of size n^2 is placed in a $n \times n$ matrix and then all the individuals within the same row and the same column are gathered for batched testing.

3.1 Hierarchical Algorithms

Dorfman's procedure is just one example of a wider family of pooled testing procedures called hierarchical algorithms. Some improvements to his work have been proposed (cf. [4, 16, 17]) by dividing positive pools into smaller subpools until eventually all positive specimens are individually tested.

A multistage hierarchical algorithm is an algorithm that generalizes Dorfman's procedure to more than two stages, that is, a sample is divided at each stage into smaller nonoverlapping groups until eventually all positive specimens are individually tested. At each stage, subsamples from the samples tested positively are retested. For practical reasons, only two or three stages are usually performed. Let us consider a hierarchical algorithm with s stages and let n_i denote the number of individuals at the ith stage. At the last stage, when the classification problem is considered, we have $n_s = 1$. However, this condition might not be fulfilled, when we just want to estimate the prevalence rate, and the condition verified is just $n_1 > \cdots > n_s \geq 1$ (cf. [2, 6, 10]). For low prevalence rates, the use of $n_s > 1$ for achieving a greater efficiency may be justified if a positive outcome when testing a pooled sample of size n_s at the last stage means (almost surely) that only one of the individuals is infected (cf. [14]). Under this assumption, it is now easy to compute the proportion of infected individuals.

Table 1 shows all possible scenarios at the last stage of a hierarchical algorithm indicating what scenarios correspond, for estimation purposes, to a correct/wrong classification ($\checkmark/\times$) of an individualX_i.

One of the less intuitive classifications shown in Table 1 is to have a correct decision when the sample at the sth stage is infected and the test outcome is positive. This is almost 100 % true as it means (almost surely) that only one individual is infected and that the other individuals are not. Therefore, concerning the estimation problem, all the individuals are (almost surely) well classified as one infected and

Table 2 Comparing the efficiency of the different methodologies

Methodology	RC($\mathscr{M}$)	$\varphi_e^{\mathscr{M}}$	$\varphi_s^{\mathscr{M}}$
Individual test	1	0.9900	0.9000
$MA2(49:7:1)$	0.34	0.9995	0.6810
$MA2(100:10:1)$	0.31	0.9991	0.6596
$MA2(100:10)$	0.22	0.9990	0.7290

$n_s - 1$ non-infected. Hence, although we may not be able to identify who is infected it is now straightforward to compute the ML estimator presented in (6). However, the given estimate may not be a ML estimate since, although unlikely, it is possible to have two infected individuals at the sth stage.

3.2 Array-Based Group Testing Algorithms

Array-based group testing is an alternative to hierarchical group testing that uses overlapping pools. In its simplest two-stage version (square array), denoted by $A2\ (n:1)$, a sample of size n^2 is placed in a $n \times n$ matrix in the following way. Each individual is allocated to one and only one matrix position. Then, all the individuals within the same row and the same column are gathered for batched testing. This process involves at least $2n$ tests as subsequent individual tests are performed to the samples lying in a row and/or column that tested positively. A variant of this methodology consists in performing a priori pooled sample test on all the n^2 individuals (master pool). If the master pool test result is negative no further testing is needed as the individuals are all classified as negative. This methodology with a master pool will be represented by $MA2\left(n^2:n:1\right)$. The performance of subsequent individual tests is required to avoid ambiguities. For instance, it is possible to have a row tested positive but all columns tested negative (the number of infected individuals can be any integer from zero to n) or to have two positive rows and columns (the number of infected individuals may be 2, 3 or 4). To obtain a greater efficiency we suggest the dropping of the subsequent individual testing because it is not necessary to determine who exactly the infected individuals are when dealing with an estimation problem. A square array-based group testing with no individual tests will be represented by $MA2\left(n^2:n\right)$ or $A2\left(n^2:n\right)$ depending on the performance or not of a master pool test. Let us look for a simple example of a square array procedure with two lines (with or without) a master pool (Table 2).

Example 1 ([9]) compares the operating characteristics of two square array procedures with a master pool: $MA2\ (49:7:1)$ and $MA2(100:10;\,1)$ when screening for a disease in Malawi with prevalence rate 0.045. We computed the operating characteristics of the last procedure without any individual tests: $MA2(100:10)$. In this case, an individual is classified as positive if and only if both "row" and "column" tested positive.

Dropping the performance of individual tests ($MA2(100:10)$) results on a great reduction of the relative cost from more than 0.3 to 0.22. Moreover, concerning the methodology specificity, $MA2(100:10)$ methodology performs as well as the others and it also has a greater sensitivity than the other array based group testing methodologies.

Hence, if all the columns (rows) tested negative and a row (column) tested positive, all the individuals are classified as negative. This approach although much more efficient than the others has a great drawback. It almost surely underestimates the prevalence rate!

As it is not possible to use the proportion of defective individuals without avoiding an underestimation of the prevalence rate, we propose the computation of a ML estimate, using a proper script. This will combine a greater efficiency with the computation of an accurate estimate.

When the number of rows and columns of the two-dimensional array is low it is possible to compute the exact value of the likelihood function for a given prevalence rate p_0. For an array with two rows and two columns it is easy although tedious to write a script to compute the ML function for any value. Hence, a proper iterative process gives the ML estimate.

The inputs of the script must be the test sensitivity φ_s, the test specificity φ_e and the number of arrays that have $i-1$ positive rows and $j-1$ positive columns for $i=1,2,3$ and $j=1,2,3$. These values may be inserted in a 3×3 matrix O. The matrix O resumes the experimental results.

To compute the ML function at p_0 one is also required to compute the probability of observing $i-1$ positive rows and $j-1$ positive columns, where $i=1,2,3$ and $j=1,2,3$, given p_0 and taking into account the test sensitivity φ_s and the test specificity φ_e. Suppose these values are recorded in a matrix P. For instance, if $\varphi_s=\varphi_e=0.95$ (consider that the individual test sensitivity is equal to the pooled sample sensitivity) and $p_k=0.1$. The matrix P_0 is

$$P_0=\begin{pmatrix}0.5351 & 0.0689 & 0.0029\\ 0.0689 & 0.2477 & 0.0277\\ 0.0029 & 0.0277 & 0.0183\end{pmatrix}$$

As the matrix of a square array is always symmetric it can be written as an upper triangular matrix

$$P=\begin{pmatrix}0.5351 & 0.1378 & 0.0058\\ 0 & 0.2477 & 0.0554\\ 0 & 0 & 0.0183\end{pmatrix}$$

In this case, the matrix of the frequencies of the number of positive lines and positive columns O must be also rewritten as $O(i,j)=O(i,j)+O(j,i)$ for $j>i$. Approximately 14 % of arrays are expected to have only one positive row (column) and no positive columns (rows). In the traditional application of a square array methodology, this would require the performance of individual tests. The ML function for p_0 is given by

$$ML(p_0) = \prod_{i=1}^{3} \prod_{j=i}^{3} P(i, j)^{O(i,j)}. \tag{11}$$

Example 2 To assess the MSE of the estimator, for a prevalence rate $p = 0.1$, 100 replicates of a 2×2 array ($A2(2:1)$) were simulated in software MatLab R2011 and the ML estimate was computed. This procedure was repeated 1000 times to produce 1000 prevalence rate estimates. The matrix O was set to be equal to the matrix P. Although, in practice, the matrix O only admits integer values, it also works for any non-negative numbers.

The mean value of the estimates was 0.1189 with standard error 0.0120. The 5 and 95 % percentiles are, respectively, 0.1029 and 0.1408. Thus, by (10), an estimate for the mean square error of the estimator is

$$\text{MSE}\,(ML) = 5.01 \times 10^{-4}. \tag{12}$$

To evaluate the MSE of the estimator, we will compare these results with the ones obtained using Dorfman's procedure. The optimal batch size for $p = 0.1$ is $n = 4$. By (8)–(10), the mean square error is given by

$$\text{MSE}\,(\widehat{p}) = 0.015671^2 + \frac{0.102291}{400} = 5.01 \times 10^{-4}. \tag{13}$$

The MSE is the same for both methods. Moreover both present a problem of overestimation due to the test sensitivity and specificity.

However, when the number of rows and columns is just 3 or more it is not easy to use the previous method to compute a value of the ML function. In this case, we suggest the computation of an estimate for the ML function value for a given prevalence rate in the following way.

1. Record in a matrix O of size $r \times c$ the number of two-dimensional arrays with $i-1$ positive rows and $j-1$ positive columns where $i = 1, \ldots, r$ and $j = 1, \ldots, c$.
2. For some possible prevalence rate values p, chosen in some logical sequence (for instance, 0, 0.1, 0.2, …, 1) simulate a reasonable number of replicates rep of the possible matrices.
3. Compute the probability of observing $i-1$ positive rows and $j-1$ positive columns for each replicate (taking into account the test sensitivity φ_s and the test specificity φ_e, and store that value in the position $(i-1, j-1)$ of the matrix P). Add the probabilities computed for all the replicates and multiply P by $1/rep$.
4. Compute the ML function for the matrix O using the values of P.

$$ML(p_0) = \prod_{i=1}^{r}\prod_{j=1}^{c} P(i, j)^{O(i,j)}.$$

5. Compare the ML function for each prevalence rate estimate and chose the two estimates with the highest value of the ML function.
6. Repeat the process from step two until the difference between the ML function at the two points chosen in step five be lower than some prefixed tolerance.
7. The estimate is the weighted mean value between those two points, say p_1 and p_2, using as weights $ML\,(p_1)$ and $ML\,(p_2)$, i.e.,

$$\widehat{p} = \frac{ML\,(p_1) \times p_1 + ML\,(p_2) \times p_2}{ML\,(p_1) + ML\,(p_2)}.$$

$P(i, j)$ is an estimate of the probability of having $i - 1$ positive rows and $j - 1$ positive columns in an array. In practice, the values for p in step two don't have to span the entire interval [0, 1] as the use of pooled samples is advised only for prevalence rates lower than about 1/3.

Let us look at the following example.

Example 3 Consider a matrix O generated by simulating 1000 replicates of a square array $A2(4 : 1)$ for a prevalence rate $p = 0.01$ and $\varphi_s = \varphi_e = 0.99$ using software MatLab R2011.

$$O = \begin{pmatrix} 76 & 4 & 0 & 0 & 0 \\ 4 & 14 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The total proportion of infected individuals of this simulation was, by chance, equal to the prevalence rate $p = 0.01$.

For computing a ML estimate for the prevalence rate given this matrix O, 100 square arrays were simulated to compute each matrix P. Hence, the sample size is $4^2 \times 100 = 1600$. Then, 100 estimates were, independently, computed. The mean value of the estimates was 0.0148 with standard error 0.0027. The 5 and 95 % percentiles are, respectively, 0.0113 and 0.0183. Thus, an estimate for the MSE of the estimator is

$$\text{MSE}\,(ML) = 3.02 \times 10^{-5}. \tag{14}$$

Once again, in order to evaluate the MSE of the estimator we will compare these results with the ones obtained using Dorfman's procedure. The optimal batch size for $p = 0.01$ is $n = 11$. The MSE is given by

$$\text{MSE}\,(\widehat{p}) = 0.00177^2 + \frac{0.011636}{1600} = 1.04 \times 10^{-5}. \tag{15}$$

The MSE of both estimators are similar. Thus, the estimates given by this algorithm seem to be reliable. However, we are not performing a formal comparison between the two methods as there is no way, at least to our knowledge, to find the optimal array-based group testing design for a given estimation problem (unless one supposes there are no test errors, cf. [9]).

4 Final Remarks

The main achievement of this work is the dropping of the individual tests when we just want to determine a prevalence rate estimate.

When one is dealing with a hierarchical method, an estimate very close to the real proportion of infected elements can be achieved even without performing any individual tests.

The use of a square array methodology is only possible with the advent of robotic pooling. These methods can be very efficient if no individual tests, are performed. Furthermore, the iterative method for computing a ML estimate allows the use of these kinds of strategies and the computational cost does not have to be very high in order to obtain accurate estimates (comparing to Dorfman's procedure). One problem that still is unsolved is to find a method to easily identify the best array to use in a given situation. This issue will be dealt in a future work. A generalization of this iterative method to higher dimensional arrays is straightforward. More details on the use of arrays with dimensions higher than two is discussed by [9].

Acknowledgments Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal — FCT under the project (PEst-OE/MAT/UI0006/2014).

References

1. Bilder, C.R., Zang, B., Schaarschmidt, F., Tebbs, J.M.: binGroup: a package for group testing. R J. **2**, 56–60 (2010)
2. Chen, C.L., Swallow, W.H.: Using group testing to estimate a proportion, and to test the binomial model. Biometrics **46**, 1035–1046 (1990)
3. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Stat. **14**, 436–440 (1943)
4. Finucan, H.M.: The blood testing problem. Appl. Stat. **13**, 43–50 (1964)
5. Garner, F.C., Stapanian, M.A., Yfantis, E.A., Williams, L.R.: Probability estimation with sample compositing techniques. J. Off. Stat. **5**, 365–374 (1989)
6. Hung, M., Swallow, W.H.: Robustness of group testing in the estimation of proportions. Biometrics **55**, 231–237 (1999)
7. Hwang, F.K.: Group testing with a dilution effect. Biometrika **63**, 671–673 (1976)
8. Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., Pilcher, D.: Comparison of group testing algorithms for case identification in the presence of testing errors. Biometrics **63**, 1152–1163 (2007)

9. Kim, H., Hudgens, M.: Three-dimensional array-based group testing algorithms. Biometrics **65**, 903–910 (2009)
10. Lancaster, V.A., Keller-McNulty, S.: A review of composite sampling methods. J. Am. Stat. Assoc. **93**, 1216–1230 (1998)
11. Liu, S.C., Chiang, K.S., Lin, C.H., Chung, W.C., Lin, S.H., Yang, L.C.: Cost analysis in choosing group size when group testing for Potato virus Y in the presence of classification errors. Ann. Appl. Biol. **159**, 491–502 (2011)
12. Loyer, M.W.: Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. **37**, 57–59 (1983)
13. Martins, J.P., Felgueiras, M., Santos, R.: Meta-analysis techniques applied in prevalence rate estimation. Discuss. Math. Probab. Stat. **33**, 79–97 (2013)
14. Santos, R., Pestana, D., Martins, J.P.: Extensions of Dorfman's theory. In: Oliveira, P.E., Graa, M., Henriques, C., Vichi, M. (eds.) Studies in Theoretical and Applied Statistics, Recent Developments in Modeling and Applications in Statistics, pp. 179–189. Springer, New York (2013)
15. Sobel, M., Elashoff, R.M.: Group testing with a new goal, estimation. Biometrika **62**, 181–193 (1975)
16. Sterret, A.: On the detection of defective members of large populations. Ann. Math. Stat. **28**, 1033–1036 (1957)
17. Wein, L.M., Zenios, S.A.: Pooled testing for HIV screening: capturing the dilution effect. Oper. Res. **44**, 543–569 (1996)
18. Zenios, S., Wein, L.: Pooled testing for HIV prevalence estimation exploiting the dilution effect. Stat. Med. **17**, 1447–1467 (1998)

On Intra-individual Variations in Hair Minerals in Relation to Epidemiological Risk Assessment of Atopic Dermatitis

Tomomi Yamada, Todd Saunders, Tsuyoshi Nakamura, Koichiro Sera and Yoshiaki Nose

Abstract We have conducted a cohort study of 834-mother-infant pairs to determine the association between hair minerals at one month and the onset of atopic dermatitis (AD) at ten months after birth. Thirty-two minerals were measured by PIXE (particle induced X-ray emission) method. (Yamada et al., J. Trace Elem. Med. Bio. 27, 126-131, 2013, [11]) described a logistic model with explanatory variables Selenium (Se), Strontium (Sr) and a family history of AD whose performance in predicting the risk of AD was far better than that of any similar study. However, as discussed in (Saunders et al., Biometrie und Medizinische Informatik Greifswalder Seminarberichte, 18, 127-139, 2011, [9]), intra-individual variations in those minerals were large and could have degraded the regression coefficients of Sr and Se in the logistic model. Therefore, (Yamada et al., Biometrie und Medizinische Informatik Greifswalder Seminarberichte, 2013, [12]) examined the intra-individual variations of Sr levels in the mothers (Mother-Sr) assuming log-normality and obtained a regression

T. Yamada (✉)
Graduate School of Medicine, Osaka University, 2-2 Yamadaoka, Suita,
Osaka 565-0871, Japan
e-mail: t-yamada@doc.medic.mie-u.ac.jp, t-yamada@stat.med.osaka-u.ac.jp

T. Saunders
Graduate School of Medicine, Nagasaki University, 1-12-4 Sakamoto,
Nagasaki 852-8523, Japan
e-mail: styer2000@hotmail.com

T. Nakamura
Graduate School of Science and Engineering, Chuo University,
1-13-27 Kasuga, Bunkyo, Tokyo 112-0003, Japan
e-mail: naka@nagasaki-u.ac.jp

K. Sera
Cyclotron Research Center, Iwate Medical University,
348-58 Tomegamori, Takizawa, Iwate 020-0173, Japan
e-mail: ksera@iwate-med.ac.jp

Y. Nose
Graduate School, Kumamoto Health Science University,
325 Izumimachi Kita-ku, Kumamoto 861-5533, Japan
e-mail: nosyosi@oct-net.ne.jp

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_9

coefficient of Mother-Sr corrected for the variations. This paper addresses Sr levels in the babies (Baby-Sr) which are not distributed as log-normal and require more sophisticated modeling of the variations. Here we elaborate on the “true-equivalent sample” (TES) method, developed in (Yamada et al., Biometrie und Medizinische Informatik Greifswalder Seminarberichte, 2013, [12]) and determine the distribution of Baby-Sr. The revised TES method presented here will be useful for determining the distribution type for minerals whose distributions are zero-inflated, thereby obtaining a risk estimate corrected for the intra-individual variations. This will allow hair mineral data to play a more important role in medical and epidemiological research.

Keywords Atopic dermatitis · Risk assessment · Epidemiology · Hair minerals · Intra-individual variation

1 Introduction

In 2005 part one of a multi-stage cohort study was started to determine what minerals, and in what amounts, could be measured in human hair, and the relationship of these mineral varieties and volumes to the risk of atopic dermatitis (AD).

Hair mineral concentrations have been used as biomarkers to measure such diverse substances as tobacco [1], mercury [8] and cortisol [7]. Mineral concentrations in hair are regarded as ideal biomarkers to measure individual exposure to elements and the Environmental Protection Agency (US EPA) considers scalp hair a suitable biological sample for estimating the body burden of trace elements [6], because hair incorporates elements from the blood at a relatively constant rate and its composition reflects the concentration of elements in blood at the time of formation [2, 5].

We took hair samples from 842 infants and their mothers in Fukuoka at the national one-month health checkup to measure hair minerals using proton induced X-ray emission (PIXE) method. Association between the hair mineral measurements at the one-month checkup and the onset of AD diagnosed by pediatricians at the ten-month checkup was assessed to obtain a model for detecting infants at high risk of developing AD with the goal of primary prevention of the disease.

Of the 32 minerals measured only Selenium (Se) and Strontium (Sr) showed statistically significant associations with the onset of AD. These mineral amounts together with individual AD family history were incorporated into a logistic model to predict the risk of AD development, which provided far better performance than any models presented in the literature to date [11]. However, large intra-individual variations among individual strands and measurement location along the strand can degrade the association to the null [9]. Currently no correction method has been developed for logistic models with covariates subject to continuous measurement errors [3] making clinical use of hair minerals problematic.

Yamada et al. [12] randomly sampled 86 mothers’ hair strands from the original cohort sample ($n = 834$) and Sr was measured at two points by PIXE for each subject. They found that the distribution of Mother-Sr was approximately log-normal, and the reliability index was approximately equal to 0.6. They also estimated a true distribution of Mother-Sr after removing the intra-individual variations. To obtain these results, they developed a method termed "true-equivalent sample (TES) method". In 2012, to perform a similar study on Baby-Sr, we resampled 208 then 6-year-old children from our original cohort sample. The objective of this report is to describe the methods and results of the study, which, we believe, will allow hair mineral data to be more effectively used in medical and epidemiological research.

2 Methods

2.1 Intra-individual Variance

In October 2012, we sent the results of part one of the cohort study to the 834 mothers with a letter requesting hair samples from their then six-year olds. We received hair strand samples from 208 children. We call this sample "the validation sample" to distinguish it from the original cohort sample. Each child's hair strands were divided into two specimens for PIXE analysis to obtain two independent measurements, X_{1i} and X_{2i}, from ith child. Since X_1's and X_2's were measured at virtually the same periods, we consider a simple random effects model:

$$X_{1i} = Z_i + \varepsilon_{1i} \quad \text{and} \quad X_{2i} = Z_i + \varepsilon_{2i} \quad \text{with} \quad \varepsilon_{1i},\ \varepsilon_{2i} \sim N(0, \sigma_e^2) \tag{1}$$

where Z_i, termed *true* or *exact* value, denotes the mineral amount averages from all locations among the hair strands of ith subject and σ_e^2 the intra-individual variance, or the measurement error variance. MLE obtained from applying the logistic model $logit(X) = \alpha + \beta X$ to X will be denoted by β_X and that from $logit(Z) = \alpha + \beta Z$ to Z by β_Z. β_X and β_Z are usually referred to as the naive and corrected regression coefficients, respectively [3].

The proportion of the variance of the true value to that of the observed value

$$\lambda = V(Z)/V(X) = V(Z)/\{V(Z) + \sigma_e^2\},$$

is usually termed as "the reliability index" [3, 10]. An estimate for the intra-individual variance and the reliability index are calculated from the validation sample as

$$\sigma_e^2 = \sum_{i=1}^{m} (X_{1i} - X_{2i})^2/2m \quad \text{and} \quad \lambda = \{V(X) - \sigma_e^2\}/V(X),$$

respectively. $V(X)$ is estimated as a pooled variance from the two samples X_1 and X_2. The value of λ of a random sample from a population is the same for that of the population (Appendix). This property is used to transport the estimate for the measurement error variance σ_e^2 to that of the cohort sample, say σ_E^2.

Let V and σ_E^2 denote the variance and the intra-individual variance of the measurements of the 834 infants at one month, respectively. As previously mentioned, it is expected that the reliability index λ obtained from the validation sample is approximately equal to that from the cohort sample. This consideration leads to the equation

$$\sigma_e^2 \, : \, V(X) \, = \, \sigma_E^2 \, : \, V \tag{2}$$

Assigning the estimates of $V(X)$, V and σ_e^2, we obtain an estimate of σ_E^2.

2.2 *Simulation*

2.2.1 True-Equivalent Sample and Observed-Equivalent Sample

Sr was larger than 1 for almost all mothers and distributed as approximately lognormal for the cohort sample (n = 834). This does not hold, however, for the sample from the children. In fact, about 21.5 % of them had less than 1 ppm. Since 0.1 ∼ 1 is regarded as the detection limit of PIXE method, we replaced Sr-levels less than 0.1 with 0.1 and then applied the square root transformation. The resulting figures are used as X 's in (1). Then, as performed for mothers' Sr [12], a true-equivalent sample was obtained as follows.

Hereafter, for the sake of brevity, Sr-level will be omitted when considered understandable. We denote the variance of a sample by *var* and that of a population by V. Let Z_A and Z_N denote the unobserved true values and X_A and X_N the observed error-prone values for AD and non-AD respectively. Since $var(X_A) = var(X_N)$ approximately holds, we assume

$$V(Z_A) = V(Z_N) = V \quad \text{and} \quad X = Z + \sigma_E \varepsilon$$

for AD and non-AD, that is "non-differential" error assumption usually employed in cohort data analysis [3]. It follows that

$$E(Z_A) = E(X_A), E(Z_N) = E(X_N), \quad \text{and} \quad V(X) = V(Z) + \sigma_E^2$$

for both AD and non-AD. Assigning observed values of $E(X_A)$, $E(X_N)$, $V(X)$ and σ_E^2 to the equations will result in estimates for $E(Z_A)$, $E(Z_N)$ and $V(Z) = V(X) - \sigma_E^2$ for both AD and non-AD samples.

Since X appears to follow approximately the Weibull distribution with two parameters α, scale, and β, shape, we obtain the parameter values of the best fit Weibull model for each sample. It is a practical advantage in the following analysis that the Weibull model is determined by the mean E and variance V as well as the values of α and β.

Weibull random numbers with $E(Z_A)$ and $V(Z)$ and also those with $E(Z_N)$ and $V(Z)$ are generated for AD and non-AD samples respectively. Each sample will be referred to as a "true-equivalent" sample (TES) and denoted by Z^*. Then, $X^* = Z^* + \sigma_E \varepsilon$, termed an "observed-equivalent" sample, was generated for both AD and non-AD samples. The distribution of X should be approximately equal to that of X^* if the TES method is valid.

Let β_{Z^*} denote the MLE obtained from applying the logistic model

$$logit(Z^*) = \alpha + \beta Z^*$$

and

$$logit(X^*) = \alpha + \beta X^*$$

to a true-equivalent sample Z^* and an observed-equivalent sample X^*, respectively. If an observed equivalent sample X^* is in fact approximately equal to X, it is regarded as evidence that the distribution of Z is approximately equal to that of Z^*. In that case, β_{X^*} and β_{Z^*} should be approximately equal to β_X and β_Z, respectively. Hereafter, β_{Z^*} is termed as a TES estimate for β_Z.

2.2.2 SIMEX

We also obtained the SIMEX estimate for β_Z [4, 12] to compare as follows. First, we generate further contaminated surrogates $X(\theta)$:

$$X(\theta) = X + \theta \sigma_E \varepsilon, \quad \varepsilon \sim N(0, 1) \tag{3}$$

where $\theta > 0$ is a pre-assigned constant. The values of θ are usually 0.5, 1, 1.5 and 2, but may depend on the case. Applying the logistic model

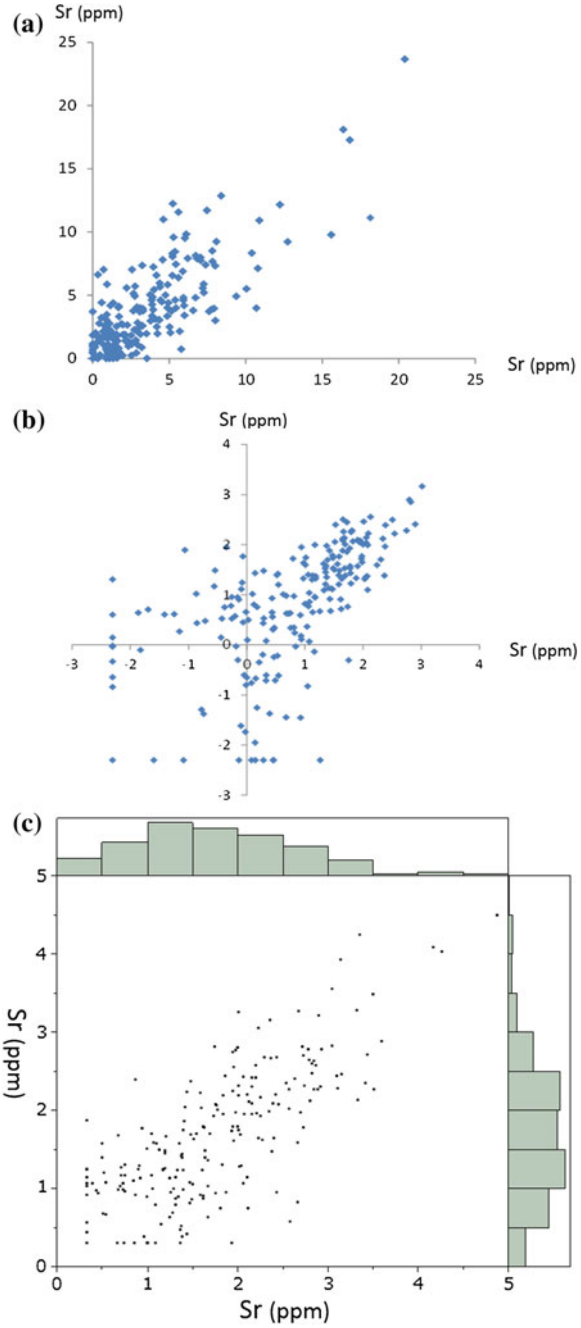
$$logit(X(\theta)) = \alpha + \beta X(\theta),$$

we obtain the MLE for β. This step is iterated 400 times so that we obtain 400 MLE's for β for each θ. The average of them will be denoted by β_θ. Then the so-called "Extrapolation step" makes a scatter plot for β_θ versus θ to determine the functional relationship between them and extrapolates it to $\theta = -1$. The value of β corresponding to $\theta = -1$, denoted by β_{-1}, is a SIMEX estimate for β_Z.

3 Results and Discussion

Figure 1a presents a scatter plot for the 1st versus 2nd Sr measurements of the validation sample (n = 208). This shows a tendency for differences between the two measurements to be greater for larger measurements, When this phenomenon is

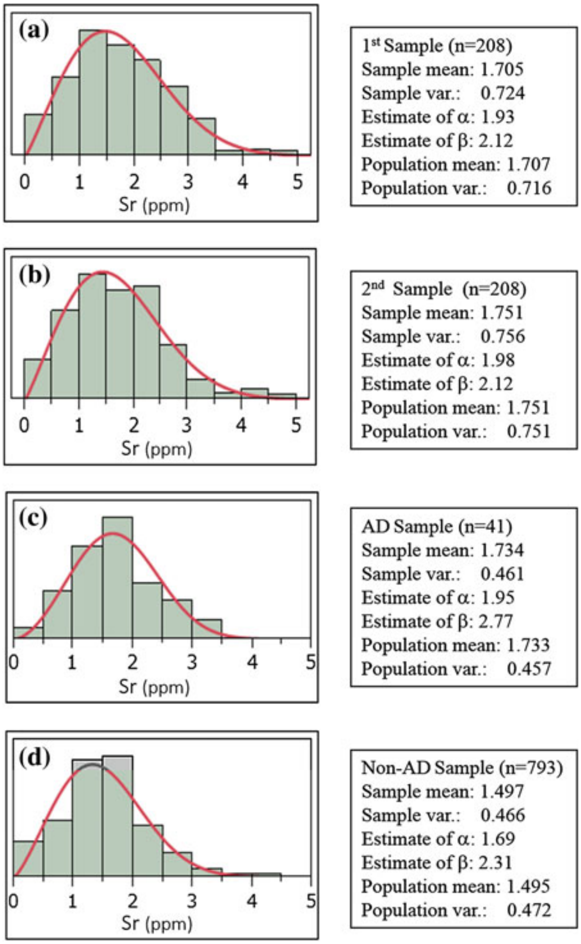
Fig. 1 Scatter plot for two independent measurements of Sr from 208 children. The X-axis and Y-axis represent 1st and 2nd measurements of Sr, respectively (**a**), that of log(Sr) (**b**) and that of $\sqrt{\text{Sr}}$ with histogram (**c**)



observed, it is customary to use log-transformed values, which indeed work fine with mothers' Sr [11]. However, Fig. 1b reveals that log(Sr) is not appropriate for the current data since small differences in Sr < 1 are exaggerated when compared to those for Sr > 1 and a large number of measurements are less than 1.

On the other hand, the square-root transformation $\sqrt{\mathrm{Sr}}$ substantially corrects the defects of log(Sr) (Fig. 1c). Hereafter, X denotes $\sqrt{\mathrm{Sr}}$ for each subject, while Z will denote the average of X over all hair strands of each subject. Z is considered as the *true* or *exact* value, as mentioned in Methods.

We discovered that a normal distribution did not fit well to $\sqrt{\mathrm{Sr}}$. Instead, the two-parameter Weibull model appears to fit reasonably well. Figure 2 shows the histograms with a fitted Weibull density function with estimated parameter values and sample means and variances for each X_1 and X_2. The population mean and variance of the Weibull model determined by the estimated parameter values α and



1st Sample (n=208)
Sample mean: 1.705
Sample var.: 0.724
Estimate of α: 1.93
Estimate of β: 2.12
Population mean: 1.707
Population var.: 0.716

2nd Sample (n=208)
Sample mean: 1.751
Sample var.: 0.756
Estimate of α: 1.98
Estimate of β: 2.12
Population mean: 1.751
Population var.: 0.751

AD Sample (n=41)
Sample mean: 1.734
Sample var.: 0.461
Estimate of α: 1.95
Estimate of β: 2.77
Population mean: 1.733
Population var.: 0.457

Non-AD Sample (n=793)
Sample mean: 1.497
Sample var.: 0.466
Estimate of α: 1.69
Estimate of β: 2.31
Population mean: 1.495
Population var.: 0.472

Fig. 2 Results of fitting Weibull model to the 1st (**a**) and 2nd (**b**) validation samples, and the cohort samples AD (**c**) and non-AD (**d**)

β agree well with the samples' means and variances calculated directly from the observed values for the samples of Fig. 2. It is thus concluded that those four samples are approximately distributed as Weibull.

The intra-individual variance σ_e^2 was estimated to be 0.191. As described in Fig. 2, $var(X_1) = 0.724$, $var(X_2) = 0.751$, the pooled variance of X_1 and X_2 is 0.746 and therefore, the reliability index $\lambda = (0.746 - 0.191)/0.746 = 0.744$. On the other hand, the pooled variance of AD (n = 41) and non-AD (n = 793) is 0.466 and thus it follows from (2) that

$$\sigma_E^2 \, : \, 0.466 = 0.191 \, : \, 0.744.$$

Solving the equation yields $\sigma_E^2 = 0.120$.

Summarizing the results, the mean of the AD and non-AD samples are 1.734 and 1.497 respectively, and their common variance is estimated as $0.466 - 0.12 = 0.346$. Then, we generated a true-equivalent sample for AD (n = 41) and for non-AD (n = 793) following the Weibull distribution determined from the means 1.734 and 1.497, respectively, with the same variance 0.346.

To confirm the validity of the true-equivalent sample, we generated an observed-equivalent sample by adding $\sigma_e \varepsilon$, where $\varepsilon \sim N(0, 1)$, to each subject of the true-equivalent AD and non-AD samples. The chosen sample size is large (n = 2000) since the sample distribution should be close to the population distribution. Figure 3 presents a best fit Weibull model for each sample. The estimated parameter values of the observed-equivalent sample for AD and non-AD in Fig. 3 are approximately equal to those of the observed samples in Fig. 2, respectively. The results support the validity of the TES method for the applications.

As for the regression coefficients, the naive estimate β_X is 0.465. We generated 500 independent true-equivalent samples Z^* to obtain 500 independent β_{Z^*}. The

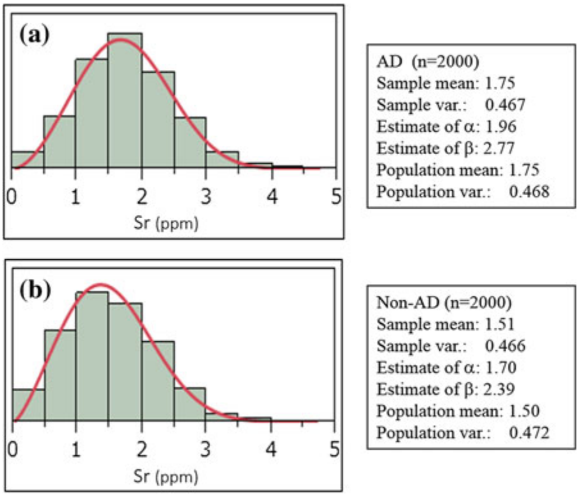


Fig. 3 Fitting the Weibull model to the observed-equivalent samples (n = 2000) for AD (**a**) and non-AD (**b**)

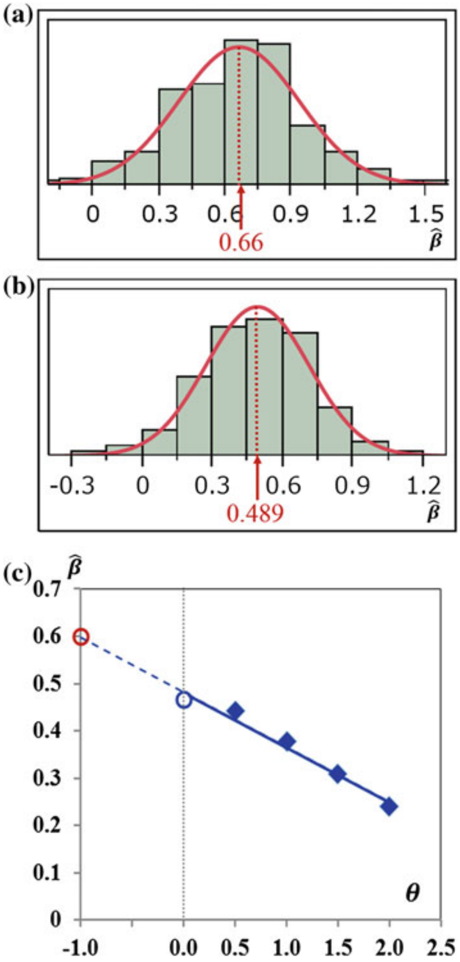


Fig. 4 Histogram of 500 β_{Z^*} whose average 0.66 is an estimate for the corrected estimate β_Z (**a**), histogram of 500 β_{X^*} whose average 0.489 corresponds to the naive estimate β_X (**b**). Figure (**c**) illustrates the SIMEX estimate $\beta_{-1} = 0.60$

histogram is shown in Fig. 4a and the average 0.66 is the TES estimate for β_Z. Similarly, we obtained 500 independent β_{X^*} that result in Fig. 4b. The average 0.489 is close to $\beta_X = 0.465$, supporting the validity of the TES method developed for this study. Figure 4c shows that the SIMEX estimate β_{-1} for β_Z is 0.60, again slightly conservative as discussed in Cook et al. [4, 12]. It is concluded that the TES method developed for the study is useful in determining the distribution type of minerals which is a crucial issue toward using hair minerals for medical and epidemiological research.

Acknowledgments We are grateful to the pediatricians and gynecologists who created and participated in this project. Thanks are also due to the mothers and infants who donated hair strands. We would also like to thank Ms. Tomoko Maeda and Ms. Yurika Kondo for preparing the target samples, and Mr. Satoshi Kinoshita for programming.

Appendix

Transportation of Measurement Error Variance of Sr

Let X_{ki} denote an observed value of Sr of kth experiment for ith subject. We assume the following random effect model:

$$X_{ki} = \tau_k + \rho_k(Z_i + \varepsilon_{ki}) = \tau_k + \rho_k Z_i + \rho_k \varepsilon_{ki}, \quad k = 1, 2\,; \quad i = 1, \cdots, n$$

Z_i is the true value of ith subject,
τ_k, ρ_k represent calibration effects for the kth experiment,
ε's are independent random variables with $E(\varepsilon_{ki}) = 0$ and $V(\varepsilon_{ki}) = \sigma_e^2$,
where ε's represent intra-individual variations

The ratio of the variance of X explained by the variation of Z to the variance of X is defined as a reliability index of X and usually denoted as λ [3, 10].

As for the kth experiment, since

$$Var(X_{ki} \mid k) = \rho_k^2 Var(Z) + \rho_k^2 \sigma_e^2,$$

the variance of X due to Z is $\rho_k^2 Var(Z)$ and that due to ε, or the measurement error variance, $\rho_k^2 \sigma_e^2$. Therefore,

$$\lambda = \rho_k^2 Var(Z)/Var(X_k) = Var(Z)/\{Var(Z) + \sigma_e^2\}$$

Thus, λ is independent of the calibration effect τ_k and ρ_k. That is, λ is a parameter intrinsic to the sample determined by the inter and intra individual variance $Var(Z)$ and σ_e^2, respectively. Hereafter, for the sake of notational simplicity, $Var(X_{ki} \mid k)$ and $E(X_{ki} \mid k)$ will be simply denoted by $Var(X_k)$ and $E(X_k)$, respectively.

It is straightforward to show that

$$E(X_k) = \tau_k + \rho_k \overline{Z},$$

$$X_{ki} - E(X_k) = \rho_k(Z_i - \overline{Z} + \varepsilon_{ki})$$

and

$$(\rho_2/\rho_1)^2 = Var(X_2)/Var(X_1).$$

Define

$$X_{1i}^{\alpha} = E(X_2) + (\rho_2/\rho_1)\{X_{1i} - E(X_1)\}.$$

Then,

$$\begin{aligned} X_{1i}^{\alpha} - X_{2i} &= E(X_2) + (\rho_2/\rho_1)\{X_{1i} - E(X_1)\} - X_{2i} \\ &= (\rho_2/\rho_1)[\rho_1\{Z_i - \overline{Z} + \varepsilon_{1i}\} - \rho_2(Z_i - \overline{Z}) - \varepsilon_{2i}] \\ &= \rho(\varepsilon_{1i} - \varepsilon_{2i}). \end{aligned}$$

Let $D^{\alpha} = \sum_i (X_{1i}^{\alpha} - X_{2i})^2/2m$, then

$$E(D^{\alpha}) = \rho_2^2 E(\sum_i (\varepsilon_{1i} - \varepsilon_{2i})^2/2m) = \rho_2^2\sigma_e^2$$

Thus, D^{α} is an unbiased estimate of the measurement error variance of X_k. Therefore, let

$$X_{1i}^{*} = \overline{X}_2 + \{Var(X_2)/Var(X_1)\}^{1/2}(X_{1i} - \overline{X}_1)$$

then

$$D^{*} = \sum_i (X_{1i}^{*} - X_{2i})^2/2m$$

is an asymptotically unbiased estimate for $\rho_2^2\sigma_e^2$.

References

1. Al-Delaimy, W.K.: Hair as a biomarker for exposure to tobacco smoke. Tob Control **11**, 176–182 (2002)
2. Bland, J.: Hair Tissue Mineral Analysis: An Emergent Diagnostic Technique. Thorson Publishers Inc., New York (1984)
3. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement Error in Nonlinear Models. Chapman and Hall, London (1995)
4. Cook, J.R., Stefanski, L.A.: Simulation-extrapolation estimation in parametric measurement error models. J. Am. Stat. Assoc. **89**(428), 1314–1328 (1994)
5. Gibson, R.S.: Hair as a biopsy material for the assessment of trace element status in infancy. A review. J. Hum. Nutr. **34**(6), 405–416 (1980)
6. Jenkins, D.W.: Biological monitoring of toxic trace metals. US Environmental Protection Agency. EPA-6000/S3-80-90:10 (1981)
7. Karlén, J., Ludvigsson, J., Frostell, A., Theodorsson, E., Faresjö, T.: Cortisol in hair measured in young adults-a biomarker of major life stressors? BMC Clin. Pathol. **11**, 12 (2011)
8. Li, Y.F., Chen, C., Li, B., Wang, J., Gao, Y., Zhao, Y., Chai, Z.: Scalp hair as a biomarker in environmental and occupational mercury exposed populations: suitable or not? Environ. Res. **107**(1), 39–44 (2008)
9. Saunders, T., Kuroda, S., Makie, T., Sera, K., Yamada, T., Nakamura, T., Nose, Y.: Effect of biological variability incidental to PIXE-Hair minerals on risk analysis of atopic dermatitis. Biometrie und Medizinische Informatik Greifswalder Seminarberichte **18**, 127–139 (2011)

10. Snedecor, G.W., Cochran, W.G.: Statistical Methods, 6th edn. Iowa State University Press, Ames (1967)
11. Yamada, T., Saunders, T., Kuroda, S., Sera, K., Nakamura, T., Takatsuji, T., Hara, T., Nose, Y.: Assoc, Fukuoka, Obstetr. Gynecol. and Pediatr. Cohort study for prevention of atopic dermatitis using hair mineral contents. J. Trace Elem. Med. Biol. **27**, 126–131 (2013)
12. Yamada, T., Saunders, T., Sera, K., Nakamura, T., Nose, Y.: On intra-individual variations in hair minerals measured by PIXE in relation to epidemiological risk assessment of atopic dermatitis. Biometrie und Medizinische Informatik Greifswalder Seminarberichte. (2013) (in press)

Assessing Risk Factors for Periodontitis Using Multivariable Regression Analysis

J.A. Lobo Pereira, Maria Cristina Ferreira and Teresa A. Oliveira

Abstract Risk is associated with all areas of Life, and studies designed to decrease it play a key role, particularly in what concerns Individual Health. Considering Epidemiological Research, the identification of Risk Factors is crucial to select prevention actions in order to improve Public Health Systems. The aim of this work is to identify the main Risk Factors for periodontal disease, using Multivariate Statistical Methods, since according to the literature these are the most important tools to assess associations and interactions between different putative risk factors and a given health condition. An application of Generalized Linear Models (GLM) with probit link function was performed to assess the impact of socio-demographic, biochemical and behavioural factors on periodontal status. We analysed data collected from a sample of 79 individuals with chronic periodontal disease, attending the clinic of Porto Dentistry School. We found a significant association between extensive periodontitis and decreased levels of high density lipoproteins (HDL). We believe public health efforts on prevention, including education of the population at risk, are highly recommended in order to decrease early causes of the illness.

Keywords Periodontitis · Risk factors · Multivariate regression

J.A. Lobo Pereira (✉)
Department of Periodontology, FMUP and Universidade Aberta, Lisbon, Portugal
e-mail: lobopereiramail@gmail.com

M.C. Ferreira
Master of MEMC, Universidade Aberta, Lisbon, Portugal
e-mail: cristina@gmail.com

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

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_10

1 Introduction

In healthcare studies, harm is defined as a negative safety and health consequence and hazard is an existing situation consisting of any source of harm or adverse effect on the individual under certain conditions. Risk is a potential harm anticipated in the future, a form of probability that an individual will experience an unwanted health effect when exposed to a certain hazard. The important thing is that in many situations it can be reduced and even avoid, under some prevention actions, and for this it is crucial to identify the main risk factors under de particular study circumstances. In Health Sciences, a risk factor is usually a variable associated with increase likelihood of developing a disease or another adverse health outcome. Some examples of important risk factors in this area are poor hygiene, excess weight, unsafe sex, high blood pressure, smoking and alcohol consumption.

In [13] the authors using regression, linear and logistic models, assessed the relevance of potential risk factors for periodontal disease, such as: Age, Gender, Diabetic Status, Education, Smoking Status and Plaque Index. The study was based on a sample of real data, collected as part of an investigation carried out in the area of Dental Medicine at the Faculty of Dental Medicine of University of Porto, Portugal. In this work the authors performed an application of Generalized Linear Models (GLM), with probit link function, in order to assess the impact of socio-demographic, biochemical and behavioral factors on periodontal status.

Periodontitis is a multifactorial inflammatory disease that affects the supporting tissues of the tooth and is characterized by destruction of alveolar bone and loss of attachment, which is influenced by genetic and environmental risk factors [12, 15].

Periodontal condition results from interaction between host and microbiologic factors that can be deleterious or protective. The factors generally accepted as influent in periodontitis initiation, progression and severity, considering its nature, can be classified in systemic, behavioral, socioeconomic, and microbiologic [6].

The clinical assessment of periodontal support level is usually made by estimating the distance from cement enamel junction (CEJ) to the periodontal probe tip located in the bottom of the sulcus/pocket near the adherence (A) (Fig. 1).

This estimate is the clinical adherence level (CAL) and it is one the most used surrogate markers of periodontal disease to estimate bone loss, (alveolar bone level). Under pristine periodontal conditions (side I of Fig. 1), the distance (CEJ-CAL) comprises the connective tissue attachment and the totality or part of the epithelial attachment, with a 2–3 mm width [18].

Periodontal disease can be characterized by extension, severity and progression rate. Extension is defined by the number of teeth periodontally affected or by the percentage of sites with bone destruction. Severity is related with the proportion of vertical bone loss around a tooth (CAL over total root length). The progression rate is the speed at which vertical bone loss occurs. However the presence of a CAL larger than 4 mm strongly suggests periodontitis (considering normal CEJ-CAL2-3 mm), but does not necessarily correlated with periodontitis severity.

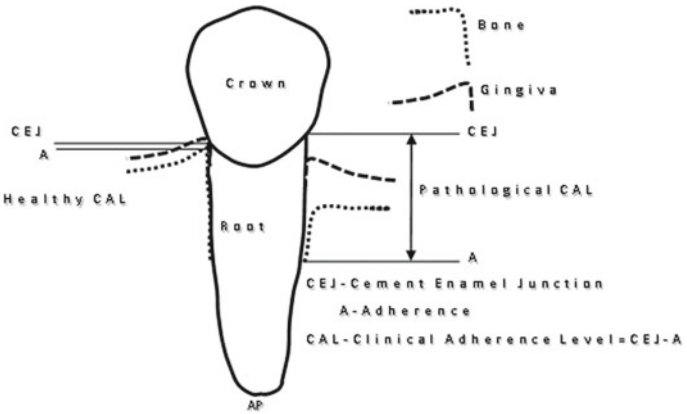


Fig. 1 Tooth representation

Considering the etiology of periodontitis is a complex combination of several conditions, multivariable statistical analysis is a useful tool to find out associations and interactions between different potential risk/protective factors and periodontitis [11, 20]. In this context the choice of multivariable analysis is properly grounded in theoretical and epidemiologic knowledge, in order for the researcher to attain a correct perception of the different factors considered.

Generalized linear models (GLMs) represent a class of multivariable regression models which allows generalization of the linear regression approach to accommodate many types of response variables and distributions, see [9, 10, 14].

The three main principles of GLMs are the existence of a sample of independent responses variables $Y_1, \dots Y_n$, from an exponential family; a linear predictor $\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i$ of the response variable Y; and the mean response which depends on the linear predictor through a link function g. Considering the described principles of GLMs three components are necessary: a random or stochastic, a systematic and a link function. The random component is the response and the associated probability distribution $(g(\mu) = \mu)$. The systematic component, which includes explanatory variables (x_i) combined linearly with the coefficients β to form the linear predictor (η) and relationships among them the link function, specifies the relationship between the systematic component or linear predictor and the mean response $\mu = \beta_0 + \beta_1 x_1 + \dots + \beta_i x_i$. It is the link function $(g(\mu))$ that allows generalization of the linear models for count, binomial and percent data thus ensuring linearity and constraining the predictions to be within a range of possible values, see [8].

Logistic regression is a specific branch of GLM applicable in several fields such as epidemiology, medical research, banking, market research and social research. One of its advantages is that the interpretation of the measure is possible through the Odds Ratios (OR), functions of the model parameters.

The aim of this study is to assess the association between some variables of interest and the extension of periodontitis through GLM models.

2 Methods

Participants

Seventy nine systemically healthy individuals (38 males and 41 females) with chronic periodontal disease varying from localized to generalized forms, who attended the clinic of Porto Dentistry School, participated in this study. All participants were Caucasian with mean age of 50.81 years and standard deviation 15.73.

Materials

A periodontal probe marked in millimeters was used to assess CAL (according to Fig. 1) in 6 sites per tooth, excluding third molars. During this procedure the hemorrhagic index was also estimates. The O'Leary Plaque index was calculated using a plaque disclosure tablet.

Procedures

Informed consent was obtained from all participants. Data on Social Demographies factors and risk behaviours for periodontitis were collected through a questionnaire fill out by the patients. The evaluation of periodontal status was performed by the same experienced periodontologist, under good clinical conditions and the body mass index (BMI) was assessed at the same appointment. Biochemical data was obtained from patients files.

The research protocol was approved by the ethical committee of Porto Dentistry School.

Design and Analyses

This retrospective study was designed to model the effect of age (Age), high density lipoproteins (HDL), tobacco smoking (Tob), and body mass index (BMI), explanatory variables considered biologically significant and associated with periodontitis by epidemiological studies [4, 7, 17, 19]. Once in Portugal research in this area is very scarce, this work is therefore justified. The dependent variable (Cond4_25) was defined as the percentage of sites equal or greater than 25 % with attachment loss (AL) equal or greater than 4mm. All variables, qualitative and quantitative, were dichotomized according to Table 1.

The collinearity among pairs of independent variables was assessed through a Chi-squared test with Yates' continuity correction.

Logistic regression with a link function logit was chosen due to the binary type of response and explanatory variables. Moreover we present the results as odds ratios

Table 1 Stratification of variables

Cond4_25		Age (years)		HDL (mg/dl)		Tob		*BMI* (>25 Kg/m^2)	
$\geq$ 25 % of sites with AL $\geq$ 4 mm	1	≥ 50	1	≥ 50	1	Smoker	1	Overweight	1
$<$ 25 % of sites with AL $\geq$ 4 mm	0	< 50	0	< 50	0	Non smoker	0	Under or normal weight	0

(ORs) and respective confidence intervals. The OR were obtained by exponentiation the logits estimates $(exp(\beta))$.

In order to find the model which would best approximate reality given the data and minimize the loss of information, our approach to model building followed the three principles highlighted by [2]: several working hypotheses, simplicity and parsimony and strength of evidence [2].

According to the principle of different working hypotheses, a series of GLM models were built based on current knowledge on periodontitis epidemiology. We started to build the full model with interactions between independent variables. However they were removed from the model due to their absence of significance and bad quality of the models obtained.

A number of reduced models with different combinations of covariates were obtained by deleting a term (Age, HDL, Tob and BMI), already in the model or adding terms.

From the set of models obtained in this phase, we selected those which represent the better compromise between model bias and variance, where bias corresponds refers to the difference between the obtained MLEs estimate values of the parameter $(\widehat{\beta_i})$ and the respective unknown true value (β_i), and variance reflects the precision of these estimates. As pointed out by [3], a model with too many variables will have low precision whereas a model with too few variables will be biased [3] highlighting the importance of the parsimony in a balanced model building.

To measure of the strength of evidence for each model we used Akaike's information criterion (AIC) defined as $-2(log - L + 2k)$ where k is the number of estimated parameters (including β_0) and L the model and likelihood [1], establish a relationship between the maximum likelihood, which is the estimation method used in this statistical analysis, and the Kullback-Leibler divergence, information that represent the loss of information when approximating reality. This information-theoretic approach allow us to manage the three principles described in the model building phase. In order to obtain a set of statistical parsimonious models the selection of terms for inclusion or deletion was based on AIC . Furthermore the AIC approach yields consistent results and is independent of the order in which the models are computed [2, 3].

Each model was compared with the best model (minimum AIC) by the value of delta AIC (Δ_i), computed as follows: $\Delta_i = AIC_i - minAIC$, where AIC_i is the AIC value for model i. As suggested by [3], values of $\Delta_i < 2$ indicate substantial evidence for the model, values between 3 and 7 indicate the model is considerably less likely, and $\Delta_i > 10$ indicates the model is very unlikely.

Usually the model acceptance is based only on the raw AIC values, making it difficult to unambiguously interpret the observed AIC differences in terms of a continuous measure such as probability.

To avoid this difficulty we used the Akaike weights (w_i) computed as: $w_i = \frac{exp(\frac{-\Delta_i}{2})}{\sum_{r=1}^{R}(-\frac{\Delta_r}{2})}$, where w_i represents the ratio of delta AIC (Δ_i) values for each model relative to the whole set of R candidate models. With w_i, we could directly the directly conditional probabilities for each model.

We also compared the likelihoods of our reduced models: $(L_1(\beta_1; y), L_2(\beta_2; y), \ldots, L_m(\beta_m; y))$ $(m = 1, 2, \ldots, m)$ with the fitted saturated model $(L_s(\Psi; y))$ or equivalently $l_S(\Psi) \equiv logL_S(; y)$ and $L_m(\beta_m) \equiv logL_m(\beta_m; y)$ to test of the link function fit and linear predictor (adequacy of the model), $L_S(\Psi; y) \geq L_m(\beta_m; y)$ because the model under study is a special case of the saturated model.

For suitable models the condition $l_S(\Psi) \approx l_m(\beta)$ would be expected. The deviance or likelihood test ratio statistic D used in our work is defined as $D = 2[l_S(\widehat{\Psi}) - l(\widehat{\beta}_m)]$ where $\widehat{\Psi}$ and $\widehat{\beta}_m$ are maximum likelihood estimates of the saturated model and each m proposed model, respectively. After calculated, models' deviances were compared by analysis of variance (ANOVA).

Residuals were also controlled for over or under dispersion. The residuals' deviance index was obtained dividing the residual deviance by the model degrees of freedom. For a binomial distribution of errors acceptable values of dispersion should be close to 1.

The implementation and evaluation of models was made with the open-source statistical *package* R, with 'MASS' and 'car' packages for GLM and diagnostics [5, 16, 19].

3 Results

In this case-control study 19 (24.05 %) patients were diagnosed positive for Cond4_25. The number of negative and positive cases for each covariate. The description of the data is presented in Table 2.

No significant statistical collinearity among covariates was found by the Chi-squared test with Yates' continuity correction (p-value > 0.05) (Table 3).

Table 2 Number of positive (1) and negative (0) cases per covariate

	Cond4_25	Age (years)	HDL (mg/dl)	Tob	*BMI* (>25 Kg/m^2)
0	60	33	40	61	40
1	19	18	39	18	39

Table 3 A Chi-squared test with Yates' continuity correction and respective (p-value)

Covariables	HDL (mg/dl)	Tob (no/yes)	*BMI* (>25 Kg/m^2)
Age (years)	0.0679 (0.7944)	0.2847 (0.5937)	2.9924 (0.08365)
HDL (mg/dl)	–	0.362 (0.5474)	0.6534 (0.4189)
Tob (no/yes)	–	–	0 (1)

*Pearson's Chi-squared test with Yates' continuity correction

Table 4 Association between covariates and dependent variable

	Age (years)		HDL (mg/dl)		Tob		*BMI* (>25 Kg/m^2)	
	0	1	0	1	0	1	0	1
Cond4_25 0	27	33	14	46	46	14	34	26
Cond4_25 1	6	13	10	9	15	4	6	13
[a]Chi-squared test	0.5881		4.5532		0		2.6992	
p-value	0.4432		0.0329		1		0.1004	

[a]Pearson's Chi-squared test with Yates' continuity correction; $p<0.05$

Table 5 Summary of the logit models under analysis

	Estimates for models intercepts and covariables and respective (p-values)					
Models	Intercept	Age = 1	HDL = 1	Tob = 1	BMI = 1	Res. Disp.
Na4_25_1	−1.0442 (0.1284)	0.3697 (0.5365)	−1.2484 (0.0284)	−0.2305 (0.7343)	0.891 (0.127)	1.056
Na4_25_2	−1.1233 (0.0832)	0.3869 (0.5160)	−1.2286 (0.0297)		0.896 (0.1250)	1.044
Na4_25_3	−0.9177 (0.1008)		−1.2266 (0.0295)		0.9594 (0.0956)	1.036

Table 6 Akaike Criterion Information analysis

Models	AIC	Δ_i	w_i
Na4_25_1	88.167	3.454	0.109
Na4_25_2	86.284	1.571	0.279
Na4_25_3	84.713	0	0.612

The association between Cond4_25 and each covariate was assessed with the Pearson's Chi-squared test with Yates' continuity correction and showed a significant relationship between low levels of HDL and Cond25 (Table 4) ($p\text{-}value = 0.0329$).

Starting with a model (Na4_25_1) with the four selected covariates we get an AIC and residual deviance dispersion of 1.056. In the second model (Na4_25_2) obtained by dropping the variable Tob, an improvement of AIC and residual deviance dispersion values was observed (86.284 and 1.044 respectively). Further improvement in AIC and residual deviance dispersion values was achieved with model Na4_25_3. The results are shown in Table 5.

The value of $\Delta_i = 1{,}571$ for model Na4_25_2 indicates substantial evidence, and $\Delta_i = 3{,}454$ for Na4_25_1 indicates that the model is considerably less likely (Table 6).

The values of w_i shows that Na4_25_3 yields a higher probability (0.612) to be the best model (Table 6).

The deviances of models were compared with ANOVA. The comparison (Table 7), yields a progressive reduction of deviance values from model Na4_25_1 to Na4_25_3.

Table 7 Models deviance comparison with ANOVA

	Residual degree of freedom	Residual deviance	Degree of freedom	Deviance	*Pr* (>*Chi*)
Na4_25_1	74	78.167			
Na4_25_2	75	78.284	−1	−0.11753	0.7317
Na4_25_3	76	78.713	−1	−0.42877	0.5126

Table 8 Odds ratio for Cond25 with respective confidence intervals

	OR	95 % CI
(Intercept)	0.399	0.123–1.145
$HDL \geq 50\,\text{mg/dl}$	0.293	0.095–0.883
$BMI > 25\,\text{Kg/m}^2$	2.610	0.869–8.561

The growth of residual deviance is marginal and the simultaneous increment of residual degree of freedom leads to a reduction of residual dispersion (Table 5).

Models Na4_25_3 and Na4_25_2 show no significant statistical difference. By interpreting the model selection criteria we considered Na4_25_3 the best model. The maximum likelihood estimates for the intercept and slopes (HDL and BMI) are $\widehat{\beta}_0 = -0.9177$, $\widehat{\beta}_1 = -1.2266$ and $\widehat{\beta}_2 = 0.9594$, which yields the following estimated logistic regression model:

$Na4\_25\_3 = -0.9177 - 1.2266 * HDL + 0.9594 * BMI.$

The estimate of HDL is negative and significant at confidence level of 0.05 indicating a protective role for Cond25. The BMI estimate is positive but only marginally significant suggesting that excess weight is a potential risk factor for Cond25.

The odds ratios (ORs) obtained from estimates, $\widehat{\beta}_1$ and $\widehat{\beta}_2$ (Table 6) show individuals with HDL over 50 mg/dl have approximately one third of chances to be positive for Cond4_25 indicating HDL levels $\geq$ 50 mg/dl can be a protective factor. The OR for BMI suggests a marginal association with Cond25, meaning that overweigh people are approximately 2.6 times more prone to develop Cond25 (Table 8).

4 Conclusions

The selection of variables and statistical methods allied to a judicious and well-grounded selection of models is of crucial importance to provide good quality scientific knowledge. The parameters included in the model must be of biological relevance to the research and assessment of collinearity between them should be conducted carefully.

The process of scientific evidence production must result from team work in order to identify quantitative and qualitative issues in research. Among the most important issues, we highlight the identification of research questions and hypotheses, unit of analysis, random variables (outcomes), proximity between the methodology and an original research question.

From our analysis we conclude that the extension of periodontitis seems to be related to HDL levels, higher levels of which being protective against extensive periodontitis. The opposite effect is suggested to BMI for over weight people.

In future research we intend to explore comparisons of behaviour between the Portuguese reality and other countries, in what concerns the disease main risk factors and prevention.

Acknowledgments Research partially sponsored by national founds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the projects:
(PEst-OE/MAT/UI0006/2011 and PEst-OE/MAT/UI0006/2014).

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in Statistics. Foundations and Basic Theory. Springer Series in Statistics, Perspectives in Statistics, vol. 1, pp. 610–624. Springer, New York (1992)
2. Burnham, K.P., Anderson, D.R.: Kullback-Leibler information as a basis for strong inference in ecological studies. Wildl. Res. **28**, 111–119 (2001)
3. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer, New York (2002)
4. Chaffee, B.W., Weston, S.J.: Association between chronic periodontal disease and obesity: a systematic review and meta-analysis. J Periodontol. **81**(12), 1708–1724 (2010). doi:10.1902/jop.2010.100321
5. Fox, J.: Companion to Applied Regression R Foundation for Statistical Computing. Vienna (2007)
6. Genco, R.J., Borgnakke, W.S.: Risk factors for periodontal disease. Periodontology 2000 **62**(1), 59–94 (2013)
7. Griffiths, R., Barbour, S.: Lipoproteins and lipoprotein metabolism in periodontal disease. Clin. Lipidol. **5**(3), 397–411 (2010)
8. Guisan, A., Edwards, T., Hastie, C.: Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. **157**, 89–100 (2002)
9. Hilbe, J.: Generalized linear models. Am. Stat. Assoc. **48**, 255–265 (1994)
10. Hoffmann, J.P.: Generalized Linear Models: An Applied Approach. Pearson, Boston (2004)
11. Kinane, D.F., Marshall, G.J.: Periodontal manifestations of systemic disease. Aust. Dent. J. **46**(1), 2–12 (2001)
12. Kornman, K.S.: Mapping the pathogenesis of periodontitis: a new look. J Period. **79**(Suppl.), 1560–1568 (2008)
13. Lobo Pereira, J.A., Ferreira, M.C., Oliveira, T.A.: Assessing risk factors for periodontitis using regression. In: Proceedings of ICNAAM 2013, Rhodes Island - Greece, 21–27 September 2013 (in Press)
14. Nelder, J., Wedderburn, R.: Generalized linear models. J. R. Stat. Soc. A. **135**, 370–384 (1972)
15. Page, R.C., Kornman, K.S.: The pathogenesis of human periodontitis: an introduction. Periodontology **2000**(14), 9–11 (1997)

16. RTeam: R Development Core Team. R: a language and environmental for statistical computing. Version 2.8.0. Vol. R Foundation for Statistical Computing. Vienna (2008)
17. Streckfus, C.F., Parsell, D.E., Streckfus, J.E., Pennington, W., Johnson, R.B.: Relationship between oral alveolar bone loss and aging among African-American and Caucasian individuals. Gerontology **45**, 110–114 (1999)
18. Vacek, J.S., Gher, M.E., Assad, D.A., Richardson, A.C., Giambarresi, L.I.: The dimensions of the human dentogingival junction. Int J Periodontics Restor. Dent. **14**(2), 154–165 (1994)
19. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)
20. Ojima, M., Hanioka, T.: Destructive effects of smoking on molecular and genetic factors of periodontal disease. Tob. Induc. Dis. **8**(4), (2010)

COPD: On Evaluating the Risk for Functional Decline

F. Rodrigues, I. Matias, J. Oliveira, S. Vacas and A. Botelho

Abstract Chronic obstructive pulmonary disease (COPD) is an important cause of chronic morbidity and mortality worldwide. It is characterized by chronic pulmonary and extrapulmonary manifestations with great impact on patients' health, functional impairment, decrease in quality of life and need for prolonged assistance as well as the risk of becoming dependent on others. The aim of this study is to identify COPD patients with the risk of becoming dependent on others to perform activities of daily living (ADL), in order to provide them early intervention and assistance. The study is longitudinal, observational, quantitative and correlational. An intentional sample was used, consisting of patients diagnosed with COPD, clinically stable for at least 3 weeks, who were or had been on a pulmonary rehabilitation program at the Pulmonary Rehabilitation Unit of Hospital Pulido Valente. The IMPALA score is obtained through a questionnaire of self-reported performance for 20 Activities of Daily Living (ADL), assessing the dependent/independent status and four possible early signs of risk of dependence: taking longer to do the activities, reporting difficulty in doing them, having to take breaks while doing them or doing them less frequently. This score was compared with sociodemographic factors, pulmonary function testing (FEV_1), the 6-min walking test (6 MWT) and a disease health-related quality of life score (CAT score). Statistical analysis was performed using exploratory data analysis, visualization techniques and correlation analysis using R. With respect to

F. Rodrigues (✉) · I. Matias
Serviço de Pneumologia, Centro Hospitalar Lisboa Norte, Lisbon, Portugal
e-mail: fatima.rodriguesed@gmail.com

I. Matias
e-mail: becas.mattias@sapo.pt

J. Oliveira · S. Vacas
Centro Hospitalar Lisboa Norte, Lisbon, Portugal
e-mail: joana.a.oliveira@hotmail.com

S. Vacas
e-mail: saravacas@hotmail.com

A. Botelho
NOVA Medical School, Lisbon, Portugal
e-mail: amalia.botelho@fcm.unl.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_11

disease characteristics and ADL performance (IMPALA score), COPD Grade D patients showed the worst ADL performance at basal time and a substantial variation at 6 months. Grades A, B and C had most ADL performances close to full capacity and showed little variation after 6 months. ADL performance after 6 months was worse in patients with frequent exacerbations and, although there was no significant correlation to age, older patients tended to improve ADL performance after 6 months. We found a weak correlation between the IMPALA score and exercise functional capacity, but a good correlation with basal health-related quality of life (CAT score). In conclusion, IMPALA score seems to be an additional disease marker evaluating the impact on current functional capacity, well suited to show early risk of incapacity in this group of COPD patients.

Keywords COPD · Health-status · Activities of daily living · Functional capacity

1 Introduction

Chronic obstructive pulmonary disease (COPD) is a preventable and treatable disease that represents a major public health problem, for which prevalence tends to rise due to increases in exposure to risk factors and longevity [2, 5, 9].

COPD is an important cause of chronic morbidity and mortality worldwide, causing patients to suffer for a long time and dying prematurely from it or its complications, and is currently the fourth leading cause of death in developed countries [7, 15]. It is a complex disease characterized by chronic pulmonary and extrapulmonary manifestations that impose a great impact on the patients' health, leading to progressive functional impairment, a decrease in quality of life and the need for prolonged assistance.

The loss of capacity to perform activities of daily living (ADL) is a worrying factor of this disease, since it causes a substantial burden on the patients' independence, on their caretakers and on health systems [3, 4].

It is now recognized that no single measure can adequately reflect the nature or severity of COPD [10, 11]. Scientific search of a comprehensive knowledge of COPD morbidity and prognosis led to combining variables such as airway obstruction (FEV_1-forced expiratory volume in 1 s), number of exacerbations (<2 or ≥ 2 hospital visits for respiratory reasons, such as respiratory infections, with or without the need of hospital admission), health status (CAT-COPD Assessment Test) and symptoms (mMRC-modified Medical Research Council dyspnea) deriving the recent multidimensional GOLD classification (categories A, B, C and D) [9] or combining variables such as airway obstruction (FEV_1), mMRC-dyspnea, body mass index and exercise (6 min walking distance in meters) deriving the BODE prognosis index [6].

COPD burden on the patients' ADL performance outlines the need for an early identification of patients who are at risk of becoming dependent on others.

2 Aim

The aim of this study is to identify COPD patients who are at risk of becoming dependent to perform their Activities of Daily Living, using a score to evaluate the patients' current ADL performance.

3 Methods

3.1 Study Design

This is a preliminary study, which is still running, and it is longitudinal, observational and correlational. The study is being conducted at the Pulmonary Rehabilitation Unit/Day-care Hospital for Respiratory Failure Patients' of the Hospital Pulido Valente—Lisbon. The evaluations were applied twice, basal and at 6 months follow up. Exacerbations during that period, requiring hospital assistance, were also tracked. Data collection was undertaken by health care professionals—one pulmonologist and one nurse—and two at the time medical students, who received previous training on the application of the various questionnaires and measures.

3.2 The Sample

An intentional sample was used, consisting of patients diagnosed with COPD, clinically stable for at least 3 weeks, who were or had been on a pulmonary rehabilitation program where the study took place and who consented to participate in the study.

3.3 The IMPALA Score and Questionnaire

The study questionnaire was based on self-reported performance for 20 Activities of Daily Living, namely walking, self-care activities [12, 14] and instrumental activities of daily living [13]. The patients report being dependent or independent on others to perform each ADL and 4 possible early signs of risk of dependence are assessed for each ADL: (1) taking longer, (2) reporting difficulty, (3) having to take breaks, or (4) doing it less frequently. The Impact on Life Activities (IMPALA) Score, is obtained by giving the patients 0 points for each ADL they are dependent on others to perform and 1 point for each ADL they report performing independently. For the patients who are able to perform a task independently 0.2 points are subtracted for each of the four signs of risk of dependence that they report in that ADL. By summing the points for each ADL and multiplying the result by 5 we obtain a score that varies from a minimum of 0 to a maximum of 100. The difference of the 6 month interval values was called IMPALA Score 6-month variation, with a possible positive variation (a raise ≥1 point), a negative variation (decrease ≥1 point) and a neutral variation (<1 point).

This ADL score, and its 6-month variation, were then compared with sociodemographic factors and measures obtained through standardized instruments, namely, pulmonary function testing (FEV_1), the 6-min walking test (6 MWT) and a health-related quality of life score (CAT Score).

Ethical procedures were followed, informed consent was obtained from all the participants and the trial conduction was approved by the Ethics Committee of the NOVA Medical School (01/2014/CEFCM) and by the Ethics Committee and the administration board of the Centro Hospitalar Lisboa Norte (DIRCLIN-22/05/2014-151).

Data analysis was performed using software R®, making use of descriptive statistics and visualization techniques, exploratory data analysis and non-parametric tests.

4 Results

4.1 Sample Description

The sample's size is 34 patients, all of them caucasian, approximately 15 % female and 85 % male, distributed by age as represented in Table 1.

The *Age* average was 68.4 ± 8.8 years old. Regarding *Education* level, 18 % of the sample reported they could neither read nor write and almost half of the sample (44 %) had a Basic Education level.

Table 1 Sample distribution by age and sex

Variables	Age					
	[45–54]	[55–64]	[65–74]	[75–84]	[85–94]	Total
Sex						
Male	2	6	11	9	1	**29**
Female	0	2	1	2	0	**5**
Total	**2**	**8**	**12**	**11**	**1**	**34**

Table 2 Common comorbidities in the sample

System	%	**Endocrine and metabolic disorders**	
Respiratory system		Type 2 Diabetes mellitus	15
Chronic Respiratory Failure	71	Dyslipidemia	9
Bronchiectasis	27	**Prostatic disease**	21
TB sequelae	18	**Psychiatric disorders**	
Obstructive sleep apnea syndrome	15	Alchoolism	12
Pulmonary thromboembolism	6	Depression	9
Cardiovascular system		**Nutrition disorders**	
Hypertension	62	Excess weight	29
Chronic heart failure	21	Obesity	15
Chronic atrial fibrillation	12	Malnourishment	12
Chronic cor pulmonale	12	Low weight	9
Pulmonary hypertension	12	**Ophtalmology disorders**	12
Ischemic cardiopathy	9	**Osteoarticular disorders**	18
		Gastrointestinal disorders	21

Concerning *F*amily, 58.8 % (n = 20) of the patients were married or in civil union, 18 % were divorced, 15 % were widowed and 9 % were single. About 21 % (n = 7) of the patients reported living alone, 71 % (n = 5) of which were 65 years or above.

As far as *S*moking Habits, we verified that almost all of patients—97 % (n = 33)—were current or former smokers, with an average smoking burden of 66.7 ± 38.9 pack-years and with 65 % (n = 22) of them having smoked 50 pack-years or above.

In terms of *C*omorbidities, we found Chronic Respiratory Failure to be the most common (71 %, n = 24), closely followed by Systemic Hypertension (62 %, n = 21). Most common comorbidities are shown in Table 2.

Of the total sample, only 25 patients completed the 6-month re-evaluation, since 5 of them died in this period (4 of which directly related to respiratory disease), 3 of them haven't yet completed the 6-month period and one was lost to follow-up.

4.2 Combined COPD Assessment (GOLD)

When we applied the Combined COPD Assessment (GOLD) to our sample, we verified that most patients—62 % (n = 21)—were Grade D of the disease, as can be seen in Fig. 1.

Fig. 1 Sample distribution by COPD grade, according to the Combined COPD Assessment (GOLD)

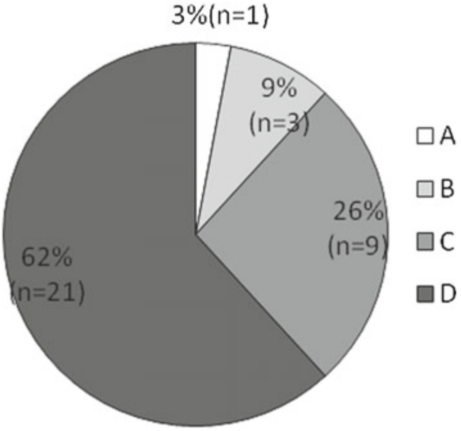
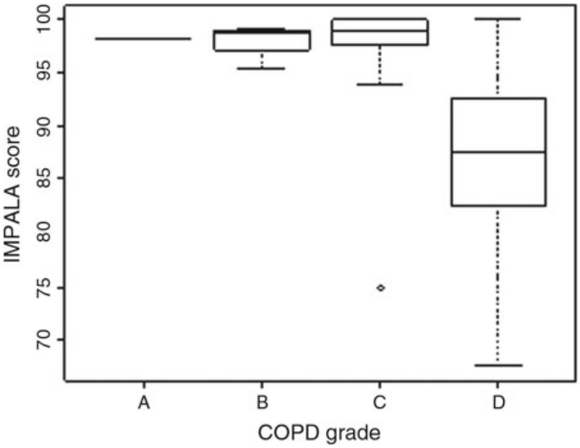


Fig. 2 IMPALA score versus COPD grade



We compared the COPD Grade according to GOLD to IMPALA Score and noticed that Grade D patients had the worst performance, the other categories being near full capacity (Fig. 2). We then compared COPD Grade with IMPALA Score 6 month variation and obtained a scarce variation distribution (Fig. 3).

Within the most severe COPD Grade (D) patients, there was a substantial variation of self-reported functional capacity, either at basal evaluation or after 6 months.

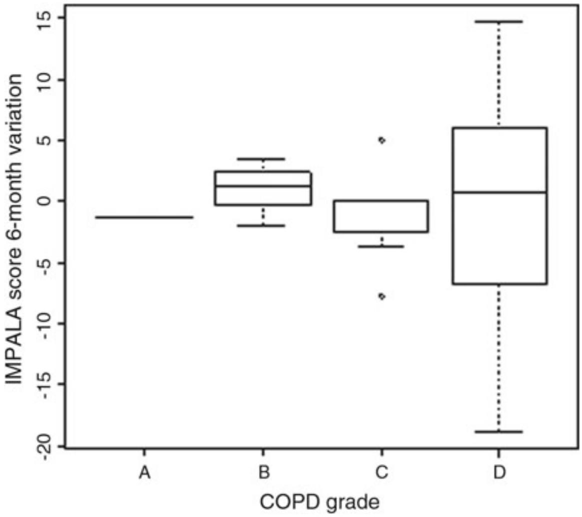


Fig. 3 IMPALA score 6-month variation versus COPD grade

4.3 The IMPALA Score 6-Month Variation

After 6 months, 33 % of the patients improved their ADL performance (IMPALA score), while 37 % had a negative variation. Five (17 %) patients died. 13 % of them had no variation comparing to the first evaluation (Fig. 4).

We tried to identify possible correlations between the IMPALA Score 6-month variation with the variables age, sex, family status, smoking habits and COPD Grade, but found no significant correlation.

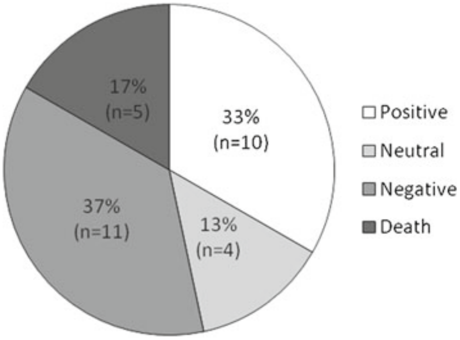


Fig. 4 IMPALA score 6-month variation

4.4 IMPALA Score 6-Month Variation Versus Exacerbations

55 % of the patients with <2 exacerbations in the previous year showed either a positive or neutral variation of the ADL performance (IMPALA score) after 6 months (Fig. 5), 36 % had a negative variation and 9 % died. However, frequent exacerbators (≥2 exacerbations in previous year) showed worse outcomes, with only 22 % experiencing positive or neutral 6-month variation, 33 % having a negative variation and 45 % (n = 4) resulting in death (Fig. 6).

Fig. 5 The IMPALA score 6-month variation in patients with <2 exacerbations

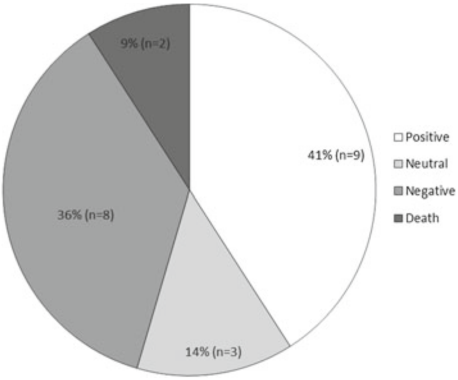
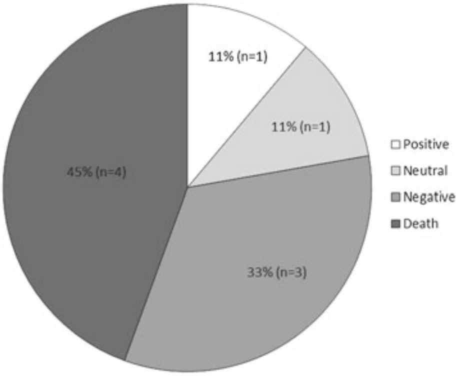


Fig. 6 The IMPALA score 6-month variation in patients with ≥2 exacerbations



4.5 IMPALA Score Versus Age

There was a wide distribution of the IMPALA score at basal time, irrespective of patients age, resulting in a correlation coefficient of −0.15 (Fig. 7). The IMPALA Score 6-month variation according to age (Fig. 8) showed older patients had more positive variations than younger patients, with a correlation coefficient of 0.40, as shown at Fig. 8.

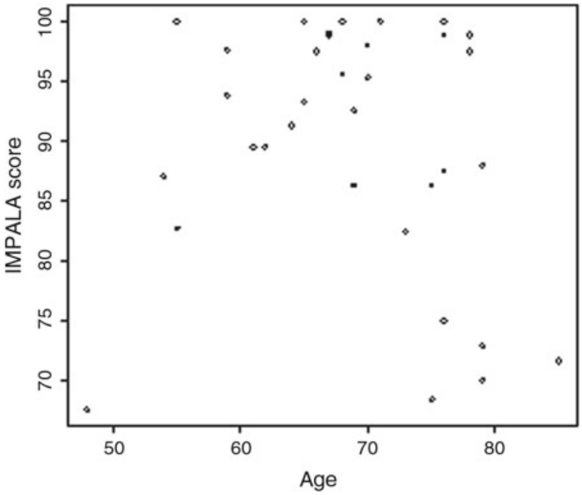


Fig. 7 The IMPALA score according to age, at basal time

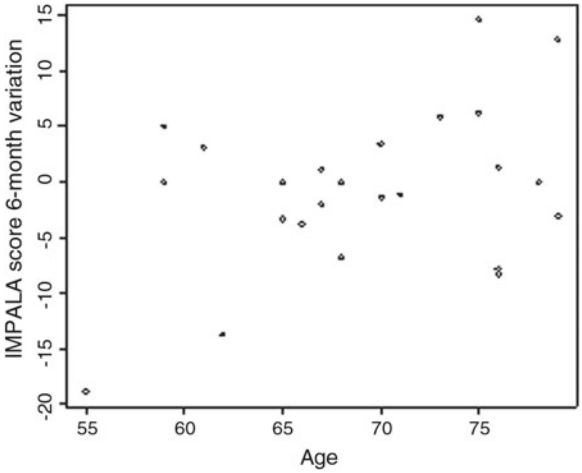


Fig. 8 The IMPALA score according to age, at 6 months

4.6 IMPALA Score Versus 6 MWT

We found a weak correlation between ADL performance (IMPALA score) and exercise functional capacity (6 MWT)—correlation coefficient = 0.29 (Fig. 9). Likewise, after 6 months the variations of both parameters were not related (correlation coefficient = −0.25), see Fig. 10.

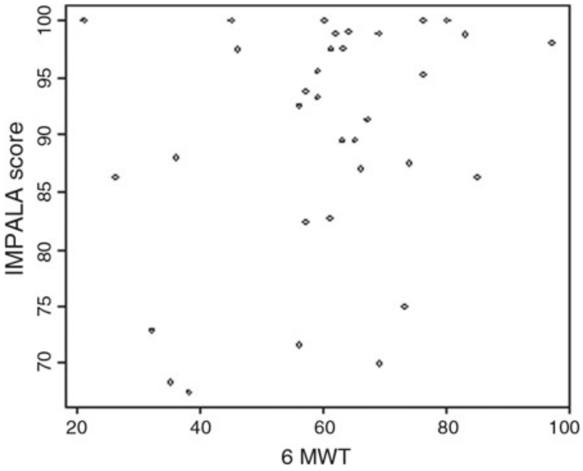


Fig. 9 The IMPALA score according to 6 MWT (m), at basal time

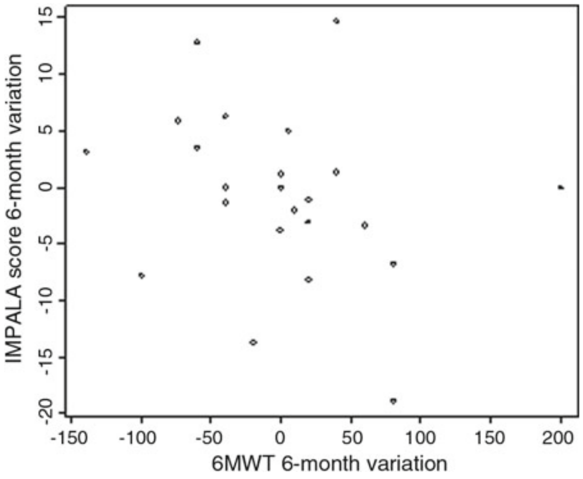


Fig. 10 The IMPALA score 6-month variation according to the 6 MWT 6-month variation

4.7 IMPALA Score Versus CAT Score

We found a good correlation (−0.53) between basal ADL performance (IMPALA score) and health-related quality of life (CAT score), as shown at Fig. 11. After 6 months, variation of both parameters were not related (correlation coefficient = 0.05), see Fig. 12.

Fig. 11 The IMPALA score according to the CAT score, at basal time

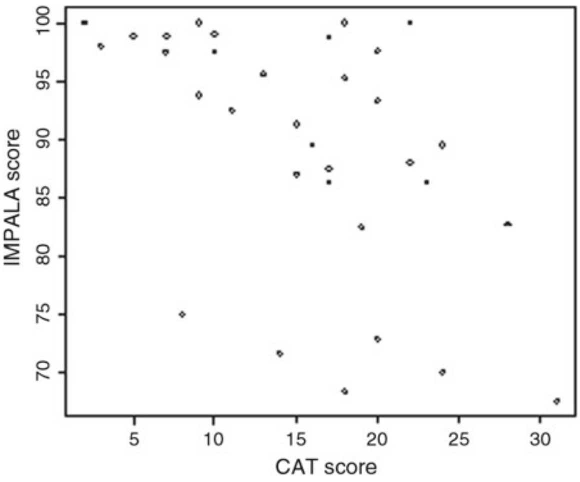
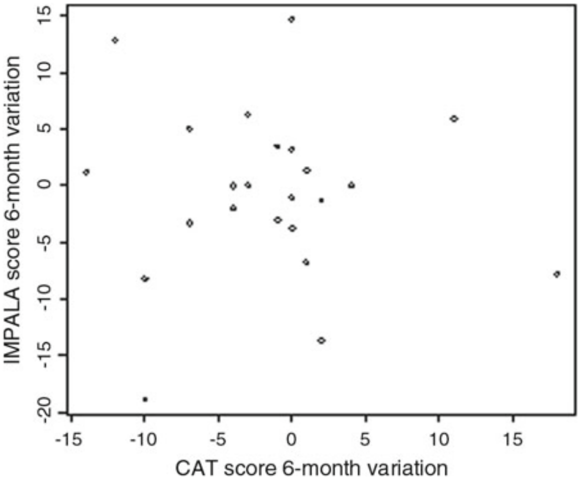


Fig. 12 The IMPALA score according to the CAT score—6-month variation



5 Discussion

The present study, in COPD stable patients engaged in a pulmonary rehabilitation program, was able to identify those at risk of becoming dependent in current daily activities.

With respect to COPD patient's grade, most of them (62 %) were GOLD COPD category D, the most severe one, and showed a significant variation on impact of disease reported activities of daily living. Other authors also evidenced a substantial heterogeneity in COPD patients. In ECLIPSE study [1], the severity of airflow limitation in COPD patients was poorly related to the degree of breathlessness, health status, presence of co-morbidity, exercise capacity and number of exacerbations reported in the year before the study. The distribution of these variables within each GOLD stage was wide. Even in subjects with severe airflow obstruction, a substantial proportion did not report symptoms, exacerbations or exercise limitation. The clinical manifestations of COPD are highly variable and the degree of airflow limitation does not capture the heterogeneity of the disease [1].

From the COPD disease markers, exacerbations are important contributors to accelerate health status decline and increase health related costs [16]. In line with these data, in our study, frequent exacerbator patients (≥2 exacerbations in previous year) showed a predominantly negative variation on the functional capacity after 6 months, while more than half of infrequent exacerbators showed either a positive or neutral variation.

There was a wide variation of ADL performance (IMPALA score) irrespective of patient's age, which reinforces the notion that age is no more than one of the factors contributing to the overall patient's status. Nevertheless, older patients tended to improve ADL performance after 6 months, which is consistent with the recognized benefits of physical activity that counteracts the aging progressive reduction of maximum abilities [8]. As such, ageing COPD patients are suitable candidates for pulmonary rehabilitation, with improvement of domestic function and physical activity [17].

The correlation analysis between the IMPALA Score 6-month variation and other variables like age, sex, family status, smoking habits and COPD Grade might be limited by the reduced number of observations. Additional studies with larger samples could allow further investigation of these correlations.

In this preliminary study, ADL performance (IMPALA score) showed a stronger association with health-related quality of life (CAT score) than with exercise functional capacity (6 MWT). All these variables translate COPD impact on patient's health and wellbeing. Probably they constitute different dimensions of this clinical entity, ADL performance being related to current low demanding tasks, thus complementing the holistic evaluation in each patient.

6 Conclusion

Self-reported ADL performance (IMPALA score), based on 20 ADL tasks performance limitations, seems to be an additional disease marker, eliciting the impact on current functional capacity and having a good correlation with basal health-related quality of life, as measured by CAT score.

In COPD patients, ADL performance (IMPALA score), as a marker of current functional capacity showed: (1) heterogeneity among the most severe GOLD grade D patients; (2) negative impact of frequent exacerbations; (3) age as not specifically related; (4) complementary information gathered by the 6 MWT that evaluates submaximal exercise capacity.

In conclusion, we believe that a simple and easy to implement evaluation of ADL performance (IMPALA score)—is well suited for early detection of risk of incapacity in this group of COPD patients.

References

1. Agusti, A.: COPD heterogeneity in the ECLIPSE cohort. Respir. Res. **11**, 122 (2010)
2. Bárbara, C., Rodrigues, F., Dias, H., Cardoso, J., Almeida, J., Matos, M.J., Simão, P., Santos, M., Ferreira, J.R., Gaspar, M., Gnatiuc, L., Burney, P.: Prevalência da doença pulmonar obstrutiva crónica em Lisboa, Portugal: estudo Burden of Obstructive Lung Disease. Rev. Port. Pneumol. **19**(3), 96–105 (2013)
3. Botelho, A.: Autonomia Funcional em Idosos. Caracterização multidimensional em idosos utentes de um centro de saúde urbano (1ł ed.). Laboratórios Bial, Porto (2000)
4. Botelho, A.: Método de Avaliação Biopsicossocial—um instrumento de detecção em saúde. In: Fernandes, L., Pereira, M.G., Pinto, L.C., Firmino, H., Leuschner, A. (eds.) Jornadas de Gerontopsiquiatria 1ł Edição, pp. 87–89 (2011)
5. Buist, A.S., McBurnie, M.A., Vollmer, W.M., Gillespie, S., Burney, P., Mannino, D.M. et al.: BOLD Collaborative Research Group. International variation in the prevalence of COPD (the BOLD Study): a population-based prevalence study. Lancet **370**, 741–750 (2007)
6. Celli, B.R., Cote, C.G., Marin, J.M., Casanova, C., de Montes, O.M., Mendez, R.A., Pinto-Plata, V., Cabral, H.J.: The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. N. Engl. J. Med. **350**, 1005–1012 (2004)
7. Chapman, K.R., Mannino, D.M., Soriano, J.B., Vermeire, P.A., Buist, A.S., Thun, M.J. et al.: Epidemiology and costs of chronic obstructive pulmonary disease. Eur. Respir. J. **27**, 188–207 (2006)
8. Darren, E.R., Warburton, D.E., Nicol, C.W., Bredin, S.S.D.: Health benefits of physical activity: the evidence. CMAJ **174**(6) (2006). doi:10.1503/cmaj.051351
9. Global Initiative for Chronic Obstructive Lung Disease—GOLD: global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease. Available on the GOLD. www.goldcopd.org (2013)
10. Jones, P.W.: Health status and the spiral of decline. COPD J. Chronic Obstr. Pulm. Dis. **6**, 59–63 (2009)
11. Jones, P., Miravitlles, M., Molen, T., Kulich, K.: Beyond FEV1 in COPD: a review of patient-reported outcomes and their measurement. Int. J. Chronic Obstr. Pulm. Dis. **7**, 697–709 (2012)
12. Katz, S., Ford, A.B., Moskowitz, R.W., Jackson, B.A., Jaffe, M.W.: Studies of illness in the aged: the index of ADL; a standardized measure of biological and psychosocial function. JAMA **185**, 914–919 (1963)

13. Lawton, M.P., Brody, E.M.: Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist **9**, 179–186 (1969)
14. Mahoney, F.I., Barthel, D.W.: Functional evaluation: the Barthel index. Md. State Med. J. **14**, 61–65 (1965)
15. Raherison, C., Girodet, P.O.: Epidemiology of COPD. Eur. Respir. Rev. **18**, 213–221 (2009)
16. Seemungal, T.A., Hurst, J.R., Wedzicha, J.A.: Exacerbation rate, health status and mortality in COPD—a review of potential interventions. Int. J. Chronic Obstr. Pulm. Dis. **4**, 203–223 (2009)
17. Sewell, L., Singh, S.J., Williams, J.E., Collier, R., Morgan, M.: Can individualized rehabilitation improve functional independence in elderly patients with COPD? Chest **128**, 1194–1200 (2005)

Microarray Experiments on Risk Analysis Using R

Teresa A. Oliveira, Amílcar Oliveira and Andreia A. Monteiro

Abstract The microarray technique is a powerful biotechnological tool, expanding in a interesting way the vision with which issues in medicine are studied. Microarray technology, allows simultaneous evaluation of the expression of thousands of genes in different tissues of a given organism, and in different stages of development or environmental conditions. However, experiments with microarrays are still substantially costly and laborious, and as a consequence, they are usually conducted with relatively small sample sizes, thereby requiring a careful experimental design and statistical analysis. This paper adopts some applications of microarrays in risk analysis using R statistical software.

Keywords ANOVA · Bioinformatics · Design of experiments · DNA · Microarrays technology · Software R

1 Introduction

Bioinformatics is a field of the biological sciences that is fast growing and is being developed to address need to manipulate large amounts of genetic and biochemical data. These data have originated from the individual effort of many researchers are interrelated through a common origin: cells of living organisms. To understand the relationship between these fragmented information coming from different areas of biology (such as molecular biology, structural biochemistry, enzymology, molecular biology, physiology and pathology), Bioinformatics uses computational power,

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

A. Oliveira
e-mail: Amilcar.toliveir@uab.pt

A.A. Monteiro
MEMeC - Universidade Aberta, Lisbon, Portugal
e-mail: andreiaforte@portugalmail.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_12

mathematics and statistics to catalog, organize and structure this information into a comprehensive entity. Bioinformatics has as its main purpose to unravel the large amount of data that are obtained from sequences of DNA and proteins.

Genomics and proteomics are two keywords from Biology of the new millennium. The reductionist attitude to analyze genes one by one dominant during the last decades of the twentieth century gave way to a much more comprehensive approach. The current objective is to determine the set of all genes of an organism (i.e., genome) and understand the functional networks established between the proteins encoded by them (i.e. the proteome).

A technological development that has revolutionized the ability to collect information in the field of functional genomics is called DNA microarrays. Once the genome of an organism has been sequenced, the DNA microarrays allow us to obtain the complete profile of the genes that are expressed in any cell type of that body. This technology arose in the mid-1990s.

The microarray technology, allows the simultaneous evaluation of the expression of thousands of genes, in different tissues of a particular organism, at different developmental stages or environmental conditions. However, experiments with microarrays are still considerably expensive and cumbersome, consequently, they are usually conducted with relatively small sample sizes. Such experiments involve a series of laboratory procedures, which invariably introduce different levels of additional variation data so, to complete a successful microarray experiment, several factors must be addressed. The driving test with microarrays requires then a very careful experimental design and statistical analysis. The strong connections between this area and risk analysis are evident in the application of microarray techniques on the classification and progression of a disease which plays a key role in the identification of high risk groups and in response to the treatment of the disease. Microarray analysis is important to study complex and multigenic diseases, as for example Alzheimer's or Parkinson's diseases. The great challenge in understanding the genetics of such disorders is to identify which are the responsible genes that increase the risk of an individual to develop a particular disease. To better understand how huge is this challenge, we refer to the paper by Mora et al. [12] where the authors refer to the construction of 497 multigenic disease groups from the database OMIM (Online Mendelian Inheritance In Man).

2 Introduction to Microarrays Technology

The prerequisite for any type of DNA array is the existence of an address for each component of the collection, or an individual position for each component of the arrangement. Each of these addresses in the arrangement is called a spot, and contains a small quantity of DNA immobilized properly called probe. Each of these probes tend to bind only to their complementary sequence of nucleotides by a process called hybridization [8]. This complementary sequence, usually a complementary

DNA (cDNA) produced from a messenger RNA (mRNA) represents a single gene of the genome and is called the target.

Microarray technology is the use of a slide (slide or microarray) in which the probes (DNA sample) were immobilized in accurately defined positions and quantities (spots) to make the hybridization with a pool of mRNAs extracted from biological samples (targets), that were previously labeled with fluorescent dyes.

So that the results obtained in different experiments could be compared and used as a basis for further research on the same topic, organizations such as the Functional Genomics Data Society—FGED and the European Bioinformatics Institute—EBI have established guidance documents that assist researchers to plan and implement their experiences with microarrays. One such guide is the Minimum Information About a Microarray Experiment (Miame), which contains a number of recommendations and standards for collection and analysis of data from microarray experiments.

Another interesting attempt is the sharing of raw data. For this, we point out two main databases:

(i) NCBI GEO—http://www.ncbi.nlm.nih.gov/geo;

GEO—the Gene Expression Omnibus is an initiative of the National Center for Biotechnology Information (NCBI) and ArrayExpress, maintained by the EBI. Note that to analyse microarray data that have been published in GEO, it is also highly recommended to visit GEO2R, http://www.ncbi.nlm.nih.gov/geo/geo2r, which is a very interesting tool to compare two or more groups of samples, in order to identify genes that are expressed differentially across experimental conditions.

(ii) ArrayExpress—http://www.ebi.ac.uk/microarray;

ArrayExpress—database of functional genomics experiments, maintained by the EBI. It can be queried and the data downloaded and it includes gene expression data from microarray and high throughput sequencing studies.

From the design of experiments point of view and the relative statistical analysis of microarray experiments, the distinction between the different technologies refers to the number of samples on each slide hybridized. Microarrays can be classified as single channel (one colour) or multiple channel (two colour) microarrays. In a single channel, a probe is hybridized with target DNA labelled with one colour fluorochrome. In a multiple channel it is possible to analyse genes from different samples in a single test. In this case, each target gene is labelled with fluorochrome having different fluorescence emission and the sample is allowed to hybridize in a single test with a single microarray probe. A two-color microarray experiment, showing how gene expression can be altered by a disease such as cancer, is illustrated in Fig. 1 and it was obtained from: http://www.people.vcu.edu/~mreimers/OGMDA/image.html.

Hybridized arrays, like the one in Fig. 1, are scanned to produce high resolution tiff files. The goal is to produce a large matrix or data frame of the expression data where, in general, the genes are represented by the rows, while the conditions are represented by the columns. It is important to notice that in the multiple channel microarrays there is the disadvantage that different samples may interfere with each other. Single channel microarrays need to conduct different test to quantify different gene expression, however they seem to give more accurate results.

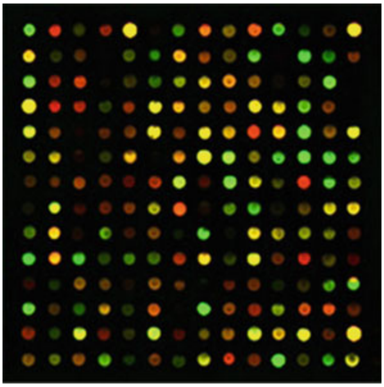


Fig. 1 Illustration of a two color microarray

3 Designs for Microarrays

Design of Experiments can be defined as a set of techniques which, when correctly chosen and applied, can make an experience more efficient and achieve maximum information with the minimum use of resources, effort and time. To get a successful experience, the experimenter must be aware of the three basic principles: Blocking, Replication and Randomization.

An appropriate experimental design is vital to the success of any experiment with microarrays. In recent years statistical works presenting ideas on these topics have arisen, such as Kerr and Churchill [9], and Yang and Speed [5]. After the selection of individuals to be used in experiments with microarrays, the definition of the experimental scheme that will be used is crucial. The experimental schemes commonly used in microarray experiments are particular block designs, namely reference sample designs and loop designs, which are illustrated in Figs. 2 and 3.

In the Reference Sample Design (RSD), as represented in Fig. 2, one sample (termed reference sample, represented by Rsd) is hybridized to each sample of each treatment, or experimental group (A, B and C).

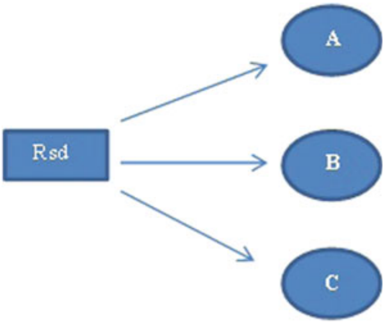


Fig. 2 Reference sample design

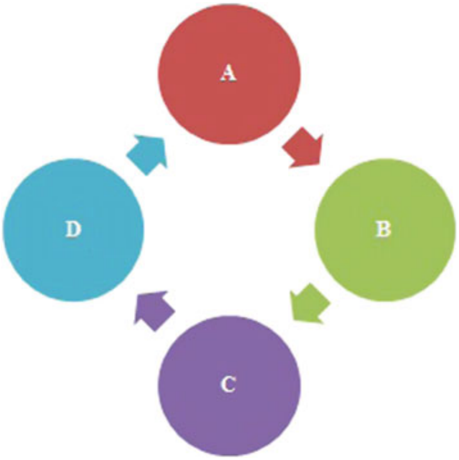


Fig. 3 Loop design

The comparison between treatments occurs indirectly. For example, comparison of treatments A and B is estimated using information relating to the difference of the contrasts between the A treatment and the reference, and the B treatment and the reference, that is, (A-Rsd)–(B-Rsd). The main advantage of designs with reference sample is that they are simpler to conduct in the laboratory and other samples can be added later to be compared with existing ones. The disadvantage relates to the fact that half of observations refer to the reference sample, which is not of direct interest.

In Loop Designs (LD) instead of using the reference sample, each sample is compared with the next so as to create a circular shape see Fig. 3. Thus, differences between treatments are estimated by combinations between direct and indirect comparisons. The circular structures are generally more efficient than designs with reference samples, see [5, 9, 19].

The use of different levels of biological replicates and of technical replicates, as different ways to mark the samples in each blade, generates a large variety structures within these two basic designs, RSD and LD. Rosa et al. [15] present three alternative designs using reference sample and a two alternative loop design, which we represent respectively in Figs. 4 and 5. Each letter (A, B, C) represents an experimental

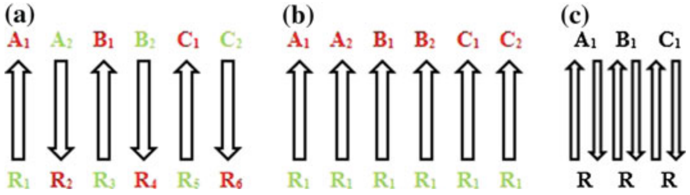


Fig. 4 Three alternative microarray experimental layouts for reference designs

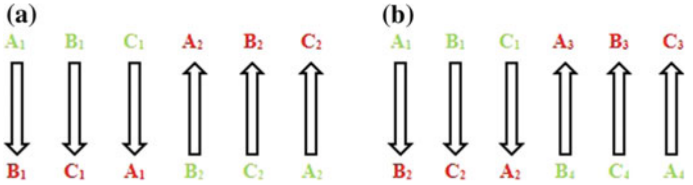

(a)
A_1 B_1 C_1 A_2 B_2 C_2
B_1 C_1 A_1 B_2 C_2 A_2
(b)
A_1 B_1 C_1 A_3 B_3 C_3
B_2 C_2 A_2 B_4 C_4 A_4

Fig. 5 Two alternative loop designs

group or a reference sample (R), and each arrow represents a slide, which connects the two samples co-hybridized on it. Furthermore, the indexes represent biological replications and the head and tail of each arrow indicate the samples labelled with Cy5 and Cy3, which are the most popular cyanine dyes used, typically combined for 2 color detection. Cy3 dyes fluoresce yellow-green (∼550 nm excitation, ∼570 nm emission), while Cy5 is fluorescent in the red region (∼650/670 nm) but absorbs in the orange region (∼649 nm).

There are advantages and disadvantages to each of these designs, but from a statistical perspective it is generally recommended to prioritize biological replicates. For a more detailed comparison in terms of efficiency and statistical power of these alternative designs as well as the resilience of these outlines on the loss of blades see for example [14, 18, 19].

Optimal Experimental Designs aim to optimize the information content of the experimental data to allow identification of the mathematical model which better fits the process under study. The precision of the model parameters depends essentially of the quality of experimental data that will be used in the identification algorithm, see [11] among others.

For microarray experiments, there are a limited number of available arrays, as well as certain amount of RNA. The challenge is to find what samples should be placed and in which arrays, in order to maximize the accuracy of the estimated parameters, which is related to the choice of the optimality criterion.There are several optimality criteria, and the most used criterion are D, A and L. The D optimality criterion aims at minimizing the determinant of the covariance matrix of the parameter estimates. The A-optimal design is that design for which the average variance of the estimated parameter is minimal. The L criterion defines as optimal the design that minimizes the average variance of estimates of various parameters of linear functions. The choice of an appropriate criteria depends on the purpose of the experiment, Sivey [17].

Clearly there are other possible structures of designs for experiments with microarrays, including general structures of row-column designs. This type of design allows control of two causes of variation which can lead to major reductions in experimental error. The most familiar example of this type of design is the Latin Square. The row-column designs apply naturally to systems of two colors: (two-color microarray) it has two rows and the columns represent the blades. In extensive studies, this type of design can be more efficient than the block designs, however it is also more complex.

Due to constraints in terms of biological material and number of blades available, search algorithms can be used to obtain optimal designs (or near optimal) for certain specific objectives of the experiments. Wit et al. [20] applied an optimization strategy based on metaheuristic simulated annealing to search for optimal designs for almost any number of treatments and any number of blades. Sacan et al. [16] also applied a strategy based on metaheuristic "Hill Climbing" to seek nearly optimal designs, additionally provided a tool available in http://www.db.cse.ohio-state.edu/MicroarrayDesigner which allows users to consult different types of design and different optimality criteria or perform the upload of new designs.

Thus, the design of microarray experiments can be treated as an optimization problem and to find the best design meeting certain criteria, algorithms including metaheuristics, can be explored.

4 Analysis of Microarray Data

After the exploratory analysis and preprocessing of the data the stage of data analysis is crucial. In this step one can take different approaches depending on the mathematical and statistical purpose of the experiment.

For example, cluster analysis is widely used both to group together genes as samples, in order to discover groups of genes or groups of samples with similar expression patterns.

Discriminant analysis is also quite common in medical studies, using samples of healthy and sick patients, to develop classification models for use in diagnostic tests. Another very common procedure with microarray data refers to significance tests for the detection of differentially expressed genes, in samples from different experimental groups. In this context, methodologies relative to linear models, such as ANOVA models and mixed models, are the most frequently used.

4.1 Analysis of Variance

The linear model with fixed effects allows us to compare more than two groups or to control other fixed effects, as effects of groups (varieties), genes, dyes (Cy3 and Cy5 labeling) and arrays. When these additional terms are associated with the response, the variance of the error (residual) can be substantially reduced. In microarray experiments, taking into account known causes of variation, increases the power of the experiment to observe significant differences in expression levels for a given gene, [2].

Depending on what are the most important causes of variation, different models can be differently useful for a particular study.

Let us consider the example where the choice of model includes parameters of the main effects for four factors, A, D, V and G, and second order interactions with G (effects of interest), variation spot to spot and interactions between dyes and genes:

$$y_{ijgk} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg} \quad (1)$$

y_{ijgk} is the logarithm of the intensity value of expression observed in the array i (A_i) , in the dye j (D_j), in variety k (V_k), in gene g (G_g), μ is general average, $(AG)_{ig}$ is the interaction (array × gene), is the interaction (dye × gene) and ε_{ijkg} are the residuals which are assumed to be random variables, independent, identically distributed with zero mean and variance σ_g^2 . Homogeneity of variances is assumed for observations from the same gene, but heterogeneity is allowed in different expression level of genes. For hypothesis testing it is necessary that the errors follow a normal distribution.

For a similar problem, Kerr et al. [10] proposed a model that includes the interaction (array × dye), $(AD)_{ij}$, given by:

$$y_{ijgk} = \mu + A_i + D_j + (AD)_{ij} + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg} \quad (2)$$

Wolfinger et al. [21] suggested separating the previous model in two, where the first contains the "global" component and may be seen as a standardization model:

$$y_{ijgk} = \mu + A_i + D_j + (AD)_{ij} + r_{ijkg} \quad (3)$$

The residuals of the first model are used as input to the model gene:

$$r_{ijgk} = G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \varepsilon_{ijkg} \quad (4)$$

i, j, g, k vary according to the experience.

A major advantage of the ANOVA model is that it can be extended to more complex designs of microarrays. George [5] proposes an approach which considers a microarray experiment as an experiment in split plot. RNA samples are considered the treatments applied to plots and genes are the treatments applied to the subplots. Assuming the arrays as blocks, the model is:

$$y_{ijg} = \mu + A_i + T_j + \varepsilon_{ij} + (G_g + (AG)_{ig} + (TG)_{jg} + \delta_{ijk} \quad (5)$$

y_{ijg} is the response variable, μ is the effect of the overall mean, $(A)_i$ is the effect of array i, (T_j) is the effect of treatment j applied to the plot, ε_{ij} is the effect of the error due to the plot, G_j is the effect of the gene g, $(AG)_{ig}$ is the interaction (array × gene), $(TG)_{jg}$ is the interaction between treatment and gene and δ_{ijk} is the effect of the error due to the subplots.

4.2 Fixed Effects Model

Fixed effects models include the effects of factors, but are based on assumptions of independence between observations and homogeneous variation in the levels of expression of the same gene. The presence of random effects in experiments on microarrays allow the introduction of correlation between expression levels. The mixed models include both fixed and random effects. In matrix notation the model is defined as:

$$y = X\beta + Zu + \varepsilon \tag{6}$$

where:
y is a vector of observed logarithm in base 2 of response variables;
X is a design matrix of known constants for the fixed effects;
β is a vector of parameters of unknown fixed effects;
Z is a design matrix of known constants for the random effects;
u is a vector of parameters of unknown random effects;
ε is a vector of random errors.

The assumptions on u and ε are:$E[u] = 0, E[\varepsilon] = 0, var[u] = \sigma_u^2$ e $var[\varepsilon] = \sigma_\varepsilon^2$, where u and ε are uncorrelated. For a detailed discussion and application we refer to Draghici [3].

5 The Software R: Useful Packages on Microarray Analysis

Microarrays produce enormous amounts of data, and the analysis of such data can be quite complex. The huge volume of data usually requires special software besides a database in which to store both the measurements and the results of the analyses.

R is a computational language adequate for mathematical and statistical research, similar to the S language, but free. Due to its diversity and to the fact of being free, R has become one of the most popular tools for analysis of microarrays.

Gentleman et al. [4] presented Bioconductor, a project to develop free software for the analysis of genomic data including microarray data, which is available at http://www.bioconductor.org. However, a disadvantage of this project is that most of the packages were developed using specific data structures, thus making communication between different procedures implemented in different packages difficult. A big challenge is to keep these packages updated and to develop them extending the possible choices of available structures. For a detailed look at the potential of this project a visit to both sites is highly recommended

http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual
http://www.bioconductor.org/help/workflows/oligo-arrays/

Table 1 Packages from R/Bioconductor used in the analysis of genomic data

Package	Description
Marray, limma	Spotted cDNA array analysis
affy	Affymetrix array analysis
vsn	Variance stabilization
ctest	Statistical tests
Genefilter, limma, multtest, siggenes	Gene filtering
Mva, cluster, clust	Clustering
Class, rpart, nnet	Classification
OLIN	Normalization
sigPathway	Performs pathway analysis on microarray data
MAMA, metaMA	Meta-analysis of microarray
IsoGene	Analysis dose-response studies in microarray experiments
maigesPack	Analyze cDNA microarray data
SurvJamda	Predict patients' survival and risk assessment
BioNet	Integrated analysis of protein-protein interaction networks and the detection of functional modules

Table 1 presents some selected packages from R/Bioconductor currently used in the analysis of genomic data.

Database GEO2R allows the possibility of producing an R script that can be imported and executed in real time and allows one conduct a microarray analysis of the groups of interest, of all the groups that you selected from the GSE. Furthermore, if the produced R script is developed specifically for published GSE data in GEO, it is possible to modify the R script so to adapt it to any data.

Applications of using R in problems involving microarrays and risk analysis are currently an unexplored research area. Some interesting examples are shown in the papers by Beisser et al. [1], Pramana et al. [13], Yasrebi [3].

- Pramana et al. [13] introduces the IsoGene package, R package for the analysis of dose-response microarray experiments to identify gene or subsets of genes with a monotone relationship between the gene expression and the doses. Illustrative examples of analysis using this package are also provided in this paper;
- Yasrebi [3] present SurvJamda (Survival prediction by joint analysis of microarray data) an R package that utilizes joint analysis of microarray gene expression data to predict patients' survival and risk assessment. Joint analysis can be performed by merging datasets or meta-analysis to increase the sample size and to improve survival prognosis;
- Beisser et al. [1] present the BioNet package for the analysis of biological networks. They apply this package to gene expression data from diffuse large B-cell lymphomas (DLBCL) and survival data with a human protein-protein interaction network based on human protein reference database.

6 Considerations and Conclusion

The microarray technique is a current and powerful biotechnological tool, increasing challenges in Health Sciences and Medical research. The application of this technique on the prognosis of disease is crucial to identify risk groups and to improve the response to disease treatment. Looking for answers, scientists search in risk factors, those characteristics or attributes that appear to be linked to the development of a disease. In the presence of the risk factors, there is an increased chance that the disease will develop, but not a certainty. Risk factors are characteristics like lifestyle, environment, and genetic background which contribute to the likelihood of getting a disease. It is very important to identify the risk factors so as to make better lifestyle choices and to help in reducing the risk of developing diseases. Some risk factors can be changed, like blood pressure or diabetes level; others cannot be changed, like genetic makeup.

Microarray analysis is very important in multigenic diseases research, namely to identify the responsible genes which increase the risk of an individual to develop a particular disease. This is a challenging area with many open research questions, given the large number of known multigenic disease groups.

Beyond the technology, planning, analysis and interpretation of data, microarray experiments also have some obstacles and challenges. Such experiments generate an enormous amount of data, with unprecedented dimensions and complexity. Thus, a careful design of such experiences is crucial to the success of the research involving microarrays. Due to the great importance of this stage of the analysis, several studies have been carried out comparing different design types. Once the microarray experiment has been conducted, the next challenge relates to data analysis. Recently various statistical methodologies have been developed, such as robust models or designs with assumptions best suited for particular cases.

Acknowledgments The authors gratefully acknowledge the references for their constructive advices to improve this paper. This research was partially sponsored by national founds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the projects (PEst-OE/MAT/UI0006/2011 and PEst-OE/MAT/UI0006/2014).

References

1. Beisser, D., Klau, G., Dandekar, T., Müller, T., Dittrich, T.: BioNet: an R-package for the functional analysis of biological networks. Bioinformatics **26**(8), 1129–1130 (2010)
2. Coffey, C.S., Cofield, S.S.: Parametric linear models. In: Allison, D.B. et al.: DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments. Chap. 12, pp. 223–243, Chapman & Hall/CRC, Boca Raton (2006)
3. Draghici, S.: Data analysis tools for DNA microarrayus. Chapmann and Hall/CRC Press,Boca Raton (2003)
4. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, C., Smith, C., Smyth, G., Tierney, L.,

Yang, J.Y.H., Zhang, J.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**, R80 (2004)

5. George Casella: Statistical Design. Springer eBooks (2008)
6. Jaluria, P., Konstantopoulos, K., Betenbaugh, M.: BioNet: a perspective on microarrays: current applications, pitfalls, and potential uses. Microb. Cell Factories **13**(8), Article 4 (2007)
7. Kerr, M.K., Churchill, G.A.: Experimental design for gene expression microarrays. Biostatistics **2**, 183–201 (2001)
8. Kerr, M.K., Afshari, C.A., Bennett, L., Bushell, P., Martinez, J., Walker, N.J., Churchill, G.A.: A statistical analysis of a gene expression microarray experiment with replication. Statistica Sínica, Taipei **12**(2), 203–217 (2002)
9. Kitsos, C.P.: Optimal Experimental Design for Non-Linear Models. Springer, Berlin (2013)
10. Mora, A., Michalickova, K., Donaldson, I.M.: A survey of protein interaction data and multigenic inherited disorders. BMC Bioinform., 14–47 (2013)
11. Pramana, S., Lin, D., Haldermans, P., Shkedy, Z., Verbeke, T., Göhlmann, H., Bondt, A., Talloen, W., Bijnens, L.: IsoGene: an R package for analyzing dose-response studies in microarray experiments. R J. **2/1** (2010)
12. Rosa, G.J.M., Steibel, J.P., Tempelman, R.J.: Reassessing design and analysis of two-colour microarray experiments using mixed effects models. Comp. Funct. Genomics **6**(3), 123–131 (2005)
13. Rosa, G.J.M., Rocha, L.B., Furlan, L.R.: Microarray gene expression studies: experimental design, statistical data analysis, and applications in livestock research. Revista Brasileira de Zootecnia **36**, (Special Supplement), 185–209 (2007)
14. Sacan, A., Ferhatosmanoglu, N., Ferhatosmanoglu, H.: Microarray designer: an online search tool and repository for near-optimal microarray experimental designs. BMC Bioinform. **10**, 304–310 (2009)
15. Sivey, S.D.: Optimal Design. Chapman and Hall, London (1980)
16. Steibel, J.P., Rosa, G.J.M.: On reference designs for microarray experiments. Stat. Appl. Genet. Mol. Biol. **4**(1), Article 36 (2005)
17. Tempelman, R.J.: Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models. Vet. Immunol. Immunopathol. **105**, 175–186 (2005)
18. Wit, E., Nobile, A., Khanin, R.: Near-optimal designs for dual-channel microarrays studies. Appl. Stat. **54**(5), 817–830 (2005)
19. Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennet, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R.S.: Assessing gene significance from cDNA midroarray expression data via mixed models. J. Comput. Biol. **8**(6), 625–637 (2009)
20. Yang, Y.H., Speed, T.: Design issues for cDNA microarray experiments. Nat. Rev. Genet. **3**, 579–588 (2002)
21. Yasrebi, H.: SurvJamda: an R package to predict patients' survival and risk assessment using joint analysis of microarray gene expression data. Bioinformatics **27**(8), 1168–1169 (2011)

Risk Assessment of Complex Evolving Systems Involving Multiple Inputs

A.G. Rigas and V.G. Vassiliadis

Abstract When monitoring complex evolving systems a question that often arises is to detect causal chains of events. Do particular inputs force the system to produce new events or prohibits them? This can be also considered as a risk assessment for the systems response. In this work we present two methods of estimating the effect of multiple inputs on a complex neurophysiological system. Both the response and the stimuli are very long binary time series. The first approach is a non-parametric one and describes the linear and the non-linear association of the stationary point processes by estimating the second- and third-order cumulant density functions. The second approach is a parametric one and the association between the inputs and the response of the system is described by a logistic regression model which takes into account the system's internal processes.

Keywords Cumulant densities · Muscle spindle · Penalized likelihood function · Periodogram · Randomized quantile residuals · Stationary point process

MSC 2010: 60G55 · 62M15 · 62F12 · 62G20 · 92C55

1 Introduction

The ways of measuring the association of stochastic point processes (p.p.)[1] through linear or quadratic models by using second- and third-order cumulant densities are presented in several papers [22, 24]. The cumulant densities come from the product densities by subtracting the lower contributions of the involved p.p. In a recent work

[1] The word 'stochastic' is implied.

A.G. Rigas (✉) · V.G. Vassiliadis
Department of Electrical and Computer Engineering,
Democritus University of Thrace, GR67100, Xanthi, Greece
e-mail: rigas@ee.duth.gr

V.G. Vassiliadis
e-mail: bvasil@ee.duth.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_13

an approach of estimating the second- order cumulant density is presented and the asymptotic distribution is developed [28]. Although the distribution of the estimate is asymptotically normal, its variance does not allow the construction of confidence intervals. In the first part of this work a direct and simple method of constructing confidence intervals is described that generalizes the work of Brillinger [3] in the case of estimates of higher-order cumulant densities for p.p. This method is based on the estimation of the second-order modified periodogram of the increments of the p.p.

In the second part of this work a logistic regression model is used in order to study the behavior of a stochastic system with one output and two inputs. The problem of the quasi-complete separation which appears during the estimation of the models parameters is solved by using a penalized likelihood function.

Our motivating example deals with the identification of the neuromuscular system called muscle spindle when it is affected by two stimuli simultaneously. The response of the system is recorded from the Ia sensory axon. The stimuli are two motoneurones; a gamma and an alpha. In this case the parent muscle that contains the muscle spindle is held at a fixed length. The data sets are very long and consist of binary sequences of zeroes and ones.

2 Cumulant Densities

The components of a vector-valued stochastic point process are random, non-negative, integer-valued variables. In particular, let $N(I, \omega)$, $I \in B_R$, $\omega \in \Omega$, be a r vector-valued stochastic point process with B_R the σ-algebra of Borel sets of the real line and (Ω, B_R, P) the basic probability space. If N_a denotes the ath component of N, then $N_a(I, \omega)$ presents the number of points of type a that fall in the interval I for the realization ω and $a = 1, 2, \dots r$. In the following the dependence of N on ω is omitted and it will be referred as $N_a(I)$, where $I = (0, t]$ and $-\infty < t < \infty$. The differential notation $dN(t) = N(t, t + dt]$ will also be used for small dt [2, 7]. It will be assumed that the point process satisfies the following conditions:

1. It is stationary, that is the distribution of $\{N(I_1), \dots, N(I_k)\}$ is the same as the distribution of $\{N(I_1 + \tau), \dots, N(I_k + \tau)\}$, where $I_i = (0, t_i]$ and $I_i + \tau = (\tau, t_i + \tau]$, $i = 1, 2, \dots, k$.
2. It will be orderly, that is the probability of having more than one event in a small interval will be negligible.
3. It is strong mixing, that is the increments of the point process become independent as their distance becomes large.

These conditions are satisfied approximately in practice [2, 8]. We proceed now to define certain parameters of the point process in the time domain. The mean intensity of the component N_a is defined by

$$p_a = \lim_{h \to 0^+} \text{Prob}\{1 \text{ event of type } a \text{ occurs in } (t, t + h]\}/h \tag{1}$$

The mean intensity does not depend on t because the p.p. is stationary. It also follows from the condition of orderliness that

$$p_a dt = \mathrm{E}\{dN_a(t)\}, \tag{2}$$

for $a = 1, \dots, r$. In general, the product density of order k is defined by

$$\begin{aligned} &p_{a_1 \cdots a_k}(u_1, \dots, u_{k-1}) = \\ &\quad \lim_{h_1,\dots,h_k \to 0^+} \text{Prob}\{1 \text{ event of type } a_j \text{ occurs in} (u_j, u_j + h_j], \\ &j = 1, \dots, k-1, \text{ and 1 event of type } a_k \text{ occurs in } (0, 0 + h_k]\}/(h_1 \cdots h_k) \end{aligned} \tag{3}$$

for $u_1, \dots, u_{k-1}, 0$ distinct, $a_1, \dots, a_k = 1, \dots, r$ and $k = 1, 2, \dots$.

Since the point process is orderly it follows that

$$\begin{aligned} &p_{a_1 \cdots a_k}(u_1, \dots, u_{k-1}) dt du_1 \cdots du_{k-1} = \\ &= \mathrm{E}\{dN_{a_1}(t + u_1) \cdots dN_{a_{k-1}}(t + u_{k-1}) dN_{a_k}(t)\}. \end{aligned} \tag{4}$$

The product densities satisfy various limiting relationships such that

$$\lim_{u \to \infty} p_{ab}(u) = p_a p_b \text{ and } \lim_{u \to \infty} p_{abc}(u, v) = p_a p_{bc}(v) \tag{5}$$

for $a, b, c = 1, \dots, r$. These results follow from the strong mixing condition. The cumulant density of order k is defined by:

$$\begin{aligned} &q_{a_1 \cdots a_k}(u_1, \dots, u_{k-1}) dt du_1 \cdots du_{k-1} = \\ &= \mathrm{Cum}[dN_{a_1}(t + u_1), \dots, dN_{a_{k-1}}(t + u_{k-1}), dN_{a_k}(t)] \end{aligned} \tag{6}$$

for $u_1, \dots, u_{k-1}, 0$ distinct. These are obtained from the product densities in the following way

$$q_a = p_a \tag{7}$$

$$q_{ab}(u) = p_{ab}(u) - p_a p_b \tag{8}$$

$$q_{abc}(u, v) = p_{abc}(u, v) - p_{ab}(u - v) p_c - p_{ac}(u) p_b - p_{bc}(v) p_a + 2 p_a p_b p_c \tag{9}$$

for $u, v \neq 0$, $u \neq 0$, $v \neq 0$. The limiting relationships of (5) lead to

$$\lim_{u \to \infty} q_{ab}(u) = 0 \text{ and } \lim_{u \to \infty} q_{abc}(u, v) = 0. \tag{10}$$

The cumulant densities have significant symmetries. Specifically, for the third-order cumulant density the following equalities hold

$$q_{aaa}(u,v) = q_{aaa}(v,u) = q_{aaa}(-v,u-v) = q_{aaa}(v-u,-u)$$
$$= q_{aaa}(u-v,-v) = q_{aaa}(-u,v-u) \tag{11}$$

and

$$q_{aaa}(-u,-v) = q_{aaa}(u,v) \tag{12}$$

It is clear from expressions (11) and (12) that there are six regions of symmetry for the third-order cumulant density. However, the boundaries of the six regions create problems. These follow from the definition of the cumulant density, since in (6) there is a restriction, the quantities $u_1, \ldots, u_{k-1}, 0$ must be distinct.

In the boundaries of the six regions of symmetry the cumulant densities are defined as

$$\text{Cum}[dN_a(t+u), dN_b(t)] = q_{ab}(u)dudt, \quad a \neq b \tag{13}$$

$$\text{Cum}[dN_a(t+u), dN_a(t)] = q_{aa}(u)dudt + q_a\delta(u)dudt \tag{14}$$

$$\text{Cum}[dN_a(t+u), dN_b(t+v), dN_c(t)] = q_{abc}(u,v)dudvdt, \quad a, b, c \text{ distinct} \tag{15}$$

$$\text{Cum}[dN_a(t+u), dN_a(t+v), dN_b(t)] =$$
$$= q_{aab}(u,v)dudvdt + \delta(u-v)q_{ab}(u)dudvdt, \ a \neq b \tag{16}$$

and

$$\text{Cum}[dN_a(t+u), dN_a(t+v), dN_a(t)] = q_{aaa}(u,v)dudvdt$$
$$+ (\delta(u) + \delta(v) + \delta(u-v))q_{aa}(u)dudvdt + \delta(u)\delta(v)q_a dudvdt \tag{17}$$

where $\delta(a)$, $-\infty < a < \infty$ is the Dirac Delta function.

At this point it must be stressed that the point process $\{N(t)\}$ is defined on a continuous parametric space, since $-\infty < t < \infty$. These point processes are called continuous with respect to t. In practice, however, the point processes will have to become discrete for further analysis in the time or in the frequency domain. Thus, if the sampling rate is $1/b$ points per unit time, then the discrete point process that follows will be $\{\mathbf{N}(t_j);\ t_j = jb,\ j = 0, \pm 1, \ldots\} = \{N_1(t_j), \ldots, N_r(t_j);\ t_j = jb,\ j = 0, \pm 1, \ldots\}$. In this case the notation $\varDelta \boldsymbol{N}(t_j) = \boldsymbol{N}(t_j, t_{j+1}] = \boldsymbol{N}(bj, b(j+1)]$ is used and the definition of the product and the cumulant densities are modified accordingly.

3 Estimation of the Cumulant Densities

In this section a way of estimating the cumulant densities of second- and third-order is presented. Some of the asymptotic results that are discussed below require the stationary point process (s.p.p.) to satisfy the following assumption.

Assumption 1 The s.p.p. $\{N(t)\}$ is such that

$$\int \cdots \int (1 + |u_j|)|q_{a_1 \cdots a_l}(u_1, \ldots, u_{l-1})|du_1 \cdots du_{l-1} < \infty \tag{18}$$

for $j = 1, 2, \ldots, l - 1$, $a_1, \ldots, a_l = 1, \ldots, r$ and $l = 2, 3, \ldots$. This assumption implies that well-separated (in time) values of the point process are weak dependent.

3.1 The Periodogram-Based Estimate

Let $\{\mathbf{N}(t_j);\ t_j = jb,\ j = 0, \pm 1, \ldots\}$ be a discrete vector-valued s.p.p. obtained by sampling a continuous s.p.p. with sampling rate $1/b$ points per unit time. Moreover, it is assumed that the discrete s.p.p. consists of T/b points $\mathbf{N}(1/b), \ldots, \mathbf{N}(T/b)$ where T is the observed time interval. The modified (corrected by subtracting the mean intensity) finite Fourier–Stieltjes transform can be approximated by the sum

$$\hat{d}_a^{(T/b)}(\lambda) \approx \sum_{j=0}^{T/b-1} \exp(-i\lambda t_j)[N_a(t_j + b) - N_a(t_j) - \hat{p}_a b], \tag{19}$$

where $\hat{p}_a = N_a(T)/T$ is the estimate of the mean intensity and $N_a(T)$ is the number of events of the component N_a in the interval (0, T] [27].

The modified periodogram of second-order is now defined by

$$\hat{I}_{a_i a_j}^{(T)}(\lambda) = \frac{1}{2\pi T} \hat{d}_{a_i}^{(T)}(\lambda) \hat{d}_{a_j}^{(T)}(-\lambda) \tag{20}$$

for $b = 1$. Then the estimates of the cumulant densities of second-order are obtained by

$$\hat{q}_{a_i a_j}(u) = \frac{2\pi}{T} \sum_{k=0}^{T/b-1} W_T(\lambda_k)\big(\hat{I}_{a_i a_j}^{(T)}(\lambda_k) - \delta_{ij} \frac{\hat{p}_{a_i}}{2\pi}\big) e^{i\lambda_k u}, \tag{21}$$

where $W_T(\lambda) = W(b_T \lambda)$ is a convergence factor, $\lambda_k = 2\pi k/T$ and δ_{ij} is Kronecker delta [2, 5]. The quantity b_T is called the bandwidth of the convergence factor [26]. It can be shown that

$$\sqrt{b_T T}\big(\hat{q}_{a_i a_j}(u) - q_{a_i a_j}(u)\big) \sim Normal\left(0, \frac{p_{a_i a_j}(u) \int W^2(\lambda) d\lambda}{2\pi}\right), \tag{22}$$

for $b_T T \to \infty$ as $T \to \infty$ and $b_T \to 0$ [28].

This result suggests that the variance of the estimate depends on the unknown quantity $p_{a_i a_j}(u)$ and therefore the construction of a confidence interval for the cumulant density is problematic. Now, under the hypothesis that the components N_{a_i} and N_{a_j} are independent it holds $p_{a_i a_j} = p_{a_i} p_{a_j}$, but this is not always true in practice. By extending the results presented above we can obtain estimates for the cumulant densities of third-order as well.

3.2 The Direct Method of Estimating the Cumulant Densities

We can construct a new process as follows

$$X_k(t_l) = \frac{[\Delta N_{a_1}(t_l + u_1) - \hat{p}_{a_1} b] \cdot [\Delta N_{a_2}(t_l + u_2) - \hat{p}_{a_2} b] \cdots [\Delta N_{a_k}(t_l) - \hat{p}_{a_k} b]}{b^k}$$

The mean value of the new process is given by

$$E[X_k(t_l)] = q_{a_1 \cdots a_k}(u_1, u_2, \ldots, u_{k-1}) = \mu.$$

Now, the problem of constructing confidence intervals for the cumulant densities of kth order is related to the problem of constructing confidence intervals for mean values [3]. If

$$q^T_{a_1 \cdots a_k}(u_1, u_2, \ldots, u_{k-1}) = T^{-1} \sum_{t=0}^{T-1} X_k(t) = \mu^T \tag{23}$$

is the estimate of the mean value of $X_k(t)$, then

$$\mu^T \sim Normal\big(\mu, T^{-1} 2\pi f_{XX}(0)\big), \tag{24}$$

where $f_{XX}(0)$ denotes the power spectrum of $X_k(t)$ at zero frequency [5].

An estimate of $f_{XX}(0)$ can be obtained as follows

$$f^{(T)}_{XX}(0) = L^{-1} \sum_{s=1}^{L} I^{(T)}_{XX}\left(\frac{2\pi s}{T}\right), \tag{25}$$

where L is the number of the components of the periodogram used in the estimation. It can be shown [5] that the estimate of $f_{XX}(0)$ is asymptotically distributed as $f^T_{XX}(0) \sim f_{XX}(0) \frac{\mathbf{X}^2_{2L}}{2L}$. Thus,

$$\frac{\mu^T - \mu}{\sqrt{T^{-1}2\pi f_{XX}^T(0)}} \sim \mathbf{t}_{2L}, \tag{26}$$

and a $100\beta\%$ confidence interval for μ can be constructed as follows

$$\mu^T - \mathbf{t}_{2L}\left(\frac{1+\beta}{2}\right)\sqrt{T^{-1}2\pi f_{XX}^T(0)} < \mu < \mu^T + \mathbf{t}_{2L}\left(\frac{1+\beta}{2}\right)\sqrt{T^{-1}2\pi f_{XX}^T(0)} \tag{27}$$

4 Logistic Regression Model

In this section a parametric method is presented based on a logistic regression model. Both the dependent and the explanatory variables are the stationary point processes that can be considered as long binary time series consisting of zeroes and ones. A zero means that there is no an event in a small time interval, whereas a one means that there is an event. Before the formulation of a logistic regression model, we need to define the likelihood function. Actually, the penalized likelihood function will be used because the phenomenon of quasi-complete separation occurs.

4.1 Penalized Likelihood Function

Let $y_t,\ t = 1, \ldots, n$ be the binary responses of n random variables Y_t, where $Y_t \sim B(1, \pi_t)$ and $\mathbf{x}_t^T$ a $1 \times k$ row-vector of measurements corresponding to covariates and dummy variables corresponding to factor levels. Then the logistic regression model is given by

$$\pi_t = \{1 + \exp(-\mathbf{x}_t^T \cdot \beta)\}^{-1}, \tag{28}$$

where β is the parameter $k \times 1$ column-vector and $\pi_t = \text{Prob}\ \{y_t = 1\}$ [9, 14, 23]. Maximum likelihood estimates of the parameters $\beta_j,\ j = 1, \ldots, k$ and consequently of the probabilities π_t, are obtained by maximizing the likelihood function

$$L(\beta|y) = \prod_{t=1}^{n} \pi_t^{y_t} \cdot (1-\pi_t)^{1-y_t}, (y_t = 0, 1),$$

or the log-likelihood function $l(\beta|y) = \log L(\beta|y)$,

$$l(\beta|y) = \sum_{t=1}^{n}\left[y_t \cdot \log\left[\frac{1}{1+\exp\left(-\mathbf{x}_t^T \cdot \beta\right)}\right] + (1-y_t) \cdot \log\left[1 - \frac{1}{1+\exp(-\mathbf{x}_t^T \cdot \beta)}\right]\right], \tag{29}$$

using a Newton–Raphson algorithm or the method of scoring. When the method of scoring is used, the estimates of the parameters β_j are obtained as the solutions to the score equations

$$\frac{\partial l(\beta|y)}{\partial \beta_j} \equiv U(\beta_j) = \sum_{t=1}^{n} (y_t - \pi_t) \cdot x_{tj} = 0, \quad (j = 1, \ldots, k). \tag{30}$$

In order to remove the small sample bias $O(n^{-1})$ of the maximum likelihood estimates, Firth suggested the use of modified score functions, given by

$$U(\beta_j)^* \equiv U(\beta_j) + 1/2\text{trace}[I(\beta)^{-1} \cdot \{\partial I(\beta)/\partial \beta_j\}] = 0, \tag{31}$$

$(j = 1, \ldots, k)$, where $I(\beta)^{-1}$ is the inverse of the information matrix evaluated at β [11–13]. This modification suggests that a penalized likelihood function is used, i.e. $L(\beta)^* = L(\beta) \cdot |I(\beta)|^{1/2}$. The penalty function $|I(\beta)|^{1/2}$ is known as Jeffreys invariant prior for this problem [18]. When the logistic regression model of (28) is assumed, then the modified score functions of (31) become

$$U(\beta_j)^* = \sum_{t=1}^{n} (y_t - \pi_t + h_t \cdot (1/2 - \pi_t)) \cdot x_{tj} = 0, \tag{32}$$

$(j = 1, \ldots, k)$, where h_t is the tth diagonal element of the 'hat' matrix

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2},$$

and $W = \text{diag}\{\pi_t \cdot (1 - \pi_t)\}$. Then, the maximum likelihood estimates can be found using iterations until convergence is obtained. At the mth iteration, the vector of the estimates will be given by

$$\beta^{(m)} = \beta^{(m-1)} + I(\beta^{(m-1)})^{-1} \cdot U(\beta^{(m-1)})^*. \tag{33}$$

The above procedure eliminates the problem of separation. In Heinze [16] it is declared that "only those problems of estimation remain which can also occur with the general linear model, for example, problems due to multicolinearity or nearly degenerate risk factor distributions". Other methods have also been proposed for removing the small sample bias and/or dealing with the separation problem. The above technique has been proved superior to its alternatives [31].

It is known that the resulting profile likelihoods for the coefficients are often asymmetrical, since they are close to boundary conditions. Thus, it is suggested that the penalized likelihood ratio test should be used instead of a possible misleading Wald-type inference statistic. In addition, the profile penalized likelihood confidence intervals are in these cases more accurate than the asymptotic ones. In this paper the profile penalized likelihood confidence intervals are computed using the algorithm of Venzon and Moolgavkar [15, 16, 29]. The computations were carried out in MATLAB(R2009a).

4.2 The Proposed Logistic Model

The logistic regression model that is proposed for the description of the stochastic neurophysiological system is given by

$$log\left(\frac{\pi_t}{1-\pi_t}\right) = SF_t + V_t - \beta_0. \tag{34}$$

The function SF_t is called the summation function and is defined by

$$SF_t = SF_{1,t} + SF_{2,t} = \sum_{u=1}^{u_M} \beta_{1,u} x_{1,t-u} + \sum_{v=1}^{v_M} \beta_{2,v} x_{2,t-v}, \tag{35}$$

where $x_{p,t-t'}$ is the observation of the pth explanatory variable at time $t-t'$ ($p = 1, 2$ and $t' = u$ or v) and $\{\beta_{1,u}, u = 1, \ldots, u_M; \beta_{2,v}, v = 1, \ldots, v_M\}$ are unknown coefficients.

The function V_t is called the recovery function and it can be described by a polynomial function of order s which is given by

$$V_t = \begin{cases} \sum_{i=1}^{s} \beta_{3,i}(\tau_t - \zeta - 1)^i, & \text{if } \tau_t \geq \zeta + 1 \\ 0, & \text{if } \tau_t < \zeta + 1 \end{cases}, \tag{36}$$

where $\{\beta_{3,i}\}$ are the coefficients of the function and ζ is the minimum inter-spike interval of the dependent variable. In addition, τ_t denotes the time elapsed since the last event of the dependent variable. The β_0 is an unknown constant value and its role will be explained in the following section with the example [6, 19].

It must be pointed out that separation is not only a problem related with small or medium sized data sets. When a logistic regression model like the one given by (34) needs to be fitted to binary data which come from a counting procedure the quasi-complete separation phenomenon is expected to occur even in large samples. As in the case of the complete separation, the quasi-complete separation also results in infinite maximum likelihood estimates when binary covariates are involved. Moreover, an insufficient memory problem should also be expected when evaluating the 'hat' matrix H, since its dimension is $n \times n$. Thus, a modification of the *logistf* routine should be made in order to store only the diagonal elements of the 'hat' matrix. This can be done as a two step procedure. First, compute the $W^{1/2}X(X^TWX)^{-1}$ part of the multiplication and then by using a loop evaluate only the diagonal elements needed.

4.3 The Randomized Quantile Residuals

After fitting the model to the observed data, it is necessary to check if the fitted model is valid, i.e. that the fitted values are correctly estimated and are adequately "close" to the observed values. A common technique used to examine the adequacy of the fitted model is based on the residuals, which can be thought of as measurements of agreement between the observed and the fitted response values.

Suppose that the logistic regression model given by (28) is fitted to the n binary responses $y_t,\ t = 1, \ldots, n$. Let $F(y_t; \pi_t) = \Pr(Y_t \leq y_t) = \sum_{m=0}^{\lfloor y_t \rfloor} \pi_t^m (1-\pi_t)^{1-m}$ be the cumulative binomial distribution of the tth binary response, where $\lfloor y_t \rfloor$ is the 'floor' under y_t, i.e. the greatest integer less than or equal to y_t. Then the *randomized quantile residuals* are defined by

$$r_{rq,t} = \Phi^{-1}\{u_t\}, \tag{37}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian, and u_t is a uniform random variable on the interval

$$(a_t, b_t] = \big(\lim_{y \uparrow y_t} F(y; \hat{\pi}_t),\, F(y_i; \hat{\pi}_t)\big] \approx \big[F(y_t - 1; \hat{\pi}_t),\, F(y_i; \hat{\pi}_t)\big].$$

The $r_{rq,t}$ follow the standard normal distribution, apart from sampling variability in $\hat{\pi}_t$. The randomized quantile residuals were first defined by Dunn and Smyth and can be similarly used for any discrete distributed response [10].

5 The Neurophysiological Example

In this section the neurophysiological system of the muscle spindle is studied when it is affected simultaneously by an alpha (α) and a gamma motoneurone (γ). The muscle spindle is an element of the neuromuscular system and plays a important role in the initiation of movement and the maintenance in posture. It is also a transducer which responds to different stimuli applied on it. Most skeletal muscles contain a number of these transducers, which lie parallel with the fibers of the muscle (known as extrafusal fibers). The fibers within a muscle spindle, known as intrafusal muscle fibers, are considerably shorter than the extrafusal fibers. These are three different types of intrafusal fibers, the dynamical nuclear-bag (DNB), the static nuclear-bag (SNB) and the nuclear chain (NC). The effect of a stimulus on the muscle spindle is transmitted to the spinal cord by the terminal branches of the axons of sensory neurons which are wrapped round all of the intrafusal fibers.

When a muscle is held at a fixed length, the sensory axons from the muscle spindle produce nerve impulses at a constant rate that depends upon the muscle length. The nerve impulse is a localized voltage change that occurs across the membrane surrounding the nerve cell and axon. Its amplitude is approximately 100 mV, and its duration 1 ms. Nerve impulses are known as action potentials or, because of their relatively short duration, spikes.

There are several classes of nerve cells which lie within the spinal cord in groups called nuclei some of which may contain as many as 2000 cells. One of these groups, called alpha motoneurones, have long axons which leave the spinal cord to innervate the extrafusal muscle fibers forming the main mass of the muscles responsible for generating forces or changes of length. The axons of the alpha motoneurones normally conduct nerve impulses from the cell body to the extrafusal muscle fibers. When a nerve impulse reaches the junction between the axon and the muscle fiber a sequence of electro-chemical events occurs which leads to the contraction of the entire muscle fiber. Another group of cells, called gamma motoneurones, lie within the spinal cord and in the neighborhood of the alpha motoneurones. These cells are considerably smaller in diameter than the alpha motoneurones and their long axons innervate the intrafusal muscle fibers. When a gamma motoneurone affects the muscle spindle by transmitting nerve impulses to the intrafusal muscle fibers, the response of the muscle spindle sensory axons, called the Ia response, is modified. It has been suggested that activity in the gamma motoneurone axons may modify the mechanical properties of the intrafusal fibers to discharge nerve impulses in the sensory axon [1, 20, 21, 25, 30].

5.1 The Nonparametric Approach

The aim here is to construct 95 % confidence intervals for the cumulant densities of the second- and third-order in order to study the behavior of the neuromuscular system of the muscle spindle. In particular, the response of the primary sensory axon (Ia) is studied under the simultaneous effect of an alpha (α) and a gamma (γ) motoneurone. The point processes have been recorded in a time interval $T = 11360$ ms with a sampling rate 1 point per ms. The number of events in each component of the point process is $N_\gamma(T) = 691$, $N_\alpha(T) = 163$ and $N_{Ia}(T) = 358$.

In Fig. 1 the estimates of the second- and third-order cumulant densities $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u, u-v)$ are presented for the case of the effect of an alpha motoneurone on the Ia sensory axon in the presence of a gamma motoneurone. It becomes clear from previous work that the presence of the gamma motoneurone reduces the effect of the alpha motoneurone on the muscle spindle and produces a second region of positive interaction about the 30th ms. The patterns in the $q_{Ia,\alpha,\alpha}(u, v)$ are in agreement with the graphs of the second-order cumulant densities.

In Fig. 2 the estimates of the cumulant densities of second- and third-order $q_{Ia,\gamma}(u)$, $q_{Ia,\alpha}(u - v)$ and $q_{Ia,\alpha,\gamma}(u, u - v)$ for the case of the simultaneous effect of an alpha and a gamma motoneurone are presented. Firstly, a positive interaction is observed between the gamma motoneurone and the Ia response of the sensory axon in the region where the alpha motoneurone blocks the response of the system. Secondly, it is seen that the cumulant density of third-order contains significant information, because the correspondence with values of the second-order cumulant

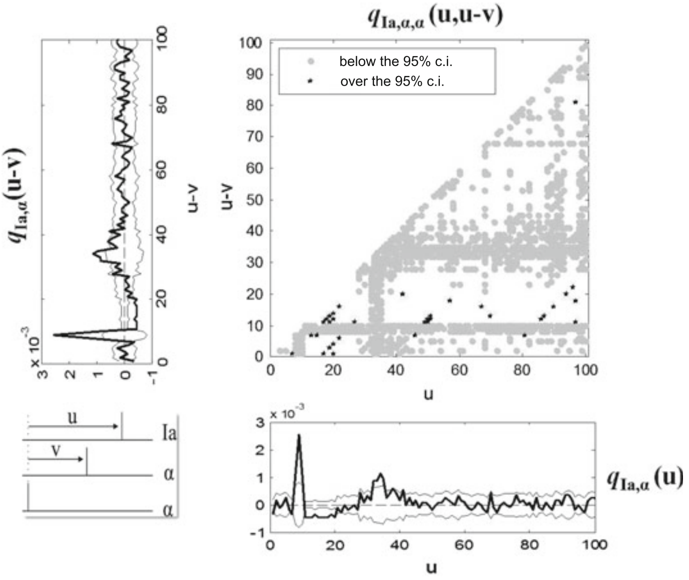


Fig. 1 The estimates of the second- and third-order cumulant densities $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u, u - v)$ for the case of the effect of an alpha motoneurone on the Ia sensory axon in the presence of a gamma motoneurone.

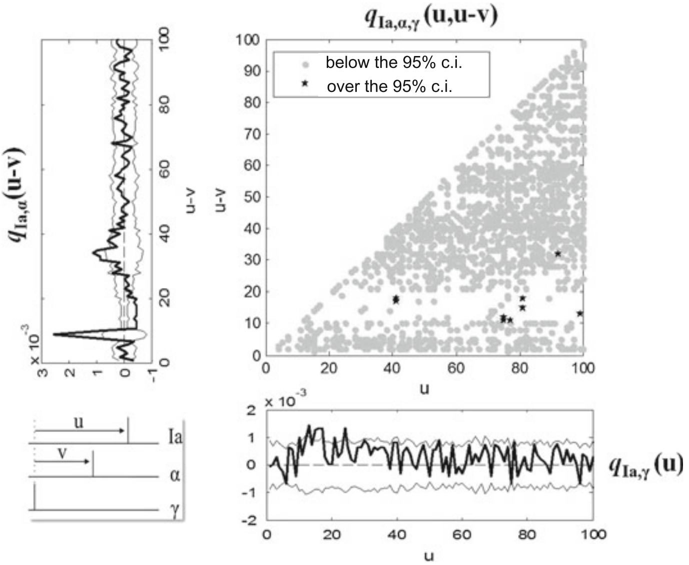


Fig. 2 The estimates of the second- and third-order cumulant densities $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u)$, $q_{Ia,\alpha}(u, u - v)$ for the case of the effect of an alpha motoneurone on the Ia sensory axon in the presence of a gamma motoneurone.

densities has been destroyed , whereas there are only eight points with significant positive values for $10 \leq u - v \leq 20$.

5.2 The Parametric Approach

The logistic regression model given by (34) is used here in order to study the behavior of the muscle spindle by recording its response from the Ia sensory to the combined effect of an alpha and a gamma motoneurone $(\gamma + \alpha)$. The axon (Ia) of the sensory nerve fires when the potential of the membrane exceeds a critical level called threshold. The potential of the membrane is influenced both by external and internal processes. Let Y_t denote the firing process of the Ia sensory axon which is associated with the muscle spindle. By choosing the time sampling h, the observations of the output can be written as follows:

$$y_t = \begin{cases} 1, \text{ when an output spike occurs in } (t, t+h] \\ 0, \text{ otherwise} \end{cases}, \tag{38}$$

where $t = h, \ldots, nh$ and $N = nh$ is the time interval in which the time series is observed. In our case we choose $h = 1$ ms. Let $X_{p,t}$, $p = 1, 2$ denote the inputs of the system imposed by $(\gamma + \alpha)$ stimulus. They consist of the observations $x_{p,t}$, $p = 1, 2$ which are binary time series defined as the y_ts.

Let θ_t denote the threshold level at the trigger zone at time t, given by $\theta_t = \theta_t^* + \varepsilon_t$, where ε_t is the noise process which includes contributions of unmeasured terms that influence the firing of the system. There is experimental evidence and theoretical verification that ε_t follows approximately a normal distribution [4, 17]. θ_t^* is a function of t, which represents the form of threshold at time t. We assume that $\theta_t^* = \beta_0$ i.e. an unknown constant.

The function representing the external processes that influence the potential of the membrane at the trigger zone are the summation functions $SF_{1,t}$ and $SF_{2,t}$. The internal processes are responsible for possible spontaneous firing of the system and they are described by the recovery function V_t. The minus before the β_0 in the relation (34) which symbolizes the threshold indicates that the strength of the external and the internal processes must exceed the level of threshold in order to get an output event

All the estimates presented here are obtained by using the penalized method discussed in Sect. 4. The iterations were stopped when the sum of the distance between the estimated parameters in two successive steps was lower than 1E-05.

In our case, which involves two inputs and one output, 138 parameters were estimated (1 constant, 7 for the recovery function, 50 for the coefficients of the (γ) summation function and 80 for the coefficients of the (α) summation function). The estimated penalized log-likelihood function was $l_{01} = -1062.9$ and the estimated penalized log-likelihood function of the null model, described as in the (γ) case, was $l_{00} = -1360.2$. In this case $2(l_{01} - l_{00})$ was 594.6 with a p-value equal to 0. The 138 parameters (without the profile likelihood confidence intervals) were estimated after

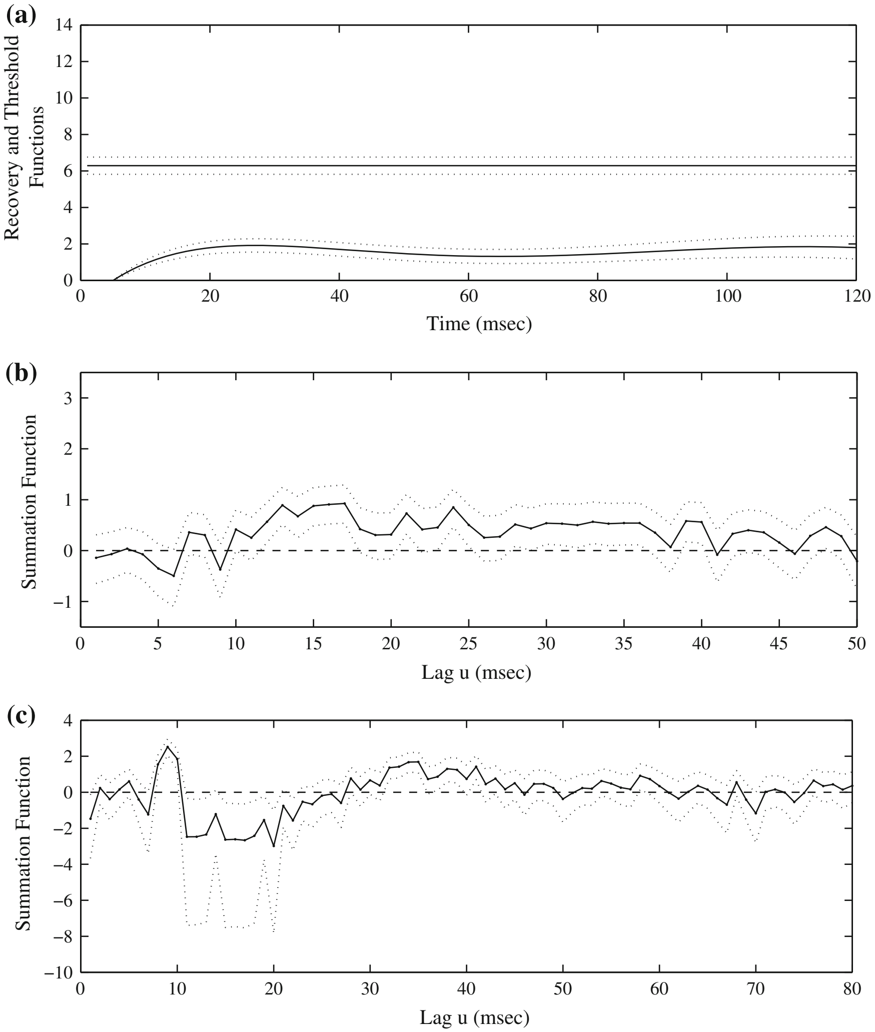


Fig. 3 The $(\gamma + \alpha)$ case. **a** The estimated recovery function and the constant threshold, **b** the penalized maximum likelihood estimates of the coefficients of the (γ) part of the summation function and **c** the penalized maximum likelihood estimates of the coefficients of the (α) part of the summation function for the $(\alpha + \gamma)$ case. The *dotted lines* above and below correspond to the 95 % profile likelihood confidence intervals

18 iterations in 7.27 s. Figure 3a shows the recovery function described by a seventh order polynomial, the constant threshold and the 95 % profile likelihood confidence intervals. It is clear that there is an increase in the estimate of the recovery function for about 40 ms but afterwards it stabilizes below the threshold. The presence of the gamma motoneurone seems to increase the distance from threshold again and

the system does not fire spontaneously when both the motoneurones are present. The seventh order polynomial was selected for this case, since the penalized log-likelihood for the eighth order polynomial was $l_1 = -1042.8$ and the restricted penalized log-likelihood for the sixth order polynomial was $l_0 = -1143.0$ suggesting a p-value equal to 0.5271.

Figure 3b shows the estimated coefficients for the (γ) summation function together with the 95 % profile likelihood confidence intervals when both the motoneurones are present. We observe that between 11 and 40 ms, the coefficients of the summation function are decreased by the presence of the alpha motoneurone, but they remain positive. This suggests that in the interval between 11 and 40 ms the response is still accelerated but the odds of firing are reduced by the presence of the alpha motoneurone.

Figure 3c shows the estimated coefficients for the (α) summation function together with the 95 % profile likelihood confidence intervals when both the motoneurones are present. The estimated coefficients of the summation function remain positive and unaffected for a very short period in the beginning, indicating that the acceleration of the system's firing is clearly a characteristic due to the effect of the alpha motoneurone. After this we observe that the duration of the blockage has been reduced from almost 40 ms to 10 ms. Moreover the increase in the summation function has started about 30 ms earlier.

Figure 4a shows the Q–Q plot of the randomized quantile residuals of the fitted model. The 5 % rejection regions were computed after 1000 Monte Carlo simulations. Only 0.7 % of the 11,327 residuals lie outside the 5 % rejection regions with a small deviation at large values. In addition, the Anderson–Darling statistic is 0.5643 with a p-value 0.1442.

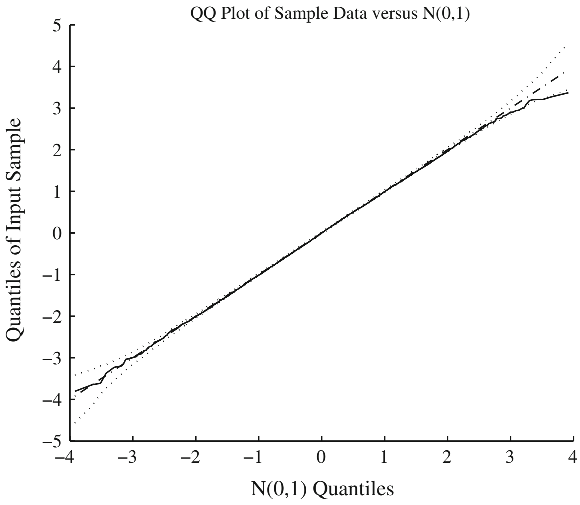


Fig. 4 The Q–Q plot of the randomized quantile residuals of the proposed fitted model, for the $(\gamma + \alpha)$ case.The 5 % rejection regions after 1000 Monte Carlo simulations are presented by the *dotted lines*

6 Conclusions

Two different approaches have been used in order to identify a stochastic system involving stationary point processes (s.p.p.). The first approach is non-parametric and is based on the second- and third-order cumulant densities. Estimates of these densities are obtained by using the modified periodogram of the increments of the s.p.p. Confidence intervals are also constructed which can reveal possible non-linearities of the system. The second approach is parametric and is based on a logistic model. The problem of the quasi complete separation which appears in the estimation of the models parameters is solved very fast and efficiently. The model involves two inputs and one output and describes a neuromuscular system called muscle spindle. The behavior of the system is studied when it is affected by two stimuli, a gamma and an alpha motoneurone. The presence of the alpha motoneurone reduces the effect of the gamma motoneurone on the neuromuscular system, while the presence of the gamma motoneurone restricts the blockage of the alpha motoneurone on the system but it creates a second effect.

References

1. Boyd, I.A.: The isolated mammalian muscle spindle. Trends Neurosci. **3**(11), 258–265 (1980)
2. Brillinger, D.R.: Estimation of product densities. In: Frane, J.W. (ed.) Computer Science and Statistics: 8th Annual Symposium, pp. 431–438. Los Angeles, U.C.L.A. (1975)
3. Brillinger, D.R.: Confidence intervals for the crosscovariance function. Selecta Statistica Canadiana **5**, 3–16 (1979)
4. Brillinger, D.R.: Nerve cell spike train data analysis: a progression of technique. JASA **87**(418), 260–271 (1992)
5. Brillinger, D.R.: Time Series: Data Analysis and Theory, vol. 36. SIAM, Philadelphia (2001)
6. Brillinger, D.R.: Some statistical methods for random process data from seismology and neurophysiology. In: Guttorp, P., Brillinger, D. (eds.) Selected Works of David Brillinger, Selected Works in Probability and Statistics, pp. 89–142. Springer, New York (2012)
7. Cox, D.R., Lewis, P.A.W.: The Statistical Analysis of Series of Events. Wiley, London (1966)
8. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure. Springer, New York (2007)
9. Dobson, A.J., Barnett, A.J.: An Introduction to Generalized Linear Model, 3rd edn. CRC Press, Boca Raton (2008)
10. Dunn, P.K., Smyth, G.K.: Randomized quantile residuals. J. Comput. Graph. Stat. **5**(3), 236–244 (1996)
11. Firth, D.: Bias reduction, the Jeffreys prior and GLIM. Advances in GLIM and Statistical Modelling, pp. 91–100. Springer, New York (1992)
12. Firth, D.: Generalized linear models and Jeffreys priors: an iterative weighted least-squares approach. In: Dodge, Y., Whittaker, J. (eds.) Computational Statistics, vol. 1, pp. 553–557. Physica-Verlag, Heidelberg (1992)
13. Firth, D.: Bias reduction of maximum likelihood estimates. Biometrika **80**(1), 27–38 (1993)
14. Hardin, J.W., Hilbe, J.: Generalized Linear Models and Extensions. Stata Press, College Station (2007)
15. Heinze, G., Ploner, M.: A SAS macro, S-plus library and r package to perform logistic regression without convergence problems. Technical report 2/2004, Medical University of Vienna, Vienna (2004)

16. Heinze, G., Schemper, M.: A solution to the problem of separation in logistic regression. Stat. Med. **21**(16), 2409–2419 (2002)
17. Holden, A.V.: Models of the Stochastic Activity of Neurones. Springer, Berlin (1976)
18. Jeffreys, H.: An invariant form for the prior probability in estimation problems. Proc. R. Soc. A **186**(1007), 453–461 (1946)
19. Karavasilis, G.J., Kotti, V.K., Tsitsis, D.S., Vassiliadis, V.G., Rigas, A.G.: Statistical methods and software for risk assessment: applications to a neurophysiological data set. Comput. Stat. Data Anal. **49**(1), 243–263 (2005)
20. Kotti, V.K., Rigas, A.G.: Identification of a complex neurophysiological system using the maximum likelihood approach. J. Biol. Syst. **11**(02), 189–204 (2003)
21. Kotti, V.K., Rigas, A.G.: Logistic regression methods and their implementation. In: Edler, L., Kitsos, C.P. (eds.) Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment, pp. 355–369. Wiley, New York (2005)
22. Lindsay, K.A., Rosenberg, J.R.: Linear and quadratic models of point process systems: contributions of patterned input to output. Prog. Biophys. Mol. Biol. **109**(3), 76–94 (2012)
23. MacCullagh, P., Nelder, J.A.: Generalized Linear Models, vol. 37, 2nd edn. CRC Press, Boca Raton (1989)
24. Marmarelis, V.Z.: Nonlinear Dynamic Modeling of Physiological Systems. Wiley-IEEE Press, New York (2004)
25. Matthews, P.B.C.: Evolving views on the internal operation and functional role of the muscle spindle. J. Physiol. **320**, 1–30 (1981)
26. Priestley, M.B.: The role of bandwidth in spectral analysis. J. R. Stat. Soc. C **14**(1), 33–47 (1965)
27. Rigas, A.G.: Spectral analysis of stationary point processes using the fast Fourier transform algorithm. J. Time Ser. Anal. **13**(5), 441–450 (1992)
28. Tsitsis, D.S., Karavasilis, G.J., Rigas, A.G.: Measuring the association of stationary point processes using spectral analysis techniques. Stat. Methods Appl. **21**(1), 23–47 (2012)
29. Venzon, D.J., Moolgavkar, S.H.: A method for computing profile-likelihood-based confidence intervals. Appl. Stat. **37**, 87–94 (1988)
30. Windhorst, U.: Muscle proprioceptive feedback and spinal networks. Brain Res. Bull. **73**(4), 155–202 (2007)
31. Zorn, C.: A solution to separation in binary response models. Polit. Anal. **13**(2), 157–170 (2005)

Monitoring Environmental Risk by a Methodology Based on Control Charts

Helton Saulo, Victor Leiva and Fabrizio Ruggeri

Abstract We propose a methodology based on control charts when the contaminant concentration follows a Birnbaum-Saunders distribution, which is implemented in the R software. We investigate the performance of this methodology through Monte Carlo simulations. An example with real-world data is given as an illustration of the proposed methodology.

Keywords Birnbaum-Saunders distribution · Contaminant concentration · Maximum likelihood and moment estimation · Monte Carlo simulation · R software · X-bar charts

1 Introduction

Control charts are popularly used tools for quality monitoring because these are simple to interpret and easy to be updated whenever further data are available; see Montgomery [27] and Figueiredo and Gomes [10]. These charts provide an earlier alert when a process is going to be out-of-control, so that an action can be taken to bring back it to the in-control state. Although such charts originated from industry applications, their use has been extended to monitoring of environmental and health risk; see Grigg and Farewell [11], Woodall [35], Morrison [26], and Manly [24].

The first ingredient of our methodology for monitoring environmental risk is the control chart. International guidelines regulate dangerous concentrations of contaminants, which are often modeled by statistical distributions. By means of these regulations and distributions, administrative targets and environmental alerts can be

H. Saulo
Instituto de Matemática e Estatística, Universidade Federal de Goiás, Goiania, Brazil

V. Leiva (✉)
Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez,
Avda. Padre Hurtado 750, Oficina A-215, Viña del Mar, Chile
e-mail: victorleivasanchez@gmail.com
URL: http://www.victorleiva.cl

F. Ruggeri
CNR IMATI, Milano, Italy

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_14

established. Such alerts allow human health to be protected, because they address episodes of extreme contamination that need corrective measures. These episodes are manifested by an increment of the incidence and severity of diseases; see Marchant et al. [25]. Thus, environmental risk monitoring is important and can be conducted by control charts.

Statistical distributions used for modeling contaminant concentrations are often asymmetrical (with positive skewness) and have support in the set of real numbers for values greater than zero. Consequently, the normal distribution is not suitable for this type of modeling.

Often environmental researchers transform their data, using for example the Box-Cox power transformations [7, 14], to eliminate asymmetry, so that the normal distribution can be used. However, it has been shown that analyses performed under an inappropriate data transformation reduce the power of the study; see Huang and Qu [12], Leiva et al. [18] and references therein. In any case, even when an appropriate transformation is used, a problem of data interpretation still remains. An alternative way to avoid the data transformation is to model them directly through a suitable distribution. For this purpose, a number of researchers have utilized the lognormal (LN) distribution for modeling environmental data, mainly due to its physical arguments [29] and its relationship with the normal distribution. However, also the beta, exponential, extreme values, gamma, inverse Gaussian, Johnson SB, log-logistic, Pearson and Weibull distributions have been used for analyzing this kind of data, although without theoretical arguments see Marchant et al. [25].

Our second ingredient is an asymmetric distribution named Birnbaum-Saunders (BS), which has attracted considerable attention; see, e.g. Johnson et al. [13, pp. 651–663], and Fierro et al. [9]. This is due to its good properties, its relation with the normal distribution and its applications in diverse fields including environmental sciences; see Leiva et al. [16, 17, 19], Vilca et al. [33, 34], Ferreira et al. [8], Marchant et al. [25], and Saulo et al. [32]. However, the most important aspect of the BS distribution is that it has physical arguments and statistical properties similar to the LN distribution to model this type of phenomena (e.g. asymmetry and an inverse bathtub shaped or unimodal hazard rate -HR-). Nevertheless, the BS distribution has further properties that the LN one does not have. This allows us to postulate the BS distribution as a candidate to model contaminant concentrations. The BS distribution is implemented in a statistical software named `R` (www.R-project.org) by the `gbs` package [4]. This package allows us to obtain probabilities, estimate parameters, generate random numbers and conduct goodness-of-fit. Applications of the BS distribution to quality monitoring tools can be found in Balakrishnan et al. [3], Lio and Park [22], Lio et al. [23], Leiva et al. [20] and references therein. Despite the good properties that the BS distribution has, it does not share the reproductive property. Thus, sum of random variables (RVs) with BS distribution (BSsum in short) does not have a BS distribution. Raaijmakers [30, 31] found the BSsum distribution and we present here its implementation in an `R` package called `bssum`.

The main objective of this paper is to propose a methodology based on control charts for monitoring environmental risk when the contaminant concentration follows a BS distribution. In particular, we analyze the pH levels of five rivers from the South

Island of New Zealand. We develop X-bar control charts with the help of the BSsum distribution.

The rest of the paper is organized as follows. In Sects. 2 and 3, we introduce BS and BSsum distributions and their properties, features and implementations in the `R` software. In Sects. 4 and 5, we propose the mentioned methodology and investigate its performance through Monte Carlo (MC) simulations. In Sect. 6, we apply it to real-world environmental data to illustrate its potential. In Sect. 7, we discuss some conclusions of this work.

2 Birnbaum-Saunders Distribution

In this section, we provide some probabilistic, statistical and computational aspects of the BS distribution, including several features and properties, which are useful for developing the methodology proposed in Sect. 4.

The BS distribution has the following characteristics. It has two parameters, one of shape α and another of scale β, with β being also a position parameter because it corresponds to its median. In addition, the BS distribution is asymmetrical with positive skewness and unimodality. It allows us to model data that take values greater than zero. Furthermore, the BS distribution is closely related to the normal distribution, so that it inherits several of its good properties. Moreover, the BS distribution has a HR with several shapes including increasing and unimodal, which are particularly useful in environmental data analyses; see Vilca et al. [33].

When a RV X follows a BS distribution with parameters $\alpha > 0$ and $\beta > 0$, the notation $X \sim \mathrm{BS}(\alpha, \beta)$ is used. BS and standard normal distributions are related by means of the RVs

$$X = \beta\big(\alpha Z/2 + ((\alpha Z/2)^2 + 1)^{1/2}\big)^2 \quad \text{and} \quad Z = \frac{1}{\alpha}\xi\,(X/\beta) \sim \mathrm{N}(0, 1), \tag{1}$$

where $\xi(u) = \sqrt{u} - 1/\sqrt{u} = 2\sinh(\log(\sqrt{u}))$. Relation in (1) allows us to obtain

$$W = \frac{1}{\alpha^2}\xi^2\,(X/\beta) \sim \chi^2(1). \tag{2}$$

Result given in (2) is useful for establishing goodness-of-fit by probability plots and detecting atypical data by the Mahalanobis distance. If $X \sim \mathrm{BS}(\alpha, \beta)$, then its probability density (PDF) and cumulative distribution (CDF) functions are respectively expressed as

$$f_X(x; \alpha, \beta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\alpha^2}\xi^2\left(\frac{x}{\beta}\right)\right) \frac{x^{-3/2}(x+\beta)}{2\alpha\sqrt{\beta}}, \tag{3}$$

$$F_X(x; \alpha, \beta) = \mathrm{P}(X \leq x) = \Phi\left(\frac{1}{\alpha}\xi\left(\frac{x}{\beta}\right)\right), \quad x > 0, \tag{4}$$

where $\Phi(\cdot)$ is the N(0,1) CDF. The quantile function (QF) of X ($q \times 100$th quantile) is

$$x(q;\alpha,\beta)) = F_X^{-1}(q;\alpha,\beta) = \beta\big(\alpha z(q)/2 + ((\alpha z(q)/2)^2 + 1)^{1/2}\big)^2, \quad 0 < q < 1, \tag{5}$$

where $F_X^{-1}(\cdot)$ is the inverse CDF of X and $z(q)$ is the N(0,1) $q \times 100$th quantile. Thus, if $q = 0.5$, then $x(0.5;\alpha,\beta) = \beta$ and, as mentioned, β is the median of the BS distribution. As also mentioned, this distribution is unimodal and its mode, denoted by x_m, may be computed as the solution of $\omega_f((\xi(x_m/\beta)/\alpha)^2) = (\alpha^2\,\beta\,x_m\,(x_m + 3\beta))/(2(x_m - \beta)(x_m + \beta)^2)$, where $\omega_f = f_X'/f_X$, with f_X being the PDF of X given in (3) and f_X' denoting its derivative.

The HR of a RV X is defined in general by $h_X(x) = f_X(x)/(1 - F_X(x))$, so that it can be easily obtained from (3) and (4) for the BS distribution. Then, the HR of $X \sim \mathrm{BS}(\alpha,\beta)$ is given by

$$h_X(x;\alpha,\beta) = \frac{\phi\left(\frac{1}{\alpha}\xi\,(x/\beta)\right)x^{-3/2}(x+\beta)}{2\alpha\sqrt{\beta}\,\Phi\left(-\frac{1}{\alpha}\xi(x/\beta)\right)}, \quad x > 0, \tag{6}$$

where $\phi(\cdot)$ is the N(0,1) PDF. Note that the normal distribution has an increasing HR, whereas gamma and Weibull distributions have increasing and decreasing HRs. Of course, the exponential distribution with constant HR is obtained from any of these two distributions. However, the LN distribution, often used to model environmental data, has a non-monotonic HR, because it is initially increasing until its change point and then it decreases to zero, that is, the LN distribution has a unimodal HR. The BS distribution has a HR that behaves similarly to that of the LN distribution. Nevertheless, the BS HR decreases to a positive constant value, such as it occurs with real environmental data, and not to zero, as in the LN distribution. In real environmental data, the tail of the concentration distribution behaves as the tail of an exponential distribution. This means that if one samples repeatedly until exceeding a threshold, then the amount of exceedance has the same distribution regardless of the chosen threshold. This is one of the aspects that supports the use of the BS distributions to model contaminant concentration data instead of other distributions that are often considered for analyzing this type of data. From the BS HR expressed in (6), note that increasing and unimodal shapes are obtained. For more details about the BS HR, the reader can see Kundu et al. [15] and Azevedo et al. [2], whereas details about environmental hazard analysis based on the HR can be found in Vilca et al. [33]. A simple way to characterize the HR is by the scaled total time on test (TTT) function. By using this function, we can detect the type of HR that the data have and then choose a suitable distribution. The TTT function is given by

$$W_X(u) = H_X^{-1}(u)/H_X^{-1}(1), \quad 0 \le u \le 1,$$

where $H_X^{-1}(u) = \int_0^{F_X^{-1}(u)}(1 - F_X(y))\,\mathrm{d}y$. The plot of the points $(r/n, W_n(k/n))$, with

$$W_n(r/n) = \frac{\sum_{i=1}^{r} x_{(i)} + (n-r)x_{(r)}}{\sum_{i=1}^{n} x_{(i)}}, \quad r = 1, \ldots, n,$$

and $x_{(i)}$ being the ith observed order statistic, allows us to approximate $W_X(\cdot)$; see Aarset [1] and Azevedo et al. [2]. The change point of the BS HR, $h_X(x; \alpha, \beta)$ say, denoted by x_c, is obtained as the solution of the equation

$$\Phi\left(-\frac{1}{\alpha}\xi(x_c/\beta)\right) = -\frac{\alpha\,\beta^{1/2}\,x_c^{1/2}(x_c+\beta)^2 f(a_{x_c})}{2\,\omega_f(\frac{1}{\alpha^2}\xi^2(x/\beta))\,(x_c-\beta)(x_c+\beta)^2 + (x_c+3\beta)\alpha^2\,\beta\,x_c},$$

whereas its limit value is $1/(2\alpha^2\beta)$; see Kundu et al. [15].

From the BS QF given in (5), a random number generator for $X \sim \mathrm{BS}(\alpha, \beta)$ is given by Algorithm 1.

Algorithm 1 Random number generator for the BS distribution

```
1: Obtain a random number $z$ from $Z \sim \mathrm{N}(0, 1)$.
2: Set values for $\alpha$ and $\beta$ of $X \sim \mathrm{BS}(\alpha, \beta)$.
3: Compute a random number $x$ from $X \sim \mathrm{BS}(\alpha, \beta)$ using (5).
4: Repeat Steps 1 to 3 until the required number of data has been generated.
```

Some properties of the BS distribution are: (i) $c\,X \sim \mathrm{BS}(\alpha,\ c\beta)$, for $c > 0$, and (ii) $1/X \sim \mathrm{BS}(\alpha,\ 1/\beta)$. These properties indicate that the BS distribution belongs to the scale and closed under reciprocation families, respectively. These properties are useful for diverse aspects related to estimation and modeling. From these properties, for example, X/β and β/X have the same distribution. From (1), note that

$$Y = \frac{1}{2}\xi\,(X/\beta) \sim \mathrm{N}\left(0,\ \alpha^2/4\right), \tag{7}$$

that is, Y follows a normal distribution with mean 0 and variance $\alpha^2/4$, which implies $X = \beta(1 + 2Y^2 + 2Y(1 + Y^2)^{1/2}) \sim \mathrm{BS}(\alpha, \beta)$. Using the transformation given in (7), the rth moment of X can be shown to be

$$\mathrm{E}(X^r) = \beta^k \sum_{j=0}^{k} \binom{2r}{2j} \sum_{i=0}^{j} \binom{j}{i} \frac{(2r-2j+2i)!}{2^{r-j+i}(r-j+r)!} \left(\frac{\alpha}{2}\right)^{2(r-j+i)}, \quad r = 1, 2 \ldots \tag{8}$$

From (8), it is possible to note that the mean and variance of X are respectively

$$\mu = \mathrm{E}(X) = \beta\left(1+\alpha^2/2\right) \quad\text{and}\quad \mathrm{V}(X) = \beta^2\left(\alpha^2+5\alpha^4/4\right), \tag{9}$$

whereas the coefficients of variation (CV), skewness (CS) and kurtosis (CK) are

$$\mathrm{CV}(X) = \frac{\sqrt{4\alpha^2+5\alpha^4}}{2+\alpha^2}, \mathrm{CS}(X) = \frac{44\alpha^3+24\alpha}{(4+5\alpha^2)^{3/2}} \text{ and } \mathrm{CK}(X) = 3+\frac{558\alpha^4+240\alpha^2}{(4+5\alpha^2)^2}.$$

Note that $\mathrm{CS}(X) \to 0$ and $\mathrm{CK}(X) \to 3$, as $\alpha \to 0$, that is, when α is small, the skewness and kurtosis of the BS distribution are similar to the skewness and kurtosis of the normal distribution, respectively. The CV, CS and CK are invariant under scale, that is, these coefficients are independent functions of the scale parameter β. Also, if X has a BS distribution with parameters α and β, then $1/X$ has a BS distribution with parameters α and $1/\beta$, and we have that

$$\mathrm{E}\,(1/X) = \left(2+\alpha^2\right)/(2\beta). \tag{10}$$

Several methods have been proposed for estimating the parameters of the BS distribution. However, in all these methods, it is not possible to find explicit expressions for its estimators, so that numerical procedures must be used.

Ng et al. [28] introduced a method of modified moments (MM) based on (10) for estimating the BS parameters, which provides easy analytical expressions to compute them. Specifically, let $X_1, \ldots, X_n$ be a random sample of size n from $X \sim \mathrm{BS}(\alpha, \beta)$ and $x_1, \ldots, x_n$ denote the observed data. Then, estimates of α and β are obtained by using the MM method. This conducts to equating (9) and (10) to their corresponding sample moments as

$$s = \beta\left(1+\alpha^2/2\right) \quad\text{and}\quad 1/r = \left(1+\alpha^2/2\right)/\beta, \tag{11}$$

where $s = (1/n)\sum_{i=1}^n x_i$ and $r = 1/((1/n)\sum_{i=1}^n(1/x_i))$ are the arithmetic and harmonic means of $x_1, \ldots, x_n$, respectively. Solving equation in (11) for α and β, the MM estimates of α and β are obtained as $\widehat{\alpha} = (2((s/r)^{1/2}-1))^{1/2}$ and $\widehat{\beta} = (s\,r)^{1/2}$. These MM estimates can be used as starting values for the numerical procedures in the maximum likelihood (ML) estimation method.

In Sects. 5 and 6, we perform simulated and real contamination data analyses by using the `R` software, with the help of the `gbs` package; for more details about how using this package, see Barros et al. [4]. Table 4 (see Appendix A1) provides examples of some commands that allow us to work with the BS distribution by using the `gbs` package. Exploratory data analysis (EDA), including graphical tools, can also be conducted with this package. In addition, ML and MM estimates of the parameters of the BS distribution can be obtained. Furthermore, goodness-of-fit of the BS distribution to contamination data can be performed by Anderson-Darling (AD) and Kolmogorov-Smirnov (KS) tests and probability plots; see Barros et al. [5].

3 Approximated Forms of the BSsum Distribution

In this section, we provide an approximation for the PDF and CDF of the BSsum distribution, which are useful for developing the proposed methodology. In fact, knowing the BSsum distribution is particularly important to implement X-bar control charts following a BS distribution, because then one can accurately obtain lower (LCL) and upper (UCL) control limits for this chart. Shape analysis of the BSsum distribution and an implementation in the R software are also discussed.

Recall $f_X(\cdot; \alpha, \beta)$ denotes the BS PDF with parameters α and β given in (3). Without loss of generality, the scale parameter can be considered as $\beta = 1$ and then the Laplace transform can be applied to this PDF obtaining

$$Lf_X(s; \alpha, 1) = \frac{\exp\left(1/\alpha^2\right)}{2\alpha\sqrt{2\pi}} \left(\sqrt{\frac{\pi}{s+a}} + \sqrt{\frac{\pi}{a}}\right) \exp(-2\sqrt{(s+a)a}), \tag{12}$$

where $a = 1/(2\alpha^2)$. Define the function

$$q(s) = \frac{1}{2}\left(1 + 1/\sqrt{s}\right) \exp\left((1-\sqrt{s})/\alpha^2\right). \tag{13}$$

After some calculations and using (13), we have $Lf_X(s; \alpha, 1) = q(1+2\alpha^2 s)$. Then,

$$Q(s; \alpha, k) = q(s)^k = \frac{\exp(2ka)}{2^k} \sum_{i=0}^{k} \binom{k}{i} \frac{\exp(-2ka\sqrt{s})}{s^{i/2}}, \tag{14}$$

where a is defined in (12). Now, the Laplace transform of a function $\mu_i(\cdot)$ is

$$L\mu_i(s) = \frac{\exp(-\sqrt{s})}{s^{i/2}}. \tag{15}$$

More details of the function $\mu_i(\cdot)$ can be found in Raaijmakers [30, 31]. Substituting (14) and (15) in the Laplace transform of the function $z(\cdot)$, we have

$$Lz(s) = Q(s; \alpha, k) = \frac{\exp(2ka)}{2^k} \sum_{i=0}^{k} \binom{k}{i} (2ka)^i \, L\mu_i((2ka)^2\, s),$$

which implies

$$z(s) = \frac{\exp(2ka)}{2^k} \sum_{i=0}^{k} \binom{k}{i} (2ka)^{i-2} \mu_i\left(\tfrac{s}{(2ka)^2}\right).$$

Thus, after using some theorems on the Laplace transform, the PDF of the sum of k RVs BS, $Y = \sum_{i=1}^{k} X_i$ say, is

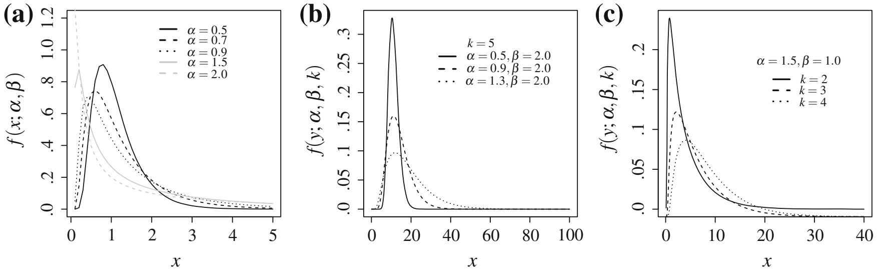


Fig. 1 PDF of $X \sim \text{BS}(\alpha, \beta = 1.0)$ (**a**) and of $Y \sim \text{BSsum}(\alpha, \beta = 1.0, k)$ (**b**), (**c**) for the indicated values

$$f_Y(y;\alpha,\beta=1,k)=\frac{a}{2^k}\exp(2ka-ay)\sum_{i=0}^{k}\binom{k}{i}(2ka)^{i-2}\mu_i\left(\tfrac{y}{4k^2a}\right),\quad y>0. \tag{16}$$

Then, when a RV Y follows a BSsum distribution with parameters $\alpha > 0$, $\beta > 0$ and $k = 1, 2, \ldots$, the notation $Y \sim \text{BSsum}(\alpha, \beta, k)$ is used. Note that the PDF given in (16) corresponds to a RV $Y \sim \text{BSsum}(\alpha, \beta = 1, k)$, but the case for $\beta \neq 1$ can be easily obtained, because it is a scale parameter. Also, similarly to the relation between the Erlang and gamma distributions, one can extend the BSsum distribution to a more general setting for $k > 0$. However, this extension will be studied in a future work. The CDF of $Y \sim \text{BSsum}(\alpha, 1, k)$ is given by $F_Y(y; \alpha, 1, k) = \text{P}(Y \leq y) = \int_0^y f_Y(u; \alpha, 1, k)\,\mathrm{d}u$, for $y > 0$, where $f_Y(\cdot; \alpha, 1, k)$ is as given in (16). Thus,

$$F_Y(y;\alpha,1,k)=\frac{1}{2^k}\exp(2ka-ay)\sum_{i=2}^{k}\frac{l_i\,\mu_i(y/(4k^2a))}{(2ka)^{2-i}}+\Phi\left(\tfrac{\varphi(y;k)}{\alpha}\right),\quad y>0, \tag{17}$$

where $l_i = l_{i+2} - \binom{k}{i}$, with $l_{k+2} = l_{k+1} = 0$, for $i = 2, \ldots, k$, and $\varphi(y; m) = \sqrt{y} - m/\sqrt{y}$, for $m = 1, 2, \ldots$, which is a generalization of the function $\xi(\cdot)$ given in (1) because $\xi(u) = \varphi(u; 1)$. The QF of $Y \sim \text{BSsum}(\alpha, \beta, k)$ must be obtained from (17) by using an iterative numerical method for solving $y(q; \alpha, \beta, k) = F_Y^{-1}(q; \alpha, \beta, k)$, with $0 < q < 1$.

Figure 1(a) shows some shapes of the PDF of the BS distribution for some values of the shape parameter α. Note that, as α decreases, the BS PDF is approximately symmetrical. Plots for different values of β have not been considered, because it is a scale parameter so that $\beta = 1$ is used without loss of generality. Figures 1(b) and (c) show displays plots of the PDF of the sum of k RVs with BS distribution for different values of α. Notice that, as k decreases, the shape of the BSsum PDF becomes more skewed to the right.

The `bssum` package is implemented for a sum of k RVs with BS distribution, for $k = 1, \dots, 24$, because for values of $k > 24$ one can use the central limit theorem as approximation. Table 4 (see Appendix A1) summarizes the main functions of this package. To estimate parameters of the BSsum distribution, we consider the value of k as fixed, so that only α and β must be estimated.

4 X-Bar Control Charts Under the BS Distribution

In this section, we introduce X-bar control charts to detect environmental risk for contaminant concentration data following a BS distribution. We calculate the control limits LCL and UCL by using the BSsum QF obtained from (17). Thus, to construct a control chart for the mean concentration based on the BS distribution, we propose Algorithm 2. Note that in Step 5 of this algorithm we use the median of the data in order to add the central control limit (CCL). It is noteworthy to point out that the median is considered as a better measure of central tendency than the mean for asymmetrical and heavy-tailed distributions; see, e.g. Bhatti [6] and Leiva et al. [21].

Algorithm 2 Construction of the X-bar control chart under the BS distribution

```
1: Collect the observations $x_1, \dots, x_k$ at $k$ sampling points and compute their arithmetic mean $\bar{x} = 1/k \sum_{i=1}^{k} x_i$.
2: Estimate the parameters $\alpha$ and $\beta$ of the BSsum distribution based on the sample obtained from the RV sum $Y_i = \sum_{j=k[i-1]+1}^{ik} X_j$, for $i = 1, \dots, n$.
3: Set the false alarm rate (FAR) $\gamma$ corresponding to the probability of declaring a situation as out-of-control when it is actually in-control.
4: Calculate the LCL and UCL of the BS X-bar control chart based on the estimates obtained in Step 2 of Algorithm 2 and the $(\gamma/2) \times 100$th and $(1-\gamma/2) \times 100$th quantiles of the BSsum distribution divided by $k$.
5: Add the CCL to the BS X-bar control chart using the median of the data as reference.
6: Declare a situation as out-of-control if the sample mean $\bar{x}$ is outside of the interval [LCL, UCL] and announce it. Otherwise, declare it as in-control.
7: Repeat Steps 1 to 6 $m$ times, where $m$ is the number of groups to be analyzed.
```

5 Simulation

In this section, we conduct two simulation studies, one for evaluating the behavior of the ML estimators of the BSsum distribution, and another one for analyzing the performance of the X-bar control chart for data following a BS distribution.

5.1 Estimation

We conduct a MC simulation study for evaluating the performance of the ML estimators of the BSsum distribution. The setting of this study considers samples of size $n \in \{10, 25, 50, 100\}$, vector of true parameters $(\alpha, \beta, k) \in \{(0.5, 1.0, 2.0), (0.5, 1.0, 3.0), (0.5, 1.0, 5.0)\}$, and a number of MC replications equal to 1000. The BSsum samples are generated by the `rbss()` function detailed in Table 4. We report the empirical mean, CS, CK, relative bias (RB) in absolute value and root of the mean squared error ($\sqrt{\text{MSE}}$) of the ML estimators in Table 1, for each parameter, sample size and some values of k. Note that the RB is defined as $\text{RB}[\widehat{\theta}] = |(\text{E}[\widehat{\theta}] - \theta)/\theta|$, where $\widehat{\theta}$ is the ML estimator of a parameter θ, and the sample CS and CK are respectively calculated by

Table 1 Summary statistics from simulated BSsum data for the indicated estimator, n and k

	n							
	10	25	50	100	10	25	50	100
	$\widehat{\alpha}$				$\widehat{\beta}$			
	$k = 2$							
True value	0.5000	0.5000	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
Mean	0.4659	0.4911	0.4982	0.5022	1.1261	1.1351	1.1358	1.1343
CS	0.4030	0.0886	0.1069	0.1581	0.1475	0.1432	0.0986	0.1097
CK	3.2099	2.9349	3.0166	2.8593	2.7175	2.8960	3.0089	3.2466
RB	0.0683	0.0179	0.0036	0.0043	0.1261	0.1351	0.1358	0.1343
$\sqrt{\text{MSE}}$	0.1323	0.0734	0.0528	0.0377	0.1794	0.1598	0.1471	0.1408
	$k = 3$							
True value	0.5000	0.5000	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
Mean	0.4478	0.4722	0.4798	0.4832	1.0923	1.0856	1.0879	1.0876
CS	0.1653	0.1050	0.0928	0.0326	0.2567	0.2026	0.1601	0.1015
CK	3.0909	3.1401	2.9474	3.2031	3.2802	2.9131	2.8473	2.6788
RB	0.1045	0.0556	0.0404	0.0337	0.0923	0.0856	0.0879	0.0876
$\sqrt{\text{MSE}}$	0.1466	0.0837	0.0613	0.0470	0.1389	0.1048	0.0991	0.0932
	$k = 5$							
True value	0.5000	0.5000	0.5000	0.5000	1.0000	1.0000	1.0000	1.0000
Mean	0.4228	0.4614	0.4684	0.4696	1.0607	1.0602	1.0592	1.0565
CS	−2.6994	−3.9311	−0.0389	0.0365	0.3524	0.1512	−0.0093	0.1427
CK	14.6201	53.4802	3.1518	3.0956	3.2242	3.2454	3.1560	2.6596
RB	0.1544	0.0772	0.0631	0.0607	0.0607	0.0602	0.0592	0.0565
$\sqrt{\text{MSE}}$	0.2167	0.1072	0.0761	0.0675	0.1008	0.0779	0.0682	0.0613

$$\mathrm{CS}[x] = \frac{\sqrt{n[n-1]}}{[n-2]}\,\frac{n^{-1}\sum_{i=1}^{n}[x_i-\bar{x}]^3}{\left[n^{-1}\sum_{i=1}^{n}\{x_i-\bar{x}\}^2\right]^{3/2}} \quad \text{and} \quad \mathrm{CK}[x] = \frac{n^{-1}\sum_{i=1}^{n}[x_i-\bar{x}]^4}{\left[n^{-1}\sum_{i=1}^{n}\{x_i-\bar{x}\}^2\right]^{2}},$$

with $x_1, \ldots, x_n$ denote as before the observations from a sample. A glance at the estimates in Table 1 shows that, as the sample size increases, the RB and $\sqrt{\text{MSE}}$ of all of the estimators decrease, tending them to be unbiased, as expected. Also, $\widehat{\alpha}$ and $\widehat{\beta}$ remain close to a normal distribution in terms of their CSs and CKs. Table 1 suggests better results for the ML estimator α (β) as k decreases (increases).

5.2 Control Charts

We now conduct MC simulations to analyze the behavior of the X-bar control chart for data following a BS distribution. We compute the average LCL and UCL as well as their standard deviations (SDs). The simulation setting assumes $m = 20$ subgroups each of size $k = 5$, shape parameter $\alpha \in \{0.5, 1.0, 2.0\}$, scale parameter $\beta = 1$ and FAR $\gamma \in \{0.1, 0.01, 0.0027\}$. The average UCL and LCL and their corresponding SDs are computed in the following manner: $m \times k$ observations are generated from a BS distribution with scale and shape parameters α and β, respectively, and m arithmetic means from k data for each group are calculated, simulating Step 1 of Algorithm 2. Then, we follow Steps 2–6 of this algorithm. Such a procedure is repeated 500 times and the average LCLs and UCLs are computed from these generated values of LCLs and UCLs, respectively. The SDs are also computed as the average from the respective 500 values. Table 2 reports the results of these MC simulations. Note that as α increases, the SDs tend to increase as well. Note also that as the FAR γ decreases, the limits become farther apart.

Table 2 Average LCL and UCL and their SD for the indicated values of α and FAR, with $\beta = 1$

α	$\gamma = 0.1$		$\gamma = 0.01$		$\gamma = 0.0027$	
	LCL	UCL	LCL	UCL	LCL	UCL
Average						
0.5	0.8169	1.6128	0.6750	1.9550	0.6156	2.1283
1.0	0.8487	2.9127	0.5979	3.9983	0.5066	4.6395
2.0	1.1802	8.8674	0.6610	13.5990	0.5267	16.0198
SD						
0.5	0.0520	0.1358	0.0557	0.2037	0.0560	0.2289
1.0	0.1012	0.4321	0.0875	0.6933	0.0863	0.8730
2.0	0.2234	2.1196	0.1749	3.2660	0.1546	3.9838

6 Example

In this section, we provide an example with real-world environmental data to illustrate the methodology proposed in this work. Specifically, we use X-bar control charts based on the BS distribution for monitoring pH concentration levels. Our statistical analysis of the data consists of (i) determination of autocorrelation, (ii) descriptive statistics and graphical tools for proposing a suitable distribution, (iii) detection of the fitting of the model to the data, and (iv) application of our methodology to monitor environmental risk using BS X-bar control charts.

6.1 The Data

We analyze data of the pH concentration level of five rivers from the South Island of New Zealand described in Manly [24, pp. 135–138]. These data are displayed in Table 5 (see Appendix A2) and correspond to monthly values from January 1989 to December 1997.

6.2 Data Analysis

Figure 2 shows graphical plots of the autocorrelation function (ACF) and partial ACF for River 1 data set. From this figure, one notes the absence of serial correlation. Therefore, data can be modeled as coming from a random sample, that is, assuming independent identically distributed RVs, which supports the use of our methodology. A similar behavior is detected for the other data sets (omitted here).

Table 3 provides descriptive statistics of the data using the functions `quantile()` and `descriptiveSummary()` of the `basics` and `gbs` 2.0 packages, respectively. In particular, Table 3 presents empirical (sample) values for quantiles (Q_p, $0 < p \leq 100$), minimum and maximum ($x_{(1)}$ and $x_{(n)}$, respectively), usual ($\text{Rg} = x_{(n)} - x_{(1)}$) and interquartile ($\text{IQR} = Q_{75} - Q_{25}$) ranges, and CV, CS and

Fig. 2 Plots of autocorrelation (**a**) and partial autocorrelation (**b**) functions for River 1 data set

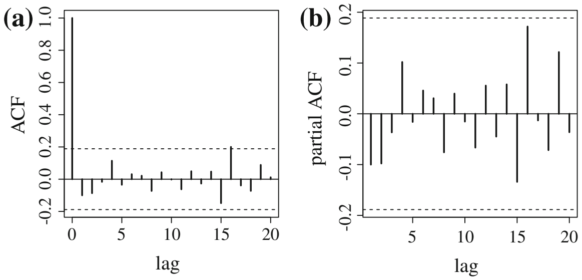


Table 3 Descriptive statistics of the pH level for the set and indicated river (Ri), with $i = 1, \ldots, 5$

Set	$x_{(1)}$	$Q_{13.5}$	Q_{25}	Q_{50}	$\bar{x}$	Q_{75}	$Q_{99.87}$	$x_{(108)}$	SD	CV	CS	CK	IQR	Rg
R1	6.94	7.248	7.493	7.650	7.654	7.850	8.487	8.52	0.307	4.013	0.039	2.814	0.358	1.58
R2	6.96	7.239	7.450	7.695	7.664	7.853	8.692	8.73	0.327	4.263	0.175	2.750	0.403	1.77
R3	6.93	7.254	7.418	7.680	7.634	7.853	8.502	8.53	0.319	4.184	−0.047	2.696	0.435	1.60
R4	6.96	7.243	7.450	7.710	7.644	7.833	8.279	8.29	0.308	4.028	−0.351	2.512	0.383	1.33
R5	6.87	7.163	7.390	7.695	7.603	7.800	8.233	8.24	0.312	4.098	−0.318	2.176	0.41	1.37

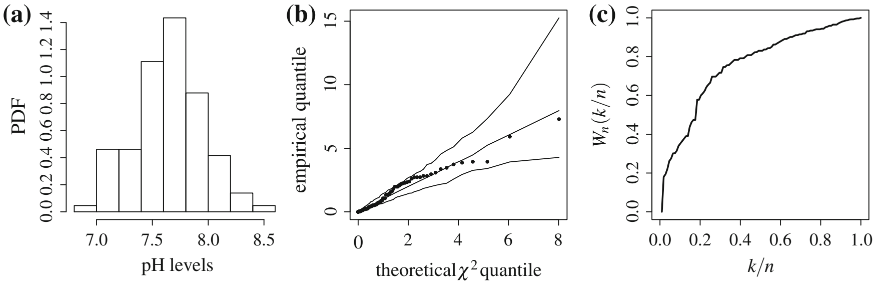


Fig. 3 Histogram (**a**) probability plot with envelope (**b**) and TTT plot (**c**) for River 1 data set

CK. From these statistics, we note that a non-normal distribution can be reasonably assumed for modeling these data, due to their asymmetric nature and their level of kurtosis.

In order to evaluate adequacy of the model to the pH data of River 1, we apply the KS test by using the function `ksbs()` of the `gbs` 2.0 package. The result of this application is presented next:

```
One-sample Kolmogorov-Smirnov test
data: river 1
D = 0.068142 p-value = 0.078677
alternative hypothesis: two-sided
```

It does not provide statistical evidence for indicating that the data do not follow a BS distribution (p-value $= 0.078677$). The histogram of the data and a probability plot with simulated envelope, generated with the function `envelopeBS()` of the `gbs` 2.0 package and shown in Fig. 3(a)-(b), support this conclusion. The TTT plot suggests an increasing HR for these data; see Fig. 3(c). Similar results are obtained for the remaining data sets (omitted here).

6.3 BS Control Charts

Once we verify that the BS distribution can be used to model the pH data, we then apply the X-bar BS control chart to monitor the mean pH levels. Specifically, this chart is used to look for changes in the average value of pH levels through time and so monitoring environmental risk. Table 5 provides these mean pH levels. The control chart is constructed in the following way by using Algorithm 2. First, we estimate the parameters of the BSsum distribution. This is done with the `estbssum()` function detailed in Table 4. Then, Steps 3–6 of Algorithm 2 are followed. As mentioned in Step 3 of this algorithm, the limits are determined by the $(\gamma/2) \times 100$th and $(1-\gamma/2) \times 100$th quantiles of the BSsum distribution. In order to have a false alarm

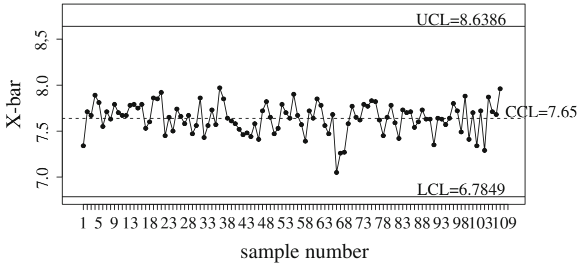


Fig. 4 BS X-bar control chart for the mean of pH level data

only in 27 of each 10,000 cases, we assume $\gamma = 0.0027$. Thus, the LCL is 6.7849, the UCL is 8.6386 and the CCL is 7.65. The corresponding BS X-bar control chart is drawn in Fig. 4. From this figure, we see that no points outside of the control limits are detected, indicating that the contamination levels are in-control and therefore it is not needed to declare an environmental emergency.

7 Concluding Remarks

In this paper, we have proposed a methodology based on control charts for monitoring environmental risk when the contaminant concentration follows a Birnbaum-Saunders distribution. We have discussed and implemented the distribution of the sum of random variables with BS distribution, which is not Birnbaum-Saunders. We have implemented this methodology in the R software and investigated its performance through Monte Carlo simulations. The results of this simulation study indicate the good performance of the methodology. We have analyzed real-world environmental data illustrating it, which have shown its potential.

Acknowledgments The authors wish to thank the Editors of the volume "Risk Assessment Challenges: Theory and Practice" of Springer Proceedings in Mathematics and Statistics, Dr. Christos P. Kitsos and Dr. Teresa A. Oliveira, and three anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. This research was partially supported by FONDECYT 1120879 grant from Chile, and by CAPES, CNPq and FACEPE grants from Brazil. H. Saulo thanks Universidade Federal de Goiás from Brazil for supporting this research.

Appendix

A1: Basic Functions of the bssum and gbs packages (see Table 4)

Table 4 Basic functions of the indicated package

Function	Instruction	Result
gbs package		
EDA	descriptiveSummary(x)	It provides a summary with the most important descriptive statistics
TTT	TTT(x)	It displays the TTT plot to detect the shape of the HR
PDF	dgbs(1.0, alpha = 0.5, beta = 1.0)	0.798
CDF	pgbs(1.0, alpha = 0.5, beta = 1.0)	0.500
QF	qgbs(0.5, alpha = 0.5, beta = 1.0)	1.000
Numbers	rgbs(n = 100, alpha = 1.0, beta = 1.0)	It generates 100 BS(1, 1) random numbers
MME	mmegbs(x)	It estimates the BS parameters by the MM method using the data x
MLE	mlegbs(x)	It estimates the BS parameters by the ML method using the data x
Histogram	histgbs(x, boxPlot = T, pdfLine = T)	It produces a histogram and a boxplot with estimated BS PDF using the data x
Envelope	envelopegbs(x)	It produces a probability plot with envelope using the data x
AD test	adgbs(x)	It computes AD p-value for the data x
KS test	ksgbs(x, graph = T)	It computes KS p-value and plots estimated theoretical BS and empirical CDF using the data x
bssum package		
PDF	dbss(1, k = 2, alpha = 0.5, beta = 1)	0.121
CDF	pbss(1, k = 2, alpha = 0.5, beta = 1)	0.016
QF	qbss(0.5, k = 2, alpha = 0.5, beta = 1)	2.118
Numbers	rbss(10, k = 2, alpha = 0.5, beta = 1)	It generates 10 BSsum(0.5, 1, 2) random numbers
MLE	estbss(k, x)	It estimates the BSsum parameters by the ML method using the data x and k fixed

A2: pH levels for 5 Rivers (Ri) in New Zealand (see Table 5)

Table 5 pH levels for the indicated river and date in New Zealand pp.135–138 [24]

Year	Month	R1	R2	R3	R4	R5	Mean	Year	Month	R1	R2	R3	R4	R5	Mean
1989	Jan	7.27	7.10	7.02	7.23	8.08	7.34	1994	Jan	8.13	8.09	8.01	7.76	7.24	7.85
	Feb	8.04	7.74	7.48	8.10	7.21	7.71		Feb	7.23	7.89	7.81	8.12	7.83	7.78
	Mar	7.50	7.40	8.33	7.17	7.95	7.67		Mar	7.08	7.92	7.68	7.70	7.40	7.56
	Apr	7.87	8.10	8.13	7.72	7.61	7.89		Apr	7.55	7.50	7.52	7.64	7.14	7.47
	May	7.60	8.46	7.80	7.71	7.48	7.81		May	7.75	7.57	7.44	7.61	8.01	7.68
	Jun	7.41	7.32	7.42	7.82	7.80	7.55		Jun	6.94	7.37	6.93	7.03	6.96	7.05
	Jul	7.88	7.50	7.45	8.29	7.45	7.71		Jul	7.46	7.14	7.26	6.99	7.47	7.26
	Aug	7.88	7.79	7.40	7.62	7.47	7.63		Aug	7.62	7.58	7.09	6.99	7.06	7.27
	Sep	7.78	7.73	7.53	7.88	8.03	7.79		Sep	7.45	7.65	7.78	7.73	7.31	7.58
	Oct	7.14	7.96	7.51	8.19	7.70	7.70		Oct	7.65	7.63	7.98	8.06	7.51	7.77
	Nov	8.07	7.99	7.32	7.32	7.63	7.67		Nov	7.85	7.70	7.62	7.96	7.13	7.65
	Dec	7.21	7.72	7.73	7.91	7.79	7.67		Dec	7.56	7.74	7.80	7.41	7.59	7.62
1990	Jan	7.66	8.08	7.94	7.51	7.71	7.78	1995	Jan	8.18	7.80	7.22	7.95	7.79	7.79
	Feb	7.71	8.73	8.18	7.04	7.28	7.79		Feb	7.63	7.88	7.90	7.45	7.97	7.77
	Mar	7.72	7.49	7.62	8.13	7.78	7.75		Mar	7.59	8.06	8.22	7.57	7.73	7.83
	Apr	7.84	7.67	7.81	7.81	7.80	7.79		Apr	7.47	7.82	7.58	8.03	8.19	7.82
	May	8.17	7.23	7.09	7.75	7.40	7.53		May	7.52	7.42	7.76	7.66	7.76	7.62
	Jun	7.79	7.46	7.13	7.83	7.77	7.60		Jun	7.61	7.72	7.56	7.49	6.87	7.45
	Jul	7.16	8.44	7.94	8.05	7.70	7.86		Jul	7.30	7.90	7.57	7.76	7.72	7.65
	Aug	7.74	8.13	7.82	7.75	7.80	7.85		Aug	7.75	7.75	7.52	8.12	7.75	7.78
	Sep	8.09	8.09	7.51	7.97	7.94	7.92		Sep	7.77	7.78	7.75	7.49	7.14	7.59
	Oct	7.20	7.65	7.13	7.60	7.68	7.45		Oct	7.79	7.30	7.83	7.09	7.09	7.42
	Nov	7.81	7.25	7.80	7.62	7.75	7.65		Nov	7.87	7.89	7.35	7.56	7.99	7.73
	Dec	7.73	7.58	7.30	7.78	7.11	7.50		Dec	8.01	7.56	7.67	7.82	7.44	7.70

(continued)

Table 5 (continued)

Year	Month	R1	R2	R3	R4	R5	Mean	Year	Month	R1	R2	R3	R4	R5	Mean
1991	Jan	8.52	7.22	7.91	7.16	7.87	7.74	1996	Jan	7.29	7.62	7.95	7.72	7.98	7.71
	Feb	7.13	7.97	7.63	7.68	7.90	7.66		Feb	7.50	7.50	7.90	7.12	7.69	7.54
	Mar	7.22	7.80	7.69	7.26	7.94	7.58		Mar	8.12	7.71	7.20	7.43	7.56	7.60
	Apr	7.62	7.80	7.59	7.37	7.97	7.67		Apr	7.64	7.75	7.80	7.72	7.73	7.73
	May	7.70	7.07	7.26	7.82	7.51	7.47		May	7.59	7.57	7.86	7.92	7.22	7.63
	Jun	7.66	7.83	7.74	7.29	7.30	7.56		Jun	7.60	7.97	7.14	7.72	7.72	7.63
	Jul	7.97	7.55	7.68	8.11	8.01	7.86		Jul	7.07	7.70	7.33	7.41	7.26	7.35
	Aug	7.86	7.13	7.32	7.75	7.08	7.43		Aug	7.65	7.68	7.99	7.17	7.72	7.64
	Sep	7.43	7.61	7.85	7.77	7.14	7.5		Sep	7.51	7.64	7.25	7.82	7.91	7.63
	Oct	7.77	7.83	7.77	7.54	7.74	7.73		Oct	7.81	7.53	7.88	7.11	7.50	7.57
	Nov	7.84	7.23	7.64	7.42	7.73	7.57		Nov	7.16	7.85	7.63	7.88	7.66	7.64
	Dec	8.23	8.08	7.89	7.71	7.95	7.97		Dec	7.67	8.05	8.12	7.38	7.77	7.80
1992	Jan	8.28	7.96	7.86	7.65	7.49	7.85	1997	Jan	7.97	7.04	7.48	7.88	8.24	7.72
	Feb	7.23	7.11	8.53	7.53	7.78	7.64		Feb	7.17	7.69	8.15	6.96	7.47	7.49
	Mar	7.68	7.68	7.15	7.68	7.85	7.61		Mar	7.52	7.84	8.12	7.85	8.07	7.88
	Apr	7.87	7.20	7.42	7.45	7.96	7.58		Apr	7.65	7.14	7.38	7.23	7.66	7.41
	May	7.94	7.35	7.68	7.50	7.12	7.52		May	7.62	7.64	8.17	7.56	7.53	7.70
	Jun	7.80	6.96	7.56	7.22	7.76	7.46		Jun	7.10	7.16	7.71	7.57	7.15	7.34
	Jul	7.39	7.12	7.70	7.47	7.74	7.48		Jul	7.85	7.62	7.68	7.71	7.72	7.72
	Aug	7.42	7.41	7.47	7.80	7.12	7.44		Aug	7.39	7.53	7.11	7.39	7.03	7.29
	Sep	7.91	7.77	6.96	8.03	7.24	7.58		Sep	8.18	7.75	7.86	7.77	7.77	7.87
	Oct	7.59	7.41	7.41	7.02	7.60	7.41		Oct	7.67	7.66	7.87	7.82	7.51	7.71
	Nov	7.94	7.32	7.65	7.84	7.86	7.72		Nov	7.57	7.92	7.72	7.73	7.47	7.68
	Dec	7.64	7.74	7.95	7.83	7.96	7.82		Dec	7.97	8.16	7.70	8.21	7.74	7.96

(continued)

Table 5 (continued)

Year	Month	R1	R2	R3	R4	R5	Mean	
1993	Jan	7.55	8.01	7.37	7.83	7.51	7.65	
	Feb	7.30	7.39	7.03	8.05	7.59	7.47	
	Mar	7.80	7.17	7.97	7.58	7.13	7.53	
	Apr	7.92	8.22	7.64	7.97	7.18	7.79	
	May	7.70	7.80	7.28	7.61	8.12	7.70	
	Jun	7.76	7.41	7.79	7.89	7.36	7.64	
	Jul	8.28	7.75	7.76	7.89	7.82	7.90	
	Aug	7.58	7.84	7.71	7.27	7.95	7.67	
	Sep	7.56	7.92	7.43	7.72	7.21	7.57	
	Oct	7.19	7.73	7.21	7.49	7.33	7.39	
	Nov	7.60	7.49	7.86	7.86	7.80	7.72	
	Dec	7.50	7.86	7.83	7.58	7.45	7.64	

References

1. Aarset, M.V.: How to identify a bathtub hazard rate. IEEE Trans. Reliab. **36**, 106–108 (1987)
2. Azevedo, C., Leiva, V., Athayde, E., Balakrishnan, N.: Shape and change point analyses of the Birnbaum-Saunders-*t* hazard rate and associated estimation. Comput. Stat. Data Anal. **56**, 3887–3897 (2012)
3. Balakrishnan, N., Leiva, V., López, J.: Acceptance sampling plans from truncated life tests from generalized Birnbaum-Saunders distribution. Commun. Stat. Simul. Comput. **36**, 643–656 (2007)
4. Barros, M., Paula, G.A., Leiva, V.: An R implementation for generalized Birnbaum-Saunders distributions. Comput. Stat. Data Anal. **53**, 1511–1528 (2009)
5. 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)
6. Bhatti, C.R.: The Birnbaum-Saunders autoregressive conditional duration model. Math. Comput. Simul. **80**, 2062–2078 (2010)
7. Box, G.E., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. B **26**, 211–246 (1964)
8. Ferreira, M., Gomes, M.I., Leiva, V.: On an extreme value version of the Birnbaum-Saunders distribution. Revstat Stat. J. **10**, 181–210 (2012)
9. Fierro, R., Leiva, V., Ruggeri, F., Sanhueza, A.: On a Birnbaum-Saunders distribution arising from a non-homogeneous Poisson process. Stat. Probab. Lett. **83**, 1233–1239 (2013)
10. Figueiredo, F., Gomes, M.I.: The skew-normal distribution in SPC. Revstat Stat. J. **11**, 83–104 (2013)
11. Grigg, O.A., Farewell, V.T.: A risk-adjusted sets method for monitoring adverse medical outcomes. Stat. Med. **23**, 1593–1602 (2004)
12. Huang, S., Qu, Y.: The loss in power when the test of differential expression is performed under a wrong scale. J. Comput. Biol. **13**, 786–797 (2006)
13. Johnson, N.L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions. Wiley, New York (1995)
14. Kelmansky, D., Martinez, E., Leiva, V.: A new variance stabilizing transformation for gene expression data analysis. Stat. Appl. Genet. Mol. Biol. **12**, 653–666 (2013)
15. Kundu, D., Kannan, N., Balakrishnan, N.: On the hazard function of Birnbaum-Saunders distribution and associated inference. Comput. Stat. Data Anal. **52**, 2692–2702 (2008)
16. Leiva, V., Barros, M., Paula, G.A., Sanhueza, A.: Generalized Birnbaum-Saunders distributions applied to air pollutant concentration. Environmetrics **19**, 235–249 (2008)
17. Leiva, V., Sanhueza, A., Angulo, J.M.: A length-biased version of the Birnbaum-Saunders distribution with application in water quality. Stoch. Environ. Res. Risk Assess. **23**, 299–307 (2009)
18. Leiva, V., Sanhueza, A., Kelmansky, S., Martinez, E.: On the glog-normal distribution and its association with the gene expression problem. Comput. Stat. Data Anal. **53**, 1613–1621 (2009)
19. Leiva, V., Vilca, F., Balakrishnan, N., Sanhueza, A.: A skewed sinh-normal distribution and its properties and application to air pollution. Commun. Stat. Theory Method **39**, 426–443 (2010)
20. Leiva, V., Marchant, C., Saulo, H., Aslam, M., Rojas, F.: Capability indices for Birnbaum-Saunders processes applied to electronic and food industries. J. Appl. Stat. **41**, 1881–1902 (2014)
21. Leiva, V., Saulo, H., Leao, J., Marchant, C.: A family of autoregressive conditional duration models applied to financial data. Comput. Stat. Data Anal. **79**, 175–191 (2014)
22. Lio, Y.L., Park, C.: A bootstrap control chart for Birnbaum-Saunders percentiles. Qual. Reliab. Eng. Int. **24**, 585–600 (2008)
23. Lio, Y.L., Tsai, T.R., Wu, S.J.: Acceptance sampling plans from truncated life tests based on the Birnbaum-Saunders distribution for percentiles. Commun. Stat. Simul. Comput. **39**, 1–18 (2008)
24. Manly, B.F.J.: Statistics for Environmental Science and Management. Chapman & Hall, Boca Raton (2009)

25. Marchant, C., Leiva, V., Cavieres, M.F., Sanhueza, A.: Air contaminant statistical distributions with application to PM10 in Santiago, Chile. Rev. Environ. Contam. Toxicol. **223**, 1–31 (2013)
26. Morrison, L.W.: The use of control charts to interpret environmental monitoring data. Nat. Areas J. **28**, 66–73 (2008)
27. Montgomery, D.: Introduction to Statical Quality Control. Wiley, New York (2008)
28. Ng, H.K.T., Kundu, D., Balakrishnan, N.: Modified moment estimation for the two-parameter Birnbaum-Saunders distribution. Comput. Stat. Data Anal. **43**, 283–298 (2003)
29. Ott, W.R.: A physical explanation of the lognormality of pollution concentrations. J. Air Waste Manag. Assoc. **40**, 1378–1383 (1990)
30. Raaijmakers, F.J.M.: The lifetime of a standby system of units having the Birnbaum-Saunders distribution. J. Appl. Probab. **17**, 490–497 (1980)
31. Raaijmakers, F.J.M.: Reliability of standby system for units with the Birnbaum-Saunders distribution. IEEE Trans. Reliab. **30**, 198–199 (1981)
32. Saulo, H., Leiva, V., Ziegelmann, F.A., Marchant, C.: A nonparametric method for estimating asymmetric densities based on skewed Birnbaum-Saunders distributions applied to environmental data. Stoch. Environ. Res. Risk Assess. **27**, 1479–1491 (2013)
33. Vilca, F., Sanhueza, A., Leiva, V., Christakos, G.: An extended Birnbaum-Saunders model and its application in the study of environmental quality in Santiago, Chile. Stoch. Environ. Res. Risk Assess. **24**, 771–782 (2010)
34. Vilca, F., Santana, L., Leiva, V., Balakrishnan, N.: Estimation of extreme percentiles in Birnbaum-Saunders distributions. Comput. Stat. Data Anal. **55**, 1665–1678 (2011)
35. Woodall, W.H.: The use of control charts in health-care and public-health surveillance. J. Qual. Tech. **38**, 89–104 (2006)

Risk Problems Identifying Optimal Pollution Level

George E. Halkos and Dimitra C. Kitsou

Abstract The determination of the optimal pollution level is essential in Environmental Economics. The associated risk in evaluating this optimal pollution level and the related Benefit Area (BA), is based on various factors. At the same time the uncertainty in the model fitting can be reduced by choosing the appropriate approximations for the abatement and damage marginal cost functions. The target of this paper is to identify analytically and empirically the Benefit Area (BA) in the case of quadratic marginal damage and linear marginal abatement cost functions, extending the work of (Halkos and Kitsos, Appl. Econ. 37:1475–1483, 2005, [9]).

Keywords Optimal pollution level · Risk · Benefit area

1 Introduction

Rationality in the formulation and applicability of environmental policies depends on careful consideration of their consequences for nature and society. For this reason it is important to quantify the costs and benefits in the most accurate way. But the validity of any cost–benefit analysis (hereafter CBA) is ambiguous as the results may have large uncertainties. Uncertainty in the evaluation of their effects is present in all environmental problems and this underlines the need for thoughtful policy design and evaluation. We may have uncertainty in the underlying physical or ecological processes, as well as in the economic consequences of the change in environmental quality.

As uncertainty may be due to the lack of appropriate abatement and damage cost data, we apply here a method of calibrating hypothetical damage cost estimates

G.E. Halkos (✉) · D.C. Kitsou
Department of Economics (Laboratory of Operations Research),
University of Thessaly, Volos, Greece
e-mail: halkos@uth.gr

D.C. Kitsou
e-mail: dimkitsou@yahoo.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_15

relying on individual country abatement cost functions. In this way a "calibrated" Benefit Area (BA^c) is estimated.

Specifically we try to identify the optimal pollution level under the assumptions of linear marginal abatement and quadratic marginal damage cost functions. That is, we consider another case of the possible approximations of the two cost curves improving the work in [9] by extending the number of different model approximations of abatement and damage cost functions and thus the assumed correct model eliminates uncertainty about curve fitting. The target of this paper is to develop the appropriate theory in this specific case.

2 Determining the Optimal Level of Pollution

Economic theory suggests that the optimal pollution level occurs when the marginal damage cost equals the marginal abatement cost. Graphically the optimal pollution level is presented in Fig. 1 where the marginal abatement (MAC $= g(z)$) and the marginal damage (MD $= \varphi(z)$) are represented as typical mathematical cost functions.

The intersection of the marginal abatement (MAC) and marginal damage (MD) cost functions defines the optimal pollution level (denoted as I in Fig. 1) with coordinates (z_0, k_0), $I(z_0, k_0)$. The value of z_0 describes the optimal damage reduction while k_0 corresponds to the optimal cost of attaining that. The area in $\mathbb{R}^2$ covered by the MAC and MD and the axis of cost is defined as the Benefit Area. That is, the point of intersection of the two curves, $I = I(z_0, k_0)$, reflects the optimal level of pollution with k_0 corresponding to the optimum cost (benefit) and z_0 to the optimum damage restriction. It is assumed (and we shall investigate the validity of this assumption subsequently) that the curves have an intersection and the area created by these curves (region AIB) is what we define as Benefit Area (see [15], among others), representing the maximum of the net benefit that is created by the activities of trying to reduce pollution.

Consider Fig. 1. Let A and B be the points of the intersection of the linear curves MD $= \varphi(z) = \alpha + \beta z$ and MAC $= \beta_0 + \beta_1 z$ with the "Y–axis". We are restricted to positive values. For these points $A = A(0, \alpha)$ and $B = B(0, \beta_0)$ the values of $a = \alpha$ and $b = \beta_0$ are the constant terms of the assumed curves that represent MD and MAC respectively.

Let us now assume that

$$\text{MAC}(z) = g(z) = \beta_0 + \beta_1 z, \quad \beta_1 \neq 0 \quad \text{and} \quad \text{MD}(z) = \varphi(z) = \alpha z^2 + \beta z + \gamma, \quad \alpha > 0.$$

The intersections of MD and MAC with the Y–axis are $b =$ MAC$(0) = \beta_0$ and $a =$ MD$(0) = \gamma$, see Figs. 2, 3 and 4. To ensure that an intersection between MAC and MD occurs we need the restriction $0 < \beta_0 < \gamma$. yboxAssuming $\alpha > 0$ three cases can be distinguished, through the determinant of $\varphi(z)$, say D, $D = \beta^2 - 4\alpha\gamma$; (a) $D = 0$ (see Fig. 2), (b) $D > 0$ (see Fig. 3) while the case $D < 0$ is without

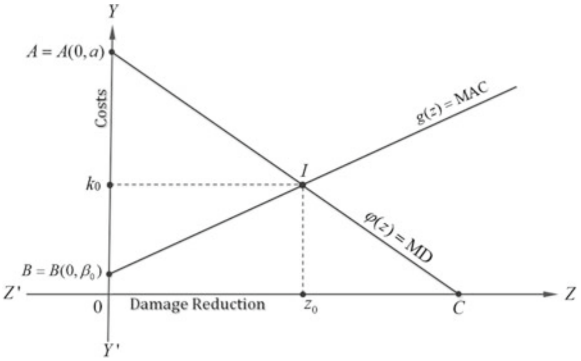


Fig. 1 Graphical presentation of the optimal pollution level (general case)

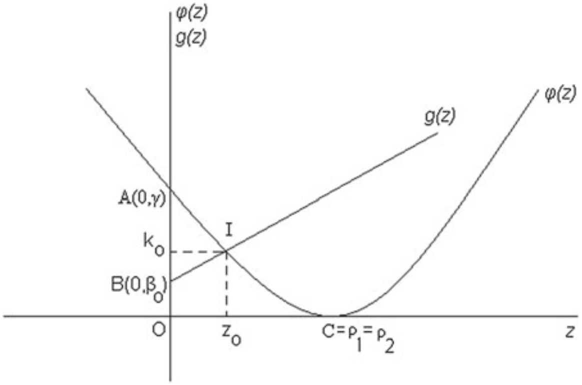


Fig. 2 $C = C\left(-\frac{\beta}{2\alpha}, 0\right), \alpha > 0$

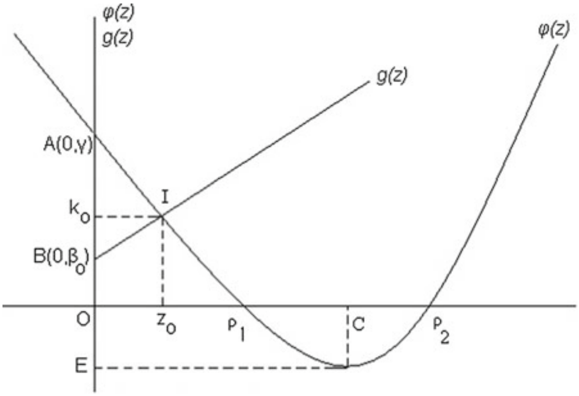


Fig. 3 $C = C\left(-\frac{\beta}{2\alpha}, 0\right), E = E\left(0, \varphi\left(-\frac{\beta}{2\alpha}\right)\right), \varphi\left(-\frac{\beta}{2\alpha}\right) = \min \varphi(z), \alpha > 0$

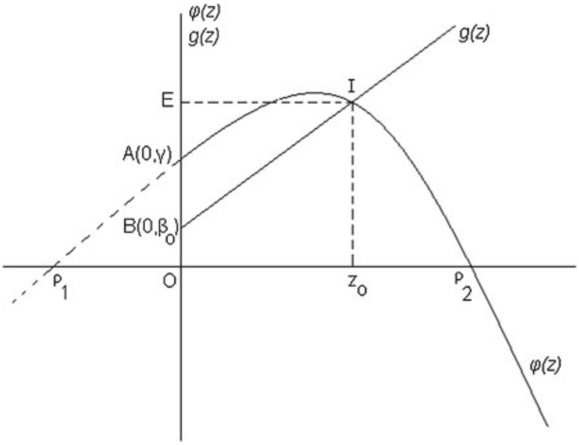


Fig. 4 $C = C\left(-\frac{\beta}{2\alpha}, 0\right)$, $E = E\left(0, \varphi\left(-\frac{\beta}{2\alpha}\right)\right)$, $\varphi\left(-\frac{\beta}{2\alpha}\right) = \min\varphi(z)$, $\alpha < 0$

economic interest (due to the complex–valued roots). Cases (a) and (b) are discussed below, while for the dual $\alpha < 0$ see Case (c). For more details see also [14].

Case (a): $\alpha > 0, \quad D = \beta^2 - 4\alpha\gamma = 0$. In this case there is a double real root for MD(z), say $\rho = \rho_1 = \rho_2 = -\frac{\beta}{2\alpha}$. We need $\rho > 0$ and hence $\beta < 0$. To identify the optimal pollution level point $I(z_0, k_0)$ the evaluation of point z_0 is the one for which

$$\mathrm{MD}(z_0) = \varphi(z_0) \Leftrightarrow g(z_0) = \mathrm{MAC}(z_0) \Leftrightarrow \alpha z_0^2 + \beta z_0 + \gamma = \beta_0 + \beta_1 z_0 \Leftrightarrow$$

$$\alpha z_0^2 + (\beta - \beta_1)z_0 + (\gamma - \beta_0) = 0. \tag{1}$$

Relation (1) provides the unique (double) solution when $D_1 = (\beta - \beta_1)^2 - 4\alpha(\gamma - \beta_0) = 0$ which is equivalent to

$$z_0 = -\frac{\beta - \beta_1}{2\alpha} = \frac{\beta_1 - \beta}{2\alpha}. \tag{2}$$

As z_0 is positive and $\alpha > 0$ we conclude that $\beta_1 > \beta$. So for the conditions are: $\alpha > 0, \beta_1 > \beta, 0 < \beta_0 < \gamma$ we can easily calculate

$$k_0 = \mathrm{MAC}(z_0) = \beta_0 + \beta_1\frac{\beta_1 - \beta}{2\alpha} > 0, \tag{3}$$

and therefore $I(z_0, k_0)$ is well defined. The corresponding Benefit Area ($\mathrm{BA_{QL}}$) in this case is

$$\begin{aligned}\mathrm{BA_{QL}} = (\mathrm{ABI}) = \int_0^{z_0} \varphi(z) - g(z)dz = \int_0^{z_0} \alpha z^2 + (\beta - \beta_1)z + (\gamma - \beta_0)dz = \\ \left[\tfrac{\alpha}{3}z^3 + \tfrac{1}{2}(\beta - \beta_1)z^2 + (\gamma - \beta_0)z\right]_{z=0}^{z_0} = \\ \tfrac{\alpha}{3}z_0^3 + \tfrac{1}{2}(\beta - \beta_1)z_0^2 + (\gamma - \beta_0)z_0.\end{aligned} \tag{4}$$

Case (b): $\alpha > 0$, $D = \beta^2 - 4\alpha\gamma > 0$. For the two roots ρ_1, ρ_2, we have $|\rho_1| \neq |\rho_2|$, $\varphi(\rho_1) = \varphi(\rho_2) = 0$ and we suppose $0 < \rho_1 < \rho_2$, see Fig. 3. The fact that $D > 0$ is equivalent to $0 < a\gamma < (\beta/2)^2$, while the minimum value of the MD function is $\varphi(-\beta/(2\alpha)) = (4\alpha\gamma - \beta^2)/(4\alpha)$.

Proposition 1 *The order* $0 < \rho_1 < \rho_2$ *for the roots and the value which provides the minimum is true under the relation*

$$\beta < 0 < \alpha\gamma < \left(\frac{\beta}{2}\right)^2. \tag{5}$$

Proof The order of the roots $0 < \rho_1 < \rho_2$ is equivalent to the set of relations:

$$D > 0, \quad \alpha\varphi\left(-\frac{\beta}{2\alpha}\right) < 0, \quad \alpha\varphi(0) > 0, \quad 0 < \frac{\rho_1 + \rho_2}{2}. \tag{6}$$

The first is valid, as we have assumed $D > 0$. For the imposed second relation from (6) we have $\alpha\varphi(-\frac{\beta}{2\alpha}) < 0 \Leftrightarrow \alpha\frac{4\alpha\gamma-\beta^2}{4\alpha} < 0 \Leftrightarrow D > 0$, which holds. As both the roots are positive $\rho_1, \rho_2 > 0$, then the product $\rho_1\rho_2 > 0$ and therefore $\frac{\gamma}{\alpha} > 0 \Leftrightarrow \alpha\gamma > 0$. The third relation $\alpha\varphi(0) = \alpha\gamma > 0$, in (6) is true already and $0 < \frac{\rho_1+\rho_2}{2} \Rightarrow 0 < -\frac{\beta}{2\alpha}$ equivalent to $\beta < 0$. Therefore we get $\beta < 0 < \alpha\gamma < (\frac{\beta}{2})^2$.

We can then identify the point of intersection $I(z_0, k_0)$, z_0 : $\mathrm{MAC}(z_0) = \mathrm{MD}(z_0)$ as before. Therefore under (5) and $\beta_1 > \beta_0$ we evaluate k_0 as in (3) and the Benefit Area $\mathrm{BA_{QL}}$ can be evaluated as in (4).

Case (c): $\alpha < 0$, $D = \beta^2 - 4\alpha\gamma > 0$. Let us now consider the case $\alpha < 0$. Under this assumption the restriction $D = 0$ is not considered, as the values of $\varphi(z)$ have to be negative.

Under the assumption of Case (c), the value $\varphi(-\frac{\beta}{2\alpha}) = \frac{4\alpha\gamma-\beta^2}{4\alpha}$ corresponds to the maximum value of $\varphi(z)$. We consider the situation where $\rho_1 < 0 < -\frac{\beta}{2\alpha} < \rho_2$ (see Fig. 4) while the case $0 < \rho_1 < -\frac{\beta}{2\alpha} < \rho_2$ has no particular interest (it can be also considered as in Case (b), see Fig. 3).

Proposition 2 *For the Case (c) as above we have:* $\rho_1 < 0 < -\frac{\beta}{2\alpha} < \rho_2$ *when* $\alpha\gamma < 0$.

Proof The imposed assumption is equivalent to $\alpha\varphi(0) < 0 \Leftrightarrow \alpha\gamma < 0$ as $\rho_1\rho_2 < 0$, $\alpha\varphi(-\frac{\beta}{2\alpha}) < 0 \Leftrightarrow \alpha\gamma < (\frac{\beta}{2})^2$. Therefore the imposed restrictions are $\alpha\gamma < 0 < (\frac{\beta}{2})^2$ (compare with (5)). Actually, $\alpha\gamma < 0$.

Case (c) requires that $\beta_0 < \gamma$ and $\beta_1 > 0$. To calculate z_0 we proceed as in (1) and z_0 is evaluated as in (2). Therefore, with $\alpha < 0$ we have $\beta_1 - \beta < 0$, i.e. $\beta_1 < \beta$. Thus for $\beta_1 < \beta$, $\alpha\gamma < 0$, the BA as in (4) is still valid.

3 An Empirical Application

In the empirical application, regression analysis is adopted to estimate the involved parameters. The available data for different European countries are used, as derived and described by [3, 4].

The abatement cost function measures the cost of reducing tonnes of emissions of a pollutant, like sulphur (S), and differs from country to country depending on the local costs of implementing best practice abatement techniques as well as on the existing power generation technology. For abating sulphur emissions various control methods exist with different cost and applicability levels, see [3–6].

Given the generic engineering capital and operating control cost functions for each efficient abatement technology, total and marginal costs of different levels of pollutant's reduction at each individual source and at the national (country) level can be constructed. According to [3, 4, 8], the cost of an emission abatement option is given by its total annualized cost (TAC) calculated by the addition of fixed and variable operating and maintenance costs. For every European country a least cost curve is derived by finding the technology on each pollution source with the lowest marginal cost per tonne of pollutant removed in the country and the amount of pollutant removed by that method on that pollution source.

Specifically the abatement cost curves were derived for all European countries after considering all sectors and all available fuels with their sulphur content for the year 2000. See [2], for technical details on deriving an abatement cost curve and on using pre-during and post–combustion desulphurization techniques. Figure 5 shows the marginal cost curve in the case of Austria and for the year 2000.

For analytical purposes, it is important to approximate the cost curves of each country by adopting a functional form extending the mathematical models described above to stochastic models, [7]. At the same time, the calculation of the damage function $\varphi(z)$ is necessary as proposed in [9, 13, 14]. The only information available is to "calibrate" the damage function, on the assumption that national authorities act independently (as Nash partners in a non-cooperative game with the rest of the world) taking as given deposits originating in the rest of the world, see [10].

The results are presented in Table 1, where Eff is the efficiency of the benefit area, in comparison with the maximum evaluated from the sample of countries under investigation and can be estimated using as measure of efficiency the expression as defined in [9]:

$$\mathrm{Eff} = \left(\frac{\mathrm{BA}}{\max \mathrm{BA}}\right) \times 100.$$

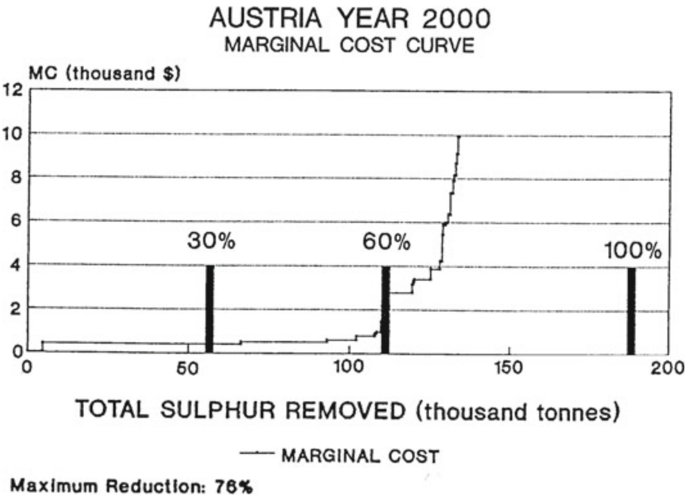


Fig. 5 The marginal abatement cost curve for Austria and for the year 2000 (*Source* Modified from [2])

Table 1 Coefficient estimates in the case of quadratic MD and MAC functions

Countries	c_0	c_1	c_2	b_0	b_1	b_2	R^2
Albania	0.7071	0.01888	0.0001397	−3.3818	0.015	0.0048	0.819
Austria	8.57143	0.055012	0.0001145	3.274	−0.221	0.004	0.748
Belgium	2.2424	0.03869	0.0001688	0.497	−0.124	0.003	0.851
Former Czech.	37.794	0.100323	0.000059	11.241	0.2358	0.00018	0.723
Denmark	10.0	0.1923	0.0060811	−2.49	0.099	0.0053	0.923
Finland	4.021	0.0781	0.0001459	2.343	−0.098	0.0046	0.583
France	33.158	0.277352	0.000197	42.374	−0.053	0.0018	0.945
Greece	3.7373	0.034133	0.0000491	−1.614	0.342	0.0006	0.998
Hungary	5.101	0.031488	0.0000417	2.506	0.216	0.0004	0.923
Italy	21.01	0.030036	0.0000191	12.5	0.36	0.0003	0.689
Luxembourg	0.421	0.3161	0.0272381	−0.7272	0.01	0.09234	0.883
Netherlands	8.353	0.19513	0.0035144	−6.18	0.41	0.0009	0.794
Norway	1.421	0.07852	0.0001701	0.94	−0.244	0.0164	0.878
Poland	6.212	0.023153	0.000071	−8.023	0.324	0.00009	0.77
Romania	9.091	0.011364	0.0000624	5.502	0.19	0.0001	0.81
Spain	11.7	0.007288	0.0049742	10.21	−0.021	0.00014	0.992
Sweden	2.4	0.06423	0.0000932	4.074	−0.252	0.004	0.854
Switzerland	2.4	0.56027	0.002803	5.7543	−1.6289	0.11203	0.912
Turkey	14.9	0.01781	0.0000122	8.0622	0.011	0.00036	0.932
UK	19.1	0.06879	0.0000467	15.54	0.0264	0.0003	0.884

Table 2 Calculated "calibrated" benefit areas (BA^c)

Countries	Linear–Quadratic					
	D	z_0	$g(z_0)$	$G(z_0)$	BA	Eff
Albania	0.0785	29.594	−3.38	−52.05	81.24	3.5872
Austria	0.1609	84.649	3.3	294.1	628.6	27.756
Belgium	0.0474	63.406	0.5	37.2	182.8	8.0715
Former Czech.	0.0378	160.988	11.24	5119.8	2264.6	100
Denmark	0.2735	58.138	−2.5	369.72	536.7	23.698
Finland	0.0619	46.182	2.4	154.72	114.3	5.0453
France	0.0428	149.22	42.4	7726.3	309.2	13.65
Greece	0.1076	16.83	−1.62	22.23	45.5	2.0095
Hungary	0.0381	13.66	2.51	54.72	17.9	0.7901
Italy	0.1191	25.22	12.5	431.19	108.1	4.7726
Luxembourg	0.5178	5.56	−0.73	1.4	5.8	0.2572
Netherlands	0.0985	54.98	−6.18	329.7	424.4	18.741
Norway	0.1356	21.056	0.94	16.75	30.6	1.3508
Poland	0.0956	46.67	−8.03	−18.57	333.7	14.734
Romania	0.0333	19.87	5.5	147.1	35.8	1.5803
Spain	0.0016	245.43	10.2	2563.2	527.8	23.305
Sweden	0.0732	73.35	4.1	147.1	201.7	8.9075
Switzerland	3.2893	17.87	5.56	55.8	76.5	3.378
Turkey	0.0099	147.82	8.1	1698.5	698.65	30.851
UK	0.0061	200.5	15.6	4452.1	759.9	33.551

Looking at Table 2 is worth mentioning that large industrial upwind counties seem to have a large benefit area. Looking at the European Monitoring and Evaluation Program (EMEP) and the provided transfer coefficients matrices with emissions and depositions between the European countries it can be seen that the countries with large benefit areas are those with large numbers on the diagonal indicating the significance of the domestic sources of pollution [1]. At the same time the large off–diagonal transfer coefficients show the influence of one country on another in terms of the externalities imposed by the Eastern European countries on the others and the transboundary nature of the problem. In the same lines, near to the sea countries may face small benefit areas as the damage caused by acidification depends on where the depositions occur. In the case of occurrence over the sea it is less likely to have much harmful effect, as the sea is naturally alkaline. Similarly if it occurs over sparsely populated areas with acid tolerant soils then the damage is low, [10].

4 Conclusions and Policy Implications

The typical approach defining the optimal pollution level has been to equate the marginal (of an extra unit of pollution) damage cost with the corresponding marginal abatement cost. An efficient level of emissions maximizes the net benefit, that is, the difference between abatement and damage costs. Therefore the identification of this efficient level shows the level of benefits maximization, which is the resulting output level if external costs (damages) are fully internalized.

In this paper the corresponding optimal cost and benefit points were evaluated analytically. We shown that the optimal pollution level can be evaluated only under certain conditions. From the empirical findings is clear that the evaluation of the "calibrated" Benefit Area, as it was developed, provides an index to compare the different policies adopted from different countries. In this way a comparison of different policies can be performed. Certainly the policy with the maximum Benefit Area is the best, and the one with the minimum is the worst. Clearly the index BA^c provides a new measure for comparing the adopted policies.

It is clear that due to the model selection, the regression fit of the model, the undergoing errors and the propagation create a Risk associated with the value of the Benefit Area. This Associated Risk is that we try to reduce, choosing the best model, and collecting the appropriately data.

Policy makers may have multiple objectives with efficiency and sustainability being high priorities. Environmental policies should consider that economic development is not uniform across regions and may differ significantly, [12]. At the same time reforming economic policies to cope with EU enlargement may face problems and this may in turn affect their economic efficiencies, [11].

Acknowledgments This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund. D.K would like to express its gratitude to Heracleitus II for the financial support.

Thanks are also due to three anonymous reviewers for their helpful and constructive comments in an earlier draft of our paper as well as to the participants in the 5th International Conference on Risk Analysis (Biomedicine, Environmetrics, Economics, Finance & Reliability) in the Polytechnic Institute at Tomar, Portugal (30 May-1 June 2013) for their encouraging comments.

An extended version of this presentation with the consideration of all combinations of linear, quadratic and exponential abatement and damage cost functions as well as the relation to the existing relative literature can be fount in [10].

References

1. EMEP.: Airborne transboundary transport of sulphur and nitrogen over Europe: model description and calculations. EMEP/MSC-W reports (various years)
2. Halkos, G.: Economic perspectives of the acid rain problem in Europe. Ph.D. thesis, Department of Economics and Related Studies, University of York (1992)

3. Halkos, G.: An evaluation of the direct costs of abatement under the main desulphurisation technologies. MPRA paper 32588, University Library of Munich, Germany (1993)
4. Halkos, G.: An evaluation of the direct cost of abatement under the main desulfurization technologies. Energy Sources **17**(4), 391–412 (1995)
5. Halkos, G.: Incomplete information in the acid rain game. Empirica J. Appl. Econ. Econ. Policy **23**(2), 129–148 (1996)
6. Halkos, G.: Modeling optimal nitrogen oxides abatement in Europe. MPRA paper 33132, University Library of Munich, Germany (1997)
7. Halkos, G.E.: Econometrics: Theory and Practice. Giourdas Publications, Athens (2006)
8. Halkos, G.: Construction of abatement cost curves: the case of F-gases. MPRA paper 26532, University Library of Munich, Germany (2010)
9. Halkos, G., Kitsos, C.P.: Optimal pollution level: a theoretical identification. Appl. Econ. **37**, 1475–1483 (2005)
10. Halkos, G., Kitsou, D.: Uncertainty in optimal pollution levels: modeling the benefit area. J. Environ. Plan. Manag. **58**(4), 678-700 (2015). doi:10.1080/09640568.2014.881333
11. Halkos, G.E., Tzeremes, N.G.: Economic efficiency and growth in the EU enlargement. J. Policy Model. **31**(6), 847–862 (2009)
12. Halkos, G.E., Tzeremes, N.G.: Measuring regional economic efficiency: the case of Greek prefectures. Ann. Reg. Sci. **45**(3), 603–632 (2010)
13. Hutton, J.P., Halkos, G.: Optimal acid rain abatement policy in Europe: an analysis for the year 2000. Energy Econ. **17**(4), 259–275 (1995)
14. Kitsou, D.: Estimating damage and abatement cost functions to define appropriate environmental policies. Ph.D. thesis, University of Thessaly (2014)
15. Kneese, A.V.: Rationalizing decisions in the quality management of water supply in urban-industrial areas. In: Edel, M., Rothenberg, J. (eds.) Readings in Urban Economics. The MacMillan Company, New York (1972)

Part II
Risk Methods for Management and Industry

Finite Populations Sampling Strategies and Costs Control

Dinis Pestana, Maria Luísa Rocha and Fernando Sequeira

To guess is cheap, to guess cheaply can be wrong and expensive.

Abstract Excellent data analysis methodologies fail to produce good results when using bad data. Bad data arise from inadequate strategies at the collecting stage, that are responsible for bias, or insufficient to produce accurate estimates of parameters of interest. Sampling is the statistical subfield that uses randomness as an ally in data gathering, the gold standard in ideal situations being to collect samples without replacement (thus each item bringing in new information, and as a consequence the estimator having reduced variance when compared to the corresponding sampling with replacement estimator). A quick overview of sampling strategies is presented, showing how they deal with cost control in non-ideal circumstances. Comments on the use of immoderately large samples, on the reuse of samples, and on computational sample augmentation, and other critical comments on misuse of statistics are

D. Pestana (✉)
CEAUL and DEIO–FCUL, Instituto de Investigação Científica Bento da Rocha Cabral,
Universidade de Lisboa, Lisbon, Portugal
e-mail: dinis.pestana@fc.ul.pt

M.L. Rocha
Departamento de Economa e Gestão and CEEAplA, Universidade dos Açores,
Azores, Portugal
e-mail: lrocha@uac.pt

F. Sequeira
CEAUL and DEIO–FCUL, Universidade de Lisboa, Lisbon, Portugal
e-mail: fjsequeira@fc.ul.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_16

registered, in the hope that these alerts improve the obtention of statistical findings, so often blurred because sophisticated statistical analysis is useless when it uses bad data.

Keywords Sampling · Sample size · Sampling strategies · Accuracy

1 Introduction

Statistics is a privileged tool in building up knowledge from raw information, since the aim of statistics is to infer from an observed sample what is likely to be true for the whole population, i. e. to provide a formal standing to inductive reasoning. With infinite populations dealing with incomplete information collected by sampling is the unique option. In finite populations, although to observe entire populations via a census is conceptually possible, sampling is in general more advisable, since a census is always more time and resources consuming than sampling, and the cost when dealing with large populations is indeed very high. For instance, the USA 2010 population census cost $42 per capita, a total of 13 billion and, at the end, bitter controversies follow, as in *PS: Political Science and Politics* **33**, namely [1, 5, 9], or *Statistical Science* **9**(4), namely [4, 7, 30], just to cite a few papers debating the issue. Moreover, there exist members of the population that try to elude census (illegals and homeless people, for instance), no census is perfect, and the cost of trying to reach this thin slice of the population is unaffordable.

It is even arguable that the use of a large number of poorly trained temporary census officers can endanger the quality of data gathering, and that appropriate sampling (i.e., representative since a correct methodology is used, and the sample size ensures high probability that the population heterogeneity is being taken into account), much less expensive, could provide better results, since it uses a much smaller number of high quality trained professional officers. The complementarity of census and samples has also been used for instance in the 2010 U.K. population census, see the Office for National Statistics report [27]. Observe also that when carrying out a census some population units try to avoid observation, and it is more complicated to deal with this kind of missing observations than with non-response in sampling.

No one would dispute that in all fields knowledge is based on partial information. At the end of the XIXth century, Galton was prescient of the importance of Statistics in tackling complex problems, and his well known statement meaning that Statistics is an intellectual swiss jackknife in cutting through the layers of difficult problems

> [Statistics are] the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man. (quoted by Pearson, in The Life and Labours of Francis Galton)

was indeed prophetic.

Yet, although the role of Statistics as the core body of the experimental method theory has been undisputed since Fisher developed Experimental Design, and Neyman and Pearson shaped the theory of statistical testing, *circa* 1930, three decades ago Statistics was seldom used in many branches of Science. Nowadays we have the reverse situation, no serious experimental science journal publishes a paper devoid of statistical analysis supporting the building of knowledge from information.

This progress is however controversial, in the sense that in many situations the data analysis is performed using bad data, and hence produces bad science. We fear that in the near future the public will mistrust science, since fraud tied to economic interests [14], mere incompetence, and the drive to publish hastily, together with a naïve trust in bibliometrics evaluation [2] leads often to the publication of false conclusions [19]. Perusing the documentation [18] on Ig Nobel prizes (that first make laugh, then [sometimes] think) is recommended, to laugh and to have a critical appraisal of the ways of modern academics. The ravaging effect of bad data are obvious from the above references, or from editorials in outstanding journals such as *The Lancet* or *The New England Journal of Medicine*,

Sampling theory has been developed to ascertain how to collect data. Gathering data is a crucial step in knowledge building, and Fisher's joke on diagnosis and autopsy in his Presidential Address to the First Indian Statistical Congress, in 1938

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

cleverly points out one of the principal causes of blunders in science.

Saving costs is a sensible goal, but unfortunately it is often done with the most inadequate strategy: saving the expense of consulting a specialist, instead of saving cleverly using sampling theory.

The main goal of sampling theory is to provide tools to obtain representative samples, cost being a natural concern, since all rigorous sampling operations are expensive. A representative sample must indeed reflect the heterogeneity of the population units, implying that the sample size must be large enough to reveal variability; but sample size, which is obviously an increasing function of the population dispersion and, in finite populations, of size, depends also on the sampling design, that must take into account questions such as:

- In sampling from finite populations, do we have a list of items in the target population?—If the answer is negative, systematic sampling can be an interesting alternative.
- Is there the need to infer for separate subgroups of the population?—If this is the case, stratified sampling should be used, or post-stratification considered.
- Is it affordable to get a sample of the size needed with the optimal design?—Non-response is an important issue, and eventual bias must be investigated, or multiple imputation techniques [35] used. New developments of Statistics, namely resampling techniques and computational augmentation of samples, and meta-analysis of statistical evidence based on multiple samples, can contribute towards solving this problem.

In the following sections we describe some of the most common designs, and circumstances advocating their use. The idea is always: control everything you can, and what cannot be controlled must be randomised, since chance is the ultimate ally to avoid bias in data gathering. Collect as much data as you need, but not much more than that, since with too many data irrelevant differences become significant very often. If you are tempted to simulate samples, investigate first whether the computational augmentation of samples is truly advisable.

2 Sampling from Finite Populations: The Gold Standard

Most monographs on sampling focus on design-based sampling from finite populations. Design-based, as opposed to model-assisted based, indicates that randomness intervenes through the probability of selection of items from the population to the sample (while in model-assistedl based sampling [32] the assumption of some stochastic model for the population has some bearing on the constitution of samples; for a very elementary example assuming Poisson randomness [33] on quadrats sampling: assuming that the population scatter has Poisson regularity, and that the observation of appropriately—i. e., randomly—chosen “unit” plots, renamed quadrats, provides enough information to estimate the density of the population, the population size is estimated as the product of the estimated density by the proportion of sampled quadrats).

When sampling from a finite population of size N in order to estimate some parameter of interest—often the mean value (and observe that a proportion p is the mean value of a Bernoulli model), or a function of the mean value μ such as the total $\tau = N\mu$—, an important step is to choose a sampling strategy, and from that to decide the sample size needed to estimate the parameter with the degree of accuracy needed, at a given confidence level.

The simplest situation arises when the selection cost per unit, for planning purposes, is assumed to be the same for each possible item, and in addition we are not interested in subpopulations.

Under that assumption, the *gold standard* is *srswr—simple random sampling without replacement*. Observe that sampling without replacement implies a mild form of dependency (exchangeability), and hence approximate confidence intervals rely on asymptotic results for sums of mildly dependent random variables.

This sampling strategy is unique in the sense that the probability of selecting any sample of size n is $\frac{1}{\binom{N}{n}}$. Observe also that selecting without replacement has an interesting consequence: any observation brings in new information, with the ultimate effect of reducing the estimator variance.

In fact, when estimating the mean value μ from a simple random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, with equal selection probabilities $\pi_s = \frac{1}{N}$, either with or without replacement, via the estimator $\widetilde{\mu} = \frac{1}{n}\sum_{k=1}^{n} X_k$ (unbiased minimum variance linear), the variance when using replacement is $\frac{\sigma^2}{n}$, while without replacement there is a finite population correction reflecting the sampling fraction $\frac{n}{N}$, and we obtain smaller variance $\frac{\sigma^2}{n}\,\frac{N-n}{N-1}$ (which can be unbiasedly estimated by $\widetilde{V}(\widetilde{\mu}) = \frac{S^2}{n}\,\frac{N-n}{N}$, where S^2 is the sample variance).

Therefore, in case we wish to estimate μ with an error bound B at a $(1-\alpha)\times 100\,\%$ confidence level from a population whose standard deviation is σ, denoting $z_{1-\frac{\alpha}{2}}$ the $(1-\frac{\alpha}{2})$-quantile of the standard normal,

- when sampling *with* replacement, the requirement $z_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}} < B$ implies that we should choose a sample size $n_w > \frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{B^2}$.
- when sampling *without* replacement, the requirement $z_{1-\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} < B$ implies that we should choose a sample size

$$n_{w^*} > \frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2\,\frac{N}{N-1}}{B^2 + z^2_{1-\frac{\alpha}{2}}\,\frac{\sigma^2}{N-1}} = \frac{N}{1+\frac{B^2}{\frac{z^2_{1-\frac{\alpha}{2}}\,\sigma^2}{N-1}}} \approx \frac{N}{1+\frac{(N-1)\,B^2}{z^2_{1-\frac{\alpha}{2}}\,s^2}},$$

an expression that clearly shows that the sampling effort when sampling without replacement should increase with the population size and the population variance.

As $n_w > n_{w^*}$, the cost to achieve a fixed accuracy is reduced when sampling without replacement. For instance, for $N = 1927$, $\sigma = 5.38$, to guarantee an error bound $B = 0.5$,with confidence $95\,\%$, $n_{w^*} \geq 362$, while if sampling with replacement we should use $n_w \geq 445$. For smaller values of the standard deviation, less than $\frac{NB^2}{80}$, say, for a population size of 1927, the size reduction is however very tiny, but sampling without replacement still saves costs mainly because it is in general easier and quicker to implement.

Observe that the use of standard normal quantiles of appropriate probability in the case of sampling with replacement is a simple consequence of the classical central limit theorem, since the independence assumption is true. When sampling without replacement, the normal approximation is justified by the Erdös-Rényi [11] central limit theorem extension assuming exchangeability, an information seldom stated in sampling monographs.

In many situations, a sample of size greater than n_{w^*} is collected, the goal being to overcome non-response, which is an important source of error and bias, since there is a tacit belief that statistical inference using large samples will be more accurate. This is trivially true, but irrelevant significance is a possible side effect of immoderately large samples; hence the recommendation is: $n \geq n_{w^*}$ is a sensible guidance, much larger samples will cost more and the benefit is arguable.

3 Deviating from the Gold Standard to Save Costs

Two main reasons to deviate from the *gold standard* are unequal costs per unit, or unavoidable drawbacks such as the absence of a sampling frame. On the other hand, the sample unit cost in multi-step cluster sampling is optimal. Note that with this strategy more information than immediately needed is in general collected, with the background purpose of using it in future studies; however the re-use of samples can be a serious source of bias, and should be avoided.

On the other hand, advantages may arise using special strategies, such as pooling units, or adapting the sampling scheme during implementation, according to whether it provides good or bad results.

3.1 Stratified and Group Sampling

Stratified sampling is advisable when we can partition the population in a small number of subpopulations. It is a combination of census (of the strata) and sampling, in general *srswr* (within each stratum). Group (or cluster) sampling combines census and sampling the other way round: as this sampling strategy is fit for the case of many distinct subgroups, we first sample to choose some of the subgroups randomly (and this procedure may be repeated if needed), and then include all the observable units in the selected subgroups in our sample.

This is considered to be the sampling strategy with best return, in the sense that the cost per unit is optimal. In general, this sampling strategy is useful when within groups heterogeneity is big, while the cluster means are approximately equal.

Observe however that there might exist unaccounted dependencies biasing the results, and that on the other hand this type of data collection is often done having in mind the purpose of storing a wealth of information for future use. However, as we comment in Sect. 4, reuse of samples should be avoided.

So, although we recognise that cluster sampling cuts costs, we shall focus on stratified sampling.

Suppose that a population of size N can be classified into ν non overlapping strata, so that the numbers of items in the strata are $N_k,\ k=1,\ldots,\nu$. Suppose further that the sampling unit costs are $c_k,\ k=1,\ldots,\nu$, and that the standard deviation within each stratum is $\sigma_k,\ k=1,\ldots,\nu$. An unbiased estimate of the population mean is $\overline{x}_{st} = \frac{1}{N}\sum\limits_{k=1}^{\nu} N_k \overline{x}_k$ where the $\overline{x}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_{kj}$ are the unbiased mean strata estimators, i.e. $(x_{k1},\ldots,x_{kn_k})$ is the $srsws$ collected in the k-th stratum.

We shall consider sampling efforts within strata $w_k,\ k=1,\ldots,\nu$; in other words, if the sample total size is n, the within strata samples are of size $n_k = n\,w_k$. The approximate sample size n required to estimate the population mean with confidence

$(1-a)\times 100\,\%$ is $n=\dfrac{\displaystyle\sum_{k=1}^{\nu}\frac{N_k^2\sigma_k^2}{w_k}}{\frac{N^2B^2}{z^2_{1-\frac{\alpha}{2}}}+\displaystyle\sum_{k=1}^{\nu}N_k\sigma_k^2}$ and optimal allocations do exist to minimise costs for fixed variance strata: $n_k=n\,\dfrac{\frac{N_k\sigma_k}{\sqrt{c_k}}}{\displaystyle\sum_{j=1}^{\nu}\frac{N_j\sigma_j}{\sqrt{c_j}}}$.

Stratified sampling is highly recommended when within strata heterogeneity is small, and on the other hand heterogeneity of strata means is considerably higher.

3.2 Systematic Sampling

The unavailability of a sampling frame hinders *srsw*, but we can use alternative designs to randomly select units from the population. For instance, we may be interested in the average caloric content of lunches of university students using a campus restaurant, but there is no listing of the population, so *srswr* is out of question.

One of the most advisable ways of dealing with the problem is the following: suppose we wish to take some measurement on 5 % elements from the population, and that we can have sequential access to the members of the population. We can then select a random integer in $\{1, 2, \ldots, 20\}$, say 17, and therefore select in our sample the 17th, the 37th,..., the $(20k+17)$th element from the sequence, $k=1,2,\ldots,n$, with n determined by $N\in[17+20n, 17+20(n+1))$.

Observe that with this systematic sampling strategy, the probability of selection is the same for all individual units, but that the probabilities of selecting different samples of the same size are radically diverse, in fact 0 for most subsets of the population.

Systematic sampling is an interesting alternative in the absence of a sampling frame, since we can fix *a priori* the sampling effort, and it emulates quite well simple sampling in two extreme circumstances: when there is no structure in the sequence of population items as we observe them, or, on the other hand, when the units roughly appear in monotone order. In the hospital balance of debts, for instance, where chronological ordering of files may be correlated to costs, this sampling strategy provides rich information, since is balances the representativity of debts from several periods.

Obviously systematic sampling can be very misleading when sampling from periodic sequences, namely when $1/f$, where f stands for the sample fraction, is approximately the period. On the other hand, if $1/f$ is much smaller than the period, systematic sample can provide some interesting insights, since it artificially creates something similar to post-stracta or to clusters. For instance, many phenomena have a 12 months period, and a $f = \frac{1}{4}$ sampling effort will provide interesting quarterly data. Saving the cost of building up a rigorous and manageable sampling frame is justification enough to use systematic sampling instead of *srswr*.

3.3 Combined, Sequential, and Adaptive Sampling

Combined and sequential sampling have been developments contributing to the US II World War effort. In particular, sequential analysis developments, because of their impact in quality control, have been classified restricted information until the end of the war, so that the original Wald paper [39] publication has been delayed until 1945, when the hostilities ceased. Sample size is not fixed in advance; instead, data are evaluated as they are collected, and further sampling is stopped in accordance with a pre-defined stopping rule as soon as significant results are observed, and this of course is bound to save costs in many instances. Observe however that this is not properly a random sampling strategy, contrarily to the others that we briefly discuss in this overview. However, as it can be considered an ancestor of more modern adaptive sampling techniques, and it produces in fact important results, this technique deserves a brief mention. Observe that sequential analysis is optimistic, in the sense that it is expected that a stopping rule will act so that our purposes can be achieved with smaller size samples than recommended using for instance *srswr*.

In the early forties, the need to detect recruits with venereal diseases led Dorfman [10] to investigate the idea of pooling blood samples of several soldiers, say 10, and to check whether the analysis would return a positive result. In that case, separate 10 analyses would be done to detect the infected ones, otherwise this single analysis would give a clean bill of health to the 10 members of the group.

Suppose that the prevalence rate of the disease is p, and that in the first step we analyse the amalgamated blood of n individuals. The expected number of analyses needed to screen each group of n individuals with this pooling technique is then $n* = (1-p)^n + (n+1)\,[1-(1-p)^n]$, and in general $n* \ll n$, lowering costs.

The optimal group size can the be easily be computed, and it naturally increases with the inverse of the prevalence rate. It is worth mentioning that this technique is worth considering for $p < 0.30663$, and that there is an interesting discontinuity, in the sense that there is an abrupt change of the optimal size from $n = 1$ to $n = 3$.

The idea of pooling also occurred to Turing and the team working at Bletchley Park with the Banburismus technique: to test hypotheses about whether different messages coded by German Enigma machines should be connected and analysed together, but this and Turing's work on sequential analysis, that it seems he devised at the same time and independently of Wald, remained secret until 1975 [29].

Combined sampling has important applications in quality control, see [6] for an extensive bibliography on the subject up to 1992, since unfortunately, as far as we know, Part B of that *Annotated bibliography of composite sampling* has never been published. Santos and co-workers have been investigating composite sampling when qualitative analyses have imperfect sensibility and/or specificity, and to quantitative analyses, [24, 31] and references therein.

Adaptive sampling is a nice evolution of sequential sampling, in the sense that it shares the same type of common sense. In fact, random samples on average perform nicely, but there is in the essence of sampling the possibility of occasionally getting samples that are not representative of the whole population. In particular, in cluster sampling, there is the possibility that some of the clusters provide poor information. Adaptive sampling, in simple terms, stems out from the belief that neighbouring clusters have some similarity, and thus that poorly informative clusters should be discarded in favour of increasing sampling effort in the vicinity of clusters providing much information.

An example of adaptive sampling helps to understand why we claim that is more realistic: suppose that we want to sample zones in the Newfoundland sea to evaluate cod stocks, and that in a first step n randomly chosen square regions with a 3 marine miles diagonal are chosen in the ocean chart.

When in one of those spots the research ship sails 0.25 marine miles without cod catching, this spot is abandoned. On the other hand, if the catch of cod occurs in some zone, the 8 neighbouring spots in the chart grid are added to the sampling plan. Hence, there is a tacit recognition that some spots are useless, while others are useful and hint that the neighbouring ones are also useful.

It is obvious that this adaptive scheme is much more rewarding than a fixed scheme, that at the end of the day could eventually provide none information whatsoever for our purposes. The many facets of adaptive sampling are detailed in [34, 37].

3.4 Distance Sampling

As we said before, in a large majority of cases sampling is done with the purpose of cleverly estimating some parameter, namely a function of the mean value. This is done taking for granted that the population size is a known constant N.

However, for instance in wildlife studies, N is unknown, and the purpose of sampling is to estimate the population size. Crude but clever methods began to develop at the end of the XIXth Century, for instance Petersen's capture-recapture based estimator [28], but in fact similar methods had already been used *circa* 1650 by Bacon to estimate game abundance, and by Laplace in 1780 to estimate the population of French departments. Subsequent sampling using this estimate must envisage increased estimators variance, in the general spirit that in hierarchical models the variance is the sum of two terms: the variance of the conditional mean plus the mean of the conditional variance.

Distance sampling uses the common sense belief that detection capacity fades out with distance. Hence, if someone looks around (point transects) some chosen spots, or looks to both sides of some paths (line transects), it is expected that items on the transect are indeed observed, but that the observation of items decays with distance from the sampling agent. Sensible choice of the transects together with appropriate modelling of the decay rate of this observation function are the cornerstones to estimate from the "censored" sample that has been observed. For further details, [22, 25] and references therein.

4 Considerations

The initial stages of statistical inference were severely constrained by the incapacity of dealing with complex models. Fortunately most of the usual statistics are functions of sums, and the central limit theorem asserted that a "normal" approximation could be used for sums of random variables, under very broad circumstances. This was very fortunate, since the cumulants of the normal are 0 but for the first and the second (this is the ultimate reason why the central limit theorem holds for sums of independent identically distributed random variables with finite variance), and hence it is a pure location/scale parametrized family, linearly amenable to the very well tabulated standard normal distribution. And assuming normality, excellent exact results did follow to deal with means (Student's t), variances (chi-square) and quotient of variances (F), appropriate to deal with samples of all sizes, *including small size or moderately sized samples*.

Also in the first half of the XXth Century, important results of nonparametric statistics served to deal with important location/scale problems, distribution fitting, asymmetry, randomness, association, and many other important questions, also with *small size or moderately sized samples*.

In the mid century, the availability of computing devices capable of easily dealing with large datasets and tricky distributions brought great developments in statistics. Namely, the team led by Tukey [26, 38] advocated EDA (Exploratory Data Analysis) so convincingly that a later return to Confirmatory Data Analysis has been necessary. Other side effects of the development of computers have been the ease of use of Monte Carlo, Jacknife, Bootstrap, and many other techniques, an interesting turning point: for instance, the question of robustness was suddenly as relevant as sufficiency, or even more relevant.

Resampling techniques, and real time data collecting data, changed the state of the art abruptly: instead of dealing with small samples, very large samples needed new *data mining* techniques, in a sense devices to separate gold from ore in haphazard collected data. This has been particularly important in fields such as geophysics, astronomy, economics, where the daily automated collection of thousands of data is easily feasible.

But even in other areas, such as life sciences, in which many experiments deal with small samples, the situation changed abruptly due to two developments: on one hand, the formal computational augmentation of samples and simulation; on the other hand, the ease of communications strengthened collaboration bonds between groups investigating similar questions, and retrieval of experimental data became a recommendation, if not a standard (namely using the Cochrane Collaboration), so that systematic reviews and meta analysis, an expression first appearing in [13] and nowadays ubiquitous in research, superseded the restrictive situation of analysing small data sets.

Observe that this accumulation of data can, however, result in significance of truly minimal and irrelevant differences; cumulative meta-analysis, stopping addition of studies when the required quantity of information is reached, is an interesting development, in a sense similar in spirit to sequential analysis. On the other hand, computational sample augmentation can back-fire. For instance, due to the optimal entropy of the standard uniform among (0,1)-supported random variables decreases power when testing uniformity using computationlly augmented samples [8].

Another caveat, for those dealing with retrospective studies: observe that in many situations the data that are being analysed are a haphazard sample, and inference in that case is hazardous.

And a final caveat: gathering data is expensive, tiring, eventually boring, and this often causes a very serious blunder: the same data are reused and reused to investigate different questions. Using already collected data, eventually to do some confirmatory data analysis on hypothesis suggested by those data at the exploratory stage is indeed a source of questionable results.

Finally, a brief reference to some sources of more detailed information for those who want to lean more on sampling. *The Survey Kit* [12] provides useful practical information, and can be complemented by [23]; [21, 33] are excellent primers, as well as [3, 36], and [35] a thorough treatment of more advanced topics; having Hansen and Hurwitz as are co-authors is by itself a recommendation of [16]. The pioneering papers by Hansen and Hurwitz [15] and by Horvitz and Thompson [17] are well worth reading. Sampling rare populations [20] requires specific strategies, be it in the area of small domains estimation, or in important issues such as estimating the VaR (Value at Risk), high quantiles of great importance in the applications of extreme value theory to risk.

Acknowledgments This research has been supported by National Funds through FCT—Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011, and PTDC/FEDER.
The authors are grateful for very useful comments from Prof. Sneh Gulati, that helped them to improve the readability of the text, and to the comments and suggestions of the anonymous referees, very helpful in improving the presentation.

References

1. Anderson, M., Fienberg, S.E.: History, myth-making and statistics: a short story about the reapportionment of congress and the 1990 census. PS. Polit. Sci. Polit. **33**, 783–794 (2000)
2. Arnold, D.N.: Integrity Under Attack: The State of Scholarly Publishing. http://www.ima.umn.edu/arnold//siam-columns/integrity-under-attack.pdf
3. Barnett, V.: Sample Surveys: Principles and Methods. Arnold, London (2002)
4. Belin, T.R., Rolph, J.E.: Can we reach consensus on census adjustment? Stat. Sci. **9**, 486–508 (1994)

5. Billard, L.: The census count: who counts? How do we count? When do we count? PS. Polit. Sci. Polit. **33**, 767–774 (2000)
6. Boswell, M.T., Gore, S.D., Lovison, G., Patil, G.P.: Annotated bibliography of composite sampling Part A: 1936–92. Environ. Ecol. Stat. **3**, 1–50 (1996)
7. Breiman, L.: The 1991 census adjustment: undercount or bad data? Stat. Sci. **9**, 458–475 (1994)
8. Brilhante, M.F., Mendonça, S., Pestana, D., Sequeira, F.: Using products and powers of products to test uniformity. In: Luzar-Stiffler, V., Jarec, I. Bekic, Z. (eds.) Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces, IEEE CFP10498-PRT, pp. 509–514
9. Brunell, T.L.: Statistical sampling to estimate the U.S. population: the methodological and political debate over census 2000. PS. Polit. Sci. Polit. **33**, 775–782 (2000)
10. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Stat. **14**, 436–440 (1943)
11. Erdös, P., Rényi, A.: On a central limit theorem for samples from a finite population. Publ. Math. Inst. Hung. Acad. Sci. **4**, 49–61 (1959)
12. Fink, A. (ed.) The Survey Kit—1: The Survey Handbook. 2: How to Ask Survey Questions. 3: How to Conduct Self-Administered and Mail Surveys. 4: How to Conduct Telephone Surveys 5: How to Conduct In-Person Interviews for Surveys. 6: How to Design Survey Studies. 7: How to Sample in Surveys. 8: How to Assess and Interpret Survey Psychometrics 9: How to Manage, Analyze, and Interpret Survey Data. 10: How to Report on Surveys. Sage Publications, Thousand Oaks (2003)
13. Glass, G.V.: Primary, secondary, and meta-analysis of resealch. Educ. Res. **5**, 3–8 (1976)
14. Goldacre, B.: Bad Science, Harper Perennial. Fourth Estate, London (2009)
15. Hansen, M.M., Hurwitz, W.N.: On the theory of sampling from finite populations. Ann. Math. Stat. **14**, 333–362 (1943)
16. Hansen, M.M., Hurwitz, W.N., Madow, W.G.: Sample Survey Methods and Theory. Wiley, New York (1962)
17. Horvitz, D.G., Thompson, D.J.: A Generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. **47**, 663–685 (1952)
18. Improbable Research. http://www.improbable.com/ig/
19. Ioannidis, J.P.A.: Why Most Published Research Findings Are False. http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124
20. Kalton, G., Anderson, D.: Sampling rare populations. J. R. Stat. Soc. A **149**, 65–82 (1986)
21. Lynn, P.: Principles of sampling,In: Greenfield, T. (ed.) Research Methods for Postgraduates, pp. 185–194. Arnold, London (2002)
22. Marques, T.A., Buckland, S.T., Bispo, R., Howland, B.: Accounting for animal density gradients using independent information in distance sampling surveys. Stat. Methods Appl. (2013). doi:10.1007/s10260-012-0223-2
23. Marsden, P., Wright, J. (eds.): Handbook of Survey Research. Emerald, United Kingdom (2010)
24. Martins, J.P., Santos, R., Felgueiras, M.: A Maximum Likelihood Estimator for the Prevalence Rate Using Pooled Sample Tests Notas e Comunicações do CEAUL 27 (2013)
25. Morrison, Ml, Block, W.M., Strickland, M.D., Collier, B.A.: Wildlife Study Design. Springer, New York (2008). Sample survey strategies 137–198 Sampling strategies: applications 199–228
26. Mosteller, F., Tukey, J.W.: Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Boston (1977)
27. Office of National Statistics: Census Coverage Survey Sample Balance Adjustment. 2011 Census: Methods and Quality Report. www.ons.gov.uk/.../census/.../census...census.../ccs-sample-balance-adjus...? (2011)
28. Petersen, C.G.J.: The yearly immigration of young plaice into the Limfjord from the German Sea. Dan. Biol. St. **6**, 5–84 (1895)
29. Randell, B.: The Colossus. In: Metropolis, N., Howlett, J., Rota, G.C. (eds.) A History of Computing in the Twentieth Century, pp. 47–92. Academic Press, New York (1980)
30. Ronzio, C.R.: Ambiguity and discord in U.S. Census data on the undercount, race/ethnicity and SES: responding to the challenge of complexity. Int. J. Crit. Stat. **1**, 11–18 (2007)

31. Santos, R., Martins, J.P., Felgueiras, M.: Discrete Compound Tests and Dorfmans Methodology in the Presence of Misclassification. Notas e Comunicações do CEAUL 26 (2013)
32. Särndal, C.-E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, New York (2003)
33. Scheaffer, R.L., Mendenhall III, W., Ott, R.L., Gerow, K.: Elementary Survey Sampling. Duxbury, Belmont (2012)
34. Seber, G.A.F., Salehi, M.M.: Adaptive Sampling Designs: Inference for Sparse and Clustered Populations. Springer, New York (2012)
35. Singh, S.: Advanced Sampling Theory with Applications, How Michael 'Selected' Amy. Kluwer, Dordrecht (2003)
36. Thompson, S.K.: Sampling. Wiley, New York (2012)
37. Thompson, S.K., Seber, G.A.F.: Adaptive Sampling. Wiley, New York (1996)
38. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Boston (1977)
39. Wald, A.: Sequential tests of statistical hypotheses. Ann. Math. Stat. **16**, 117–186 (1945)

Industrial Production of Gypsum: Quality Control Charts

Luís M. Grilo, Helena L. Grilo and Cristiano J. Marques

Abstract The production of gypsum (marketed if it accomplishes the required specifications) occurs during the process of flue gas desulphurization in a Portuguese Coal Thermoelectric Central. Important variables in this process are statistically analyzed in the chemical laboratory and quality control charts are implemented to monitor the entire process. In this study individuals and moving range charts of the variable "density of gypsum slurry" are compared with the "more efficient" ones obtained after a Box-Cox transformation. This transformation is used to normalize the data, because its observations come from non-normal models—where classical control charts are considered less appropriate, since they usually exhibit rates of false alarms different from what would be expected. Although it is important to consider different statistical approaches for quality control charts, during the monitoring of an industrial process, in this case study the achieved results lead us, essentially, to similar conclusions.

Keywords Non-normality · Box-Cox transformation · Individuals and moving range charts · Robustness

L.M. Grilo (✉)
Unidade Departamental de Matemática e Física, Instituto Politécnico de Tomar,
Estrada da Serra – Quinta do Contador, 2300-313 Tomar, Portugal
e-mail: lgrilo@ipt.pt

H.L. Grilo
Centro de Sondagens e Estudos Estatísticos, Instituto Politécnico de Tomar,
Tomar, Portugal
e-mail: helenagrilo56@gmail.com

C.J. Marques
Unidade Departamental de Engenharia, Instituto Politécnico de Tomar, Tomar, Portugal
e-mail: cristiano89marques@hotmail.com

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_17

1 Introduction

In the generation of electric power, in coal-based thermoelectric power plants, greenhouse gases are released into the environment, which are harmful to living beings. Among those emissions are nitrogen oxide (NOx), sulphur dioxide (SO2) and particles (fly ashes).

In Fig. 1 shows an installed Flue Gas Desulphurization (FGD) whose main objective is to reduce the content of SO2 from flue gas in order to achieve the prescribed emission limit values. Basically, the removal of SO2 occurs by its reaction in the absorbers with a suspension of limestone (CaCO3) through a series of partial reactions (absorption/neutralization/oxidation). The final product of these reactions is gypsum (CaSO4.2H2O), which is usually, on average, 96 % pure and may be subsequently sold, mostly to the cement industry, if it satisfies the required specifications. Thus a laboratory analysis of the statistical behavior of the relevant variables is necessary. Examples of the relevant behavior are: "pH of gypsum slurry", "% moisture gypsum" and "density of gypsum slurry (DGS)."

In this case study we compare, only for the variable DGS, the classical control charts (or Shewhart) with the ones obtained excluding severe outliers and also with the control charts based on an adequate Box-Cox transformation [2]. After the Box-Cox transformation, data have approximately a normal distribution, which allows the application of statistical quality control charts for the usual normal model. These transformations lead to the construction of relatively efficient quality control charts, in terms of the false alarm rate and the time required to detect the occurred changes. Several simulation studies prove the efficiency of Box-Cox transformations for normalization, as well as the efficiency and robustness of statistical estimators for more robust control charts [1, 3–5, 7, 8].

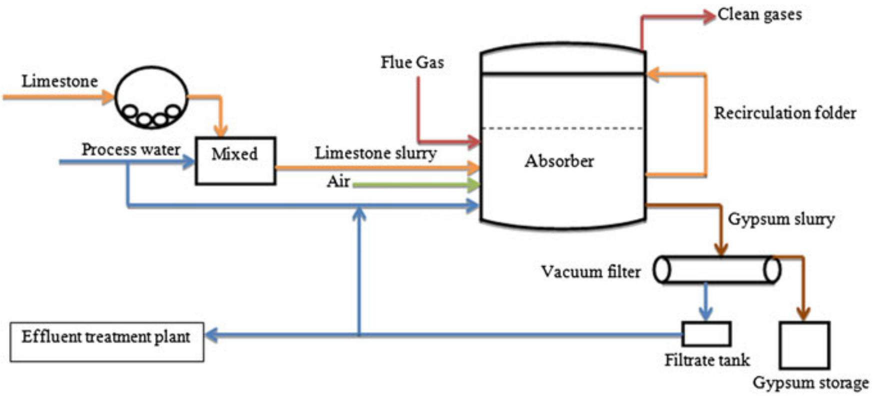


Fig. 1 Scheme of the FGD process

2 Data Analysis

Quality control charts are usually mentioned in monitoring industrial production of gypsum, but the presented variables often have empirical distributions that deviate significantly from the assumption of normality (where outliers are identified, as well as marked asymmetry and/or a high weight tail). In Fig. 2 we visualize parallel box plots for the variable DGS from two installed absorbers. The non-normality of the variable is evident from the asymmetry exhibited in the plots as well as the presence of outliers (both moderate and severe).

Table 1 shows the results of the test of normality for the variable DGS using both the Kolmogorov-Smirnov test (with Lilliefors correction) and the Shapiro-Wilk test. The obtained p-values are (nearly) equal to zero, leading us to reject the hypothesis of normality for both absorbers. Since the sample size is not very high, we believe that the Shapiro-Wilk test is more appropriate, although we also present the results of the Kolmogorov-Smirnov.

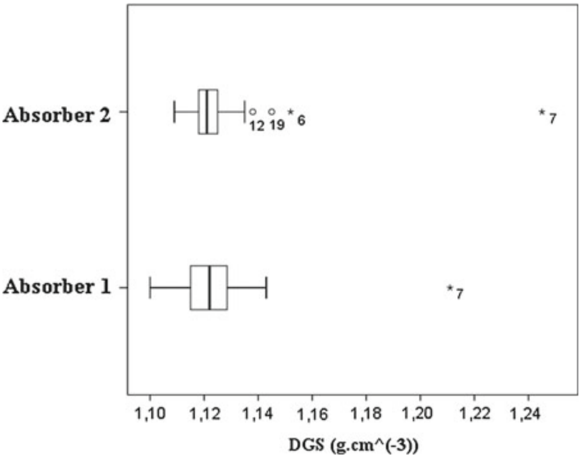


Fig. 2 Box-plot of DGS (g.cm^{-3}), in the two absorbers of FGD installation

Table 1 Results of the normality test for the variable DGS (g.cm^{-3})

Tests of normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
DGS (g.cm^{-3})	Statistic	df	p-value	Statistic	df	p-value
Absorber 1	0.212	39	0.000	0.687	39	0.000
Absorber 2	0.291	39	0.000	0.483	39	0.000

[a] Lilliefors significance correction

3 Quality Control Charts

In statistical quality control a pair of charts (individuals and moving range) is used to monitor variables data from an industrial process where a single measurement at each collection period (sample of one measurement) is available. The individuals chart displays a single measurement where each point of the process line represents an individual case. In the moving range chart each point represents the absolute difference between each value and the previous one. This chart measures the spread in terms of the range of two consecutive samples.

As with other control charts, these types of Shewhart charts also enable us to monitor a process for shifts which modify the mean or variance of the measured statistic. The procedure followed in the preparation of these control charts is also analogous to the charts of averages and ranges.

Calculation and Plotting

First, we should calculate the average of the n individual values,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the difference between data point, x_i, and its predecessor, x_{i-1} with

$$MR_i = |x_i - x_{i-1}|; \qquad i = 2, 3, \ldots, n.$$

Thus, for n individual values, we have $n - 1$ ranges with arithmetic mean

$$\overline{MR} = \frac{1}{n-1}\sum_{i=2}^{n} MR_i.$$

Individuals Control Limits

The Upper Control Limit (UCL), Central Line (CL) and Lower Control Limit (LCL) for the individual values, considering sigma level 3, are obtained with the formulas

$$UCL = \overline{X} + \frac{3}{d_2} \times \overline{MR}, \quad CL = \overline{X} \quad \text{and} \quad LCL = \overline{X} - \frac{3}{d_2} \times \overline{MR}.$$

Moving Range Control Limits

The Upper Control Limit (*UCL*), Central Line (*CL*) and Lower Control Limit (*LCL*) for the range, considering sigma level 3, are calculated with the formulas

$$UCL = D_4 \times \overline{MR}, \quad CL = \overline{MR} \quad \text{and} \quad LCL = D_3 \times \overline{MR},$$

where $D_4 = 1 + 3\frac{d_3}{d_2}$ and $D_3 = 1 - 3\frac{d_3}{d_2}$. From the usual tables given in most textbooks on statistical process control (see, among others [6]), we have, for $n = 2$,

$$d_2 = 1.128; d_3 = 0.853; D_3 = 0; D_4 = 3.267.$$

In the individuals chart all the individual data are plotted serially, following the order in which they were recorded, a line at the average value (*CL*) and lines at the *UCL* and *LCL* values are also added to this plot.

In the moving range chart the calculated ranges MR_i are plotted, a line is added for the average value (*CL*) and a second line is plotted for the range *UCL* (note that $LCL = 0$, because $D_3 = 0$).

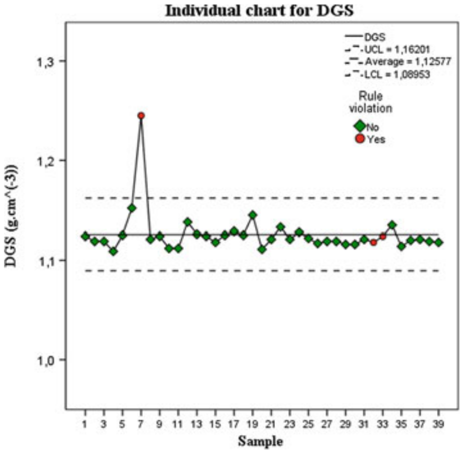
Although we have data for the "density of gypsum slurry" in the two installed absorbers, from now on we are going to consider just the analysis of this variable in the absorber 2.

4 Original Data (Under Non-normality)

In Figs. 3 and 4 we have, respectively, the individuals and moving range control charts with rule violations, obtained with original data (under non-normality). Although, the errors in decisions to declare the production process as an IN/OUT state could be higher in this case, we decided to consider these types of charts because they are a robust tool and the estimated control limits under non-normality may not be a serious problem [9–11].

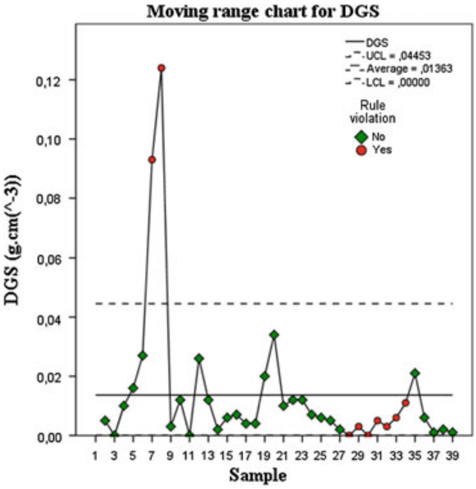
In Fig. 3 we see a point outside the 3 sigma upper control limit and two points (samples 32 and 33) for a group of 8 consecutive points below the center line. In Fig. 4 we have 9 points violating control rules and we can identify some big differences (sample ranges) between two consecutive samples.

These charts are also useful to analyze the DGS behavior and to compare them with the control charts presented next.



Rule Violations for Run	
Sample	**Violations for Points**
7	Greater than +3 sigma
7	2 points out of the last 3 above +2 sigma
32	8 consecutive points below the center line
33	8 consecutive points below the center line
3 points violate control rules.	

Fig. 3 Individuals control chart for DGS (g.cm^{-3}), with sigma level 3



Rule Violations for MR	
Sample	**Violations for Points**
7	Greater than +3 sigma
8	Greater than +3 sigma
8	2 points out of the last 3 above +2 sigma
28	8 consecutive points below the center line
29	8 consecutive points below the center line
30	4 points out of the last 5 below -1 sigma
30	8 consecutive points below the center line
31	8 consecutive points below the center line
32	4 points out of the last 5 below -1 sigma
32	8 consecutive points below the center line
33	8 consecutive points below the center line
34	8 consecutive points below the center line
9 points violate control rules.	

Fig. 4 Moving range control chart for DGS (g.cm^{-3}), with sigma level 3

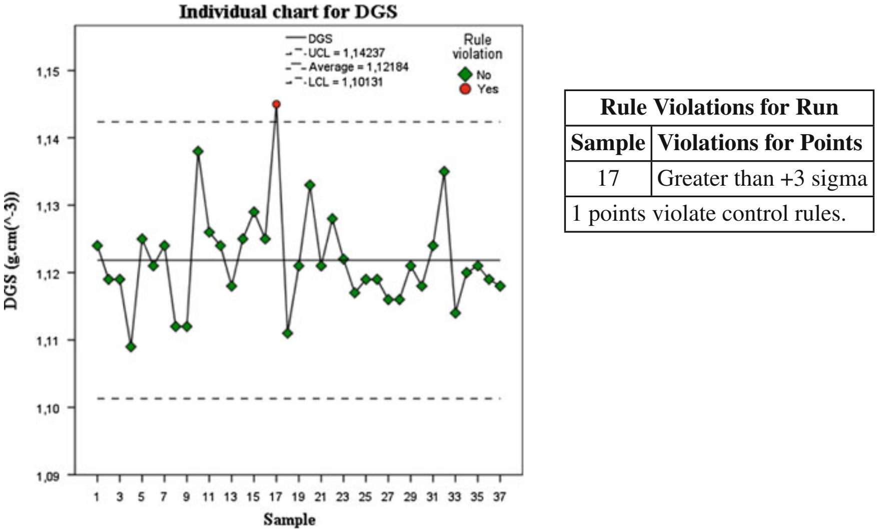
5 Data Without Two Severe Outliers

According to the laboratory technicians' opinions, severe outliers should be removed since they believe that the unusual observations were caused either by a failure of the vacuum carpets used in the dehydration of the production gypsum or some instability in the used energy. As a result we decided to remove two severe outliers and now the distribution could be considered approximately normal, for a 1 % significant level (in the last column of Table 2: since p-value $= 0.022 > 0.01$ we do not reject the null hypothesis of normality).

Table 2 Results of the normality test for the variable DGS (g.cm^{-3}), without two severe outliers

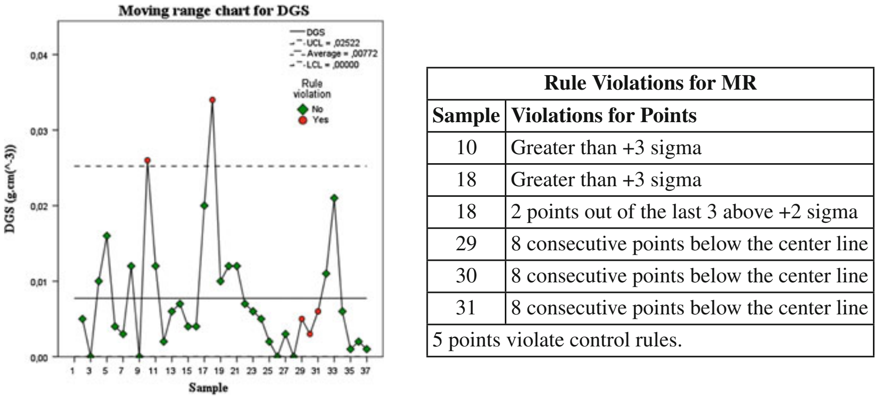
Tests of normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
DGS (g.cm^{-3})	Statistic	df	p-value	Statistic	df	p-value
Absorber 2	0.145	37	0.047	0.930	37	0.022

[a] Lilliefors significance correction



Rule Violations for Run	
Sample	Violations for Points
17	Greater than +3 sigma
1 points violate control rules.	

Fig. 5 Individuals control chart for DGS (g.cm^{-3}), with sigma level 3



Rule Violations for MR	
Sample	Violations for Points
10	Greater than +3 sigma
18	Greater than +3 sigma
18	2 points out of the last 3 above +2 sigma
29	8 consecutive points below the center line
30	8 consecutive points below the center line
31	8 consecutive points below the center line
5 points violate control rules.	

Fig. 6 Moving range control chart for DGS (g.cm^{-3}), with sigma level 3

In Figs. 5 and 6 we have, respectively, the individuals and moving range control charts with control rules violations, obtained after removing two severe outliers. Only one point violates the control rules (a "new" data point falls outside the 3 sigma limits, which represents a process change), in the Individuals chart (Fig. 5), and five points violate the control rules (points outside the 3 sigma limits and consecutive points below the center line), in the moving range chart (Fig. 6), which monitors changes in the spread of a process. This approach, without two severe outliers, gives us a better perspective of the real behavior of the variable DGS.

6 Box-Cox Transformation

Firstly we should note that, using original data and the software SigmaXL on Excel, it isn't possible to find any Box-Cox transformation that leads us to a normal data. Thus, we apply to the data, without two severe outliers, the Box-Cox transformation suggested by software SigmaXL, i.e. DGS^{-5}. In Table 3 we can see the optimal lambda equal to (-5) and the p-value (>0.05) obtained for both Anderson Darling and Shapiro-Wilk statistics, which leads us to believe that we now have a better normal approximation than before.

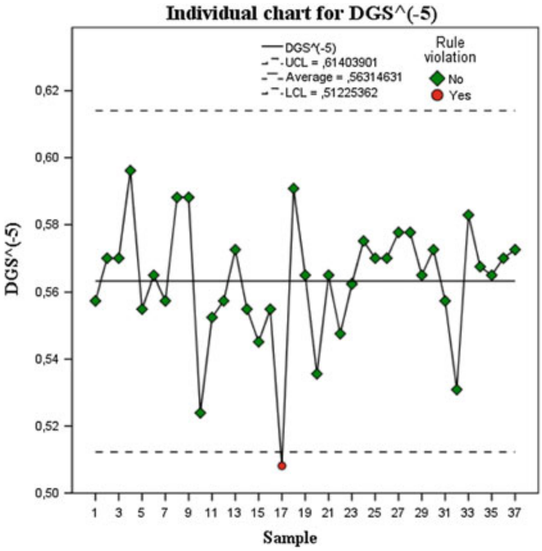
In Figs. 7 and 8 we have, respectively, the individuals and moving range control charts with control rule violations, based on data without two severe outliers and after a Box-Cox transformation. In Fig. 7 a point violates the control rules and, as expected, we have an inverse image, on a different scale, of Fig. 5. In the moving range chart (Fig. 8) we see 5 points that violate control rules. The results of this section lead to conclusions that are similar to the ones obtained before the Box-Cox transformation (see previous subsection).

Table 3 Results of the normality test for DGS^{-5}, without two severe outliers and after a Box-Cox transformation

Box-Cox Power Transformation: DGS (g.cm^{-3})	
Optimal Lambda	-5
Final (Rounded) Lambda	-5
Anderson Darling Normality Test for Transformed Data	
A-Squared	0.736
AD p-value Lambda	0.050

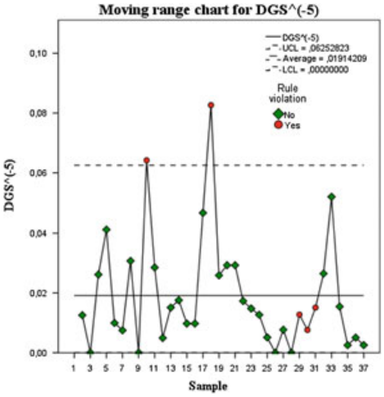
Tests of Normality						
	Kolmogorov-Smirnov [a]			Shapiro-Wilk		
DGS (DGS^{-5})	Statistic	df	p-value	Statistic	df	p-value
Absorber 2	0.137	37	0.078	0.942	37	0.055

[a] Lilliefors significance correction



Rule Violations for Run	
Sample	Violations for Points
17	Less than -3 sigma
1 points violate control rules.	

Fig. 7 Individuals control chart for DGS (g.cm^{-3}), with sigma level 3



Rule Violations for MR	
Sample	Violations for Points
10	Greater than +3 sigma
18	Greater than +3 sigma
29	8 consecutive points below the center line
30	8 consecutive points below the center line
31	8 consecutive points below the center line
5 points violate control rules.	

Fig. 8 Moving range control chart for DGS (g.cm^{-3}), with sigma level 3

7 Conclusions

In this study we use the individuals and moving range control charts to monitor industrial process in three different situations, which enable us to verify that they are sensitive in detecting changes during the process. When we compare the quality control charts, before and after removing two severe outliers, we identify some particularities (in terms of average values and dispersion) of the behavior of DGS. But, when we analyze the quality charts, with respective control rules, before and after applying a Box-Cox transformation, the results point in the same direction.

According to Wheeler's research work [9–11] a normal distribution is not required in the calculation of control limits, which made these types of charts a very robust tool. He also considers that "the transformation of the data to achieve statistical properties is simply a complex way of distorting both the data and the truth" and if we check our data for normality prior to place them on a process behavior chart we "are practicing statistical voodoo". Although we agree with Wheeler, we also believe that with highly skewed data the false alarm rate can be quite high and a Box-Cox transformation could be recommended to obtain efficient and robust estimators. As a result, a non-linear Box-Cox transformation, which in this case study seems unnecessary once we reach similar conclusions.

References

1. Borror, C.M., Montgomery, D.C., Runger, G.C.: Robustness of the EWMA control chart to non-normality. J. Qual. Technol. **31**(3), 309–316 (1999)
2. Box, G.E.P., Cox, D.R.: An analysis of transformations. J. R. Stat. Soc. **26**, 211–252 (1964)
3. Figueiredo, F., Gomes, M.I.: Estimadores robustos de localização. A Estatística em Movimento. In: Neves, M.M., Cadima, J., Martins e, M.J., Rosado, F. (eds.) Sociedade Portuguesa de Estatística, pp. 193–204 (2001)
4. Figueiredo, F., Gomes, M.I.: Transformação de dados em controlo estatístico de qualidade. Novos Rumos em Estatística. In: Carvalho, L., Brilhante e, F., Rosado, F. (eds.) Sociedade Portuguesa de Estatística, pp. 235–245 (2002)
5. Figueiredo, F., Gomes, M.I.: Monitoring industrial processes with robust control charts. Revstat Stat. J. **7**, 151–170 (2009)
6. Montgomery, D.C.: Introduction to Statistical Quality Control, 4th edn. Wiley, New York (2000)
7. Stoumbos, Z.G., Reynolds Jr, M.R.: Robustness to no-normality and autocorrelation of individuals control charts. J. Stat. Comput. Simul. **66**, 145–187 (2000)
8. Tatum, L.G.: Robust estimation of the process standard deviation for control charts. Technometrics **39**(2), 127–141 (1997)
9. Wheeler, D.J.: When can we trust the limits on a process behavior chart? Qual. Dig. (2009)
10. Wheeler D.J.: Good limits from bad data. Qual. Dig. (2009)
11. Wheeler D.J.: Do you have Leptokurtophobia? Qual. Dig. (2009)

Risk Analysis with Reference Class Forecasting Adopting Tolerance Regions

Vasilios Zarikas and Christos P. Kitsos

Abstract The target of this paper is to demonstrate the benefits of using tolerance regions statistics in risk analysis. In particular, adopting the expected beta content tolerance regions as an alternative approach for choosing the optimal order of a response polynomial it is possible to improve results in reference class forecasting methodology. Reference class forecasting tries to predict the result of a planned action based on actual outcomes in a reference class of similar actions to that being forecast. Scientists/analysts do not usually work with a best fitting polynomial according to a prediction criterion. The present paper proposes an algorithm, which selects the best response polynomial, as far as a future prediction is concerned for reference class forecasting. The computational approach adopted is discussed with the help of an example of a relevant application.

Keywords Risk analysis · Reference class forecasting · General linear regression · Predictive models · Tolerance regions

1 Introduction

Risk analysis is a field which seeks to determine and assess factors that may endanger the success of a plan/project or reaching a goal. This technique tries to find preventive tasks in order to reduce the possibility of these unwanted factors from happening. It also seeks to determine counter actions for dealing successfully with these possible obstacles. Thus risk analysis at the end proposes actions to avert negative effects on the competitiveness of the project.

V. Zarikas (✉)
Department of Electrical Engineering, Academic Institute of Technology of Central Greece,
ATEI of Central Greece, Lamia, Greece
e-mail: vzarikas@teilam.gr

C.P. Kitsos
Department of Computer Science, Technological Educational Institute of Athens, Athens, Greece
e-mail: xkitsos@teiath.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_18

Risk analysis includes topics such us: Quantitative risk analysis, Probabilistic risk assessment as an engineering safety analysis, Risk analysis for business, Risk Management, Risk management tools, Certified Risk Analyst, Food safety risk analysis etc.

In the present work we focus on to Risk analysis in business. One of the most interesting methods developed for risk assessment is Reference class forecasting (RCF) [1–15]. This is a method for future predictions about cost factors through looking at similar past situations and their outcomes. RCF estimates predictions about the outcome of a planned project based on actual outcomes in a reference class of similar projects. These methods behind RCF were developed by Daniel Kahneman and Amos Tversky. For his contribution to the theoretical support of refen mmnnp jbrence class forecasting, Kahneman was awarded the Nobel Prize in Economics.

Kahneman and Tversky [13, 14] studied human psychology and discovered that experts' judgment is optimistic most of the times due to overconfidence and positive mood to new developed projects. Managers and decision makers do not tend to consider properly and seriously available distributional information about outcomes even though they are aware of relevant data. Therefore, experts tend to underestimate the uncertainties and costs, completion times, and risks of planned actions, whereas they overestimate the various benefits and profits of the project outcomes.

Lovallo and Kahneman [12, 15] named this characteristic behavior of the experts as the "planning fallacy" and they stated that it stems from the fact that experts take an "inside view". Wrong estimations are caused by decision makers and experts taking an "inside view". The latter means that they were focused on the constituents of the specific planned project tasks and deliverables, instead of on the actual outcomes of similar ventures. The purpose of the reference class method is to utilize similar projects that have already been completed to assess the risks of the planned project. Based on this type of past information it is possible to extract distributional information. Kahneman and Tversky concluded that the absence of consideration of this distributional information, is perhaps the major source of error in forecasting. Thus they stated that "Analysts should therefore make every effort to frame the forecasting problem so as to facilitate utilizing all the distributional information that is available".

Following the terminology of the relevant literature "Taking an outside view" is the analytical task of utilizing distributional information from previous ventures, similar to the one being forecast. Thus, RCF is the methodology for taking an outside view on planned project tasks.

It has been found that wrong forecasts remain compelling even when experts are fully aware of their nature i.e. optimism bias. This is because the awareness of a perceptual or cognitive illusion does not by itself result in a better perception of reality, according to Kahneman and Tversky.

RCF overcomes human bias. This bias is of two types: optimism bias and strategic misrepresentation. This method attacks both issues by cutting directly to outcomes. In experimental tests carried out by various researchers, this method has been proved to be more accurate than conventional forecasting methods.

RCF as applied to a particular project can be described by the following four steps:

1. Identify as many members as possible of a reference class of past similar plans/projects. In order to extract correct statistical inferences the reference class should be at the same time broad enough to include many similar projects but narrow enough to be truly consistently comparable with the project under risk analysis investigation.
2. Estimate a probability distribution for all the reference classes that have been generated by the study of past data. This is not a trivial task since it needs the existence of credible, past data for many projects of each of the reference classes.
3. Decision about which reference class the project under investigation belongs to.
4. Comparison of the analysed project with the reference class distribution, in order to evaluate the most probable outcome based on the existing statistical distribution of outcomes.

RCF eliminates the various uncertain pieces of information and unknown parameters that will affect the project of study. Instead it sets the project in a statistical distribution of outcomes from the class of reference projects. From a statistical point of view the method "consists of regressing forecasters' best gue1ss toward the average of the reference class and expanding their estimate of credible interval toward the corresponding interval for the class" [14].

The idea of searching among past events patterns that closely fit the problem at hand, for weighing outcome risks is of course sensible, although picking up past similar ventures to extract outcome patterns has limitations.

The central proposal of this paper is to use the concept of tolerance regions and use the best predictive model for all the regressions required on all the reference categories used for forecasting. The adoption of the expected beta content tolerance regions, as a better approach for choosing the optimal order of a response polynomial, is an improvement since best predicts the next value within experimental region. Thus, after selecting the correct class, all the regressions based on data belonging to this class, should adopt the proposed algorithmic method. The given algorithm helps to find the best polynomial model which should be used because this is the model that best predicts on an average the future value, which lie in a certain interval with some probability. Therefore our approach contributes directly at the center of the reasoning of the RCF which is the correct prediction. This will be explained more clearly below.

Now, let us clarify the difference between a confidence interval, a prediction interval or a tolerance interval. Which of them is best for use in the method of RCF? It is useful to give a brief review of the three distinct intervals that appear in statistical analysis of data. In case of fitting a parameter to a model, the accuracy or precision can be expressed as a confidence interval, a prediction interval or a tolerance interval which are quite distinct.

The notion of Confidence intervals provides information about how well the best-fit parameter, determined by regression, has been estimated. For example taking many samples from a Gaussian distribution it is expected that about 95 % of the intervals will include the true value of the population best fit parameter. The crucial remark is that the confidence interval informs about the likely location of the true population parameter.

The definition of Prediction intervals on the other hand allows one to know where you can expect to find the next sampled data point. For many samples and assuming Gaussian distribution (a common assumption), it is expected next value to lie within that prediction interval i.e. in 95 % of the samples. The crucial point is that the prediction interval provides information about the distribution of values, not the uncertainty in determining the population parameter.

Finally the richest notion of interval is the tolerance interval. It is defined by two different ratios/percentages. The first determines "how probable" it is desired the value to be and the second expresses what fraction of the values the interval will contain. In case the first value (how sure) is set to 50 %, then a tolerance interval is the same as a prediction interval. If it is set to a higher value (i.e. 95 %) then the tolerance interval is wider. Following now a more technical statistical terminology, under (as in most applications) the assumption of a Gaussian distribution, it is usually asked the 90 % of the sampled values to lie within the the tolerance interval with probability 0.95. It will be explained in next section that as far as the "future observation" is concerned, i.e. the predicted value, the β expected tolerance interval, [16], is adopted. The main merit of a β expected tolerance interval is that remains invariant under affine transformations, [17, 18].

Statistical criteria in model selection for applications, [19] among others, are focused on finding the model that under some criteria "best" fits the data. These criteria are, in principle, functions of the Residual Sum of Squares [20]. Choosing the optimal order of a response polynomial has been treated as a multi-decision problem by Anderson in the sixties [21]. For earlier references and recent developments see for instance [22].

It is very common, albeit inappropriate, to use these best fitting models for prediction too, since the "distance" criteria are not designed for this purpose. In the present study we propose the modeling used in RCF to follow a "prediction" criterion. The chosen model is the one that best predicts on an average the future value, which lie in a certain interval with some probability. This is achieved using beta expected tolerance regions. So, while the regression oriented prediction is based on the extrapolation or interpolation of the best model fitting the data, the proposed method is based on a probabilistic reasoning and provides the model which best predicts the next value within experimental region. In Industry and Management we are more interested on predicting the future observation or to have "an interval" where this value lies. This leads us to tolerance regions.

2 Background of the Method

The definition of the general linear model (GLM), uses the expression

$$\mathrm{Y} = \mathrm{X}\theta + \sigma \mathrm{e} \tag{1}$$

where the involved matrices are $Y \in \Re^{n\times 1}$, $X \in \Re^{n\times(p+1)}$, $\theta \in \Re^{(p+1)\times 1}$ and $e \in \Re^{n\times 1}$. Y represents the observed random vector of responses while X is a matrix of known constants based on the p input variables. The quantity θ is a vector of unknown parameters that determine the polynomial model and e is an unobserved random vector of errors with $\mathrm{E}\,(\mathrm{e}) = 0$, $\mathrm{E}(\mathrm{e}\,\mathrm{e}') = \mathrm{I}$ where $0 \in \Re^{n\times 1}$ is a vector of zeros and $I_n = diag\,(1, 1, \ldots)$ is the unit matrix and variance $\sigma^2 > 0$ unknown. It is a common assumption, when statistical inference is performed, that

$$\mathrm{e} \sim \mathrm{N}(0, \mathrm{I}_n) \tag{2}$$

with mean vector 0 and covariance matrix I_n. A joint $(1-\alpha)100\,\%$ confidence region of the parameters θ can be constructed, given a realization of Y:

$$CR\,(\theta) = \left\{\theta : ps^2F(p, v; 1-\alpha) \geq (\theta - \hat{\theta})(X'X)(\theta - \hat{\theta})\right\} \tag{3}$$

where $v = n - p$, and α is the significant level and F, as usually, the F distribution, with p and v degrees of freedom and s^2 the (unbiased) estimate of σ^2. Expression (3) defines an ellipsoid which plays an important role since with the help of it is possible either to impose experimental design criteria, or to decide which input variable $X_1, X_2, \ldots, X_p \in \Re^{n\times 1}$ will participate to the model, see the pioneering work of Hocking in [23].

The concept of confidence interval uses the determination of a region which contains the parameters under investigation with a certain probability level, usually $(1-\alpha)100\,\%$ with $\alpha = 0.05$. Although this interval is very widely used in statistics, in many applications like econometric or industrial applications it is desirable to have a "region" that contains a certain portion of the production, with a predefined probability. That is the idea of the tolerance region, [24–26], seems the appropriate one. The idea of confidence interval is mainly used to see how well the model fits the data, while we adopt the idea of tolerance interval to see how well the model predicts a "future" value. We shall use the notation Q for any tolerance region, while with T we denote the one we decide as appropriate for our target: to predict the future value as well as possible.

In principle a tolerance region is a statistic $\mathrm{Q}\,(\mathrm{X}, y) = \mathrm{Q}\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}; y)$ from $\Re^n$ to the Borel σ-algebra B in $\Re$, see [16, 25] among others. It can be proved that functions $M\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})$, $\mathrm{W}\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})$ always exist such that

$$\mathrm{Q}\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n}; y) = \left(M\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})\,, \mathrm{W}\,(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{n})\right) \tag{4}$$

We also use for convenience the notation $(M\left(\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n}\right),\mathrm{W}\left(\mathrm{X}_1,\ldots,\mathrm{X}_\mathrm{n}\right)) = (M,\mathrm{W})$. Furthermore, a tolerance region, in case of sampling the random variables from a continuous cumulative distribution, is expressed as [24],

$$\mathrm{Q}\left(\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_\mathrm{n};y\right) = \left[\left(\mathrm{X}_{(\mathrm{k}_1)},\mathrm{X}_{(\mathrm{k}_1+\mathrm{k}_2)}\right)\right] \tag{5}$$

It can be proved that, see [24]

$$\mathrm{F}\left(\mathrm{X}_{(\mathrm{k}_1+\mathrm{k}_2)}\right)-\mathrm{F}\left(\mathrm{X}_{(\mathrm{k}_1)}\right) \sim \mathrm{Beta}\left(\mathrm{k}_2,\mathrm{n}-\mathrm{k}_2+1\right) \tag{6}$$

where $Beta\,(k,m)$ is the Beta distribution and $\mathrm{X}_{(j)}$ is the jth order statistics. Taking a sample from a continuous distribution function, with $\mathrm{k}_1 = r$, $r < \frac{n+1}{2}$ then it can be proved, see [27]

$$\gamma = P\{[\mathrm{F}\left(\mathrm{X}_{(\mathrm{k}_1+\mathrm{k}_2)}\right)-\mathrm{F}\left(\mathrm{X}_{(\mathrm{k}_1)}\right) \geq \beta]\} = 1-\mathrm{I}_\beta\left(n-2r+1,2r\right) \tag{7}$$

where $\mathrm{I}_\beta\left(p,q\right)$ is the Incomplete Beta distribution. It seems more appropriate in applications to be referring to β content tolerance region at confidence level γ if and only if:

$$\gamma = \mathrm{P}[\beta \leq \mathrm{P}_\mathrm{X}\left\{\mathrm{M},\mathrm{W}\right\}] \tag{8}$$

In order to extend the tolerance regions [16], Kitsos in [17] adopted a statistical invariant approach. Muller and Kitsos [18] adopted tolerance regions to construct optimum experimental designs [28]. In this paper we adopt the tolerance regions to construct the best linear econometric model.

For the General Linear Model Eq. 1, and a realization y of Y it is desirable to construct a particular region $T\,(X,y)$ such that the vector Y^* of the future observations Y_1^*, Y_2^*,...,Y_m^* will lie in $T\,(X,y) \subset \Re^m$ with a high probability. The vector Y^* of future responses will follow model Eq. 1, which is a reasonable common assumption. Thus for a given matrix X^* of the input observations we have

$$\mathrm{Y}^* = \mathrm{X}^*\theta+\sigma\mathrm{e}^* \tag{9}$$

with $e^* \sim N(0,I_m)$, $0 \in \Re^{m\times 1}$. The statistical distribution of Y^*, therefore is defined by the same parameters θ,σ from the parameter space $\Theta = \Re^P \times \Re^+$, with elements $\vartheta = (\theta,\sigma)$. Furthermore for given parameter vector θ, Y and Y^* are assumed to be independent. It should be emphasised that is impossible to construct the region $T\,(X,y)$ so that: (i) $Y^* \in T\,(X,y)$ with high probability and (ii) the above is true for every parameter vector $\vartheta = (\theta,\sigma)$ and every realization y of Y. $P_\theta[Y^* \in T\,(X,y)]$ is the probability that Y^* lies in $T\,(X,y)$. Since the tolerance region $T\,(X,y)$ cannot satisfy simultaneously (i) and (ii) as above, the average tolerance region known as β expectation tolerance region is used. Consequently, by definition, $T\,(X,y)$ satisfies

$\beta = \int f_{Y|\theta}(y) P_\theta[Y^* \in T(X, y)]dy$ for every $\theta \in \Re^p$, [25]. It can be shown that this β expectation tolerance region is also a prediction region. This is a consequence of a theorem proved in [18] which states that: the β expectation tolerance region $T(X, y)$, Classical or Bayesian, for the linear model Eq. 1 is evaluated from

$$T\ (\mathrm{X}, \mathrm{y}) = \{\omega \in \Re^{\mathrm{m}} : m(n-P)^{-1} F_{\mathrm{m,n-p_1},\beta} \geq (\omega - \mathrm{X}^*\hat{\theta})'\ \mathrm{S}\ (\mathrm{X})\ \ (\omega - \mathrm{X}^*\hat{\theta})\} \tag{10}$$

where $\hat{\theta} = (\mathrm{X}'\,\mathrm{X})^{-1}\mathrm{X}'\,\mathrm{y}$ is the Least Square Estimate (LES) of θ. Moreover, s^2 and $S(X)$ satisfy $\mathrm{s}^2 = \hat{\sigma}^2\ (\mathrm{X}, \mathrm{y}) = (\mathrm{y}-\mathrm{X}\hat{\theta})'(\mathrm{y}-\mathrm{X}\hat{\theta})$ and $\mathrm{S}\ (\mathrm{X}) = \mathrm{I_m} + \mathrm{X}^*(\mathrm{X}'\,\mathrm{X})^{-1}\mathrm{X}^{*\prime}$. Furthermore, $F_{m,n-p;\,\beta}$ is the β quantile of the F distribution with m and $n-p$ degrees of freedom. It can be also proved, see in [18] Lemma 1, that

$$\mathrm{S}^{-1}\ (\mathrm{X}) = \mathrm{I_m} - \mathrm{X}^*(\mathrm{X}'\,\mathrm{X} + \mathrm{X}^{*\prime}\,\mathrm{X}^*)^{-1}\mathrm{X}'^{*} = I_m - \Lambda(x) \tag{11}$$

where the definition of $\Lambda(X)$ is obvious.

2.1 The Best Predictive Model

Based on the above analysis it is possible now to construct the β expectation tolerance region for a given future response. Handling the volume of the "future" ellipsoid, the best predictive model can be determined with the help of the following proposed algorithm. The largest volume of the β expectation tolerance region corresponds to the worst case of the input variables set, as far as prediction is concerned. The best tolerance region is chosen using the "best amongst the worst method", in other words, amongst the regions with the max beta expected tolerance region, we choose the one with the minimum β expected tolerance. The proposed algorithm (see [29] for a detailed presentation) consists of two basic steps (Step A). Fit all possible linear models for the subsets with k variables from p, $k = 1, 2, ..., p$. For the corresponding k variables calculate the β expectation tolerance region T_{kj} (X, y) and select that k which corresponds to the largest β expectation tolerance region. (Step B). Among the max tolerance regions for the different $k = 1, 2, ..., p$ choose the minimum one. So choose the best subset of variables which corresponds to

$$k_0 = \min_{1 \leq k \leq P} \max_{j} \left\{T_{k,j}(X, y)\right\} \tag{12}$$

with $k = 1, 2, ..., P$ and $j = 1, 2$. In the present paper an algorithm (developed in Mathematica) adjusted to the case of the RCF method is presented.

3 A Specific Example and the Developed Algorithm

In [29] an algorithm developed for the best predictive polynomial model was presented and tested in various case studies of different domains and datasets. In this section a new version of the algorithm (written in Mathematica), adjusted to the RCF method will be described. In [30] a typical application of the RCF method is presented. As an example, the case studied in [30] will be used in order to demonstrate exactly, which issues our proposal tackles in order to improve RCF. This application concerns offshore wind capital cost estimation in the U.S. outer continental shelf.

Moreover, in [30], cost data from existing projects were used as a basis for analogy. For defining classes, the authors in [30] assumed that if the physical infrastructure in two regions is similar, then the project characteristics may be similar, even if other characteristics of development/installation strategies, marine vessel fleet, government regulation, etc. were different. In other words, the implicit assumption of this approach is that the commonalities of offshore wind projects, associated with the technology, infrastructure, capital intensity, complexity, and installation requirements outweigh the differences due to environmental, contractual and market conditions. This is a common philosophy for all applications of the reference class approach. Thus, this set of past projects, for which cost information was available was used to improve the accuracy of comparative cost estimation by limiting the sample to those projects that are similar to the project under analysis. After the determination of the various classes the RCF method was used to inform cost estimates and uncertainty bounds of US offshore wind farms.

The present paper proposes two type of improvements which apply to this particular example. The first improvement can be applied at the very first determination of the classes. RCF places the project in a statistical distribution of outcomes from the class of reference projects. Consequently, we suggest that the best guess uses the average of the reference class and expands the corresponding estimate of credible interval toward the corresponding interval for the class [13]. The first improvement comes into play as follows: instead of using the concept of confidence interval, as defined in refCI, in order to estimate the percentile that corresponds to a risk of cost overrun, one should use the concept of tolerance interval as defined in (10). The latter applies to all specified classes. See also Table 4, in an application concerning Transport Planning in [8].

The second improvement concerns the predictions arising from the regressions that usually appear for the chosen class. In the example of [30] regression models of normalized capital costs were constructed to investigate the physical features that impact expenditures. The present paper proposes that for all the regressions like these concerning normalised Capital Expenditures (CAPEX) versus capacity or CAPEX versus distance to shore, or CAPEX versus water depth or CAPEX versus steel price index, instead of using a conventional simple linear regression one should use the proposed best fitting polynomial according to the criterion of best prediction. The equations under investigation are

$$Capex = a_0 + a_1 Capacity + a_2 Capacity^2 + \cdots + a_n Capacity^n \tag{13}$$

$$Capex = b_0 + b_1 Distance + b_2 Distance^2 + \cdots + b_k Distance^k \tag{14}$$

$$Capex = c_0 + c_1 Depth + c_2 Depth^2 + \cdots + c_l Depth^l \tag{15}$$

$$Capex = d_0 + d_1 Steel + d_2 Steel^2 + \cdots + d_m Steel^m \tag{16}$$

Below we present a new version of the Mathematica code presented in [29] adjusted for a case of the RCF method. This code provides the correct order of the best fitting polynomial as well as its coefficients, as far as prediction with the concept of tolerance regions is concerned. Thus, it is possible to find the order n and all the values of coefficients a_i, $i = 0, \ldots, n$ of the best for prediction polynomial concerning Capacity. Similarly, the code can evaluate the order k, l, m and all the values of coefficients b_i, $i = 0, \ldots, k$, c_i, $i = 0, \ldots, l$, d_i, $i = 0, \ldots, m$ of the best for prediction polynomials concerning Distance to shore, Water Depth and European Steel Price respectively. The code is presented in Appendix A. In this code the user, can use the predefined function $BESTMODEL(n)$, where n is the desired maximum order of polynomial (up to 6). The user can also set the dataset concerning CAPEX and the datasets concerning $dataVARIABLE = (capacity, depth, distance, steel)$ at the beginning of the code.

4 Conclusions

It was demonstrated in [29] that the discussed method which finds the best general linear model for prediction adds real value and models correctly many scientific and technological applications. The applicability of the method concerns a variety of domains like investment decision making, medical estimations, business employment predictions or policy making, see [29]. For example, the method can be successfully and meaningfully be applied in cases where (i) a gynecologist is interested to find a polynomial relation between abdominal circumference and gestational age or (ii) a sociologist in order to write a study regarding employment needs to have a general linear model that relates the number of employees as a function of their annual payroll or (iii) a bank loan department in order to drive strategy regarding industrial research funding, wants to have a crude estimation regarding the relation of the number of available engineers as a function of industry R and D expenditures. Furthermore, the concrete theoretical background of the method ensures the validity of the results. It was shown that for several datasets the selected polynomial differs from the commonly selected one, if the choice respects the criterion of “the best predictive model". This obviously does not to mean that there are no cases where the two methods suggest the same polynomial.

In the present paper we have further proposed that the best predictive polynomial according to the proposed method, i.e. using the tolerance intervals, is the appropriate model for the RCF too. Since the purpose of RCF is as accurate as possible forecasting, the main reason for using the presented adopted algorithm is that after selecting the appropriate reference class, where our case belongs to, the aim is to find the best polynomial model for prediction and not a model respecting a distance criterion.

Future work will be focused in analyzing and implementing the idea of tolerance regions best predictive model in more RCF case studies in order to demonstrate the potential of our proposal. In addition as a future research, it is worth to generalize the proposed method for problems with multiple independent variables or for cases like [28]. It would be interesting to develop integration with the experiment design approach, although in most of the cases in classical experiment design theory, the models are linear [31].

Acknowledgments We would like to thank the referees for the improvement of the English as well for their valuable comments which help us to improve the paper. V. Zarikas acknowledges the support of research funding from ATEI of Central Greece.

5 Appendix

```
(* in this list we set data for matrix X. Here the user of the code has to insert data
either for Normalised Capacity, Capacity, Distance to Shore, Water depth or Euro-
pean Steel price index *)
dataVARIABLE = {, ., ., ., .};
(* in this list we set data for matrix Y which in our example concerns CAPEX *)
dataCAPEX = {, ., ., ., .};
data = Table[{xTRN[[m]], dataCAPEX[[m]]},
 {m,Dimensions[dataVARIABLE][[1]]}];
Y = Table[{dataCAPEX[[n]]},
{n, Dimensions[dataVARIABLE][[1]]}];
(* here the function tst(n) gives the t student distribution probability density
function for the relevant tolerance region *)
tst[n_]:= Sqrt[-n + (1/n*(0.05*(Sqrt[n]*
Beta[n/2, 1/2]))^(2/(1 + n)))^(-1)];
(* here the code normalises data concerning matrix X in the interval [-1,1]*)
n := Dimensions[dataVARIABLE][[1]];
A := (U + DD)/2;
B := U - A;
DD = Min[dataVARIABLE];
U = Max[dataVARIABLE];
xTRN := (dataVARIABLE - A)/B;
```

(* the coefficients of the six polynomials which will be tested with both ctiteria are structured below *)

```
X[0] := Table[{1}, {i, Dimensions[dataVARIABLE][[1]]}];
X[1] := Table[{1, xTRN[[i]]}, {i,
 Dimensions[dataVARIABLE][[1]]}];
X[2] := Table[{1, xTRN[[i]], xTRN[[i]]^2},
 {i,Dimensions[dataVARIABLE][[1]]}];
X[3] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3},
 {i,Dimensions[dataVARIABLE][[1]]}];
X[4] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3,
xTRN[[i]]^4}, {i, Dimensions[dataVARIABLE][[1]]}];
X[5] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3,
 xTRN[[i]]^4,xTRN[[i]]^5},
 {i, Dimensions[dataVARIABLE][[1]]}];
X[6] := Table[{1, xTRN[[i]], xTRN[[i]]^2, xTRN[[i]]^3,
 xTRN[[i]]^4,xTRN[[i]]^5, xTRN[[i]]^6},
 {i, Dimensions[dataVARIABLE][[1]]}];
```

(* the variable structure of the six polynomials which will be tested with both ctiteria are structured below *)

```
Xop[0] := {{1}}; Xop[1] := {{1}, {t}};
 Xop[2] := {{1}, {t}, {t^2}};
Xop[3] := {{1}, {t}, {t^2}, {t^3}};
Xop[4] := {{1}, {t}, {t^2}, {t^3}, {t^4}};
Xop[5] := {{1}, {t}, {t^2}, {t^3}, {t^4}, {t^5}};
Xop[6] := {{1}, {t}, {t^2}, {t^3}, {t^4}, {t^5}, {t^6}};
```

(* This expression should be maximised *)

```
EXPR := (Xop[i]\[Transpose].Inverse[(X[i]\[Transpose].
 X[i])].Xop[i])[[1]];
```

(* the function below finds the value of t inside the region [-1,1] that maximises EXPR *)

```
MAX[nn_] := (i = nn; NMaximize[{EXPR[[1]],
-1 <= t <= 1}, t] );
```

(* LP is the prediction criterion of the proposed method. It is the length of the tolerance region and it is evaluated for the t found before that maximises EXPR *)

```
LP := 2*tst[Dimensions[datax][[1]] - i]/
 Sqrt[Dimensions[datax][[1]] - i]
 ((INVS1p)^(1/2))*(RSSp)^(1/2);
bi:=Inverse[X[i]\[Transpose].X[i]].(X[i]\[Transpose].Y);
RSSp := Transpose[Y - X[i]. bi] . (Y - X[i]. );
```

(* this is the conventional RMS criterion for best fitting model *)

```
RMS := RSSp/(Dimensions[datax][[1]] - i - 1);
INVS1p := Inverse[1 - Xop[i]\[Transpose].
 Inverse[(X[i]\[Transpose].X[i] +
Xop[i].Xop[i]\[Transpose])]. Xop[i]];
```

(*This mathematica function is the one that the user of the programme only needs to use. He has to set as argument of this function the largest order of the polynomial to be tested. i.e. 6. This function evaluates and returns for each order of the polynomial the prediction criterion LP and the conventional criterion RMS. It also returns for each order of the polynomial the plot of the data together with the best fitting polynomial for prediction. Finally it plots also the Expression that is maximized for a certain t *)

```
BESTMODEL[q_] := (nn = q;Do[ i = k;
 Print[’’i=’’, i,’’  max’’,
 NMaximize[{EXPR[[1]], -1 <= t <= 1}, t] ];
gg = Plot[EXPR[[1]], {t, -1, 1}]; g1 = ListPlot[data];
g2= Plot[Xop[i]\[Transpose].bi, {t, -1, 1}];
Print[Xop[i]\[Transpose].bi];
asd := NMaximize[{EXPR[[1]], -1 <= t <= 1}, t] [[2]][[1]];
 Print[’’RMS=’’, RMS]; t = t /. asd; Print[’’LP=’’, LP];
 Print[Show[gg]]; Print[Show[g1, g2, PlotRange -> All]];
 t =., {k, 0, nn}]  )
```

References

1. Bent, F.: Design by Deception: The Politics of Megaproject Approval. Harvard Design Magazine. no. 22, Spring/Summer, pp. 50–59 (2005)
2. Bent, F., Bruzelius, N., Rothengatter, W.: Megaprojects and Risk: An Anatomy of Ambition, Cambridge University Press, Cambridge (2003)
3. Bent, F., Cowi.: Procedures for Dealing with Optimism Bias in Transport Planning: Guidance Document, London, UK Department for Transport (2004)
4. Bent, F., Skamris Holm, M.K., Buhl, S.L.: Underestimating costs in public works projects, error or lie? J. Am. Plan. Assoc. **68**(3), Summer, pp. 279–295 (2002)
5. Bent, F., Skamris Holm, M.K., Buhl, S.L.: What causes cost overrun in transport infrastructure projects? Transp. Rev. **24**(1), 3–18 (2004)
6. Bent, F., Skamris Holm, M.K., Buhl, S.L.: How (In)accurate are demand forecasts in public works projects? The case of transportation. J. Am. Plan. Assoc. **71**(2), 131–146 (2005)
7. Bent, F., Lovallo, D.: Delusion and Deception: Two Models for Explaining Executive Disaster (in progress)
8. Bent, F.: Curbing optimism bias and strategic misrepresentation in planning, reference class forecasting in practice. Eur. Plan. Stud. **16**, 3–21 (2008)
9. Gilovich, T., Griffin, D., Kahneman, D. (eds.) Heuristics and Biases: The Psychology of Intuitive Judgment. 19/32 Cambridge University Press (2002)
10. Gordon, P., Wilson, R.: The determinants of light-rail transit demand: an international cross-sectional comparison. Transp. Res. A **18A**(2), 135–140 (1984)
11. Kahneman, D.: New challenges to the rationality assumption. J. Inst. Theor. Econ. **150**, 18–36 (1994)
12. Kahneman, D., Lovallo, D.: Timid choices and bold forecasts, a cognitive perspective on risk taking. Manag. Sci. **39**, 17–31 (1993)
13. Kahneman, D., Tversky, A.: Prospect theory, an analysis of decisions under risk. Econometrica **47**, 313–327 (1979)

14. Kahneman, D., Tversky, A.: Intuitive prediction, biases and corrective procedures. In: Makridakis, S., Wheelwright, S.C. (ed.), Studies in the Management Sciences, Forecasting, vol. 12. Amsterdam, North Holland (1979)
15. Lovallo, D., Kahneman, D.: Delusions of success. how optimism undermines executives' decisions. Harv. Bus. Rev. pp. 56–63 (2003)
16. Ellerton, R.R.W., Kitsos, C.P., Rinco, S.: Choosing the optimal order of a response polynomical-structural approach with minimax criterion. Commun. Stat. Theory Methodol. **15**, 129–136 (1986)
17. Kitsos, C.P.: An algorithm for construct the best predictive model. In: Faulbaum, F. (ed.) Softstat 93, Advances in Statistical Software, pp. 535–539, Stuttgart, New York (1994)
18. Muller, C.H., Kitsos C.P.: Optimal design criteria based on tolerance regions. In: Bucchianico, A., Lauter, H., Wynn. H.P. (Eds.) mODa 7-Advances in Model-Oriented Design and Analysis, pp: 107–115, Physica-Verlag, Heidelberg (2004)
19. Maddala, G.: Introduction to Econometrics, 2nd edn. p. 663, Macmillan, New York (1992)
20. Stigler, S.M.: Gauss and the invention of least squares. Ann. Stat. **9**(3), 465–474 (1981)
21. Anderson, T.W.: The choice of the degree of a polynomial regression as a multiple decision problem. Ann. Math. Stat. **33**, 255–265 (1962)
22. Martins, J.P., Mendonca, S., Pestana, D.: Optimal and quasi-optimal designs. Revstat **6**, 279–307 (2008). Available via http://www.ine.pt/revstat/pdf/rs080304.pdf
23. Hocking, R.R.: The analysis of selection of variables in linear regression. Biometrics **32**, 1–49 (1976)
24. Wilks, S.S.: Mathematical Statistics. Wiley, New York (1962)
25. Guttman, I.: Construction of -content tolerance regions at confidence level for large samples for k-Variate normal distribution. Ann. Math. Stat. **41**, 376–400 (1970)
26. Boente, G., Farall, A.: Robust multivariate tolerance regions, influence function and Monte Carlo study. Technometrics **50**, 487–500 (2008)
27. Kendall, M.G., Stuart A.: The Advanced Theory of Statistics, vols. II, III. C. Griffin Ltd., London (1976)
28. Zarikas, V., Gikas, V., Kitsos, C.P.: Evaluation of the optimal design "Cosinor model" for enhancing the potential of robotic theodolite kinematic observation. Measurement **43**(10), 1416–1424 (2010)
29. Kitsos, C.P., Zarikas, V.: On the best predictive general linear model for data analysis, a tolerance region algorithm for prediction. J. Appl. Sci. **13**(4), 513–524 (2013)
30. Kaiser, M.J., Snyder, B.: Offshore wind capital cost estimation in the US outer continental shelf, a reference class approach. Marian. Policy **36**, 1112–1122 (2012)
31. Valente, V., Oliveira T. A.: Hierarchical linear models: review and applications. In: Proceedings of the 9th International Conference of Numerical Analysis and Applied Mathematics, September 19–25, Halkidiki, Greece, pp. 1549–1552 (2011)

Randomly Stopped *k*th Order Statistics

Sandra Mendonça, Dinis Pestana and M. Ivette Gomes

Abstract Randomly stopped order statistics when the stopping rule is generated by a basic count distribution are investigated. Unified expressions in terms of the subordinator are presented, extending results from geometrically thinned sequences. Using the results on limit stable distributions for max-geometric laws, and Smirnov's techniques to deal with limit laws of extreme order statistics, some results on stability of Panjer subordinated randomly stopped order statistics are discussed.

Keywords Randomly stopped order statistics · Panjer family · Basic count distributions · Order statistics · Geometric thinning

1 Introduction

Risk analysis and extreme value theory (EVT) walk naturally hand-in-hand. Let $\{X_i,\ i \in \mathbb{N}\}$, with $\mathbb{N}$ the set of positive natural numbers, be a sequence of independent and identically distributed (iid) random variables (rvs), replicas of X, an absolutely continuous rv, with a common cumulative distribution function (cdf) F_X. Let us further denote by $(X_{1:n} \leq \cdots \leq X_{n:n})$ the set of ascending order statistics associated

S. Mendonça (✉)
CCEE, Universidade da Madeira, 9000-390 Funchal, Portugal
e-mail: smfm@uma.pt

S. Mendonça
CEAUL, Universidade de Lisboa, Lisbon, Portugal

D. Pestana · M.I. Gomes
CEAUL and DEIO–FCUL, Universidade de Lisboa, Campo Grande,
1749-016 Lisbon, Portugal
e-mail: dinis.pestana@fc.ul.pt

M.I. Gomes
e-mail: ivette.gomes@fc.ul.pt

D. Pestana · M.I. Gomes
Instituto de Investigação Científica Bento da Rocha Cabral, Lisbon, Portugal

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_19

with a random sample $(X_1, \ldots, X_n)$. Classical EVT deals with the behaviour of the sequences of extreme upper and lower order statistics, and in particular with the sequence of maximum values $\{X_{n:n},\ n \in \mathbb{N}\}$. Suppose that it is possible to linearly normalise the sequence of maximum values so that we get a non-degenerate limiting rv. Then such a limiting rv has a max-stable cdf G given by

$$G(x) \equiv \begin{cases} G_1(x) = \exp\left(-x^{-\alpha}\right) I_{(0,+\infty)}(x), \\ G_2(x) = \exp\left[-(-x)^{\alpha}\right] I_{(-\infty,0)}(x) + I_{[0,+\infty)}(x), \\ G_3(x) = \exp\left[-\exp(-x)\right], \end{cases} \tag{1}$$

where $\alpha > 0$ (cf., e.g., [2]). We then say that F_X belongs to the max-domain of the attraction of G.

Rachev and Resnick (cf. [8]) considered the limiting behaviour of the maximum of the linearly normalised vector $(X_1, \ldots, X_N)$, with N a geometric rv. They have then shown that, under adequate conditions, the so-called geo-max stable limit laws $\mathscr{G}$ have a cdf related to the extreme value cdf G given in (1) through the expression

$$\mathscr{G}(x) = \frac{1}{1 - \ln G(x)}.$$

Therefore, the geo-max stable types cdfs are given by

$$\mathscr{G}(x) \equiv \begin{cases} \mathscr{G}_1(x) = \frac{1}{1+x^{-\alpha}} I_{(0,+\infty)}(x), \\ \mathscr{G}_2(x) = \frac{1}{1+(-x)^{\alpha}} I_{(-\infty,0)}(x) + I_{[0,+\infty)}(x), \\ \mathscr{G}_3(x) = \frac{1}{1+\exp(-x)}, \end{cases} \tag{2}$$

respectively known as the loglogistic, the backward loglogistic and the logistic distributions.

In this work inspired by the Rachev and Resnick theory on stable limits of randomly stopped maxima with geometric subordinator—also called geo-max stability (see [8]), we take this theory a step further, in a parallel way to classical EVT.

In Sect. 2 we briefly discuss randomly stopped order statistics, and in Sect. 3 we describe the basic count distributions, the Panjer distributions [6]—Poisson, binomial and negative binomial, that in a sense are the yardstick with unitary dispersion coefficient, underdispersed and over dispersed count variables—, and Sundt's [12] extension (logarithmic and extended negative binomial). Next, in Sect. 4, we present unified expressions for the distribution of randomly stopped order statistics when the stopping rule is generated by a Panjer class subordinator. We start Sect. 5 restating Rachev and Resnick results on limiting stable laws for geometrically thinned sequences of iid rvs. We then refer to e-Bay auctions, to claim that there are situations for which the mentioned order statistics of geometrically thinned sequences are important. Finally, in Sect. 5.3, inspired on Smirnov's work [10] on what he calls asymptotic results for variational series, we investigate stability results for order statistics of geometrically thinned sequences.

2 Randomly Stopped Order Statistics

Let N be a discrete rv with a support contained in the set of the natural numbers $\mathbb{N}_0$. Let us further assume that N is independent of any of the elements of the sequence of iid rvs, $\{X_i,\ i \in \mathbb{N}\}$.

Definition 1 Conditionally to the event $N \geq k \in \mathbb{N}$, the kth ascending and descending N-randomly stopped order statistics are respectively the conditional rvs $X_{k:N}|N \geq k$ and $X_{N-k+1:N}|N \geq k$.

Remark 1 Although redundant, it is useful to state that $X_{N:N}$ denotes the N-randomly stopped maximum and $X_{1:N}$ denotes the N-randomly stopped minimum.

The cdf of the kth N-randomly stopped order statistic is given by

$$\begin{aligned} F_{X_{k:N}|N\geq k}(x) &= \frac{\mathbb{P}\left[X_{k:N} \leq x, N \geq k\right]}{\mathbb{P}\left[N \geq k\right]} \\ &= \frac{1}{\mathbb{P}\left[N \geq k\right]} \sum_{j=k}^{+\infty} \mathbb{P}\left[X_{k:N} \leq x, N \geq k | N = j\right] \mathbb{P}\left[N = j\right] \\ &= \frac{1}{\mathbb{P}\left[N \geq k\right]} \sum_{j=k}^{+\infty} F_{X_{k:j}}(x)\, \mathbb{P}\left[N = j\right] \end{aligned} \tag{3}$$

and, although redundant, as it will be useful,

$$F_{X_{N-k+1:N}|N\geq k}(x) = \frac{1}{\mathbb{P}\left[N \geq k\right]} \sum_{j=k}^{+\infty} F_{X_{j-k+1:j}}(x)\, \mathbb{P}\left[N = j\right]. \tag{4}$$

The relation $X_{k:n} \overset{d}{=} -\left[(-X)_{n-k+1:n}\right]$ of order statistics of independent rvs is inherited by N-randomly stopped order statistics:

Proposition 1 *Given N, a discrete rv with a support contained in* $\mathbb{N}_0$, $\{X_i,\ i \in \mathbb{N}\}$, *a sequence of iid rvs, independent of N and equal in distribution to X, an absolutely continuous rv, and* $k \in \mathbb{N}$, *we have*

$$X_{k:N}|N \geq k \overset{d}{=} -\left[(-X)_{N-k+1:N}\,|N \geq k\right]. \tag{5}$$

Proof Using the equalities (3) and (4), and the well known expression

$$1 - F_{X_{k:n}}(x) = \sum_{i=0}^{k-1} \binom{n}{i} [F_X(x)]^i [1 - F_X(x)]^{n-i}$$

$\left(\text{we can also write } F_{X_{n-k+1:n}}(x) = \sum_{i=0}^{k-1} \binom{n}{i} [1 - F_X(x)]^i [F_X(x)]^{n-i}\right)$, we obtain

$$\begin{aligned} F_{X_{k:N}|N \geq k}(x) &= \sum_{j=k}^{+\infty} \left\{ 1 - \sum_{i=0}^{k-1} \binom{j}{i} [F_X(x)]^i [1 - F_X(x)]^{j-i} \right\} \frac{\mathbb{P}[N = j]}{\mathbb{P}[N \geq k]} \\ &= 1 - \sum_{j=k}^{+\infty} \sum_{i=0}^{k-1} \binom{j}{i} [F_X(x)]^i [1 - F_X(x)]^{j-i} \frac{\mathbb{P}[N = j]}{\mathbb{P}[N \geq k]} \end{aligned}$$

and

$$\begin{aligned} F_{X_{N-k+1:N}|N \geq k}(x) &= \sum_{j=k}^{+\infty} \sum_{i=0}^{k-1} \binom{j}{i} [1 - F_X(x)]^i [F_X(x)]^{j-i} \frac{\mathbb{P}[N = j]}{\mathbb{P}[N \geq k]} \qquad (6) \\ &= \sum_{j=k}^{+\infty} \sum_{i=0}^{k-1} \binom{j}{i} \left[F_{-X}(-x)\right]^i \left[1 - F_{-X}(-x)\right]^{j-i} \frac{\mathbb{P}[N = j]}{\mathbb{P}[N \geq k]} \\ &= 1 - F_{(-X)_{k:N|N \geq k}}(-x) = F_{-\left[(-X)_{k:N|N \geq k}\right]}(x), \end{aligned}$$

which lead us to (5). □

Using the equality (6), we easily prove the following theorem.

Theorem 1 *Given N, a discrete rv with support contained in $\mathbb{N}_0$, $\{X_i,\ i \in \mathbb{N}\}$, a sequence of iid rvs, independent of N and equal in distribution to X, an absolutely continuous rv, and $k \in \mathbb{N}$, we have*

$$\begin{aligned} F_{X_{N-k+1:N}|N \geq k}(x) \ &= 1 - \sum_{i=k}^{+\infty} [1 - F_X(x)]^i \sum_{j=0}^{+\infty} \frac{\mathbb{P}[N = j + i]}{\mathbb{P}[N \geq k]} \\ &\quad \times \binom{j+i}{i} [F_X(x)]^j . \qquad (7) \end{aligned}$$

Proof In fact,

$$\begin{aligned}
F_{X_{N-k+1:N}|N\geq k}(x) &= \sum_{j=k}^{+\infty} \frac{\mathbb{P}[N=j]}{\mathbb{P}[N\geq k]} \sum_{i=0}^{k-1} \binom{j}{i} [1-F_X(x)]^i [F_X(x)]^{j-i} \\
&= \sum_{j=k}^{+\infty} \frac{\mathbb{P}[N=j]}{\mathbb{P}[N\geq k]} \left\{ 1 - \sum_{i=k}^{j} \binom{j}{i} [1-F_X(x)]^i [F_X(x)]^{j-i} \right\} \\
&= 1 - \sum_{i=k}^{+\infty} \sum_{j=0}^{+\infty} \frac{\mathbb{P}[N=j+i]}{\mathbb{P}[N\geq k]} \binom{j+i}{i} [1-F_X(x)]^i [F_X(x)]^j .
\end{aligned}$$

□

3 Count Distributions

A discrete distribution with support $\mathbb{N}_0$ or $\mathbb{N}$, or any initial subsection of either $\mathbb{N}_0$ or $\mathbb{N}$, is a count distribution. Adequate count distributions are the appropriate choice for use as subordinators of randomly stopped order statistics.

Three of the most frequently used discrete models have an interesting property: a simple recursive relation for the successive probability atoms that has been several times rediscovered in different contexts (for instance, Katz [4] used it to organise a family of discrete models in the same spirit of the Pearson family), cf. [7]. An important breakthrough has been Panjer's [6] idea of using the above mentioned recursive expression to iteratively compute or approximate densities of the aggregate claim in risk theory, cf. [9].

Definition 2 We say that a discrete rv $N_{a,b}$ belongs to the Panjer family of order 0 (0-Panjer family) if its probability mass function (pmf) satisfies the relation

$$p_{n+1} = \mathbb{P}\left[N_{a,b} = n+1\right] = \left(a + \frac{b}{n+1}\right) p_n, \ \ n \in \mathbb{N}_0. \tag{8}$$

Table 1 identifies the three nondegenerate members of 0-Panjer family (cf. [6]).

Hence the 0-Panjer distributions are exactly the discrete Morris natural exponential families whose variance is at most a quadratic function of the mean value (cf. [5]).

The NegativeBinomial $(1, p)$, usually referred to as the Geometric (p) distribution, and the Poisson(p) are the most commonly used subordinators of randomly stopped variables, essentially due to the simplicity of their Panjer set of coefficients,

Table 1 Members of the 0-Panjer family

Distribution	p_n	a	b
Poisson (p), $(p > 0)$	$\exp(-p)\,\frac{p^n}{n!},\ n \in \mathbb{N}_0$	0	p
Binomial (m, p), $(m \in \mathbb{N},\ p \in (0,1))$	$\begin{cases} \binom{m}{n}\, p^n\,(1-p)^{m-n}, & n = 0, ..., m \\ 0, & n = m+1, ... \end{cases}$	$\frac{-p}{1-p}$	$\frac{m+1}{1-p}\,p$
NegativeBinomial (r, p), $(r \in \mathbb{N},\ p \in (0,1))$	$\binom{n+r-1}{n}\, p^n\,(1-p)^r,\ n \in \mathbb{N}_0$	p	$(r-1)\,p$

$(0, p)$ and $(p, 0)$, respectively. Analysis of the Panjer recursion by Pestana and Velosa [7] exhibits how cumbersome the expressions for the general Panjer coefficients (a, b) are as opposed to the elegant expressions associated with the coefficient pairs $(0, p)$ and $(p, 0)$ of the Poisson and of the geometric subordinator cases.

Remark 2 A non-recursive expression for the 0-Panjer pmf would be:

$$\begin{aligned} p_{n+1} &= \mathbb{P}\left[N_{a,b} = n+1\right] = \left(a + \frac{b}{n+1}\right) p_n \\ &= \left(a + \frac{b}{n+1}\right)\left(a + \frac{b}{n}\right)\left(a + \frac{b}{n-1}\right)\cdots\left(a + \frac{b}{n-k}\right)\cdots(a+b)\,p_0 \\ &= \frac{p_0}{(n+1)!}\prod_{k=1}^{n+1}(ka+b)\,,\ \ n \in \mathbb{N}_0. \end{aligned}$$

In short,

$$p_n = \frac{p_0}{n!}\prod_{k=1}^{n}(ka+b)\,, \quad n \in \mathbb{N}.$$

For $a = 0$ (in the Poisson case),

$$p_n = \frac{p_0}{n!} b^n, \quad n \in \mathbb{N}.$$

When in expression (8) we use instead of the condition $n \in \mathbb{N}_0$ the condition $n \geq L$ (and take $p_n = 0$, for $n < L$) we obtain an extension of the Panjer family that we name the L-Panjer distribution (cf. [11]).

For $L = 1$, two new nondegenerate Panjer distributions do exist (cf. Sundt and Jewell [12]): the Logarithmic (p), with pmf defined for $n \in \mathbb{N}$ by

$$p_n = -\frac{1}{\ln(1-p)} \frac{p^n}{n}, \quad p \in (0,1)\,, \tag{9}$$

and Engen's [1] extended negative binomial (ENB) distribution with pmf given by

$$p_n = \frac{\alpha\, \Gamma(n+\alpha)}{n!\, \Gamma(1+\alpha)} \frac{p^n (1-p)^{\alpha}}{1-(1-p)^{\alpha}}, \tag{10}$$

for $\alpha \in (-1, 0)$, $p \in (0, 1]$ and $n \in \mathbb{N}$ (cf. also [13]).

The ENB distribution is very cumbersome but useful in some ecological and population dynamics models. Observe that ENB distribution has Panjer coefficients $(p, -p(1-\alpha))$ and that letting $\alpha \to 0$ we obtain the Logarithmic(p) distribution, that has Panjer coefficients $(p, -p)$.

Although the above two new Panjer families (logarithmic and ENB) are not left truncated 0-Panjer distributions, for $L \geq 2$, Hess et al. [3] have established that any L-Panjer distribution is the left endpoint truncation of an $(L-1)$-Panjer distribution. For that reason, they called the binomial, the Poisson, the negative binomial, the logarithmic and the ENB distributions the basic count models.

4 Stopped Order Statistics with Panjer Subordinators

Randomly stopped order statistics with Panjer subordinators do have intrinsic interest, although results are cumbersome except for the Poisson, the geometric and the logarithmic cases. The expressions of their cdfs have however some similarities.

4.1 Stopped Order Statistics with 0*-Panjer Subordinator*

Theorem 2 *If $N(p)$ is a member of* 0*-Panjer family then*

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x) = 1 - \frac{\mathbb{P}\left[N(p^*) \geq k\right]}{\mathbb{P}\left[N(p) \geq k\right]},$$

with

$$p^* = \begin{cases} p\left[1 - F_X(x)\right], & \textit{if } N(p) \textit{ is Poisson or binomial distributed,} \\ p\left[1 - F_X(x)\right] / \left[1 - pF_X(x)\right], & \textit{if } N(p) \textit{ is negative binomial distributed.} \end{cases}$$

Proof We will show the result separately for each family distribution of the 0-Panjer family.

1. Let $N(p)$ be a rv with Poisson distribution with mean value $p > 0$, i.e.,

$$\mathbb{P}\left[N(p) = i\right] = \exp(-p)\frac{p^i}{i!}, \text{ for } i \in \mathbb{N}_0.$$

From (7) we know that:

$$\begin{aligned} &F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x) \\ &= 1 - \frac{1}{\mathbb{P}\left[N(p) \geq k\right]} \sum_{i=k}^{+\infty} \frac{\left[1 - F_X(x)\right]^i}{i!} \sum_{j=0}^{+\infty} \exp(-p)\frac{p^{j+i}}{(j+i)!}\frac{(j+i)!}{j!} \\ &\quad \times \left[F_X(x)\right]^j \\ &= 1 - \frac{\exp(-p)}{\mathbb{P}\left[N(p) \geq k\right]} \sum_{i=k}^{+\infty} \frac{\left[1 - F_X(x)\right]^i p^i}{i!} \sum_{j=0}^{+\infty} \frac{1}{j!}\left[pF_X(x)\right]^j \\ &= 1 - \frac{1}{\mathbb{P}\left[N(p) \geq k\right]} \sum_{i=k}^{+\infty} \exp\left\{-p\left[1 - F_X(x)\right]\right\} \frac{\left\{p\left[1 - F_X(x)\right]\right\}^i}{i!} \\ &= 1 - \frac{\mathbb{P}\left[N\left(p\left[1 - F_X(x)\right]\right) \geq k\right]}{\mathbb{P}\left[N(p) \geq k\right]}. \end{aligned}$$

2. Let $N(p) \equiv N(m, p)$ be a rv with binomial distribution with parameters $p \in (0, 1)$ and $m \in \mathbb{N}$, i.e.,

$$\mathbb{P}\left[N(p) = k\right] = \binom{m}{k} p^k (1 - p)^{m-k} \mathbf{I}_{\{0,1,\ldots,m\}}(k).$$

Let $1 \leq k \leq m$. Again from (7), we know that:

$$\begin{aligned} &F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x) \\ &= 1 - \frac{1}{\mathbb{P}\left[N(p) \geq k\right]} \\ &\quad \times \sum_{i=k}^{m} \frac{\left[1 - F_X(x)\right]^i}{i!} \sum_{j=0}^{m-i} \binom{m}{j+i} p^{j+i} (1 - p)^{m-(j+i)} \frac{(j+i)!}{j!} \left[F_X(x)\right]^j \end{aligned}$$

$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]}$$
$$\times \sum_{i=k}^{m} \frac{\{p[1-F_X(x)]\}^i}{i!} \frac{m!}{(m-i)!} (1-p)^{m-i} \sum_{j=0}^{m-i} \binom{m-i}{j} \left[\frac{pF_X(x)}{1-p}\right]^j$$
$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]}$$
$$\times \sum_{i=k}^{m} \{p[1-F_X(x)]\}^i \binom{m}{i} (1-p)^{m-i} \left(1 + \frac{pF_X(x)}{1-p}\right)^{m-i} .$$
$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \sum_{i=k}^{m} \binom{m}{i} \{p[1-F_X(x)]\}^i (1 - p[1-F_X(x)])^{m-i}$$
$$= 1 - \frac{\mathbb{P}[N(p[1-F_X(x)]) \geq k]}{\mathbb{P}[N(p) \geq k]}.$$

3. Let $N(p) \equiv N(r, p)$ be a rv with negative binomial distribution with parameters $p \in (0, 1)$ and $r \in \mathbb{N}$, i.e.,

$$\mathbb{P}[N(p) = k] = \binom{k+r-1}{k} p^k (1-p)^r \mathbf{I}_{\mathbb{N}_0}(k) .$$

Let $k \in \mathbb{N}$. From (7) we know that:

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)$$
$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]}$$
$$\times \sum_{i=k}^{+\infty} \frac{[1-F_X(x)]^i}{i!} \sum_{j=0}^{+\infty} \binom{j+i+r-1}{j+i} p^{j+i} (1-p)^r \frac{(j+i)!}{j!} [F_X(x)]^j$$
$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \sum_{i=k}^{+\infty} \frac{(1-p)^r \{p[1-F_X(x)]\}^i}{i!(r-1)!}$$
$$\times \sum_{j=0}^{+\infty} \frac{(j+i+r-1)!}{j!} [pF_X(x)]^j .$$

Noting that, for $|y| < 1$, $\frac{(i-1)!}{(1-y)^i} = \sum_{j=0}^{+\infty} \frac{(j+i-1)!}{j!} y^j$, we obtain

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)$$
$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \sum_{i=k}^{+\infty} \frac{(1-p)^r \{p[1-F_X(x)]\}^i}{i!(r-1)!} \frac{(i+r-1)!}{[1-pF_X(x)]^{i+r}}$$

$$= 1 - \frac{1}{\mathbb{P}\left[N\left(p\right) \geq k\right]} \sum_{i=k}^{+\infty} \binom{i+r-1}{i} \left\{ \frac{p\left[1 - F_X\left(x\right)\right]}{1 - pF_X\left(x\right)} \right\}^i \left(\frac{1-p}{1 - pF_X\left(x\right)} \right)^r$$

$$= 1 - \frac{1}{\mathbb{P}\left[N\left(p\right) \geq k\right]} \sum_{i=k}^{+\infty} \binom{i+r-1}{i} \left\{ \frac{p\left[1 - F_X\left(x\right)\right]}{1 - pF_X\left(x\right)} \right\}^i$$
$$\times \left(1 - \frac{p\left[1 - F_X\left(x\right)\right]}{1 - pF_X\left(x\right)} \right)^r,$$

i.e. the result in the theorem follows. □

4.2 *Stopped Order Statistics with Genuinely* 1*-Panjer Subordinator*

The Panjer families of different orders form an increasing chain. Besides the three nondegenerate distributions that belong to the 0-Panjer family, the 1-Panjer family has as nondegenerate members the Logarithmic (p) distribution defined for $n \in \mathbb{N}$ by expression (9) and the ENB distribution (cf. [13]), defined by expression (10) for $\alpha \in (-1, 0)$, $p \in (0, 1]$ and $n \in \mathbb{N}$.

Theorem 3 *If $N\left(p\right)$ is logarithmic distributed with pmf given by (9) (with $p \in (0, 1)$ and $n \in \mathbb{N}$), then*

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}\left(x\right) = 1 - \frac{\mathbb{P}\left[N\left(p^*\right) \geq k\right]}{\mathbb{P}\left[N\left(p\right) \geq k\right]} \frac{\mathbb{P}\left[N\left(p\right) = 1\right]}{\mathbb{P}\left[N\left(p^*\right) = 1\right]} \frac{p^*}{p}$$

with

$$p^* = \frac{p\left[1 - F_X\left(x\right)\right]}{1 - pF_X\left(x\right)}.$$

Proof Let $p \in (0, 1)$ and $n, k \in \mathbb{N}$. Consider $N(p)$, a rv with logarithmic distribution, i.e. pmf given by

$$p_n = -\frac{1}{\ln\left(1-p\right)} \frac{p^n}{n}. \tag{11}$$

From the equality (7),

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}\left(x\right)$$
$$= 1 - \frac{1}{\mathbb{P}\left[N\left(p\right) \geq k\right]} \sum_{i=k}^{+\infty} \frac{\left[1 - F_X\left(x\right)\right]^i}{i!} \sum_{j=0}^{+\infty} \frac{-1}{\ln\left(1-p\right)} \frac{p^{j+i}}{j+i} \frac{\left(j+i\right)!}{j!} \left[F_X\left(x\right)\right]^j$$

$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \sum_{i=k}^{+\infty} \frac{\{p[1 - F_X(x)]\}^i}{i!} \frac{-1}{\ln(1-p)}$$

$$\times \sum_{j=0}^{+\infty} \frac{(j+i-1)!}{j!} [pF_X(x)]^j$$

$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \sum_{i=k}^{+\infty} \frac{\{p[1 - F_X(x)]\}^i}{i!} \frac{-1}{\ln(1-p)} \frac{(i-1)!}{[1 - pF_X(x)]^i}.$$

Taking $p^* = \frac{p[1-F_X(x)]}{1-pF_X(x)}$ we obtain

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)$$

$$= 1 - \frac{1}{\mathbb{P}[N(p) \geq k]} \frac{\ln(1-p^*)}{\ln(1-p)} \sum_{i=k}^{+\infty} \frac{-1}{\ln(1-p^*)} \frac{(p^*)^i}{i}$$

$$= 1 - \frac{\mathbb{P}\left[N(p^*) \geq k\right]}{\mathbb{P}[N(p) \geq k]} \frac{\ln(1-p^*)}{\ln(1-p)}.$$

Since

$$p_1 = -\frac{p}{\ln(1-p)} \Rightarrow \ln(1-p) = \frac{-p}{\mathbb{P}[N(p) = 1]},$$

the result in the theorem follows. □

Theorem 4 *If $N(p)$ has an ENB distribution with pmf given by (10) with $\alpha \in (-1, 0)$, $p \in (0, 1]$ and $n \in \mathbb{N}$, then*

$$F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x) = 1 - \frac{\mathbb{P}\left[N(p^*) \geq k\right]}{\mathbb{P}[N(p) \geq k]} \left(\frac{1-p^*}{1-p}\right)^{\alpha} \frac{\mathbb{P}[N(p) = 1]}{\mathbb{P}[N(p^*) = 1]} \frac{p^*}{p} \tag{12}$$

with

$$p^* = \frac{p[1 - F_X(x)]}{1 - pF_X(x)}.$$

Proof Let $\alpha \in (-1, 0)$, $p \in (0, 1]$ and $n, k \in \mathbb{N}$. Consider $N(p)$ a rv with an ENB distribution with pmf given by

$$p_n = \frac{\alpha\,\Gamma(n+\alpha)}{n!\,\Gamma(1+\alpha)} \frac{p^n (1-p)^{\alpha}}{1 - (1-p)^{\alpha}}.$$

Again using (7), we obtain

$$\begin{aligned}
&F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)\\
&=1-\sum_{i=k}^{+\infty}\frac{[1-F_X(x)]^i}{\mathbb{P}[N(p)\geq k]\,i!}\\
&\quad\times\sum_{j=0}^{+\infty}\frac{\alpha\Gamma(j+i+\alpha)}{(j+i)!\Gamma(1+\alpha)}\frac{p^{j+i}(1-p)^\alpha}{1-(1-p)^\alpha}\frac{(j+i)!}{j!}[F_X(x)]^j\\
&=1-\frac{1}{\mathbb{P}[N(p)\geq k]}\sum_{i=k}^{+\infty}\frac{\{p[1-F_X(x)]\}^i}{i!\Gamma(1+\alpha)}\frac{\alpha(1-p)^\alpha}{1-(1-p)^\alpha}\\
&\quad\times\sum_{j=0}^{+\infty}\Gamma(j+i+\alpha)\frac{1}{j!}[pF_X(x)]^j\,.
\end{aligned}$$

Since

$$\sum_{j=0}^{+\infty}\Gamma(j+i+\alpha)\frac{1}{j!}[pF_X(x)]^j=\frac{\Gamma(\alpha+i)}{[1-pF_X(x)]^{\alpha+i}}$$

we obtain

$$\begin{aligned}
&F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)\\
&=1-\frac{1}{\mathbb{P}[N(p)\geq k]}\sum_{i=k}^{+\infty}\frac{\{p[1-F_X(x)]\}^i}{i!\Gamma(1+\alpha)}\frac{\alpha(1-p)^\alpha}{1-(1-p)^\alpha}\frac{\Gamma(\alpha+i)}{[1-pF_X(x)]^{\alpha+i}}\\
&=1-\frac{1}{\mathbb{P}[N(p)\geq k]}\frac{(1-p)^\alpha}{1-(1-p)^\alpha}\frac{1}{[1-pF_X(x)]^\alpha}\frac{1-(1-p^*)^\alpha}{(1-p^*)^\alpha}\\
&\times\sum_{i=k}^{+\infty}\frac{\alpha\Gamma(\alpha+i)}{i!\Gamma(1+\alpha)}\left(p^*\right)^i\frac{(1-p^*)^\alpha}{1-(1-p^*)^\alpha}
\end{aligned}$$

with $p^*=\frac{p[1-F_X(x)]}{1-pF_X(x)}=\frac{p-1+1-pF_X(x)}{1-pF_X(x)}=1-\frac{1-p}{1-pF_X(x)}$. Hence,

$$\begin{aligned}
&F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}(x)\\
&=1-\frac{\mathbb{P}\left[N(p^*)\geq k\right]}{\mathbb{P}[N(p)\geq k]}\frac{1}{[1-pF_X(x)]^\alpha}\frac{(1-p)^\alpha}{1-(1-p)^\alpha}\frac{1-(1-p^*)^\alpha}{(1-p^*)^\alpha}.
\end{aligned}$$

But

$$p_1=\frac{\alpha p(1-p)^\alpha}{1-(1-p)^\alpha}\Rightarrow\frac{(1-p)^\alpha}{1-(1-p)^\alpha}=\frac{\mathbb{P}[N(p)=1]}{\alpha p}$$

and hence

$$\begin{aligned}
&F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}\,(x)\\
&= 1-\frac{\mathbb{P}\left[N\left(p^{*}\right)\geq k\right]}{\mathbb{P}\left[N\left(p\right)\geq k\right]}\frac{1}{\left[1-pF_{X}\left(x\right)\right]^{\alpha}}\frac{\mathbb{P}\left[N\left(p\right)=1\right]}{\alpha p}\frac{\alpha p^{*}}{\mathbb{P}\left[N\left(p^{*}\right)=1\right]}\\
&= 1-\frac{\mathbb{P}\left[N\left(p^{*}\right)\geq k\right]}{\mathbb{P}\left[N\left(p\right)\geq k\right]}\frac{1}{\left[1-pF_{X}\left(x\right)\right]^{\alpha}}\frac{\mathbb{P}\left[N\left(p\right)=1\right]}{\mathbb{P}\left[N\left(p^{*}\right)=1\right]}\frac{p^{*}}{p},
\end{aligned}$$

and (12) follows. □

5 Order Statistics of Geometrically Thinned Sequences

Poisson thinned sequences and geometrically thinned sequences have specially nice properties. In this section we will study the particular case of the geometrically random stopped order statistics as defined by Rachev and Resnick in (cf. [8]).

5.1 *Stable Limit Laws for Maxima of Geometrically Thinned Sequences*

For $p \in (0, 1)$, define $q := 1 - p$ and let $N(p)$ be a geometrically distributed rv such that

$$\mathbb{P}\left[N(p) = k\right] = pq^{k-1},\ k \in \mathbb{N}. \tag{13}$$

Notice that this rv has the same distribution as the rv $N^{*} + 1$ where N^{*} has a NegativeBinomial$(1, 1 - p)$ distribution, as defined in Table 1.

Let $\{X_i,\ i \in \mathbb{N}\}$ be a sequence of iid rvs, replicas of X, an absolutely continuous rv, all independent of N, with a common cdf F_X. Suppose that X belongs to the max-domain of the attraction of one of the possible types of max-stable distributions with cdf given in (1), generally denoted $G(x)$. The limit laws $\mathscr{G}$, when $p \to 0$, of the maximum of the vector $(X_1, ..., X_N)$, properly normalised by functions of p, were described by Rachev and Resnick (cf. [8]). They are related to the extreme value cdf given in (1) through the expression $\mathscr{G}\left(x\right) = 1/\left(1 - \ln G\left(x\right)\right)$, and they have been provided in (2).

5.2 To Whom are the Geometrically Thinned Second Maxima Important?

In auctions such as the e-Bay the final price is a fixed increment of the second maximum biding. For interesting, fairly rare and expensive items many bidders wait until the very end before making their bid (which is called “sniping”). Since they wait till the very last moment, they do not know the others bids and are therefore bidding independently of each other. On the other hand, internet speed and communication jams preclude some bids to arrive before the auction is closed, and we shall accept that this thinning is geometric, and that for the same valuable item bids are iid.

Some sellers auction similar items at different times, and sometimes under different identities, for instance we have seen, dozens of times, auctions of a rocking mother and child or of the king and queen, by Henry Moore, without a certificate of authenticity (COA), and numbered using some cypher from 1 to 9 (Moore only cast less than 10 of those miniatures). It is of course expected that the sellers will continue to put similar items on sale on the future, and that (s)he wants to maximise the selling price.

The seller can force bids to be greater than a “reserve price”. Hence the problem is to determine the distribution of the geometrically thinned maximum, given that several previous thinned second maxima have been recorded. The seller can then choose an appropriate threshold as starting bid or as a reserve price, depending obviously on his greed and on his need to make money, conflicting issues in determining the probability of selling he wants to attain.

For instance, the same seller currently advertises at e-Bay Giacometti’s Walking Man with an auction starting price US$600 (1 day left, 0 bids), or “buy it now” for US$780. Similar items have been advertised by the same seller, effective published selling prices ranging from US$600—observe that this happens whenever there is only one bid, that can be considerably higher than the final selling price—to US$679, and therefore we estimate that the fair price the seller can expect is approximately US$675, but that the maximum bid is greater than US$679. Sales also occurred when the seller used the option “buy it now or best offer”, but for those there is only the information “Best Offer accepted” without disclosing the selling price; with the very incomplete disclosure of best offers or biddings, we cannot estimate the maximum, although we believe it is less than US$780. The seller however has complete information, and surely uses it to estimate the maximum price he can get and therefore to determine the starting bid and the immediate selling price he accepts. The fact that occasionally he uses the option “or best offer” seems to indicate that the seller guesses that US$780 is an optimistic estimate of the fair price for that item.

5.3 Limit Theorems for Geometrically Randomly Stopped Order Statistics

The randomly stopped order statistics inherit some of the simple properties of order statistics in the traditional iid scheme. The extensions of Smirnov [10] asymptotic limit distributions for the geometrically randomly stopped order statistics that follow exhibit some of those similarities.

Consider the rv $N(p)$ with pmf given by (13). In this section we will study the limit behaviour (when $p \downarrow 0$) of the distributions of $X_{N(p)-k+1:N(p)|N(p)\geq k}$ and of $X_{k:N(p)|N(p)\geq k}$, properly normalised. Let us start with $X_{N(p)-k+1:N(p)|N(p)\geq k}$ and call $\mathscr{G}^{(k)}$ the associated limit distribution. From expression (6) we know that:

$$F_{X_{N(p)-k+1:N(p)|N(p)\geq k}}(x) = \sum_{j=k}^{+\infty}\sum_{i=0}^{k-1}\binom{j}{i}\left[1-F_X(x)\right]^i\left[F_X(x)\right]^{j-i}\frac{\mathbb{P}\left[N(p)=j\right]}{\mathbb{P}\left[N(p)\geq k\right]}.$$

Replacing the expression of the pmf of $N(p)$ given by (13) and noting that $1=\sum\limits_{i=0}^{j}\binom{j}{i}\left[1-F_X(x)\right]^i\left[F_X(x)\right]^{j-i}$ we can write

$$F_{X_{N(p)-k+1:N(p)|N(p)\geq k}}(x) = p\sum_{j=k}^{+\infty}\frac{(1-p)^{j-1}}{(1-p)^{k-1}} \times\left[1-\sum_{i=k}^{j}\binom{j}{i}\left[1-F_X(x)\right]^i\left[F_X(x)\right]^{j-i}\right].$$

After some simplifications and the exchange of the order of the two sums we obtain

$$F_{X_{N(p)-k+1:N(p)|N(p)\geq k}}(x) = 1-\frac{p}{(1-p)^{k-1}}\sum_{i=k}^{+\infty}\left[1-F_X(x)\right]^i(1-p)^{i-1}\sum_{j=0}^{+\infty}\binom{j+i}{i}\left[(1-p)F_X(x)\right]^j.$$

Finally, using the equality $\sum_{j=0}^{+\infty}\left[\binom{j+i}{i}y^j\right]=\frac{1}{(1-y)^{i+1}}$, $|y|<1$, we obtain

$$F_{X_{N(p)-k+1:N(p)|N(p)\geq k}}(x) = 1-\left[1-\frac{pF_X(x)}{1-(1-p)F_X(x)}\right]^k. \tag{14}$$

Suppose that the rv $X_{N(p):N(p)}$, conveniently normalised, weakly converges to a nondegenerate rv with cdf $\mathscr{G}$, i.e., that there exists $\alpha=\alpha(p)$ and $\beta=\beta(p)>0$, such that

$$\lim_{p\downarrow 0} F_{\frac{X_{N(p):N(p)}-\alpha(p)}{\beta(p)}}\left(x\right) = \lim_{p\downarrow 0} F_{X_{N(p):N(p)}}\left(\beta\left(p\right)x + \alpha\left(p\right)\right) = \mathscr{G}\left(x\right).$$

Let $x_p = \beta\left(p\right)x + \alpha\left(p\right)$. Hence, from (14) with $k = 1$, we conclude that

$$\begin{aligned}\lim_{p\downarrow 0} F_{X_{N(p):N(p)}}\left(x_p\right) &= \lim_{p\downarrow 0}\left\{1 - \left[1 - \frac{p\,F_X\left(x_p\right)}{1-\left(1-p\right)F_X\left(x_p\right)}\right]\right\} \\ &= \lim_{p\downarrow 0}\frac{p\,F_X\left(x_p\right)}{1-\left(1-p\right)F_X\left(x_p\right)} = \mathscr{G}\left(x\right).\end{aligned}$$

For general k we have then

$$\begin{aligned}\lim_{p\downarrow 0} F_{X_{N(p)-k+1:N(p)}|N(p)\geq k}\left(x_p\right) &= \lim_{p\downarrow 0}\left\{1 - \left[1 - \frac{p\,F_X\left(x_p\right)}{1-\left(1-p\right)F_X\left(x\right)}\right]^k\right\} \\ &= 1 - \left[1-\mathscr{G}\left(x\right)\right]^k,\end{aligned}$$

proving the following theorem.

Theorem 5 *Let* $p \in (0, 1)$*, define* $q := 1 - p$ *and consider the rv* $N\left(p\right)$ *with pmf given by (13). If* $X_{N(p):N(p)}$ *properly normalised weakly converges to a nondegenerate rv with cdf* $\mathscr{G}$*, i.e., if there exists* $\alpha\left(p\right)$ *and* $\beta\left(p\right) > 0$*, such that*

$$\lim_{p\downarrow 0} F_{X_{N(p):N(p)}}\left(\beta\left(p\right)x + \alpha\left(p\right)\right) = \mathscr{G}\left(x\right),$$

then the kth geometric maximum, $X_{N(p)-k+1:N(p)}|N\left(p\right) \geq k$*, properly normalised with the same functions, i.e., the rv*

$$\frac{X_{N(p)-k+1:N(p)|N(p)\geq k} - \alpha\left(p\right)}{\beta\left(p\right)},$$

weakly converges to a nondegenerate rv with cdf $\mathscr{G}^{(k)}$ *given by*

$$\mathscr{G}^{(k)}\left(x\right) = 1 - \left[1-\mathscr{G}\left(x\right)\right]^k.$$

For the kth geometric order statistic a similar reasoning leads to the equality

$$F_{X_{k:N(p)}|N(p)\geq k}(x) = \left[\frac{F_X\left(x\right)}{p+\left(1-p\right)F_X\left(x\right)}\right]^k.$$

Taking $x_p = \beta(p)\,x + \alpha(p)$, suppose that the limit $\mathscr{L}(x)$ of $F_{X_{1:N(p)}}\left(x_p\right)$, when $p \downarrow 0$, exists:

$$\lim_{p\downarrow 0} F_{X_{1:N(p)}}\left(x_p\right) = \mathscr{L}(x) = \lim_{p\downarrow 0} \frac{F_X\left(x_p\right)}{p + (1-p)\,F_X\left(x_p\right)}.$$

Then

$$\lim_{p\downarrow 0} F_{X_{k:N(p)}|N(p)\geq k}\left(x_p\right) = \lim_{p\downarrow 0}\left[\frac{F_X\left(x_p\right)}{p + (1-p)\,F_X\left(x_p\right)}\right]^k = \mathscr{L}^k(x),$$

which leads to the following theorem.

Theorem 6 *Let $p \in (0,1)$, define $q := 1 - p$ and consider the rv $N(p)$ with pmf given by (13). If $X_{1:N(p)}$ properly normalised weakly converges to a nondegenerate rv with cdf $\mathscr{L}$, i.e., if there exists $\alpha(p)$ and $\beta(p) > 0$, such that*

$$\lim_{p\downarrow 0} F_{X_{1:N(p)}}\left(\beta(p)\,x + \alpha(p)\right) = \mathscr{L}(x),$$

then the kth geometric order statistic, $X_{k:N(p)}|N(p) \geq k$, similarly normalised, i.e., the rv

$$\frac{X_{k:N(p)|N(p)\geq k} - \alpha(p)}{\beta(p)},$$

weakly converges to a nondegenerate rv with cdf $\mathscr{L}^{(k)}$ given by $\mathscr{L}^{(k)}(x) = \mathscr{L}^k(x)$.

Acknowledgments The authors are grateful to the referees for many valuable suggestions that have been used to improve the presentation of the paper. Professor Sneh Gulati's very thorough proofreading of the text is gratefully acknowledged.

This research has been supported by National Funds through FCT – Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2011, PEst-OE/MAT/UI0006/2014, and EXTREMA, PTDC /MAT /101736/2008.

References

1. Engen, S.: On species frequency models. Biometrika **61**, 263–270 (1974)
2. Galambos, J.: The Asymptotic Theory of Extreme Order Statistics. Wiley, New York (1987)
3. Hess, K.T., Liewald, A., Schmidt, K.D.: An extension of Panjer's recursion. ASTIN Bull. **32**, 283–297 (2002)
4. Katz, L.: Unified treatment of a broad class of discrete probability distributions. In: Patil, G.P. (ed.) Classical and Contagious Discrete Distributions, pp. 175–182. Pergamon Press, Oxford (1965)

5. Morris, C.L.: Natural exponential families with quadratic variance functions. Ann. Stat. **10**, 65–80 (1982)
6. Panjer, H.H.: Recursive evaluation of a family of compound distributions. ASTIN Bull. **12**, 22–26 (1981)
7. Pestana, D., Velosa, S.: Extensions of Katz-Panjer families of discrete distributions. REVSTAT **2**, 145–162 (2004)
8. Rachev, S.T., Resnick, S.: Max-geometric infinite divisibility and stability. Commun. Stat. —Stoch. Models **7**, 191–218 (1991)
9. Rólski, T., Schmidli, H., Schmidt, V., Teugels, J.: Stochastic Processes for Insurance and Finance. Wiley, New York (1999)
10. Smirnov, N. V.: Limit distributions for the terms of a variational series. Trudy Mat. Inst. Steklov., Acad. Sci. USSR, Moscow-Leningrad **25**, 3–60 (1949)
11. Sundt, B.: On some extensions of Panjer's class of counting distributions. ASTIN Bull. **22**, 61–80 (1992)
12. Sundt, B., Jewell, W.: Further results on recursive evaluation of compound distributions. ASTIN Bull. **12**, 27–39 (1981)
13. Willmot, G.E.: Sundt and Jewell's family of discrete distributions. ASTIN Bull. **18**, 17–29 (1988)

The Role of Asymmetric Families of Distributions in Eliminating Risk

Fernanda Otília Figueiredo and Maria Ivette Gomes

Abstract Modeling is always a crucial component of the risk assessment process. The use of adequate classes of distributions to model real data sets seems sensible to accommodate specific peculiarities of the data and enable us to implement resistant procedures, less sensitive to changes in the model. Despite the practical advantages of using the normal distribution, it is recognized that most of the data from diverse areas of application, such as economics, environment, finance, insurance, meteorology, reliability and statistical quality control, among others, usually exhibit moderate to strong asymmetry and heavier tails than the normal tail. This study motivates the use of two classes of skew-normal distributions that include highly skewed and heavy-tailed distributions as well as models that are close to the Gaussian family. Some guidelines for inference on the parameters of the model are suggested, and applications to real data sets are presented.

Keywords Asymmetric families of distributions · Control charts · Data modeling · Skew-normal · SPC

1 Introduction

Although the methods for assessment of risk may differ among organizations, firms, industries and areas of research, modeling is always a crucial component of the risk assessment process. Nowadays it is well known among statisticians that the use of adequate classes of distributions to model real data sets enables the accommodation of specific peculiarities of the data. Even if confronted with more sophisticated

F.O. Figueiredo (✉)
FEP, Universidade do Porto, and CEAUL, Rua Dr Roberto Frias,
4200-464 Porto, Portugal
e-mail: otilia@fep.up.pt

M.I. Gomes
FCUL, DEIO and CEAUL, Universidade de Lisboa, Campo Grande,
1749-016 Lisbon, Portugal
e-mail: ivette.gomes@fc.ul.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_20

estimation procedures, the computational capabilities we have access to these days enable us to use more resistant procedures, i.e., less sensitive to changes in the model.

We are often confronted with data sets from asymmetric parents, with different degrees of asymmetry and tail weight. This has motivated the detailed study of several asymmetric distributions, both univariate and multivariate.

In this paper, we shall pay attention to two different families of distributions with varying asymmetry, denoted by $\mathscr{F}_1$ and $\mathscr{F}_2$, both including the Gaussian distribution as a particular case. These families are obtained through the application of different mechanisms of asymmetry to the normal distribution. Those mechanisms enable us to have access to distributions with different degrees of asymmetry and tail-weight. It is worth noting that despite the relevant role of the normal distribution in statistical quality control, most of the available data sets are from non-Gaussian processes. Particularly in the area of reliability most asymmetric distributions, like the exponential, the Weibull, the gamma and the log-normal, are used as life distributions, so that we can easily accommodate the instantaneous hazard rate behavior. And even in potentially normal processes, the distribution underlying the data often differs a lot from a normal distribution due to uncontrollable factors. Thus it makes sense to use more general families that can accommodate those possible differences. Indeed, despite the practical advantages of using the normal distribution, it is recognized that most of the data from diverse areas of application, such as economics, environment, finance, insurance, meteorology, reliability and statistical quality control, among others, usually exhibit moderate to strong asymmetry and tails heavier than the normal tail, with an exponential decay and a penultimate Weibull behavior (see Beirlant et al. [9]).

The scope of this article is sketched in the following. In Sect. 2, we introduce the families of distributions under consideration, emphasizing some of their properties, in order to show their versatility and to motivate their use in practical applications. Section 3 is devoted to the modeling of a few real data sets through the use of these asymmetric normal distributions. Finally, in Sect. 4, some general concluding remarks are put forward.

2 Asymmetric Families of Distributions Under Consideration

Let us consider the usual notation ϕ and Φ for the probability density function (p.d.f.) and cumulative distribution function (c.d.f.) of a standard normal random variable (r.v.), respectively.

In the family $\mathscr{F}_1$, introduced in O'Hagan and Leonard [15] and later on studied in more detail by several authors, among whom we mention Azzalini [3, 4] and Azzalini and Regoli [8], we have distributions with a p.d.f. given by

$$f_1(x;\lambda,\delta,\alpha)=\frac{2}{\delta}\,\phi\left(\frac{x-\lambda}{\delta}\right)\Phi\left(\frac{\alpha(x-\lambda)}{\delta}\right),\quad x\in\mathbb{R},\tag{1}$$

with $\lambda,\ \alpha\in\mathbb{R}$ and $\delta\in\mathbb{R}^+$.

In the family $\mathscr{F}_2$, introduced in Fernandez and Steel [10], we find distributions with a p.d.f. given by

$$f_2(x;\lambda,\delta,\alpha)=\begin{cases}\dfrac{2}{\delta(\alpha+1/\alpha)}\,\phi\left(\dfrac{\alpha(x-\lambda)}{\delta}\right), & \text{if } x<\lambda,\\ \dfrac{2}{\delta(\alpha+1/\alpha)}\,\phi\left(\dfrac{x-\lambda}{\alpha\delta}\right), & \text{if } x\ge\lambda,\end{cases}\tag{2}$$

with $\lambda\in\mathbb{R}$ and $\delta,\ \alpha\ \in\mathbb{R}^+$. This family has also been considered by several other authors, among whom we mention Ferreira and Steel [11] and Abtahi et al. [1]. If X is a r.v. with p.d.f. either in $\mathscr{F}_1$ or in $\mathscr{F}_2$, a linear transformation of the type $(X-\lambda)/\delta$ enables us to get the standardized versions of these families, i.e., distributions with location 0 and scale 1.

Several generalizations of these families have been recently considered in the literature. As an example, and for the multivariate case, see Azzalini [5] and Azzalini and Capitanio [7]. We further mention asymmetric distributions defined in intervals on the basis of these unimodal families, a nice alternative for modeling data from a mixture model (see Jamalizadeb et al. [13]).

Without loss of generality, and to illustrate some of their properties, we consider the standardized versions of the aforementioned families $\mathscr{F}_1$ and $\mathscr{F}_2$. We further denote their p.d.f.'s by $f_i(x;\alpha):=f_i(x;0,1,\alpha)$, for $i=1,2$. The parameter α enables us to control the asymmetry and the tail-weight of the distribution, as we shall see in Sects. 2.1 and 2.2.

2.1 Properties of the $\mathscr{F}_1$ Family

The normal distribution obviously belongs to the family $\mathscr{F}_1$. We just need to consider $\alpha=0$ in (1). For $\alpha\neq 0$ we get asymmetric distributions, either with a positive ($\alpha>0$) or a negative ($\alpha<0$) asymmetry. These distributions become more asymmetric, as the distance between α and zero increases. Such a feature is illustrated in Fig. 1, where we should take into account the property, $f_1(x;\alpha)=f_1(-x;-\alpha)$, for $x\in\mathbb{R}$ and $\forall\alpha\in\mathbb{R}$, easy to derive from the expression of the p.d.f. in (1).

An r.v. X with p.d.f. $f_1(x;\alpha)$ has finite moments of all orders, $\forall\alpha\in\mathbb{R}$. The ordinary moments of order k, with k being any positive integer, are given by

$$\mathbb{E}(X^k)=\begin{cases}1\times 3\times\cdots\times(k-1), & k=2,4,6,\ldots\\ \dfrac{\sqrt{2}\,k!\,\alpha}{2^{\frac{k-1}{2}}\sqrt{\pi}(1+\alpha^2)^{\frac{k}{2}}}\displaystyle\sum_{m=0}^{(k-1)/2}\frac{m!(2\alpha)^{2m}}{(2m+1)!\left(\frac{k-1}{2}-m\right)!}, & k=1,3,5,\ldots\end{cases}\tag{3}$$

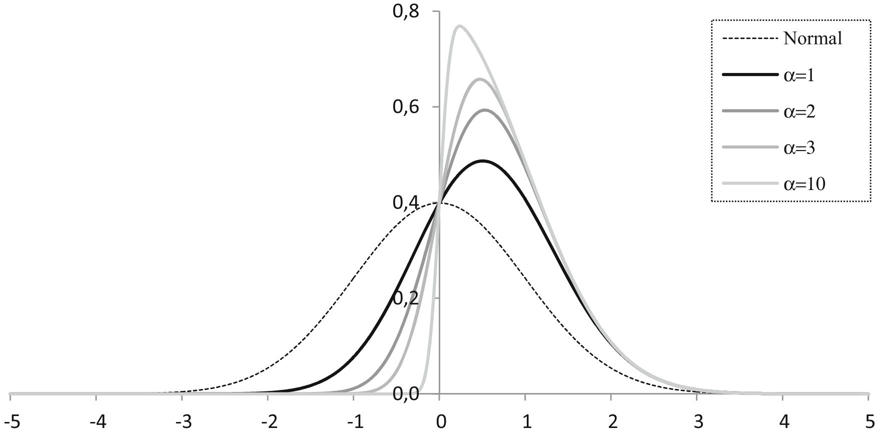


Fig. 1 Graphical representation of $f_1(x;\alpha)$ for several non-negative values of α

Consequently, the mean value and the variance of the r.v. X are respectively given by

$$\mathbb{E}(X) = \frac{\sqrt{2}\alpha}{\sqrt{\pi(1+\alpha^2)}} \quad \text{and} \quad \mathbb{V}(X) = 1 - \frac{2\alpha^2}{\pi(1+\alpha^2)}.$$

Note that in $\mathscr{F}_1$, if $\alpha \to \pm\infty$ we get a truncated version of the normal distribution. Thus, in this family of asymmetric distributions, the increase of the weight of the right ($\alpha > 0$) or the left ($\alpha < 0$) tail is accompanied by a decrease of the weight of other tail.

However, the tail weight depends strongly on α. If $\alpha > 0$ the right-tail of the distribution becomes heavier than that of a normal distribution, whereas the weight of the left-tail becomes weaker than the one of the normal. Things go the other way round if $\alpha < 0$. Indeed, when we change the sign of α the p.d.f. becomes reflected on the opposite side of the y-axis. Other properties of these models can be found in Figueiredo and Gomes [12] and Azzalini [5].

It is still relevant to mention that if we consider two r.v.'s X_1 and X_2 with a normal distribution, either independent or linearly correlated, the sample statistics $T_{\max} = \max(X_1, X_2)$ and $T_{\min} = \min(X_1, X_2)$ have an asymmetric normal distribution in $\mathscr{F}_1$. These univariate statistics $T_{\max}$ and $T_{\min}$ thus enable the implementation of control charts to simultaneously monitor two relevant quality characteristics, and are alternatives to the multivariate control charts based on Hotelling's statistic. Moreover, they can be used when we have only access to a unique observation at a certain time period. We should not obviously exclude the cases in which those variables X_1 and X_2 are averages of n observations of two crucial process variables.

We next illustrate the performance of the one-sided control chart based on the statistic $T_{\max}$, in the monitoring of a bivariate normal process (X_1, X_2), assuming without loss of generality that the r.v.'s X_1 and X_2 are independent and identically

Table 1 ARL of the one-sided $T_{\max}$-chart, based on independent $X_i \sim \mathrm{N}(\mu, \sigma)$, for $i = 1, 2$

$\delta \backslash \theta$	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.5
0.0	370.4	156.7	80.7	47.8	31.4	22.2	16.7	13.1	10.7	9.0	7.7	4.6
0.1	268.0	119.5	64.1	39.1	26.3	19.0	14.6	11.6	9.6	8.1	7.1	4.3
0.2	195.8	91.8	51.2	32.2	22.2	16.4	12.7	10.3	8.6	7.4	6.5	4.1
0.3	144.4	71.1	41.1	26.7	18.8	14.2	11.2	9.2	7.7	6.7	5.9	3.8
0.4	107.4	55.5	33.3	22.2	16.0	12.3	9.9	8.2	7.0	6.1	5.4	3.6
0.5	80.7	43.6	27.1	18.6	13.7	10.7	8.7	7.3	6.3	5.6	5.0	3.4
0.6	61.2	34.6	22.2	15.7	11.8	9.4	7.7	6.6	5.7	5.1	4.6	3.2
0.7	46.8	27.6	18.3	13.3	10.2	8.2	6.9	5.9	5.2	4.7	4.3	3.1
0.8	36.2	22.2	15.2	11.3	8.9	7.3	6.2	5.4	4.8	4.3	4.0	2.9
0.9	28.2	18.0	12.7	9.7	7.7	6.5	5.5	4.9	4.4	4.0	3.7	2.8
1.0	22.2	14.7	10.7	8.3	6.8	5.7	5.0	4.4	4.0	3.7	3.4	2.6
1.5	7.7	6.1	5.0	4.3	3.8	3.4	3.1	2.9	2.7	2.6	2.5	2.1
2.0	3.4	3.0	2.7	2.5	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.8
2.5	1.9	1.8	1.8	1.7	1.7	1.7	1.6	1.6	1.6	1.6	1.6	1.5

Under control, $\mu = \mu_0$ ($\delta = 0$) and $\sigma = \sigma_0$ ($\theta = 1$), whereas out of control, $\mu \to \mu_1 = \delta > 0$ and/or $\sigma \to \sigma_1 = \theta > 1$

distributed. When the process is under control, let us assume that those variables have a mean value $\mu = \mu_0 = 0$ and a standard deviation $\sigma = \sigma_0 = 1$. Then, the c.d.f. and the p.d.f. of the control statistic $T_{\max} = \max(X_1, X_2)$ are given by

$$F(t) = P(T_{\max} \le t) = \Phi^2(t), \quad t \in \mathbb{R}, \tag{4}$$

and

$$f(t) = 2\phi(t)\Phi(t), \quad t \in \mathbb{R}, \tag{5}$$

respectively. This shows that $T_{\max}$ has a standardized asymmetric distribution belonging to the family $\mathscr{F}_1$, with a shape parameter $\alpha = 1$. When the process is out of control, let us assume that the mean value and/or the process standard deviation change to the values $\mu = \mu_1 = \delta > 0$ and $\sigma = \sigma_1 = \theta > 1$, respectively. To assess the performance of the $T_{\max}$-chart we shall analyze its ARL-behavior, with ARL denoting the Average Run Length, i.e., the mean number of samples taken until the chart signals. The Upper Control Limit (UCL) of the one-sided $T_{\max}$-chart that leads to an ARL under control equal to 370.4, or equivalently, a false alarm rate equal to 0.0027, is the solution of the equation $\Phi^2(\mathrm{UCL}) = 1 - 0.0027$, i.e., the value UCL $= 2.99977$. The corresponding ARL values of the chart when the process is out of control are presented in Table 1. These values show the interesting performance of this T_{max}-chart. Indeed, the ARL decreases quickly when the mean value and/or the standard deviation increases. This shows the high capacity of the chart in the detection of changes in the process's parameters. For further details on the topic, see Figueiredo and Gomes [12].

Algorithm 2.1 enables the simulation of a random sample $(X_1, X_2, \ldots, X_n)$ from the distribution $f_1(x; \alpha) \in \mathscr{F}_1$.

Algorithm 2.1

```
For $i = 1$ until $n$, repeat:
1. Generate two independent observations, $Z_1$ and $Z_2$, from a standard normal distribution;
2. Fix $\alpha$;
3. If $Z_1 < \alpha Z_2$, take $X_i = Z_2$; otherwise, take $X_i = -Z_2$.
```

2.2 Properties of the Family $\mathscr{F}_2$

It can easily be seen that $\alpha = 1$ in (2) provides the normal distribution. For values of $\alpha \neq 1$, we get positive asymmetry ($\alpha > 1$) or negative asymmetry ($0 < \alpha < 1$), and distributions again more asymmetric as the distance between α and one increases. These features are shown in Fig. 2. Take also into account the fact that $f_2(x; \alpha) = f_2(-x; 1/\alpha)$, for $x \in \mathbb{R}$ and $\forall \alpha \in \mathbb{R}^+$, a property easily derived from the p.d.f. in (2).

For an r.v. X with p.d.f. $f_2(x; \alpha)$, the ordinary moments of order k, with k being any positive integer, are given by

$$\mathbb{E}(X^k) = M_k \left(\frac{\alpha^{k+1} + (-1)^k/\alpha^{k+1}}{\alpha + 1/\alpha} \right), \text{ with } M_k = \int_0^\infty 2s^k \phi(s) \mathrm{d}s. \qquad (6)$$

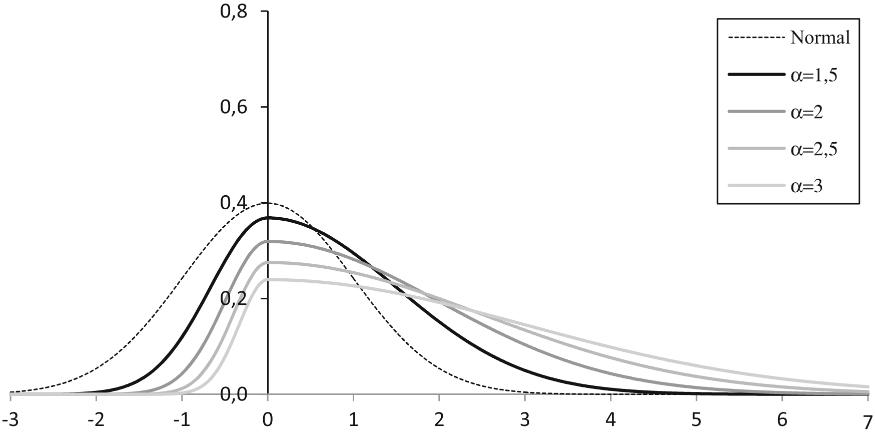


Fig. 2 Graphical representation of $f_2(x; \alpha)$ for several values of $\alpha \geq 1$

Note that M_k is the ordinary moment of order k of an r.v. S with a half-normal distribution. Indeed, the half-normal p.d.f. is given by $f(s) = 2\phi(s)$, for $s \geq 0$. In particular, $M_1 = \sqrt{2/\pi}$ and $M_2 = 1$. Consequently, the mean value and the variance of the r.v. X are given by

$$\mathbb{E}(X) = \sqrt{\frac{2}{\pi}}\left(\frac{\alpha^2-1}{\alpha}\right) \text{ and } \mathbb{V}(X) = \left(1-\frac{2}{\pi}\right)\left(\frac{\alpha^4-2\alpha^2+1}{\alpha^2}\right)+1,$$

respectively.

We have seen that the distributions in the family $\mathscr{F}_1$ have finite moments of any order, converging to finite values when $\alpha \to \pm\infty$. The same does not happen with the distributions in $\mathscr{F}_2$. Here, these moments diverge to infinity when $\alpha \to +\infty$ or when $\alpha \to 0$.

Algorithm 2.2 enables the random generation of a sample $(X_1, X_2, \ldots, X_n)$ of random values from a distribution $f_2(x; \alpha) \in \mathscr{F}_2$.

Algorithm 2.2

```
For $i = 1$ until $n$, repeat:
1. Generate a random number $U$ from a uniform distribution in (0, 1);
2. Fix $\alpha$;
3. If $U \leq 1/(1+\alpha^2)$ consider $X_i = \frac{\Phi^{-1}\left(U(1+\alpha^2)/2\right)}{\alpha}$; otherwise, consider
   $X_i = \alpha\Phi^{-1}\left(\frac{U(1+\alpha^2)}{2\alpha^2} + \frac{\alpha^2-1}{2\alpha^2}\right)$.
```

2.3 A Few Considerations on the Estimation of the Unknown Parameters

The *maximum likelihood* (ML) estimates of the unknown parameters of the distribution in any of the families $\mathscr{F}_i$, for $i = 1, 2$, can be obtained only numerically through iterative procedures. The same happens with the moment estimates. The R-package sn, implemented by Azzalini [6], enables us to easily obtain those estimates for models in the family $\mathscr{F}_1$.

A simpler estimation procedure has been proposed by Abtahi et al. [1]. They start with the use of the data *mode* (m_o) and *inter-quartile range* (IQR) as location and scale estimates, i.e., $\hat{\lambda} = m_o$ and $\hat{\delta} = \text{IQR}$. To estimate the shape parameter, they suggest the use of an asymmetry indicator defined in Arnold and Groeneveld [2], given by AG := $1 - 2\mathbb{P}(X < \text{mode(X)})$, that can be estimated by

$$\widehat{\mathrm{AG}} = 1 - \frac{2}{n} \sum_{i=1}^{n} I_{(-\infty,0)}(x_i - m_o),$$

where $I_A(.)$ denotes the indicator function of the set A. This asymmetry indicator, AG, lies between -1 and 1, taking negative (positive) values whenever the distributions are negatively (positively) asymmetric, and the value zero in the case of symmetric distributions. They thus propose the following estimates of α. For models $f_1 \in \mathscr{F}_1$, the estimate is the solution of the equation

$$\frac{1-\widehat{\mathrm{AG}}}{2} = \int_{-\infty}^{0} 2\phi(x)\,\Phi\left(\hat{\alpha}x\right)\mathrm{d}x = \frac{1}{2} - \frac{1}{\pi}\arctan(\hat{\alpha}),$$

i.e.,

$$\hat{\alpha} = \tan\left(\frac{\pi}{2}\widehat{\mathrm{AG}}\right),$$

where $\tan(\cdot)$ and $\arctan(\cdot)$ denote, as usual, respectively the tangent and the arctangent functions. For models $f_2 \in \mathscr{F}_2$, the suggested estimate of α is the solution of the equation

$$\frac{1-\widehat{\mathrm{AG}}}{2} = \frac{1}{\hat{\alpha}^2+1},$$

i.e.,

$$\hat{\alpha} = \sqrt{\frac{2}{1-\widehat{\mathrm{AG}}} - 1}.$$

3 Applications to Real Data

To illustrate the importance of these families of distributions in applied areas, we consider four real data sets to which we fit asymmetric normal distributions in the families $\mathscr{F}_i$, for $i = 1, 2$. We analyze the following data sets:

- RESV: Resistence to the opening of 1 litre glass bottles with a non-alcoholic beverage, measured in psi pressure units; $n = 100$ observations (*Source:* Montgomery [14], *Table 9.1, p. 368.*)
- TVER: Summer average temperature (153 days) in Munich, measured in Celsius degrees, in the period 1781–1988; $n = 208$ observations. (*Source: Eamonn Keogh's and StatLib databases.*)
- TNEG: Sum of winter negative temperatures (153 days) in Munich, measured in Celsius degrees, in the period 1781–1988; $n = 208$ observations. (*Source: Eamonn Keogh's and StatLib databases.*)

- ECO2: Total CO2 emissions in USA, measured in megagrams per person and per month, in the period from January 1981 until December 2003; $n = 276$ observations. (*Source: Carbon Dioxide Information Analysis Center, Tenessee USA and Department of Agricultural & Resource Economics, Oregon USA.)*

In Fig. 3 we present the histograms associated with these data sets and the estimated p.d.f.'s, $f_i \in \mathscr{F}_i$, for $i = 1, 2$. In Table 2 we provide the estimates of the parameters of the fitted models, and the *p-value* of the Kolmogorov-Smirnov (K-S) test. For sake of simplicity, the parameters of f_1 were estimated through ML, using the `R`-package `sn`, and for the parameters of f_2 we have used the estimates suggested by Abtahi et al. [1].

On the basis of the presented results, we can conclude that the family $\mathscr{F}_1$ provides the models that better describe the data sets under analysis. On the basis of the K-S goodness-of-fit test, there is no reason to reject f_1 as an adequate model to fit the data; at the most common significance levels of 1 and 5 %, the models f_2 are rejected

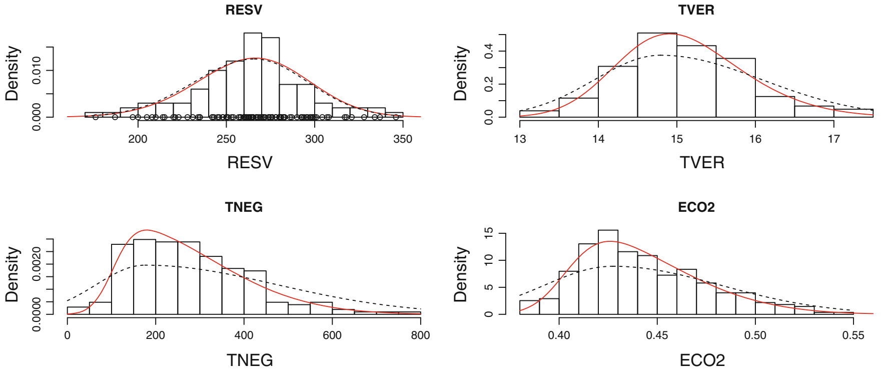


Fig. 3 Histograms and fitted f.d.p.'s, f_1 (—-) and f_2 (. . . .)

Table 2 Estimates of the parameters of the fitted models and *p-value* of the K-S goodness-of-fit test

Data	Model	$\hat{\lambda}$	$\hat{\delta}$	$\hat{\alpha}$	*p-value* (K-S test)
RESV	f_1	286.1546	38.7629	–1.0224	0.5084
	f_2	265.8621	32.0000	0.9417	>5 %
TVER	f_1	14.2940	1.1474	1.8387	0.9969
	f_2	14.7922	1.0400	1.2396	≃5 %
TNEG	f_1	103.9236	218.3129	5.6269	0.9948
	f_2	175.4237	180.7000	1.6475	<1 %
ECO2	f_1	0.4037	0.0504	3.5348	0.6628
	f_2	0.4274	0.0437	1.2568	<1 %

for the data sets TNEG and ECO2, contrary to what happens for the data sets RESV and TVER.

4 Concluding Remarks

For an adequate modeling of real data, we are fully convinced that it is better to play with flexible families of models, with different types of p.d.f.'s. This is surely more sensible than to consider the fitting to a specific and simple model, dependent only on the estimation of a location and a scale parameter, like the normal model. To model possibly asymmetric data, it is adequate to consider models in $\mathscr{F}_1$. To describe approximately normal or quasi-symmetric data, both classes $\mathscr{F}_1$ and $\mathscr{F}_2$ provide adequate models. In both of these models, the parameters can be estimated either through the ML method or through other estimation procedures, like the ones mentioned above, in Sect. 2.3.

Acknowledgments Research partially supported by National Funds through FCT—Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014, PEst-OE /MAT /UI0006 /2011 (CEAUL) and EXTREMA, PTDC /MAT /101736 /2008. The authors would like to acknowledge the helpful suggestions of the anonymous reviewers and the editor.

References

1. Abtahi, A., Towhidi, M., Behboodian, J.: An appropriate empirical version of skew-normal density. Stat. Pap. **52**, 469–489 (2011)
2. Arnold, B.C., Groeneveld, R.A.: Measuring skewness with respect to the mode. Am. Stat. **49**, 34–38 (1995)
3. Azzalini, A.: A Class of distributions which includes the normal ones. Scand. J. Stat. **12**, 171–178 (1985)
4. Azzalini, A.: Further results on a class of distributions which includes the normal ones. Statistica **XLVI**, 199–208 (1986)
5. Azzalini, A.: The skew-normal distribution and related multivariate families. Scand. J. Stat. **32**, 159–188 (2005)
6. Azzalini, A.: R package 'sn': The skew-normal and skew-t distributions (version 0.4-17). http://azzalini.stat.unipd.it/SN (2011)
7. Azzalini, A., Capitanio, A.: Statistical applications of the multivariate skew normal distributions. J. R. Stat. Soc. B **61**, 579–602 (1999)
8. Azzalini, A., Regoli, G.: Some properties of skew-symmetric distributions. Ann. Inst. Stat. Math. **64**, 857–879 (2012)
9. Beirlant, J., Caeiro, F., Gomes, M.I.: An overview and open research topics in statistics of univariate extremes. Revstat **10**(1), 1–31 (2012)
10. Fernandez, C., Steel, M.F.J.: On Bayesian modeling of fat tails and skewness. J. Am. Stat. Assoc. **93**, 359–371 (1998)
11. Ferreira, J.T.A.S., Steel, M.F.J.: A constructive representation of univariate skewed distributions. J. Am. Stat. Assoc. **101**, 823–829 (2006)
12. Figueiredo, F., Gomes, M.I.: The skew-normal distribution in SPC. Revstat. **11**, 83–104 (2013)

13. Jamalizadeb, A., Arabpour, A.R., Balakrishnan, N.: A generalized skew two-piece skew-normal distribution. Stat. Pap. **52**, 431–446 (2011)
14. Montgomery, D.C.: Introduction to Statistical Quality Control. Wiley, New York (2005)
15. O'Hagan, A., Leonard, T.: Bayes estimation subject to uncertainty about parameter constraints. Biometrika **63**, 201–202 (1976)

Parametric and Semi-parametric Approaches to Extreme Rainfall Modelling

Isabel Fraga Alves and Pedro Rosário

Abstract In a meteorological setup, considering a data set of daily rainfall in Barcelos, Portugal, a survey of possible parametric and semi-parametric approaches in Extreme Value Theory is presented, with the main goal of the analyzing high observations of records over time, since these might entail negative consequences for society. These analysis embraces estimation of several extreme value parameters, including return levels associated with T-year return periods, for large T.

Keywords Extreme value parameters · Extreme value theory · Parametric and semi-parametric approaches · Return levels · Rainfall

1 Introduction

When we are dealing with meteorological data there is the need to differentiate between two situations: the case of data set concentrated around the average, with no disastrous consequences for the society; and on the other hand, the case of data away from the center of a distribution, that can have a very negative impact and which is important to quantify. Typically, one is interested in the analysis of maximal observations and records over time, since these entail the negative consequences. Rainfall is a good example of this: the engineering structures associated with extremal precipitation levels, need to be constructed to withstand the extremal behavior of this process; for example, a reservoir must be able to store the amount of rain expected to fall in some specific location.

Extreme Value Theory (EVT) is the theory of modeling and measuring events which occur with very small probability, and has proved to be a powerful and useful tool to describe atypical situations that may have a significant impact in many

I. Fraga Alves (✉) · P. Rosário
DEIO and CEAUL, Faculty of Sciences, University of Lisbon, Lisbon, Portugal
e-mail: mialves@fc.ul.pt

P. Rosário
e-mail: parosario@fc.ul.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_21

application areas, where knowledge of the behavior of the tail of a distribution is of main interest. The classical result is the Gnedenko theorem [10]. It establishes that there are three type of possible limit distributions (max-stable) for maxima of blocks of observations, which are unified in a single representation—the Generalized Extreme Value (GEV) distribution. The second theorem in EVT is the so called Pickands-Balkema-de Haan theorem [1, 16]. Loosely speaking, it allows us to model the Generalized Pareto (GP) distribution to the excesses of high thresholds—POT approach—for distributions in the domain of a GEV distribution. Complementary to these parametric approaches, we also consider up a possible semi-parametric approach, comparing it with the previous ones. Additional information about parametric and semi-parametric inference for extreme values can be found in some overview papers (see [3, 5, 11], for instance) and reference books in the field of EVT and its real world applications (see [2, 4, 6, 7, 9, 18]).

For rainfall data in Barcelos, we will employ these two approaches to estimate a p-return levels associated with $T = 1/p$-year return periods, for small p, some extreme quantiles and the probability of exceedance of a high level. Design levels typically correspond to return periods of 100 years or more; however, time series of 100 or more years are rare. A model for extrapolation is required and here integrates with the EVT, a theory specifically designed for modelling rare events.

The paper is structured as follows: in Sect. 2 we present some parameters of rare events for real world problems, describe the rainfall data and motivate the EVT framework. Then we move on with Sects. 3 and 4, in which we shortly sketch some parametric and semi-parametric approaches, prescribed for the respective univariate data type available *a priori*. In Sect. 5 we make some final statements about how both approaches benefit a complementary statistical analysis of extreme values. Moreover, throughout the text, we will often mention useful capabilities of some libraries of R-package [17].

2 Preliminaries

In this section some preliminary concepts are presented. Denote by F the distribution function (DF) underlying the data under study and $F^{\leftarrow}$ its generalized inverse. Typical design values are:

Definition 1 (*T-year Return Level: u_T*) A value which is exceeded once in a year with a probability $1/T$

$$u_T = F^{\leftarrow}(1 - 1/T) \tag{1}$$

Definition 2 (*u_T-Return Period: T*) Average number of years between occurrences of an event of magnitude greater than a predefined high level u_T

$$T = \frac{1}{P[X > u_T]} \tag{2}$$

Definition 3 (*Small Exceedance Probability*) The probability for an exceedance of a very high level

$$p_x = P[X > x], \text{ with } x > x_{n:n} =: \text{sample maximum} \tag{3}$$

If in (1) the value $1/T =: p$ is very small, say $p < 1/n$, with n denoting the available sample size, then we are dealing with *high or extreme quantiles* and it is crucial for modelling rare events.

We cannot simply assume that these atypical values are impossible. Design levels correspond to return periods of 100 years or more and the empirical cumulative distribution function (ECDF) is not enough! It is pertinent here to quote Emil Gumbel:

> Il est impossible que l'improbable n'arrive jamais.
> Il y aura toujours une valeur qui dpassera toutes les autres.
> *Emil Gumbel (1891–1966)*

Daily rainfall in Barcelos 1932–2008 The following data set, represented in Fig. 1, is freely available from www.snirh.pt and has also been analyzed in [14, 15], including high quantiles estimation for monthly maxima.

If we 'zoom into' of the upper part of ECDF for daily rainfall in Barcelos (see Fig. 1), and aim to estimate 100-year return level, the best we can do with the ECDF is giving the sample maximum, and the same applies to any T-year return level, with $T > 75$. Consequently, extrapolation is required.

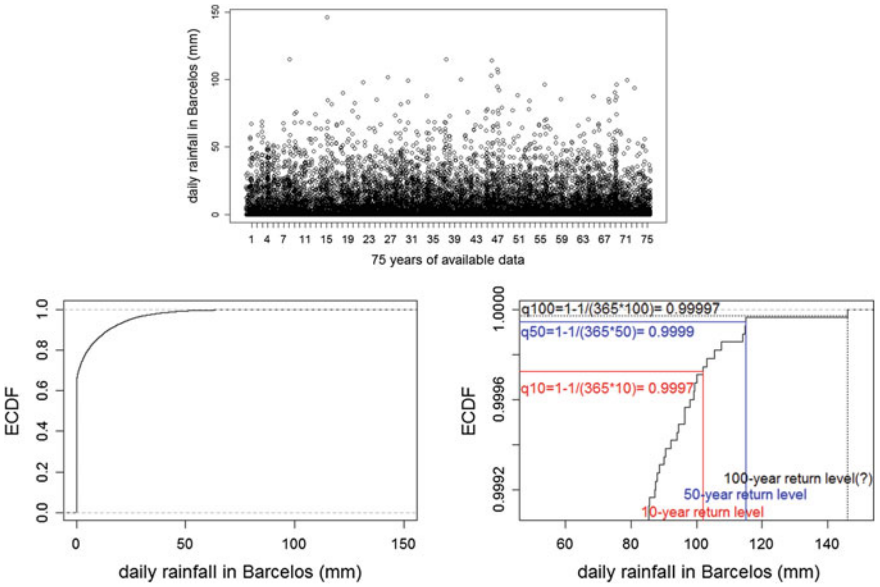


Fig. 1 Daily rainfall in Barcelos 1932–2008 (*up*). ECDF for daily rainfall in Barcelos: all data (*down left*); Top data (*down right*); q_T denotes $q_T := 1 - p_T = 1 - \frac{1}{365\times T}$

3 Parametric Approaches

3.1 Annual Maxima Approach—Gumbel Method

EVT provides limit laws for an extrapolation beyond sample. The classical result is the Gnedenko theorem [10], which establishes that there are three type of possible limit distributions max-stable for maxima of blocks of n independent and identically distributed (iid) observations with common (DF) F, M_n, which are unified in a single representation—the GEV distribution

$$G_\gamma(x) = \exp\left\{-\left[1+\gamma x\right]_+^{-1/\gamma}\right\}, \qquad \gamma \in \mathbb{R}. \tag{4}$$

[Notation: $x_+ := \max(0, x)$]. That is, if there are sequences $a_n > 0$ and b_n, such that $P\left[(M_n - b_n)/a_n \le x\right] \longrightarrow G(x)$, as $n \to \infty$, for some non-degenerated fd G, then G is of the same type of $G_\gamma(x)$ and we say that F belongs to the max-domain of attraction G_γ [Notation: $F \in \mathscr{D}(G_\gamma)$].

Consider the available data—Daily rainfall in Barcelos 1932–2008—divided in m blocks, usually years, and pick up the maximum in each block, as illustrated in Fig. 2.

A quick summary of descriptive statistics for annual maxima data (`annual_max`) is given by the following R-Package [17] command

```
> summary(annual_max)
   Min.  1st Qu.  Median    Mean  3rd Qu.   Max.
  42.00    62.25   68.40   73.99    86.50  146.00
```

The autocorrelation function (ACF) for daily and annual maxima of daily rainfall records is represented in Fig. 3, which highlights the absence of a significant dependence for the latter. We should also mention that the tendency is not significant, which was concluded from a preliminary statistical test study.

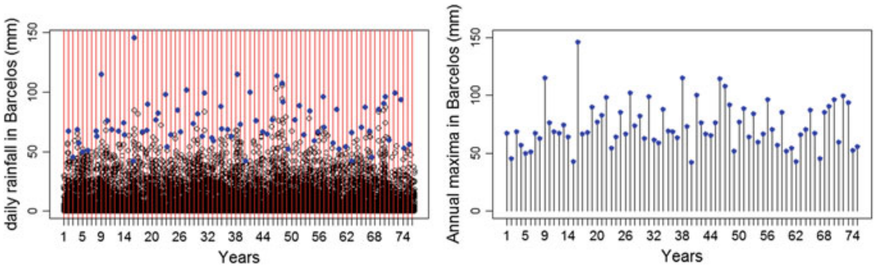


Fig. 2 Blocks of years, daily data and Annual Maxima (*left*); Annual Maxima (*right*)

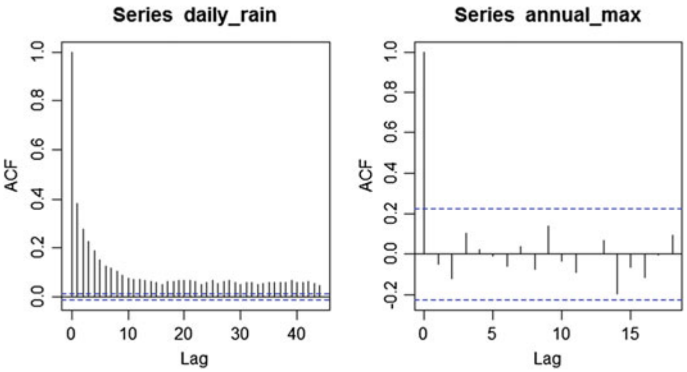


Fig. 3 ACF for daily rainfall (*left*) and for annual maxima (*right*) [R-package]

Letting Y denote the annual maximum of a random sample of n rainfall values $\max(X_1, X_2, \ldots, X_n)$, assume that our available sample consists of m iid annual maxima: $Y_1, Y_2, \ldots, Y_m$. Fit the GEV distribution $G_\gamma(x; \lambda, \delta) := G_\gamma((x - \lambda)/\delta)$, where γ denotes the Extreme Value Index (EVI), λ a real valued location parameter and δ a positive scale parameter.

Thereafter use $(\hat{\gamma}, \hat{\lambda}, \hat{\delta})$ in the associated GEV fit for Y to estimate rare events:

- Exceedance probability for high level u, $1 - G_\gamma(u; \lambda, \delta)$,
- Return period for level u, $T_u = \dfrac{1}{1 - G_\gamma(u; \lambda, \delta)}$,
- T-year return level, $G_\gamma^{\leftarrow}\left(1 - \frac{1}{T}; \lambda, \delta\right)$.

Let ML and PWM be, respectively, the Maximum Likelihood and Probability Weighted Moments estimators [12]; in last column of Table 1 we also include the probability of annual maxima of rainfall levels around Barcelos station that are above 159 mm, a level that has also been considered in [15].

With the GEV ML fit, the 95 %-CI for Return Levels, using profile log-likelihood, are (see also Fig. 4)

- MLE Return Level 100-year = 133.867 mm (118.428, 174.901),
- MLE Return Level 400-year = 152.798 mm (130.070, 226.663);

Table 1 ML and PWM estimates in annual maxima approach [R-library(fExtremes)]

	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\delta}$	$\hat{P}[Y > 159]$
ML	−0.030	65.19	16.00	0.0016
PWM	−0.027	65.08	16.18	0.0018

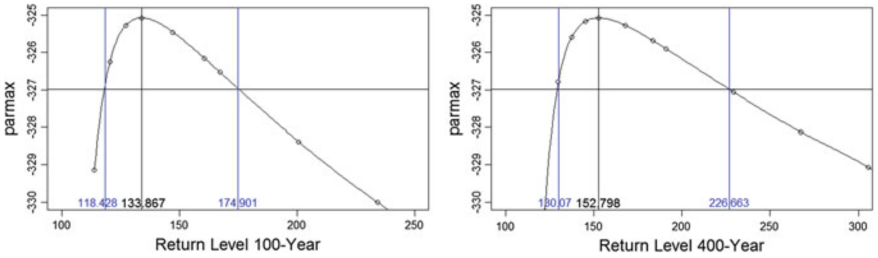


Fig. 4 95 %-CI for return levels and profile log-likelihood: 100-year (*left*); 400-year (*right*). [R-library(evir)]

3.2 TOP Annual Approach—10 Largest Observations Per Year

At this stage, we take into account the 10 largest observations per year (see Fig. 5). The TOP annual approach relies on a convenient parametric model underlying the sample of the r largest observations, observed for the m years.

Consider the limit joint model for r top order statistics (o.s.), r fixed, with joint limit density function

$$g_{1,\ldots,r}(w_1,\ldots,w_r) := G_\gamma(w_r)\prod_{i=1}^{r}\frac{g_\gamma(w_i)}{G_\gamma(w_i)},\ \text{ for } w_1 > \cdots > w_r\ , \tag{5}$$

with $g_\gamma(w) := \frac{\partial G_\gamma}{\partial w}(w)$. In statistical inference for rare events, a possible approach is to model with the above joint structure the top observations (TO) available from the sample. More precisely, $F \in \mathscr{D}(G_\gamma)$ for $a_n > 0$ and b_n iff the r-vector $\left(\frac{X_{n:n}-b_n}{a_n},\ldots,\frac{X_{n-r+1:n}-b_n}{a_n}\right)$ has joint limit density function given in (5).

In Table 2 the estimation results are summarized.

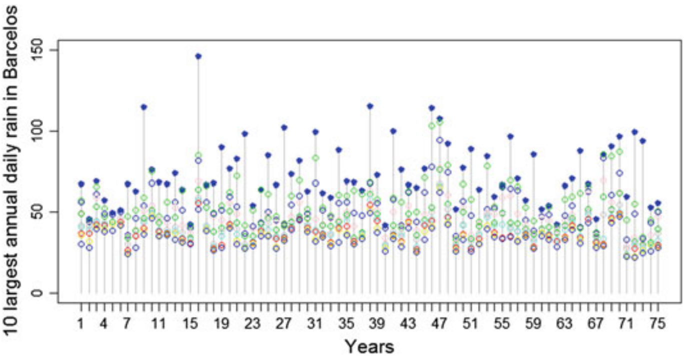


Fig. 5 Blocks of years, daily data and 10 top observations per year

Table 2 Parameters of GEV fit to annual maximum, by TO approach, with $r = 1$ (annual maxima), $r = 5$ and $r = 10$, by ML. [R-library(ismev)]

# TO	$\hat{\gamma}$	$\hat{\lambda}$	$\hat{\delta}$	$\hat{P}[Y > 159]$	rl 100-year	rl 400-year
r = 1	−0.030	65.19	16.00	0.0016	133.867	152.798
r = 5	0.013	66.38	15.60	0.0033	140.265	163.442
r = 10	0.005	66.95	15.04	0.0024	136.880	158.302

3.3 Monthly Maxima Approach

Now we consider monthly maxima data of daily rainfall (mm) in Barcelos, similar to data worked out in [14, 15], where high quantiles Q_{1-p} of monthly maxima have been estimated. Results are summarized in Table 3: the first row refers directly to estimation through GEV fit to monthly maximum, as in Sect. 3.1, but with 919 monthly blocks; in the second, third and fourth rows, since we are relying on monthly maxima, the GEV fit to the annual maximum, based on TO approach with $r = 1, 5, 10$, provides estimated high quantiles of monthly maximum by the input of annual estimated parameters $\lambda_{year}, \delta_{year}, \gamma_{year}$ at the expression $Q_{1-p} := G_{\gamma}^{\leftarrow}\left((1-p)^{12}; \lambda, \delta\right)$, relying on max-stability.

3.4 Monthly Maximum—POT Approach

The basic idea behind the Peaks Over Threshold (POT) approach is to base statistical inference for extremes on the excesses over a high threshold. In Fig. 6, we combine monthly maxima with the excesses over the threshold $u = 42$ mm, which corresponds to the minimum of annual maxima in the period under study. Note that although

Table 3 Estimates of high quantiles for monthly maximum—several approaches

$\hat{Q}_{1-p}$ (mm)	0.05	0.01	0.001	0.0001	0.00001
Month-max	73.530	104.995	153.034	205.422	262.641
Annual-max fit, $r = 1$	72.900	97.966	131.387	162.491	191.487
Annual-max fit, $r = 5$	73.972	99.824	137.330	175.864	215.517
Annual-max fit, $r = 10$	74.258	98.923	134.150	169.687	205.595
Empirical quantiles	71.28	98.90	–	–	–

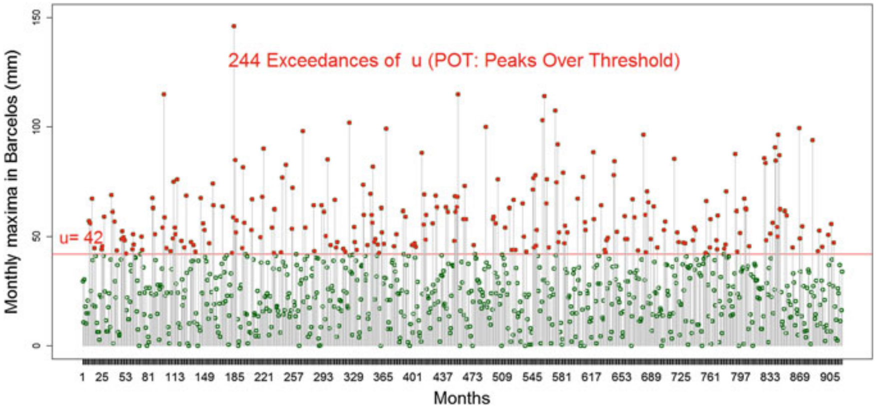


Fig. 6 Monthly maximum—POT ($u = 42\,\text{mm} = \min(annual\ maxima)$)

independence of monthly maxima is questionable, the monthly maxima excesses over $u = 42$ perform reasonably well in what concerns independence (see Fig. 7).

The theory behind modelling such a data base involves the Generalized Pareto approximation to the distribution of $(X-u)|X > u$, for a convenient high threshold u,

$$F_u(y) = P[X - u \leq y|X > u] = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad 0 \leq y \leq x^F - u,$$

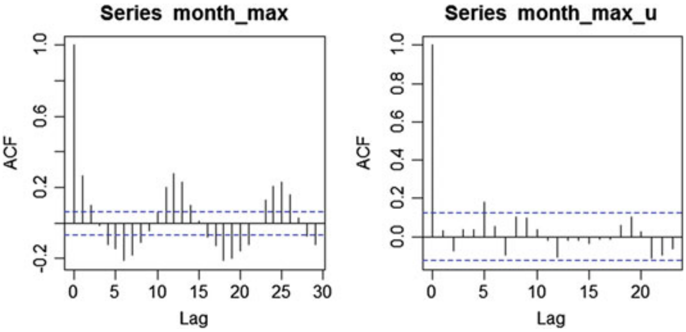


Fig. 7 ACF for monthly maxima (*left*) and for monthly maxima excesses over $u = 42$ (*right*) [R-package]

with x^F denoting the right endpoint of F. That is, we fit a Generalized Pareto DF

$$F_u(y) \approx H_\gamma(y;\sigma_u) := 1 - \left(1 + \gamma \frac{y}{\sigma_u}\right)_+^{-1/\gamma}, \quad \gamma \in \mathbb{R},\ \sigma_u > 0,$$

using the estimated parameters $H_{\hat\gamma}(y;\hat\sigma_u)$.

This is supported by Pickands-Balkema-de Haan theorem [1, 16], which states that

$$F \in \mathscr{D}(G_\gamma),\ \gamma \in \mathbb{R} \Leftrightarrow \lim_{u \to x^F} \sup_{0<y<x^F-u} \left|F_u(y) - H_\gamma(y;\sigma_u)\right| = 0.$$

High quantiles for monthly maxima of daily rainfall in Barcelos, $Q_{1-p} = F^{\leftarrow}(1-p)$, are obtained by approximating

$$\hat{Q}_{1-p} := \widehat{F}^{\leftarrow}(1-p) \approx u + \frac{\hat\sigma_u}{\hat\gamma}\left(\left(\frac{np}{N_u}\right)^{-\hat\gamma} - 1\right),$$

with threshold $u = 42$, for a number of excesses $N_u = 244$ from a total number of months $n = 919$ (Table 4). The estimated values are $\hat\sigma_u = 19.82$ and $\hat\gamma = -0.088$.

It is also possible to obtain T-year return levels, estimated by

$$u_T \approx u + \frac{\hat\sigma}{\hat\gamma}\left(\left(\frac{np}{N_u}\right)^{-\hat\gamma} - 1\right)$$

- T $=$ 100-year return level: with $p = 1/(100 \times 12)$ then $u_T \approx 131.56$ mm,
- T $=$ 400-year return level: with $p = 1/(400 \times 12)$ then $u_T \approx 147.12$ mm,

which compare well with the previous corresponding values by the Annual Maxima approach in Sect. 3.1, 133.867 and 152.798 mm, respectively.

Table 5 summarizes high quantile estimation for monthly maxima of daily rainfall in Barcelos. It is worthwhile to highlight that this estimation using POT-methodology may depend heavily on the threshold u.

In Fig. 8 we show the EVI and the 0.01-quantile estimation, against various values of k, relating also the particular value of $k = 244$ with the threshold $u = 42$. Two estimation methodologies are compared: ML and PWM (details for the latter in [13]).

Table 4 Estimates of high quantiles for monthly maximum POT approach [R-library(ismev)]

$\hat{Q}_{1-p}$	0.05	0.01	0.001	0.0001	0.00001
POT month-max ($u = 42$)	72.768	98.435	129.367	154.616	175.227

Table 5 Estimates of high quantiles for monthly maximum—parametric approaches

$\hat{Q}_{1-p}$ (mm)	0.05	0.01	0.001	0.0001	0.00001
Month-max	73.530	104.995	153.034	205.422	262.641
POT month-max ($u = 42$)	72.768	98.435	129.367	154.616	175.227
Annual-max fit, $r = 1$	72.900	97.966	131.387	162.491	191.487
Annual-max fit, $r = 5$	73.972	99.824	137.330	175.864	215.517
Annual-max fit, $r = 10$	74.258	98.923	134.150	169.687	205.595
Empirical quantiles	71.28	98.90	–	–	–

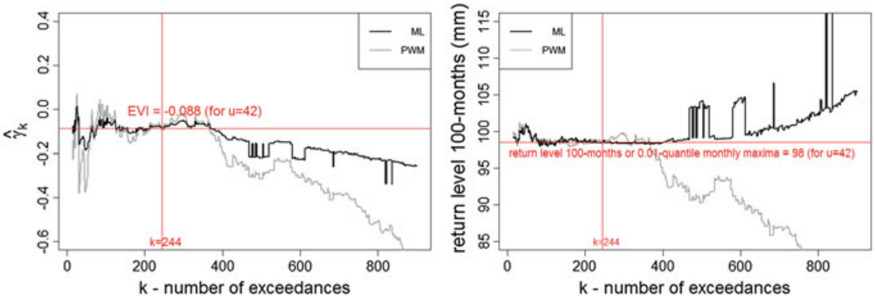


Fig. 8 Monthly maximum—POT: EVI (*left*) and 0.01-quantile (*right*) estimation, against *k*—[R-library(evir)]

4 Semiparametric Approach: $k + 1$ Top Observations

In a semiparametric setup, we do not assume that there is any parametric model underlying the data. Alternatively, we assume that the observations $X_1, X_2, \ldots, X_n$ are iid with common (DF) $F \in \mathscr{D}(G_\gamma)$, for some $\gamma \in \mathbb{R}$. The statistical inference is then pursued on the bases of the top part of the available sample

$$X_{n:n} \geq X_{n-1:n} \geq \cdots \geq X_{n-k:n},$$

where the *random threshold* $X_{n-k:n}$ is an intermediate o.s., with

$$k \equiv k_n \to \infty, \quad \text{with} \quad k/n \to 0, \quad \text{when} \quad n \to \infty.$$

It is as if the deterministic threshold u of POT Sect. 3.4 is now replaced by a random $X_{n-k:n}$, and sometimes we call this a PORT approach (*Peaks Over Random*

Threshold). EVI parameter estimation is the main goal in a first stage. In the literature, there are several estimators $\hat{\gamma}_{k,n}$ with good properties. Here we only consider the classical Moment estimator $\hat{\gamma}_{k,n}^{M}$ for $\gamma \in \mathbb{R}$, presented in [8], defined by (6)

$$\hat{\gamma}_{k,n}^{M} = M_{k,n}^{(1)} + 1 - \frac{1}{2}\left\{1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right\}^{-1} \tag{6}$$

with

$$M_{k,n}^{(r)} = \frac{1}{k}\sum_{i=1}^{k}(\log X_{n-i+1:n} - \log X_{n-k:n})^r, \quad r = 1, 2.$$

An extremal quantile estimator, under this semiparametric setup, and for $\gamma \neq 0$, is given by

$$\widehat{F}^{\leftarrow}(1-p) = X_{n-k:n} + \hat{a}\left(\frac{n}{k}\right)\frac{(\frac{k}{np})^{\hat{\gamma}} - 1}{\hat{\gamma}} \tag{7}$$

with $\hat{a}\left(\frac{n}{k}\right) = X_{n-k:n}M_{k,n}^{(1)}(1 - \hat{\gamma}_{k,n}^{-})$ and $\hat{\gamma}_{k,n}^{-} = 1 - \frac{1}{2}\left\{1 - \frac{(M_{k,n}^{(1)})^2}{M_{k,n}^{(2)}}\right\}^{-1}$.

The EVI and the extremal quantile estimates are plotted against k in Figs. 9 and 10.

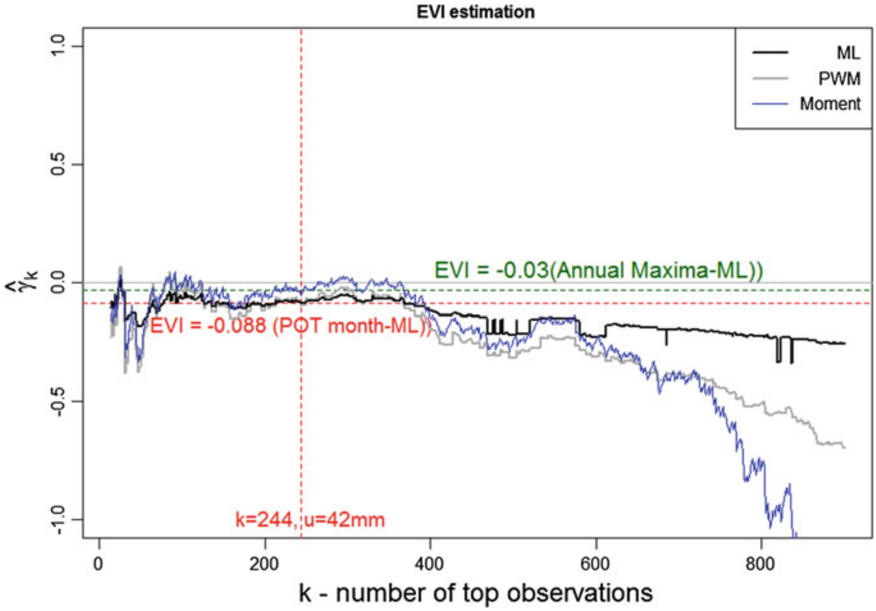


Fig. 9 Monthly maximum—semiparametrics and parametrics—EVI estimation

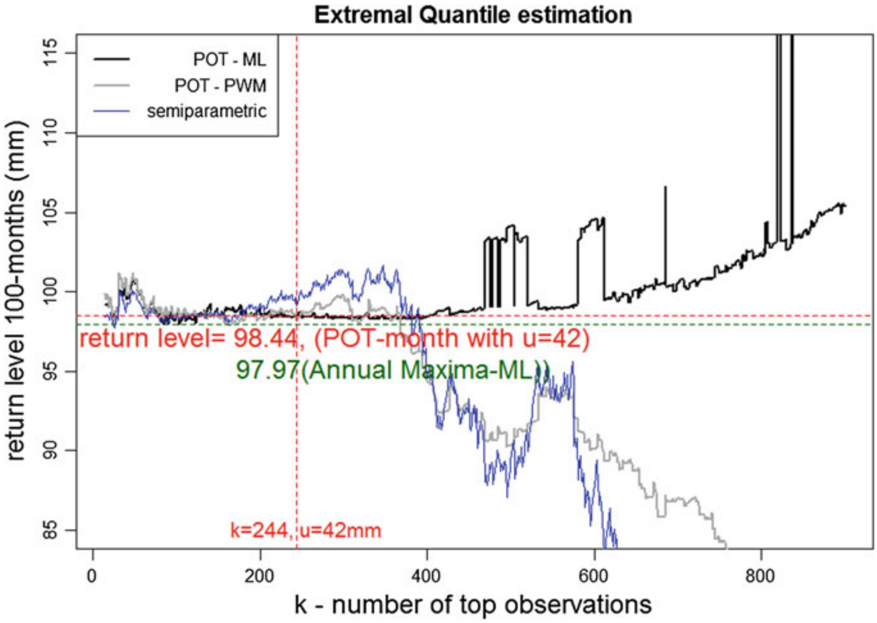


Fig. 10 Monthly maximum—semiparametrics and parametrics—extremal quantile estimation

We also represent in Fig. 9 and in Fig. 10 the sample paths of the corresponding estimates under parametric setup, POT-ML and POT-PWM of Sect. 3.4 and the threshold at $k = 244$ corresponding to $u = 42$ is also marked with a vertical dashed line. In addition, the EVI and the 100-months return level—or, equivalently, the 0.01-quantile for monthly maximum—obtained by Annual Maxima approach of Sect. 3.1 are also marked with horizontal lines (check also Tables 1, 2, 3, 4 and 5).

5 Concluding Remarks

In this paper, parametric and semi-parametric approaches have been considered to estimate parameters of rare events, such as the EVI, high return levels and small exceedance probabilities. All the aforementioned methodologies are based in approximations, which rely on asymptotic results that require that the sample size goes to infinity ($n \to \infty$). In practice, one has a sample with a finite sample size n, and the particular choice of the threshold u for POT method or the value of top number of observations k in semi-parametric setup, is not an easy task; in the first case, the theoretical results also require that the threshold goes to the right endpoint ($u = u_n \to x^F$), while in the second case, the underlying asymptotic framework considers that the top order statistic is of intermediate nature ($k = k_n \to \infty, k/n \to 0$). This rainfall case study illustrates how the different statistical approaches can contribute in practice to an overall choice of the most adequate sample region for inference in extremes. In conclusion, we should say that the parametric and semiparametric approaches do not compete. Instead, the two approaches are complementary to each other, as the last two Figs. 9 and 10 clearly show, which is validated by the flat pattern of the sample paths in a certain upper common region, where all the estimates are similar.

Acknowledgments Research partially supported by FCT: PEst-OE/MAT/UI0006/2011 and 2014, EXTREMA-PTDC/MAT/101736/2008.

References

1. Balkema, A., de Haan, L.: Residual life time at great age. Ann. Probab. **169**(2), 792–804 (1974)
2. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. Wiley, England (2004)
3. Beirlant, J., Caeiro, F., Gomes, M.I.: An overview and open research topics in statistics of univariate extreme. Revstat **10**(1), 1–31 (2012)
4. Castillo, E., Hadi, A.S., Balakrishnan, N., Sarabia, J.M.: Extreme Value and Related Models with Applications in Engineering and Science. Wiley, New York (2005)
5. Charras-Garido, M., Lezaud, P.: Extreme value analysis: an introduction. J. de la Société Française de Statistique **154**(2), 66–97 (2013)
6. Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer, London (2001)
7. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction Springer Series in Operations Research and Financial Engineering. Springer, Boston (2006)
8. Dekkers, A., Einmahl, J., de Haan, L.: A moment estimator for the index of an extreme-value distribution. Ann. Stat. **17**(4), 1833–1855 (1989)
9. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events for Insurance and Finance, 3rd edn. Springer, Berlin (2001)
10. Gnedenko, B.: Sur la distribution limite du terme maximum d'une serie aléatoire. Ann. Math. **44**(3), 423–453 (1943)
11. Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I., Pestana, D.: Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. Extremes **11**(1), 3–34 (2008)
12. Hosking, J.R.M., Wallis, J.R.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics **27**, 251–261 (1985)
13. Hosking, J.R.M., Wallis, J.R.: Parameter and quantile estimation for the generalized pareto distribution. Technometrics **29**, 339–349 (1987)
14. Nascimento, F.F.: Abordagem Bayesiana Não-paramétrica para Análise de Valores Extremos. Ph.D. Thesis. Universidade Federal do Rio de Janeiro (2009)
15. Nascimento, F.F., Gamerman, D., Lopes, H.F.: A semiparametric Bayesian approach to extreme value estimation. Stat. Comput. **22**, 661–675 (2012)
16. Pickands, J.: Statistical inference using extreme order statistics. Ann. Stat. **3**, 119–131 (1975)
17. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/
18. Reiss, R.-D., Thomas, M.: Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields, 3rd edn. Birkhauser Verlag, Boston (2007)

A Log Probability Weighted Moment Estimator of Extreme Quantiles

Frederico Caeiro and Dora Prata Gomes

Abstract In this paper we consider the semi-parametric estimation of extreme quantiles of a right heavy-tail model. We propose a new Probability Weighted Moment estimator for extreme quantiles, which is obtained from the estimators of the shape and scale parameters of the tail. Under a second-order regular variation condition on the tail, of the underlying distribution function, we deduce the non degenerate asymptotic behaviour of the estimators under study and present an asymptotic comparison at their optimal levels. In addition, the performance of the estimators is illustrated through an application to real data.

Keywords Extreme quantile · Extreme value index · Log probability weighted moment · Optimal level · Statistics of extremes

1 Introduction

Let us consider a set of n independent and identically distributed (i.i.d.), or possibly weakly dependent and stationary random variables (r.v.s), $X_1, X_2, \ldots, X_n$, with common distribution function (d.f.) F. We shall assume that $\overline{F} := 1-F$ has a Pareto-type right tail, i.e., with the notation $g(x) \sim h(x)$ if and only if $g(x)/h(x) \to 1$, as $x \to \infty$,

$$\overline{F}(x) \ \sim \ (x/C)^{-1/\gamma}, \qquad x \to \infty, \tag{1}$$

with $\gamma > 0$ and $C > 0$ denoting the shape and scale parameters, respectively. Then the quantile function $U(t) := F^{\leftarrow}(1-1/t) = \inf\{x : F(x) \geq 1-1/t\}, \ \ t > 1$ is a regularly varying function with a positive index of regular variation equal to γ, i.e.,

F. Caeiro (✉) · D. Prata Gomes
CMA and FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: fac@fct.unl.pt

D. Prata Gomes
e-mail: dsrp@fct.unl.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_22

$$\lim_{t\to\infty} U(tx)/U(t) = x^{\gamma}. \tag{2}$$

Consequentially, we are in the max-domain of attraction of the Extreme Value distribution

$$EV_{\gamma}(x) := \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), \ 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0 \\ \exp(-\exp(-x)), \qquad x \in \mathscr{R} & \text{if} \quad \gamma = 0. \end{cases} \tag{3}$$

and denote this by $F \in \mathscr{D}_M(EV_{\gamma})$. The parameter γ is called the extreme value index (EVI), the primary parameter in Statistics of Extremes.

Suppose that we are interested in the estimation of a extreme quantile q_p, a extreme value exceeded with probability $p = p_n \to 0$, small. Since $q_p = F^{\leftarrow}(1-p) \sim Cp^{-\gamma}, \ \ p \to 0$, for any heavy tailed model under (1), we will also need to deal with the estimation of the shape and scale parameters γ and C, respectively. Let $X_{n-k:n} \le \cdots \le X_{n-1:n} \le X_{n:n}$ denote the sample of the $k+1$ largest order statistics (o.s.) of the sample of size n, where $X_{n-k:n}$ is a intermediate o.s., i.e., k is a sequence of integers between 1 and n such that

$$k \to \infty \quad \text{and} \quad k/n \to 0, \quad \text{as} \quad n \to \infty. \tag{4}$$

The classic semi-parametric estimators of the parameters γ and C, introduced in [21], are

$$\hat{\gamma}^{H}_{k,n} := \frac{1}{k}\sum_{i=1}^{k}\left(\ln X_{n-i+1:n} - \ln X_{n-k:n}\right), \quad k = 1, 2, \ldots, n-1, \tag{5}$$

and

$$\hat{C}^{H}_{k,n} := X_{n-k:n}\left(\frac{k}{n}\right)^{\hat{\gamma}^{H}_{k,n}}, \quad k = 1, 2, \ldots, n-1, \tag{6}$$

respectively. The EVI estimator in (5) is the well know Hill estimator, the average of the log excesses over the high threshold $X_{n-k:n}$. The classic semi-parametric extreme quantile estimator is the Weissman-Hill estimator [23] with functional expression

$$\hat{W}^{H}_{k,n}(p) := X_{n-k:n}\left(\frac{k}{np}\right)^{\hat{\gamma}^{H}_{k,n}}, \quad k = 1, 2, \ldots, n-1. \tag{7}$$

Most classical semi-parametric estimators of parameters of the right tail usually exhibit the same type of behaviour, illustrated in Fig. 1: we have a high variance for high thresholds $X_{n-k:n}$, i.e., for small values of k and high bias for low thresholds, i.e., for large values of k. Consequently, the mean squared error (MSE) has a very peaked pattern, making it difficult to determine the optimal k, defined as the value k_0 where the MSE is minimal. For a detailed review on the subject see for instance [19] and [3].

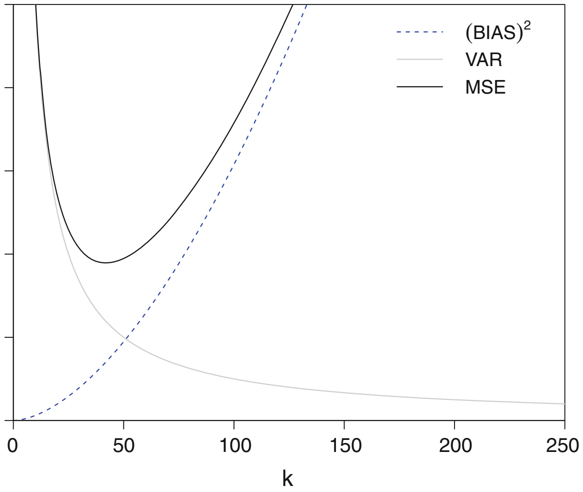


Fig. 1 Illustration of the asymptotic squared bias, variance and mean squared error patterns, as function of k, of most classical semi-parametric estimators, for a sample of size $n = 250$

Apart from the classical EVI, scale and extreme quantile estimators in (5), (6) and (7), respectively, we shall introduce in Sect. 2 the corresponding Log Pareto Probability Weighted Moment estimators. In Sect. 3, we derive their non degenerate asymptotic behaviour and present an asymptotic comparison of the estimators under study at their optimal levels.

2 Pareto Log Probability Weighted Moment Estimators

The probability weighted moments (PWM) method, introduced in [20] is a generalization of the method of moments. The PWM of a r.v. X, are defined by $M_{p,r,s} := E(X^p(F(X))^r(1-F(X))^s)$, with $p,\ r,\ s\ \in \mathscr{R}$. When $r = s = 0$, $M_{p,0,0}$ are the usual non-central moments of order p. Hosking and Wallis [22] advise the use of $M_{1,r,s}$ because the relation between parameters and moments is usually simpler than for the non-central moments. Also, if r and s are positive integers, $F^r(1-F)^s$ can be written as a linear combination of powers of F or $1-F$ and usually work with one of the moments $a_r := M_{1,0,r} = E(X(1-F(X))^r)$ or $b_r := M_{1,r,0} = E(X(F(X))^r)$. Given a sample size n, the unbiased estimators of a_r and b_r are, respectively,

$$\hat{a}_r = \frac{1}{n}\sum_{i=1}^{n-r}\frac{\binom{n-i}{r}}{\binom{n-1}{r}}X_{i:n}, \quad \text{and} \quad \hat{b}_r = \frac{1}{n}\sum_{i=r+1}^{n}\frac{\binom{i-1}{r}}{\binom{n-1}{r}}X_{i:n}.$$

The first semi-parametric Pareto PWM (PPWM) estimators for heavy tailed models appeared in [5], for the estimation of the shape and scale parameters γ and C, and in [10], for the estimation of extreme quantiles and tail probabilities. Since all those

PPWM estimators use the sample mean, they are only consistent if $0 < \gamma < 1$. Caeiro and Gomes [7] generalized the estimators in [5] with a class of PPWM estimators, consistent for $0 < \gamma < 1/r$ with $r > 0$. In order to remove the right-bounded support of the previous PPWM estimators and have consistent estimators for every $\gamma > 0$, we shall next introduce new semi-parametric estimators based on the log-moments $l_r := E((\ln X)(1 - F(X))^r)$. For non-negative integer r, the unbiased estimator of l_r is given by

$$\hat{l}_r = \frac{1}{n} \sum_{i=1}^{n-r} \frac{\binom{n-i}{r}}{\binom{n-1}{r}} \ln X_{i:n}.$$

For the strict Pareto model with d.f. $F(x) = 1 - (x/C)^{-1/\gamma}$, $x > C > 0$, $\gamma > 0$ the Pareto log PWM (PLPWM) are $l_r = \ln(C)/(1 + r) + \gamma/(1 + r)^2$.

To obtain the tail parameters estimators of γ and C of a underlying model with d.f. under (1), we need the followings results:

- $\frac{X_{n-k:n}}{C(n/k)^\gamma}$ converges in probability to 1, for intermediate k;
- the conditional distribution of $X|X > X_{n-k:n}$, is approximately Pareto with shape parameter γ and scale parameter $C(n/k)^\gamma$.

The PLPWM estimators of γ and C, based on the k largest observations, are

$$\hat{\gamma}_{k,n}^{PLPWM} := \frac{1}{k} \sum_{i=1}^{k} \left(2 - 4\frac{i-1}{k-1}\right) \ln X_{n-i+1:n}, \quad k = 2, \ldots, n, \tag{8}$$

and

$$\hat{C}_{k,n}^{PLPWM} := \left(\frac{k}{n}\right)^{\hat{\gamma}_{k,n}^{PLPWM}} \exp\left\{D_{k,n}\right\}, \quad k = 2, \ldots, n, \tag{9}$$

with $D_{k,n} := \frac{1}{k}\sum_{i=1}^{k}\left(4\frac{i-1}{k-1} - 1\right) \ln X_{n-i+1:n}$. Notice that $\hat{\gamma}_{k,n}^{PLPWM}$ is a weighted average of the k largest observations, with the weights $g_{i,k} := (2 - 4\frac{i-1}{k-1})$. Since $g_{i,k} = -g_{k-i+1,k}$, the weights are antisymmetric and their sum is zero. On the basis of the limit relation $q_p \sim Cp^{-\gamma}$, $p \to 0$, we shall also consider the following quantile estimator

$$\hat{Q}_{k,n}^{PLPWM}(p) := \left(\frac{k}{np}\right)^{\hat{\gamma}_{k,n}^{PLPWM}} \exp\left\{D_{k,n}\right\}, \quad k = 2, \ldots, n, \tag{10}$$

valid for $\gamma > 0$.

3 Asymptotic Results

3.1 Non Degenerate Limiting Distribution

In this section we derive several basic asymptotic results for the EVI estimators in (5) and (8) and for the quantiles estimators, $\hat{W}^{H}_{k,n}(p)$ and $\hat{Q}^{PLPWM}_{k,n}(p)$. Asymptotic results for the scale C-estimators are not presented but can be obtained with an analogous proof.

To ensure the consistency of the EVI semi-parametric estimators, for all $\gamma > 0$, we need to assume that k is an intermediate sequence of integers, verifying (4). To study the asymptotic behaviour of the estimators, we need a second order regular variation condition with a parameter $\rho \le 0$ that measures the rate of convergence of $U(tx)/U(t)$ to x^γ in (2) and is given by

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln x}{A(t)} = \frac{x^\rho - 1}{\rho} \Leftrightarrow \lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{11}$$

for all $x > 0$, with $|A|$ a regular varying function with index ρ and $\frac{x^\rho - 1}{\rho} = \ln x$ if $\rho = 0$.

Theorem 1 *Under the second order framework, in (11), and for intermediate k, i.e., whenever (4) holds, the asymptotic distributional representation of* $\hat{\gamma}^{\bullet}_{k,n}$, *with* $\bullet$ *denoting either H or PLPWM, is given by*

$$\hat{\gamma}^{\bullet}_{k,n} \overset{d}{=} \gamma + \frac{\sigma_\bullet Z^\bullet_k}{\sqrt{k}} + b_\bullet A(n/k)(1 + o_p(1)), \tag{12}$$

where $\overset{d}{=}$ *denotes equality in distribution,* $Z^\bullet_k$ *is a standard normal r.v.,*

$$b_H = \frac{1}{1-\rho}, \quad b_{PLPWM} = \frac{2}{(1-\rho)(2-\rho)}, \quad \sigma_H = \gamma \quad \textit{and} \quad \sigma_{PLPWM} = \frac{2}{\sqrt{3}}\gamma.$$

If we choose the intermediate level k such that $\sqrt{k}\, A(n/k) \to \lambda \in \mathscr{R}$, *then,*

$$\sqrt{k}(\hat{\gamma}^\bullet_{k,n} - \gamma) \xrightarrow{d} N(\lambda\, b_\bullet, \sigma^2_\bullet).$$

Proof For the Hill estimator, the proof can be found in [13]. For the PLPWM EVI-estimator, note that $\sum_{i=1}^{k} \left(2 - 4\frac{i-1}{k-1}\right) = 0$ and consequently

$$\hat{\gamma}^{PLPWM}_{k,n} = \frac{1}{k}\sum_{i=1}^{k}\left(2 - 4\frac{i-1}{k-1}\right) \ln \frac{X_{n-i+1:n}}{X_{n-k:n}} = \frac{1}{k}\sum_{i=1}^{k} g_{i,k} \ln \frac{X_{n-i+1:n}}{X_{n-k:n}}, \quad k < n.$$

We can write $X \overset{d}{=} U(Y)$ where Y is a standard Pareto r.v., with d.f. $F_Y(y) = 1 - 1/y$, $y > 1$. Consequently and provided that k is intermediate, we can apply Eq. (11) with $t = Y_{n-k:n}$ and $x = Y_{n-i+1:n}/Y_{n-k:n} \overset{d}{=} Y_{k-i+1:k}$, $1 \le i \le k$, to obtain

$$\ln \frac{X_{n-i+1:n}}{X_{n-k:n}} \overset{d}{=} \gamma \ln Y_{k-i+1:k} + \frac{Y_{k-i+1:k}^{\rho} - 1}{\rho} A(Y_{n-k:n})(1 + o_p(1)).$$

Then, since $nY_{n-k:n}/k \xrightarrow{p} 1$, as $n \to \infty$,

$$\hat{\gamma}_{k,n}^{PLPWM} \overset{d}{=} \frac{1}{k} \sum_{i=1}^{k} g_{i,k} \left\{ \gamma E_{k-i+1:k} + \frac{Y_{k-i+1:k}^{\rho} - 1}{\rho} A(n/k)(1 + o_p(1)) \right\},$$

where $\{E_i\}_{i \ge 1}$, denotes a sequence of i.i.d. standard exponential r.v.'s. The distributional representation of the EVI-estimator $\hat{\gamma}_{k,n}^{PLPWM}$ follows from the results for linear functions of ordinal statistics [12], i.e., $Z_k^{PLPWM} = \frac{\sqrt{k}}{\sigma_{PLPWM}} \frac{1}{k} \sum_{i=1}^{k} (g_{i,k} E_{k-i+1:k} - 1)$ is a standard normal r.v. and $\frac{1}{k} \sum_{i=1}^{k} g_{i,k} \frac{Y_{k-i+1:k}^{\rho} - 1}{\rho}$ converges in probability towards $\frac{2}{(1-\rho)(2-\rho)}$, as $k \to \infty$.

The asymptotic normality of $\sqrt{k}(\hat{\gamma}_{k,n}^{\bullet} - \gamma)$ follows straightforward from (12).

Remark 1 Notice that $\hat{\gamma}_{k,n}^{PLPWM}$ has a smaller asymptotic bias, but a larger asymptotic variance than $\hat{\gamma}_{k,n}^{H}$. A more precise comparison will be dealt in Sect. 3.2.

Remark 2 For intermediate k such that $\sqrt{k}A(n/k) \to \lambda$, finite, as $n \to \infty$, the Asymptotic Mean Squared Error (AMSE) of any semi-parametric EVI-estimator, with asymptotic distributional representation given by (12), is $AMSE(\widehat{\gamma}_{n,k}^{\bullet}) := \frac{\sigma_\bullet^2}{k} + b_\bullet^2 A^2(n/k)$, where $Bias_\infty(\widehat{\gamma}_{n,k}^{\bullet}) := b_\bullet A(n/k)$ and $Var_\infty(\widehat{\gamma}_{n,k}^{\bullet}) := \sigma_\bullet^2/k$. Let $k_0^\bullet$ denote the level k, such that $AMSE(\widehat{\gamma}_{n,k}^{\bullet})$ is minimal, i.e., $k_0^\bullet \equiv k_0^\bullet(n) := \arg\min_k AMSE(\widehat{\gamma}_{n,k}^{\bullet})$. If $A(t) = \gamma\beta t^\rho,\ \ \beta \neq 0,\ \rho < 0$ which holds for most common heavy tailed models, like the Fréchet, Burr, Generalized Pareto or Student's t, the optimal k-value for the EVI-estimation through $\hat{\gamma}_{n,k}^{\bullet}$ is well approximated by

$$k_0^\bullet = \left(\frac{\sigma_\bullet^2 n^{-2\rho}}{(-2\rho) b_\bullet^2 \gamma^2 \beta^2} \right)^{\frac{1}{1-2\rho}}. \tag{13}$$

Remark 3 The estimation of the shape second-order parameter ρ can be done using the classes of estimators in [11, 16, 17] or [8]. Consistency of those estimators is achieved for intermediate k such that $\sqrt{k}A(n/k) \to \infty$ as $n \to \infty$. For the estimation of the scale second-order parameter β, for models with $A(t) = \gamma\beta t^\rho$, $\beta \neq 0$, $\rho < 0$, we refer the reader to the estimator in [18]. That estimator is consistent for intermediate k such that $\sqrt{k}A(n/k) \to \infty$ as $n \to \infty$ and estimators of ρ such that $\hat{\rho} - \rho = o_p(1/\ln n)$. Further details on the estimation of (ρ,β) can be found in [9].

For the extreme quantile estimators in (7) and (10), their asymptotic distributional representations follows from the next, more general, Theorem.

Theorem 2 *Suppose that* $\bullet$ *denotes either H or PLPWM EVI-estimators with distributional representation given by (12). Under the conditions of Theorem 1, if* $p = p_n$ *is a sequence of probabilities such that* $c_n := k/(np) \to \infty$, $\ln c_n = o(\sqrt{k})$ *and* $\sqrt{k}A(n/k) \to \lambda \in \mathscr{R}$, *as* $n \to \infty$, *then,*

$$\frac{\sqrt{k}}{\ln c_n}\left(\frac{\hat{Q}^{\bullet}_{k,n}(p)}{q_p} - 1\right) \overset{d}{=} \frac{\sqrt{k}}{\ln c_n}\left(\frac{\hat{W}^{\bullet}_{k,n}(p)}{q_p} - 1\right) \overset{d}{=} \sqrt{k}\left(\hat{\gamma}^{\bullet}_{k,n} - \gamma\right)(1 + o_p(1)). \tag{14}$$

Proof Since $q_p = U(1/p)$, we can write

$$\frac{\hat{W}^{\bullet}_{k,n}(p)}{q_p} = \frac{X_{n-k:n}}{U(n/k)} \cdot \frac{U(n/k)}{U(nc_n/k)}(c_n)^{\hat{\gamma}^{\bullet}_{k,n}}.$$

Using the second order framework, in (11), with $t = n/k$ and $x = \frac{k}{n}Y_{n-k:n}$, results in $\frac{X_{n-k:n}}{U(n/k)} \overset{d}{=} 1 + \frac{\gamma}{\sqrt{k}}B_k + o_p(A(n/k))$ where $B_k := \sqrt{k}\left(\frac{k}{n}Y_{n-k:n} - 1\right)$ is asymptotically a standard normal random variable. Using the results in [14], Remark B.3.15 (p. 397), $\left(\frac{U(c_n n/k)}{U(n/k)c_n^{\gamma}}\right)^{-1} = 1 + \frac{A(n/k)}{\rho}(1 + o(1))$ follows. Then, since $(c_n)^{\hat{\gamma}^{\bullet}_{k,n} - \gamma} \overset{d}{=} 1 + \ln(c_n)(\hat{\gamma}^{\bullet}_{k,n} - \gamma)(1 + o_p(1))$, we get

$$\frac{\hat{W}^{\bullet}_{k,n}(p)}{q_p} \overset{d}{=} 1 + \ln(c_n)(\hat{\gamma}^{\bullet}_{k,n} - \gamma)(1 + o_p(1)) + \frac{\gamma B_k}{\sqrt{k}} + \frac{A(n/k)}{\rho}(1 + o_p(1)),$$

and the second equality in (14) follows immediately.

For the other quantile estimator, we can write

$$\hat{Q}^{\bullet}_{k,n}(p) = X_{n-k:n}\left(\frac{k}{np}\right)^{\hat{\gamma}^{\bullet}_{k,n}} \exp\{\tilde{D}_{k,n}\} = \hat{W}^{\bullet}_{k,n}(p)\exp\{\tilde{D}_{k,n}\},$$

with $\tilde{D}_{k,n} := \frac{1}{k}\sum_{i=1}^{k}\left(4\frac{i-1}{k-1} - 1\right)\ln\frac{X_{n-i+1:n}}{X_{n-k:n}}$. Then, since we have

$$\exp\{\tilde{D}_{k,n}\} \overset{d}{=} 1 + \frac{\gamma}{\sqrt{3k}}P_k - \frac{\rho A(n/k)(1 + o_p(1))}{(1-\rho)(2-\rho)},$$

with P_k a standard normal r.v., the first equality in (14) follows.

3.2 Asymptotic Comparison at Optimal Levels

We now proceed to an asymptotic comparison of the PLPWM EVI estimator in (8) with the Hill estimator in (5) and the PPWM EVI estimator in [5], at their optimal levels. This comparison is done along the lines of [6, 13], among others. Similar results hold for the extreme quantile estimators, at their optimal levels, since they have the same asymptotic behaviour as the EVI estimators, although with a slower convergence rate.

Let $k_0^\bullet$ be the optimal level for the estimation of γ through $\widehat{\gamma}_{k,n}^\bullet$ given by (13), i.e., the level associated with a minimum asymptotic mean square error, and let us denote $\widehat{\gamma}_{n0}^\bullet := \widehat{\gamma}_{k_0^\bullet,n}^\bullet$, the estimator computed at its optimal level. Dekkers and de Haan [15] proved that, whenever $b_\bullet \neq 0$, there exists a function $\varphi(n; \gamma, \rho)$, dependent only on the underlying model, and not on the estimator, such that

$$\lim_{n\to\infty} \varphi(n; \gamma, \rho) AMSE(\widehat{\gamma}_{n0}^\bullet) = \left(\sigma_\bullet^2\right)^{-\frac{2\rho}{1-2\rho}} \left(b_\bullet^2\right)^{\frac{1}{1-2\rho}} =: LMSE(\widehat{\gamma}_{n0}^\bullet). \tag{15}$$

It is then sensible to consider the following:

Definition 1 Given two biased estimators $\widehat{\gamma}_{n,k}^{(1)}$ and $\widehat{\gamma}_{n,k}^{(2)}$, for which distributional representations of the type (12) hold with constants (σ_1, b_1) and (σ_2, b_2), $b_1, b_2 \neq 0$, respectively, both computed at their optimal levels, $k_0^{(1)}$ and $k_0^{(2)}$, the Asymptotic Root Efficiency (*AREFF*) indicator is defined as

$$AREFF_{1|2} := \sqrt{\frac{LMSE\left(\widehat{\gamma}_{n0}^{(2)}\right)}{LMSE\left(\widehat{\gamma}_{n0}^{(1)}\right)}} = \left(\left(\frac{\sigma_2}{\sigma_1}\right)^{-2\rho} \left|\frac{b_2}{b_1}\right|\right)^{\frac{1}{1-2\rho}}, \tag{16}$$

with LMSE given in (15) and $\widehat{\gamma}_{n0}^{(i)} := \widehat{\gamma}_{k_0^{(i)},n}^{(i)}$, $i = 1, 2$.

Remark 4 Note that this measure was devised so that the higher the AREFF indicator is, the better the first estimator is.

Remark 5 For the PPWM EVI estimator, in [5], we have

$$b_{PPWM} = \frac{(1-\gamma)(2-\gamma)}{(1-\gamma-\rho)(2-\gamma-\rho)} \text{ and } \sigma_{PPWM} = \frac{\gamma\sqrt{1-\gamma}(2-\gamma)}{\sqrt{1-2\gamma}\sqrt{3-2\gamma}}, \quad 0 < \gamma < 0.5.$$

To measure the performance of $\hat{\gamma}_{k,n}^{PLPWM}$, we have computed the AREFF-indicator, in (16), as function of the second order parameter ρ. In Fig. 2 (left), we present the values of

$$AREFF_{PLPWM|H}(\rho) = \left((3/4)^{-\rho}\,(1-\rho/2)\right)^{\frac{1}{1-2\rho}}, \tag{17}$$

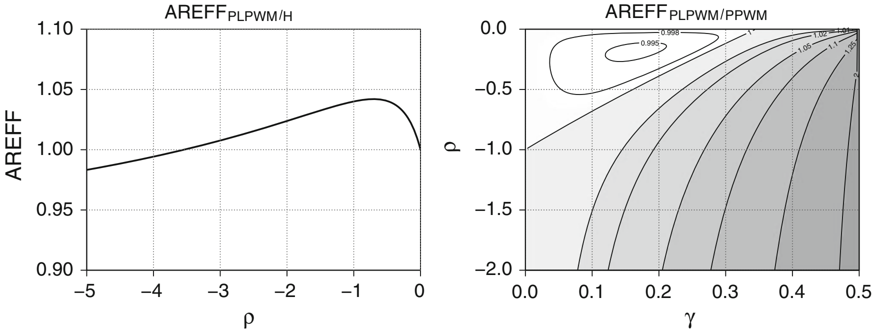


Fig. 2 *Left* Plot with the indicator $AREFF_{PLPWM|H}(\rho)$, in (17), as a function of ρ. *Right* Contour plot with the indicator $AREFF_{PLPWM|PWM}$, as a function of (γ,ρ)

as a function of ρ. This indicator has a maximum near $\rho = -0.7$, and we have $AREFF_{PLPWM|H} > 1$, if $-3.54 < \rho < 0$, an important region of ρ values in practical applications. It is also easy to check that $\lim_{\rho\to-\infty} AREFF_{PLPWM|H}(\rho) = \sqrt{3}/2 \approx 0.866$ and $\lim_{\rho\to 0} AREFF_{PLPWM|H}(\rho) = 1$.

In Fig. 2 (right) we show a contour plot with the comparative behaviour, at optimal levels, of the PLPWM and PPWM EVI-estimators in an important region of the (γ,ρ)-plane. The grey colour marks the area where $AREFF_{PLPWM|PPWM} > 1$. At optimal levels, there is only a small region of the (γ,ρ)-plane where the AREFF indicator is slightly smaller than 1. Also, the $AREFF_{PLPWM|PPWM}$ indicator increases, as γ increases and/or ρ decreases.

4 A Case Study

As an illustration of the performance of the estimators, we shall consider the analysis of the Secura Belgian Re automobile claim amounts exceeding 1,200,000 Euro, over the period 1988–2001. This data set of size $n = 371$ was already studied by several authors [1, 2, 4].

In Fig. 3, we present, at the left, the EVI estimates provided by the Hill and PLPWM EVI-estimators in (5) and (8), respectively. At the right we present the corresponding quantile estimates provided by Weissman-Hill and PLPWM estimators, in (7) and (10), with $p = 0.001$. For a fair comparison of the PLPWM estimators with the equivalent classic estimators, the PLPWM estimators are now based on the top $k + 1$ largest o.s.'s. For this dataset, we have $\hat{\rho} = -0.756$ and $\hat{\beta} = 0.803$, obtained at the level $k_1 = [n^{0.999}] = 368$ [4]. Using these values, the estimates of the optimal level, given by (13), are $\hat{k}_0^H = 55$ and $\hat{k}_0^{PLPWM} = 76$. Consequently,

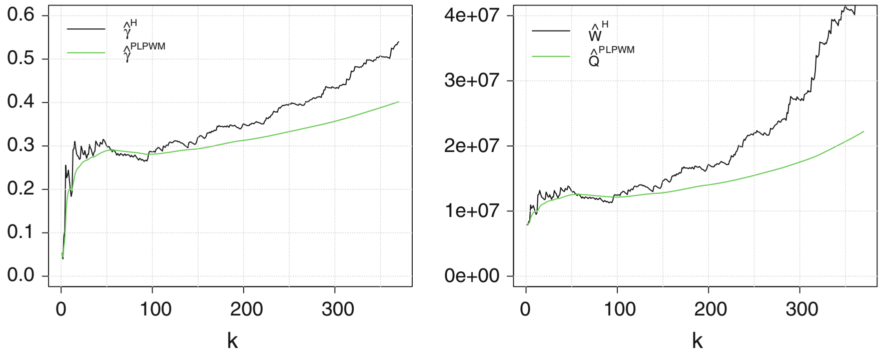


Fig. 3 *Left* Estimates of the EVI for the Secura Belgian Re data; *Right* Estimates of the quantile q_p with $p = 0.001$ for the Secura Belgian Re data

we have $\hat{\gamma}^{H}_{55,371} = 0.291$ and $\hat{\gamma}^{PLPWM}_{76,371} = 0.286$. Finally, the quantile estimates are given by $\hat{W}^{H}_{55,371}(p) = 12622248$ and $\hat{Q}^{PLPWM}_{76,371}(p) = 12373324$.

5 Some Overall Conclusions

Based on the results here presented we can make the following comments:

- Regarding efficiency at optimal levels, the new PLPWM estimators are a valid alternative to the classic Hill, Weissman-Hill and PPWM estimators. And they are consistent for any $\gamma > 0$, which does not happen for the PPWM estimators.
- The analysis of the automobile claim amounts gave us the impression that the PLPWM EVI and extreme quantile estimators have a much smoother sample pattern than the Hill and the Weissman-Hill estimators.
- It is also important to study the behaviour of the new PLPWM estimators for small sample sizes. That topic should be addressed in future research work.

Acknowledgments Research partially supported by FCT – Fundação para a Ciência e a Tecnologia, project UID/MAT/00297/2013 (CMA/UNL), EXTREMA, PTDC/MAT /101736/2008.

References

1. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes. Theory and Applications. Wiley, New York (2004)
2. Beirlant, J., Figueiredo, F., Gomes, M.I., Vandewalle, B.: Improved reduced-bias tail index and quantile estimators. J. Stat. Plan. Inference **138**(6), 1851–1870 (2008)
3. Beirlant, J., Caeiro, F., Gomes, M.I.: An overview and open research topics in statistics of univariate extremes. Revstat **10**(1), 1–31 (2012)

4. Caeiro, F., Gomes, M.I.: Computational validation of an adaptative choice of optimal sample fractions. In: 58th World Statistics Congress of the International Statistical Institute, Dublin, pp. 282–289 (2011)
5. Caeiro, F., Gomes, M.I.: Semi-parametric tail inference through probability-weighted moments. J. Stat. Plan. Inference **141**, 937–950 (2011)
6. Caeiro, F., Gomes, M.I.: Asymptotic comparison at optimal levels of reduced-bias extreme value index estimators. Stat. Neerl. **65**, 462–488 (2011)
7. Caeiro, F., Gomes, M.I.: A class of semi-parametric probability weighted moment estimators. In: Oliveira, P.E., da Graça Temido, M., Henriques, C., Vichi, M. (eds.) Recent Developments in Modeling and Applications in Statistics, pp. 139–147. Springer, Berlin (2013)
8. Caeiro, F. and Gomes, M.I.: A semi-parametric estimator of a shape second order parameter. In: Pacheco, A., Santos, R., Rosário Oliveira, M. and Paulino, C.D. (eds.) New Advances in Statistical Modeling and Applications, Studies in Theoretical and Applied Statistics, pp. 137–144, Springer (2014)
9. 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)
10. Caeiro, F., Gomes, M.I., Vandewalle, B.: Semi-parametric probability-weighted moments estimation revisited. Methodol. Comput. Appl. 16(1), 1–29 (2014)
11. Ciuperca, G., Mercadier, C.: Semi-parametric estimation for heavy tailed distributions. Extremes **13**(1), 55–87 (2010)
12. David, H., Nagaraja, H.N.: Order Statistics. Wiley, New York (2003)
13. de Haan, L., Peng, L.: Comparison of tail index estimators. Stat. Neerl. **52**, 60–70 (1998)
14. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction. Springer, New York (2006)
15. Dekkers, A., de Haan, L.: Optimal sample fraction in extreme value estimation. J. Multivar. Anal. **47**(2), 173–195 (1993)
16. Fraga Alves, M.I., Gomes, M.I., de Haan, L.: A new class of semi-parametric estimators of the second order parameter. Port. Math. **60**(2), 193–213 (2003)
17. Goegebeur, Y., Beirlant, J., de Wet, T.: Kernel estimators for the second order parameter in extreme value statistics. J. Stat. Plan. Inference **140**, 2632–2652 (2010)
18. Gomes, M.I., Martins, M.J.: "Asymptotically unbiased" estimators of the tail index based on external estimation of the second order parameter. Extremes **5**(1), 5–31 (2002)
19. Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I., Pestana, D.D.: Statistics of extremes for IID data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. Extremes **11**(1), 3–34 (2008)
20. Greenwood, J.A., Landwehr, J.M., Matalas, N.C., Wallis, J.R.: Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour. Res. **15**, 1049–1054 (1979)
21. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
22. Hosking, J., Wallis, J.: Parameter and quantile estimation for the generalized pareto distribution. Technometrics **29**(3), 339–349 (1987)
23. Weissman, I.: Estimation of parameters of large quantiles based on the *k* largest observations. J. Am. Stat. Assoc. **73**, 812–815 (1978)

A Mean-of-Order-p Class of Value-at-Risk Estimators

M. Ivette Gomes, M. Fátima Brilhante and Dinis Pestana

Abstract The main objective of *statistics of univariate extremes* lies in the estimation of quantities related to extreme events. In many areas of application, like *finance*, *insurance* and *statistical quality control*, a typical requirement is to estimate a *high quantile*, i.e. the *Value at Risk* at a level $q\,(\mathrm{VaR}_q)$, high enough, so that the chance of exceedance of that value is equal to q, with q small. In this paper we deal with the semi-parametric estimation of VaR_q, for heavy tails, introducing a new class of VaR-estimators based on a class of *mean-of-order-p* (MOP) *extreme value index* (EVI)-estimators, recently introduced in the literature. Interestingly, the MOP EVI-estimators can have a mean square error smaller than that of the classical EVI-estimators, even for small values of k. They are thus a nice basis to build alternative VaR-estimators not only around optimal levels, but for other levels too.The new VaR-estimators are compared with the classical ones, not only asymptotically, but also for finite samples, through Monte-Carlo techniques.

Keywords Heavy right tails · Semi-parametric estimation · Statistics of extremes · Value-at-risk estimation

M.I. Gomes (✉) · D. Pestana · M.F. Brilhante
Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa,
Lisbon, Portugal
e-mail: ivette.gomes@fc.ul.pt

M.I. Gomes · D. Pestana
Instituto de Investigação Científica Bento da Rocha Cabral, Lisbon, Portugal
e-mail: dinis.pestana@fc.ul.pt

M.F. Brilhante
Universidade dos Açores (DM), Ponta Delgada, Portugal
e-mail: fbrilhante@uac.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_23

1 Introduction and Scope of the Paper

A relevant situation in risk management is the risk of a big loss that occurs rarely or even very rarely. Such a risk is generally expressed as the *Value at Risk* (VaR), i.e. the size of the loss that occurred with a fixed small probability, q. We are thus dealing with a (high) *quantile*,

$$\chi_{1-q} \equiv \mathrm{VaR}_q := F^{\leftarrow}(1-q),$$

of an unknown cumulative distribution function (CDF) F, with $F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$ denoting the generalized inverse function of F. As usual, let us denote by U the generalized inverse function of $1/(1-F)$. Then, for small q, we want to estimate the parameter

$$\mathrm{VaR}_q = U\,(1/q)\,, \quad q = q_n \to 0, \quad nq_n \leq 1,$$

i.e. we want to extrapolate beyond the sample. Since in real applications one often encounters heavy tails, we shall assume that the CDF underlying the data satisfies

$$1 - F(x) \;\sim\; c\,x^{-1/\xi}, \quad \text{as } x \to \infty, \tag{1}$$

for some positive constant c. Equivalently, and for some $C > 0$,

$$U(t) \;\sim\; C\,t^{\xi}, \quad \text{as } t \to \infty, \tag{2}$$

where the notation $a(y) \sim b(y)$ means that $a(y)/b(y) \to 1$, as $y \to \infty$. The parameter ξ in either (1) or (2) is the *extreme value index* (EVI), the primary parameter of extreme (and large) events.

Generally [18], if we consider a random sample $(X_1, \ldots, X_n)$ from F and if we can find attraction coefficients (a_n, b_n), with $a_n > 0$ and $b_n \in \mathbb{R}$, such that the sequence of suitably normalized maxima, $\{(X_{n:n} - b_n)/a_n\}_{n\geq 1}$, converges to a non-degenerate random variable (RV), then such a RV is compulsory of the type of a general *extreme value* (EV) RV, with CDF

$$\mathrm{EV}_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}),\ 1+\xi x > 0, & \text{if } \xi \neq 0, \\ \exp(-\exp(-x)),\ x > 0, & \text{if } \xi = 0. \end{cases} \tag{3}$$

We then say that F is in the max-domain of attraction of EV_ξ, and use the notation $F \in \mathscr{D}_{\mathscr{M}}(\mathrm{EV}_\xi)$. If (1) holds, or equivalently (2) holds, the limit law in (3) also appears, but with $\xi > 0$.

Weissman [34] proposed the following semi-parametric VaR-estimator:

$$Q^{(q)}_{\hat{\xi}}(k) := X_{n-k:n}\left(\frac{k}{nq}\right)^{\hat{\xi}}, \tag{4}$$

where $X_{n-k:n}$ is the $(k+1)$th top order statistic (o.s.), $\hat{\xi}$ any consistent estimator for ξ and Q stands for quantile. Further details on semi-parametric estimation of extremely high quantiles for any real EVI can be found in de Haan and Rootzén [13] and Ferreira et al. [15]. For heavy right-tails, Gomes and Figueiredo [23], Matthys and Beirlant [31] Mathys et al. [32], Gomes and Pestana [24] and Caeiro and Gomes [8, 9], among others, dealt with reduced bias VaR-estimation, a topic beyond the scope of this paper.

The estimator in (4) is an *asymptotic* estimator, in the sense that it provides useful estimates when the sample size n is high. Also, and as usual in semi-parametric estimation of parameters of extreme events, we need to work with an *intermediate* sequence of integers,

$$k = k_n \to \infty, \quad k \in [1, n), \qquad k = o(n) \qquad \text{as} \quad n \to \infty. \tag{5}$$

For heavy tails, the classical EVI-estimator, usually the one which is used in (4), for a semi-parametric quantile estimation, is the Hill estimator $\hat{\xi} = \hat{\xi}(k) =: \mathrm{H}(k)$ [30], with the functional expression,

$$\mathrm{H}(k) := \frac{1}{k}\sum_{i=1}^{k} V_{ik}, \quad V_{ik} = \ln\frac{X_{n-i+1:n}}{X_{n-k:n}}, \ 1 \le i \le k. \tag{6}$$

If we plug in the Hill estimator $\mathrm{H}(k)$ in (4), we get the so-called Weissman-Hill quantile or VaR_q-estimator, with the obvious notation, $Q^{(q)}_{\mathrm{H}}(k)$. Since $Q^{(q)}_{\mathrm{H}}(k)$ is skewed (see [24]), it is advisable to work with the lnVaR estimator

$$\ln Q^{(q)}_{\hat{\xi}}(k) = \ln X_{n-k:n} + \hat{\xi}(k)\ \ln\left(\frac{k}{nq}\right), \tag{7}$$

for any consistent EVI-estimator, $\hat{\xi}(k)$. Again, if we plug $\mathrm{H}(k)$ into (7), we get the so-called Weissman-Hill lnVaR estimator, with the obvious notation $\ln Q^{(q)}_{\mathrm{H}}(k)$.

In order to be able to study the asymptotic behavior of $\ln Q^{(q)}_{\mathrm{H}}(k)$, as well as of alternative $\ln\mathrm{VaR}_q$-estimators, it is useful to impose a second-order expansion on the tail function $1 - F$ or on the function U. Here we shall assume that we are working in Hall-Welsh class of models [29], where, as $t \to \infty$ and with $C,\ \xi > 0$, $\rho < 0$ and β non-zero,

$$U(t) = Ct^{\xi}\left(1 + A(t)/\rho + o\left(t^{\rho}\right)\right), \qquad A(t) = \xi\ \beta\ t^{\rho}. \tag{8}$$

The class in (8) is a wide class of models, that contains most of the heavy-tailed parents useful in applications, like the *Fréchet*, the *Generalized Pareto* and the *Student*-t_ν, with ν degrees of freedom. Indeed, (8) implies either (1) or (2).

Since

$$\mathrm{H}(k) = \sum_{i=1}^{k} \ln \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k} = \ln \left(\prod_{i=1}^{k} \frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k}, \quad 1 \le i \le k < n,$$

we observe that the Hill estimator is the logarithm of the *geometric mean* (or *mean-of-order*-0) of $U_{ik} := X_{n-i+1:n}/X_{n-k:n},\ 1 \le i \le k < n$. More generally, Brilhante et al. [3] considered as basic statistics the *mean-of-order-p* (MOP) of U_{ik}, $1 \le i \le k$, $p \in \mathbb{R}_0^+$, i.e., the class of statistics

$$\mathrm{A}_p(k) = \begin{cases} \left(\frac{1}{k} \sum\limits_{i=1}^{k} U_{ik}^p \right)^{1/p}, & \text{if } p > 0, \\ \left(\prod\limits_{i=1}^{k} U_{ik} \right)^{1/k}, & \text{if } p = 0, \end{cases}$$

and the following class of EVI-estimators:

$$\mathrm{H}_p(k) \equiv \mathrm{MOP}(k) \equiv \hat{\xi}^{\mathrm{H}_p}(k) := \begin{cases} \left(1 - \mathrm{A}_p^{-p}(k)\right)/p, \text{ if } 0 < p < 1/\xi, \\ \ln \mathrm{A}_0(k) = \mathrm{H}(k), \text{ if } p = 0, \end{cases} \tag{9}$$

with $\mathrm{H}_0(k) \equiv \mathrm{H}(k)$, given in (6). This class of MOP EVI-estimators, studied in Brilhante et al. [3], depends now on this *tuning* parameter $p \ge 0$, and was shown to be highly flexible. Note that the restriction $0 < p < 1/\xi$ in (9) ensures the consistency of the MOP EVI-estimators.

The aim of this paper is to find the asymptotic and finite sample properties of alternative estimators for $\ln\mathrm{VaR}_q$, replacing, in (7), $\ln Q_{\hat{\xi}}^{(q)}(k)$ by the new $\ln\mathrm{VaR}_q$ estimators $\ln Q_{\mathrm{H}_p}^{(q)}(k)$ based on the MOP EVI-estimator, $\mathrm{H}_p(k)$, in (9). If we choose the value of p that provides the highest asymptotic efficiency for $\mathrm{H}_p(k)$ (see [4]), the new estimators have an asymptotic mean square error (MSE) smaller than the Weissman-Hill lnVaR-estimators for all k. Consequently, they are alternatives to the previous estimators not only around optimal levels but for all k. The outline of the paper is as follows. In Sect. 2, we briefly discuss general first and second-order frameworks under a heavy-tailed set-up. The classes of EVI and VaR-estimators under study are discussed in Sect. 3, where we also deal with asymptotic properties of the EVI and lnVaR-estimators under consideration. Section 4 is devoted to a Monte-Carlo simulation, that enables the derivation of the distributional properties of the new classes of MOP lnVaR-estimators. Finally, in Sect. 5, we provide some general remarks on the topic.

2 A Brief Review of General First and Second-Order Conditions for Heavy Right Tails

In the area of *statistics of extremes* and whenever working with large values, i.e. with the right tail of the model F underlying the data, a model F is usually said to be *heavy-tailed* whenever the right tail-function,

$$\overline{F} := 1 - F$$

is a regularly varying function with a negative index of regular variation equal to $-1/\xi$, $\xi > 0$. We then use the notation $\overline{F} \in \mathscr{R}_{-1/\xi}$. Note that a regularly varying function with an index of regular variation equal to $a \in \mathbb{R}$, i.e. an element of $\mathscr{R}_a$, is a positive measurable function $g(\cdot)$ such that for all $x > 0$, $g(tx)/g(t) \to x^a$, as $t \to \infty$ (see [2], for details on regular variation). Heavy-tailed models are thus such that $\overline{F}(x) = x^{-1/\xi} L(x)$, $\xi > 0$, with $L \in \mathscr{R}_0$, a *regularly varying* function with an *index of regular variation* equal to zero, also called a *slowly varying* function at infinity. Equivalently, with $F^{\leftarrow}(x) := \inf\{y : F(y) \geq x\}$, the *reciprocal tail quantile function* $U(t) := F^{\leftarrow}(1 - 1/t)$, $t \geq 1$, is of regular variation with index ξ [11], i.e. $U \in \mathscr{R}_\xi$. If either (1) or (2) holds, the slowly varying function $L(\cdot)$ behaves as a constant.

We thus have the validity of any of the equivalent and general first-order conditions,

$$F \in \mathscr{D}_{\mathscr{M}}^{+} := \mathscr{D}_{\mathscr{M}}\left(EV_\xi\right)_{\xi>0} \quad \Longleftrightarrow \quad \overline{F} \in RV_{-1/\xi} \quad \Longleftrightarrow \quad U \in RV_\xi. \tag{10}$$

The second-order parameter ρ (≤ 0) measures the rate of convergence in the general first-order conditions, in (10), and can be defined as the non-positive parameter in the limiting relation,

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \begin{cases} \frac{x^\rho - 1}{\rho}, & \text{if } \rho < 0, \\ \ln x, & \text{if } \rho = 0, \end{cases}$$

$x > 0$, and where $|A|$ must be of regular variation with an index ρ [17]. This condition has been widely accepted as an appropriate condition to specify the right tail of a Pareto-type distribution in a semi-parametric way and easily enables the derivation of the non-degenerate bias of EVI and VaR-estimators, under a semi-parametric framework. Further developments of the topic can be found in de Haan and Ferreira [14]. If we consider only negative values of ρ, we are then in the class of models in (8), or equivalently either (1) or (2) holds.

3 EVI and VaR-Estimators Under Heavy-Tailed Frameworks: Asymptotic Behavior

Let $\mathcal{N}(\mu, \sigma^2)$ stand for a normal RV with mean value μ and variance σ^2. In Sect. 3.1 we deal with known results on the asymptotic behavior of the EVI-estimators under consideration. A parallel but new study is performed in Sect. 3.2 for the VaR-estimators.

3.1 The EVI-Estimators

It follows from the results of de Haan and Peng [12] that in Hall-Welsh class of models in (8), and for intermediate k-values, i.e. if (5) holds,

$$\sqrt{k}\,(\mathrm{H}(k) - \xi) \overset{d}{=} \mathcal{N}\left(0, \xi^2\right) + \sqrt{k}\left(\frac{\xi\ \beta\ (n/k)^\rho}{1-\rho}\right)(1 + o_p(1)), \tag{11}$$

where the bias $\xi\ \beta\ \sqrt{k}\ (n/k)^\rho/(1-\rho)$ can be very large, moderate or small, i.e. go to infinity, constant or zero, as $n \to \infty$.

Just as proved in Brilhante et al. [3], the result in (11) can be generalized. Under the same conditions as above, for $0 \le p < 1/(2\xi)$, and with $\mathrm{H}_p(k)$ given in (9),

$$\sqrt{k}\left(\mathrm{H}_p(k) - \xi\right) \overset{d}{=} \mathcal{N}\left(0, \frac{\xi^2(1-p\xi)^2}{1-2p\xi}\right) + \sqrt{k}\left(\frac{\xi\ \beta\ (n/k)^\rho(1-p\xi)}{1-\rho-p\xi}\right)(1 + o_p(1)). \tag{12}$$

3.2 Extreme Quantile or VaR-Estimators

Under condition (8), the asymptotic behavior of $\ln Q_{\mathrm{H}}^{(q)}(k)$ is well-known [34]:

$$\frac{\sqrt{k}}{\ln(k/(nq))}\left(\ln Q_{\mathrm{H}}^{(q)}(k) - \ln \mathrm{VaR}_q\right) \overset{d}{=} \mathcal{N}\left(\frac{\lambda}{1-\rho}, \xi^2\right),$$

provided that the sequence $k = k_n$ satisfies the condition $\lim\limits_{n\to\infty} \sqrt{k}A(n/k) = \lambda \in \mathbb{R}$, finite, with $A(\cdot)$ the function in (8).

Regarding VaR-estimation, we shall here consider, as possible alternatives to the classical Weissman-Hill lnVaR-estimator, $\ln Q_{\mathrm{H}}^{(q)}(k)$, the class of estimators

$$\ln Q_{\mathrm{H}_p}^{(q)}(k) := \ln X_{n-k:n} + \mathrm{H}_p(k)\ \ln\left(\frac{k}{nq}\right), \tag{13}$$

with H_p given in (9).

As previously mentioned, for intermediate k, i.e., whenever (5) holds, we are dealing with semi-parametric $\ln \mathrm{VaR}_q$ estimators, of the type of $\ln Q^{(q)}_{\hat{\xi}}$ in (7), where $\hat{\xi} \equiv \hat{\xi}(k)$ can be any semi-parametric estimator of the tail index ξ, and here it is taken to be the MOP EVI-estimator in (9). We may state the following:

Theorem 1 *In* Hall-Welsh *class of models in* (8)*, for intermediate k, i.e. k-values such that* (5) *holds, whenever*

$$\ln\,(n\, q_n) = o\left(\sqrt{k}\right), \tag{14}$$

$\sqrt{k}\; A(n/k) \rightarrow \lambda$*, finite, possibly non-null, and for any* $p < 1/(2\xi)$

$$\frac{\sqrt{k}}{\ln(k/(nq))}\left(\ln Q^{(q)}_{\mathrm{H}_p}(k) - \ln \mathrm{VaR}_q\right) \xrightarrow[n\to\infty]{d} \mathscr{N}\left(0, \frac{\xi^2(1-p\xi)^2}{1-2p\xi}\right), \tag{15}$$

with H_p *any of the estimators in* (9).

Proof We may write

$$\ln X_{n-k:n} \overset{d}{=} \ln U(n/k) + \frac{\xi\; B_k}{\sqrt{k}} + o_p(A(n/k)),$$

with B_k asymptotically standard normal. Since

$$\ln \mathrm{VaR}_q = \ln U\left(\frac{1}{q}\right) = \ln U\left(\frac{n}{k} \times \frac{k}{nq}\right),$$

we have, with $A(t)$ the function in (8),

$$\begin{aligned}\ln Q^{(q)}_{\hat{\xi}}(k) - \ln \mathrm{VaR}_q \;&\overset{d}{=}\; -\left(\ln U\left(\frac{n}{k} \times \frac{k}{nq}\right) - \ln U\left(\frac{n}{k}\right)\right) + \frac{\xi\; B_k}{\sqrt{k}} \\ &\qquad + \hat{\xi}(k)\;\ln\left(\frac{k}{nq}\right) + o_p(A(n/k)) \\ \overset{d}{=} \left(\hat{\xi}(k) - \xi\right)\ln\left(\frac{k}{nq}\right) &+ \frac{\xi\; B_k}{\sqrt{k}} - \frac{(k/(nq))^\rho - 1}{\rho}\; A(n/k)(1+o(1)) + o_p(A(n/k)).\end{aligned}$$

Consequently, since $(k/(nq))^\rho = o(1)$,

$$\ln Q^{(q)}_{\hat{\xi}}(k) - \ln \mathrm{VaR}_q \;\overset{d}{=}\; \left(\hat{\xi}(k) - \xi\right)\ln\left(\frac{k}{nq}\right) + \frac{\xi\; B_k}{\sqrt{k}} + \frac{A(n/k)}{\rho} + o_p(A(n/k)).$$

The dominant term is thus of the order of $\left\{\ln\left(k/(nq)\right)/\sqrt{k}\right\}$, that must converge towards zero, and this is true due to condition (14). The results in (15) follow from (12).

Apart from the MOP lnVaR-estimator, in (13), we have further considered in the lnVaR-estimator in (7), the replacement of the estimator $\hat{\xi}(k)$ by one of the most simple classes of corrected-bias Hill estimators, the one in Caeiro et al. [7]. Such a class is defined as

$$\mathrm{CH}(k) \equiv \mathrm{CH}(k; \hat{\beta}, \hat{\rho}) := \mathrm{H}(k)\Big(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})\Big). \tag{16}$$

The estimators in (16) can be second-order minimum-variance reduced-bias (MVRB) EVI-estimators, for adequate levels k and an adequate external estimation of the vector of second-order parameters, (β, ρ), introduced in (8), i.e. the use of $\mathrm{CH}(k)$ can enable us to eliminate the dominant component of bias of the Hill estimator, $\mathrm{H}(k)$, keeping its asymptotic variance. Indeed, from the results in Caeiro et al. [7], we know that it is possible to adequately estimate the second-order parameters β and ρ, so that we get

$$\sqrt{k}\,(\mathrm{CH}(k) - \xi) \overset{d}{=} \mathcal{N}\left(0, \xi^2\right) + o_p\big(\sqrt{k}(n/k)^{\rho}\big),$$

i.e. $\mathrm{CH}(k)$ overpasses $\mathrm{H}(k)$ for all k. Overviews on reduced-bias estimation can be found in Chap. 6 of [33], Gomes et al. [25] and Beirlant et al. [1].

For the estimation of the vector of second-order parameters (β, ρ), we propose an algorithm of the type of the ones presented in Gomes and Pestana [24], where the authors used the β-estimator in Gomes and Martins [21] and the simplest ρ-estimator in Fraga Alves et al. [16], both computed at a level $k_1 = \lfloor n^{0.999} \rfloor$, with the notation $\lfloor x \rfloor$ standing for the integer part of x. More recent estimators of β can be found in Gomes et al. [26] and Caeiro and Gomes [5, 6]. For alternative estimation of ρ, see Goegebeur et al. [19, 20] and Ciuperca and Mercadier [10].

4 Simulated Behaviour of the lnVaR Estimators

We have implemented large-scale multi-sample Monte-Carlo simulation experiments of size 5000 × 20, essentially for the new classes of lnVaR-estimators, $\ln Q^{(p)}_{\mathrm{H}_p}(k)$, in (13), with H_p given in (9), for a few values of p. We have considered sample sizes $n = 100, 200, 500, 1000, 2000$ and 5000, and $\xi = 0.1, 0.25, 0.5$ and 1, from the following models:

1. Fréchet(ξ) model, with CDF $F(x) = \exp(-x^{-1/\xi}),\ x \ge 0\ (\rho = -1)$;
2. Extreme value model, with CDF $F(x) = \mathrm{EV}_\xi(x)$, in (3) $(\rho = -\xi)$;

3. Burr(ξ, ρ) model, with CDF $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x \geq 0$, for the aforementioned values of ξ and for $\rho = -0.25$, -0.5 and -1;
4. Generalized Pareto model, with CDF $F(x) = GP_\xi(x) = 1 + \ln \mathrm{EV}_\xi(x) = 1 - \left(1 + \xi x\right)^{-1/\xi}$, $x \geq 0$ ($\rho = -\xi$).

We have further considered

5. Student-t_ν underlying parents, with $\nu = 4$ ($\xi = 1/\nu = 0.25$; $\rho = -2/\nu = -0.5$), with probability density function

$$f(x;\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1 + x^2/\nu\right)^{-(\nu+1)/2}, \quad t \in \mathbb{R}.$$

For details on multi-sample simulation, see Gomes and Oliveira [22].

4.1 Mean Values and MSE Patterns as a Function of k

For each value of n and for each of the aforementioned models, we have first simulated the mean value (E) and root MSE (RMSE) of the lnVaR-estimators under consideration, as functions of the number of top order statistics k involved in the estimation, and on the basis of the first run of size 5000. As an illustration, we present Figs. 1 and 2, respectively associated with GP$_{0.25}$ and EV$_{0.25}$ parents. In these figures, we show, for $n = 1000$, $q = 1/n$, and on the basis of the first $N = 5000$ runs, the simulated patterns of mean value, $E[\cdot]$, and root mean squared error, $RMSE[\cdot]$, of a few RV's $\ln Q_{\hat\xi}^{(p)}(k) - \ln \chi_{1-q}$, based on the statistics $\ln Q_{\hat\xi}^{(p)}(k)$ in (7), with $\hat\xi$ replaced by both H_p, in (9), for some values of p, and CH in (16). We shall use the obvious

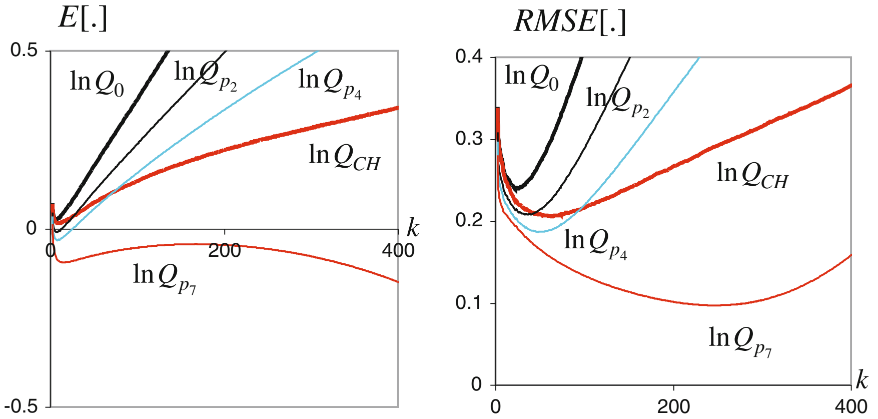


Fig. 1 Underlying *GP* parent with $\xi = 0.25$ ($\rho = -0.25$)

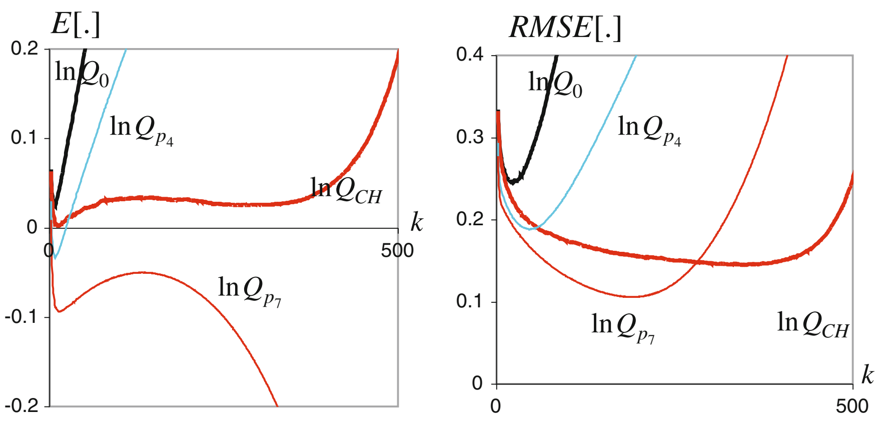


Fig. 2 Underlying *EV* parent with $\xi = 0.25$ $(\rho = -0.25)$

notations $\ln Q_p$ and $\ln Q_{\text{CH}}$. Apart from $p = 0$, associated with the Weissman-Hill lnVaR-estimator, we have considered $p = p_j = j/(10\xi)$, $j = 1, 2, 4$, all within the framework of Theorem 1, as well as $j = 7$ for which we can no longer guarantee the asymptotic normality of the new lnVaR-estimators.

We further present in Figs. 3 and 4, similar results but for two other models, a Student t_4 $(\rho = -0.5)$ and a Fréchet(1) $(\rho = -1)$, where the patterns are slightly different, from the ones obtained before for $\rho = -0.25$.

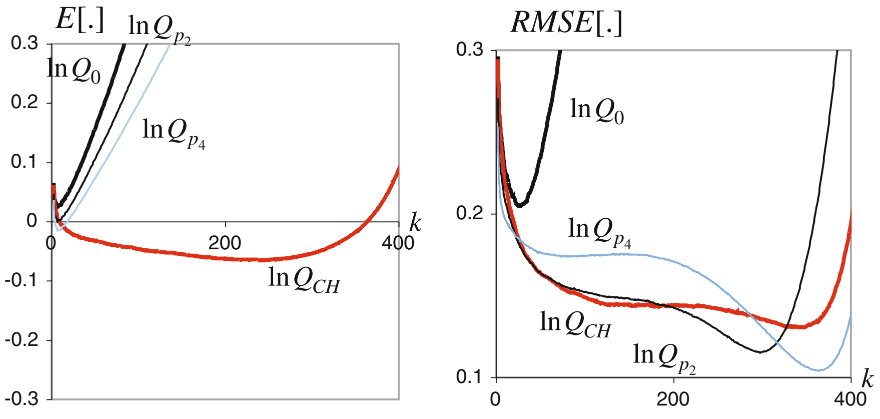


Fig. 3 Underlying *Student* t_ν parent, with $\nu = 4$ $(\xi = 1/\nu = 0.25,\ \rho = -2/\nu = -0.5)$

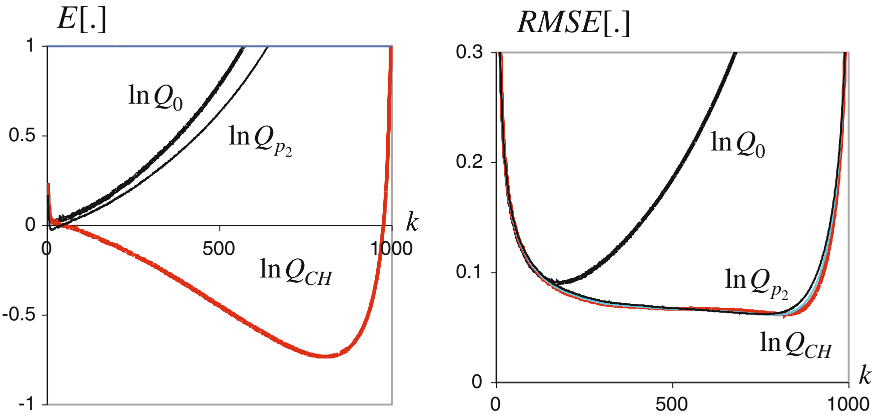


Fig. 4 Underlying *Fréchet* parent, with $\xi = 1$ $(\rho = -1)$

4.1.1 Mean Values, RMSEs and Relative Efficiency Indicators at Optimal Levels

We have further computed the Weissman-Hill lnVaR-estimator $\ln Q_{\text{H}}^{(q)}(k) \equiv \ln Q_{\text{H}_0}^{(q)}(k)$, with $\ln Q_{\hat{\xi}}^{(q)}(k)$ defined in (7), at the simulated value of $k_{0|\text{H}_0}^{(q)} := \arg\min_k \text{RMSE}\big(\ln Q_{\text{H}}^{(q)}(k)\big)$, the simulated optimal k in the sense of minimum RMSE. Such a value is not relevant in practice, but provides an indication of the best possible performance of the Weissman-Hill lnVaR-estimator. Such an estimator is denoted by $\ln Q_{00}$. We have also computed $\ln Q_{p0}$, for a few values of p. As an illustration of the bias of the new lnVaR-estimators, at optimal levels, see Tables 1 and 2. We present there, for $n = 100,\ 200,\ 500,\ 1000,\ 2000$ and 5000, the simulated mean values at optimal levels of the lnVaR-estimators under study. Information on 95 % confidence intervals (CIs), computed on the basis of the 20 replicates with 5000 runs each, is again provided. Among the estimators considered, the one providing the smallest squared bias is underlined, and written in **bold**.

We have further computed the simulated indicators,

$$\text{REFF}_{p|0} := \frac{\text{RMSE }(\ln Q_{00})}{\text{RMSE }\big(\ln Q_{p0}\big)}. \tag{17}$$

A similar REFF-indicator, $\text{REFF}_{\text{CH}|0}$ has also been computed for the lnVaR-estimator based on CH EVI-estimators, in (16).

Remark 1 This indicator has been conceived so that an indicator higher than one means a better performance than the one of the Weissman-Hill lnVaR-estimator. Consequently, the higher these indicators are, the better the associated lnVaR-estimators perform, compared to $\ln Q_{00}$.

Table 1 Simulated mean values, at optimal levels, of $T_{00} := \ln Q_{00} - \ln \chi_{1-q}$, $q = 1/n$ (first row) and REFF-indicators of $\ln Q_{\mathrm{CH}|0} - \ln \chi_{1-q}$ and $\ln Q_{p_j|0} - \chi_{1-q}$, for $p_j = j/(10\xi)$, $j = 1, 2, 4$ and 7, for GP and EV parents, with $\xi = 0.25$ ($\rho = -0.25$), together with 95 % CIs

n	100	200	500	1000	2000	5000
GP parent, $\xi = 0.25$						
H	0.077 ± 0.0058	0.078 ± 0.0040	0.081 ± 0.0056	0.084 ± 0.0051	0.083 ± 0.0043	0.083 ± 0.0037
CH	−0.093 ± 0.0041	**0.002** ± 0.0031	0.090 ± 0.0038	0.089 ± 0.0044	0.086 ± 0.0035	0.085 ± 0.0027
p_1	0.063 ± 0.0051	0.070 ± 0.0055	0.073 ± 0.0040	0.074 ± 0.0023	0.077 ± 0.0027	0.078 ± 0.0021
p_2	0.063 ± 0.0051	0.060 ± 0.0035	0.064 ± 0.0029	0.067 ± 0.0029	0.071 ± 0.0028	0.073 ± 0.0022
p_4	**−0.009** ± 0.0021	0.016 ± 0.0024	**0.011** ± 0.0013	**0.006** ± 0.0008	**0.006** ± 0.0022	**0.003** ± 0.0009
p_7	−0.153 ± 0.0046	−0.171 ± 0.0024	−0.120 ± 0.0011	−0.053 ± 0.0006	−0.009 ± 0.0004	−0.001 ± 0.0003
EV parent, $\xi = 0.25$						
H	**0.075** ± 0.0060	**0.081** ± 0.0056	0.080 ± 0.0050	0.083 ± 0.0041	0.084 ± 0.0040	0.080 ± 0.0038
CH	−0.083 ± 0.0088	−0.086 ± 0.0058	−0.067 ± 0.0040	0.028 ± 0.0034	0.084 ± 0.0032	0.084 ± 0.0027
p_1	−0.083 ± 0.0045	−0.094 ± 0.0053	−0.031 ± 0.0056	0.020 ± 0.0028	0.080 ± 0.0022	0.076 ± 0.0025
p_2	−0.103 ± 0.0041	−0.094 ± 0.0051	**−0.003** ± 0.0031	**0.012** ± 0.0021	0.065 ± 0.0012	0.073 ± 0.0017
p_4	−0.156 ± 0.0034	−0.111 ± 0.0040	−0.120 ± 0.0027	−0.114 ± 0.0013	**−0.045** ± 0.0007	**0.005** ± 0.0003
p_7	−0.238 ± 0.0026	−0.177 ± 0.0015	−0.137 ± 0.0021	−0.124 ± 0.0039	−0.122 ± 0.0028	−0.131 ± 0.0037

Table 2 Simulated mean values, at optimal levels, of $T_{00} := \ln Q_{00} - \ln \chi_{1-q}$, $q = 1/n$ (first row) and REFF-indicators of $\ln Q_{\mathrm{CH}|0} - \ln \chi_{1-q}$ and $\ln Q_{p_j|0} - \chi_{1-q}$, for $p_j = j/(10\xi)$, $j = 1, 2$ and 4, for Student t_4 ($\xi = 0.25$, $\rho = -0.5$) and Fréchet parents with $\xi = 0.25$ ($\rho = -1$), together with 95 % CIs

n	100	200	500	1000	2000	5000
Student t_4 parent, $(\xi, \rho) = (0.25, -0.5)$						
H	0.065 ± 0.0520	0.065 ± 0.0041	0.073 ± 0.0055	0.071 ± 0.0028	0.071 ± 0.0031	0.070 ± 0.0032
CH	−0.072 ± 0.0023	−0.079 ± 0.0060	−0.080 ± 0.0026	**−0.011** ± 0.0024	**0.031** ± 0.0017	**0.051** ± 0.0012
p_1	**0.052** ± 0.0052	0.062 ± 0.0051	0.064 ± 0.0035	0.066 ± 0.0030	0.068 ± 0.0029	0.067 ± 0.0018
p_2	0.053 ± 0.0062	**0.051** ± 0.0033	0.053 ± 0.0044	0.062 ± 0.0025	0.064 ± 0.0022	0.063 ± 0.0020
p_4	0.074 ± 0.0044	0.053 ± 0.0036	**0.052** ± 0.0026	0.055 ± 0.0027	0.057 ± 0.0018	0.058 ± 0.0020
Fréchet parent, $(\xi, \rho) = (0.25, -1)$						
H	0.227 ± 0.0089	0.213 ± 0.0101	0.197 ± 0.0052	0.183 ± 0.0046	0.160 ± 0.0071	0.137 ± 0.0038
CH	−0.232 ± 0.0100	−0.198 ± 0.0092	**−0.162** ± 0.0039	**−0.139** ± 0.0050	**−0.116** ± 0.0039	**−0.010** ± 0.0027
p_1	0.222 ± 0.0105	0.209 ± 0.0072	0.188 ± 0.0058	0.178 ± 0.0051	0.158 ± 0.0054	0.135 ± 0.0039
p_2	0.201 ± 0.0075	0.209 ± 0.0085	0.186 ± 0.0060	0.169 ± 0.0039	0.157 ± 0.0027	0.132 ± 0.0034
p_4	**0.191** ± 0.0060	**0.191** ± 0.0076	0.180 ± 0.0045	0.169 ± 0.0038	0.155 ± 0.0041	0.132 ± 0.0028

In the first row of Tables 3 and 4, we provide the RMSE of $\ln Q_{00}$, denoted by RMSE_0, so that we can easily recover the RMSE of all other estimators. The subsequent rows provide the REFF-indicators of the lnVaR-estimators based on CH and on H_p. The highest REFF indicator is underlined and **bolded**. Among the MOP lnVaR-estimators within the scope of Theorem 1, we place the highest value in *italic* and underlined, whenever such a value is not the highest one among all estimators under consideration.

Table 3 Simulated RMSE of $\ln Q_{00}$, $q = 1/n$ (first row) and REFF-indicators of $\ln Q_{\mathrm{CH}|0}$ and $\ln Q_{p_j|0}$, for $p_j = j/(10\xi)$, $j = 1, 2, 4$ and 7, for GP and EV parents, with $\xi = 0.25$ ($\rho = -0.25$), together with 95 % CIs

n	100	200	500	1000	2000	5000
GP parent, $\xi = 0.25$						
RMSE$_0$	0.345 ± 0.0351	0.307 ± 0.0356	0.268 ± 0.0297	0.245 ± 0.0259	0.224 ± 0.0216	0.200 ± 0.0201
CH	1.243 ± 0.0086	1.492 ± 0.0092	1.256 ± 0.0074	1.174 ± 0.0035	1.131 ± 0.0035	1.093 ± 0.0029
p_1	1.116 ± 0.0026	1.098 ± 0.0018	1.082 ± 0.0018	1.072 ± 0.0019	1.063 ± 0.0015	1.053 ± 0.0011
p_2	1.299 ± 0.0055	1.239 ± 0.0047	1.187 ± 0.0030	1.160 ± 0.0039	1.136 ± 0.0031	1.11 ± 0.0021
p_4	**2.477** ± 0.0115	**2.676** ± 0.0182	**2.442** ± 0.0147	*2.361* ± 0.0158	*2.350* ± 0.0154	*2.430* ± 0.0163
p_7	1.239 ± 0.0064	1.269 ± 0.0085	1.633 ± 0.0133	**2.516** ± 0.0166	**4.438** ± 0.0153	**7.738** ± 0.0571
EV parent, $\xi = 0.25$						
RMSE$_0$	0.353 ± 0.0358	0.311 ± 0.0364	0.270 ± 0.0276	0.246 ± 0.0257	0.224 ± 0.0216	0.200 ± 0.0209
CH	1.009 ± 0.0048	1.056 ± 0.0080	1.290 ± 0.0137	1.695 ± 0.0086	1.299 ± 0.0086	1.173 ± 0.0043
p_1	1.050 ± 0.0055	1.078 ± 0.0066	1.500 ± 0.0262	1.754 ± 0.0107	1.438 ± 0.0088	1.253 ± 0.0050
p_2	1.088 ± 0.0064	1.099 ± 0.0056	**1.751** ± 0.0180	**1.886** ± 0.0115	1.662 ± 0.0091	1.356 ± 0.0057
p_4	**1.126** ± 0.0081	**1.146** ± 0.0052	1.154 ± 0.0066	1.312 ± 0.0096	**2.849** ± 0.0230	**4.772** ± 0.0244
p_7	1.049 ± 0.0075	1.106 ± 0.0065	1.111 ± 0.0057	1.086 ± 0.0067	1.053 ± 0.0057	1.003 ± 0.0064

Table 4 Simulated RMSE of $\ln Q_{00}$, $q = 1/n$ (first row) and REFF-indicators of $\ln Q_{\mathrm{CH}|0}$ and $\ln Q_{p_j|0}$, for $p_j = j/(10\xi)$, $j = 1, 2$, and 4, for Student t_4 ($\xi = 0.25, \rho = -0.5$) and Fréchet parents with $\xi = 0.25$ ($\rho = -1$), together with 95 % CIs

n	100	200	500	1000	2000	5000
Student t_4 parent, $(\xi, \rho) = (0.25, -0.5)$						
RMSE$_0$	0.300 ± 0.0291	0.264 ± 0.0264	0.228 ± 0.0232	0.207 ± 0.0192	0.188 ± 0.0180	0.164 ± 0.0132
CH	0.991 ± 0.0170	1.030 ± 0.0058	1.164 ± 0.0092	**1.609** ± 0.0115	**1.684** ± 0.00115	**1.468** ± 0.0083
p_1	1.093 ± 0.0033	1.079 ± 0.0021	1.064 ± 0.0024	1.054 ± 0.0016	1.047 ± 0.0013	1.039 ± 0.0014
p_2	1.219 ± 0.0041	1.181 ± 0.0041	1.137 ± 0.0045	1.116 ± 0.0028	1.097 ± 0.0025	1.077 ± 0.0021
p_4	**1.524** ± 0.0078	**1.348** ± 0.0044	**1.242** ± 0.0063	*1.195* ± 0.0049	*1.154* ± 0.0044	*1.113* ± 0.0034
Fréchet parent, $(\xi, \rho) = (0.25, -1)$						
RMSE$_0$	0.683 ± 0.0757	0.600 ± 0.0642	0.501 ± 0.0435	0.432 ± 0.0368	0.372 ± 0.0308	0.303 ± 0.0307
CH	0.899 ± 0.0035	0.850 ± 0.0683	0.929 ± 0.0035	00.968 ± 0.0066	1.019 ± 0.0066	**1.144** ± 0.0058
p_1	1.033 ± 0.0010	1.028 ± 0.0017	1.023 ± 0.0011	1.021 ± 0.0011	1.019 ± 0.0012	1.018 ± 0.0009
p_2	**1.066** ± 0.0023	1.052 ± 0.0038	1.038 ± 0.0020	**1.033** ± 0.0020	**1.029** ± 0.0023	*1.026* ± 0.0022
p_4	1.101 ± 0.0045	**1.070** ± 0.0062	**1.042** ± 0.0039	1.028 ± 0.0033	1.017 ± 0.0042	1.011 ± 0.0034

4.1.2 Discussion

1. Note that the functionals T under play are functions of $\ln X$. Consequently, if X is a Fréchet(ξ) RV, denoted F_ξ, the uniform transformation enables us to write

$$\exp\left(-\mathrm{F}_\xi^{-1/\xi}\right) \overset{d}{=} U \iff \ln \mathrm{F}_\xi \overset{d}{=} -\xi \ln(-\ln(U)),$$

   i.e. $\ln \mathrm{F}_\xi / \xi$ does not depend on ξ. This also happens with a Burr(ξ, ρ) model. For such a RV, now denoted by $\mathrm{B}_{\xi,\rho}$, we get

$$\big(1 + \mathrm{B}_{\xi,\rho}^{-\rho/\xi}\big)^{1/\rho} \overset{d}{=} U \quad \Longleftrightarrow \quad \ln \mathrm{B}_{\xi,\rho} \overset{d}{=} -\xi \ln \big(U^{\rho} - 1\big)/\rho,$$

i.e. again $\ln \mathrm{B}_{\xi,\rho}/\xi$ does not depend on ξ. Due to the above mentioned reasons, the REFF-indicators, $\mathbb{E}(T/\xi)$ and RMSE_0/ξ do not depend on ξ for Fréchet and Burr(ξ, ρ) underlying parents. Also, with the same notation, $\ln \mathrm{B}_{\xi,\rho}/\xi = \ln \mathrm{B}_{1,\rho}$. Moreover, with the obvious notation GP_ξ for a GP RV, with EVI ξ, we have $\mathrm{GP}_\xi = \ln \mathrm{B}_{\xi,-\xi}/\xi$. Consequently, the equivalent tables for Burr(ξ, ρ) RVs, with $\rho = -0.25$, are trivially obtained from Table 3, GP parent. Particularly, the REFF indicators are the same.

2. Regarding bias, the MOP lnVaR–estimators often outperform the MVRB EVI-estimators whenever $|\rho| < 0.5$.
3. For values of $|\rho| \le 0.25$ the use of lnVarQ_p, with $p = p_7$, always enables a reduction in RMSE. Moreover, the bias is also reduced comparatively with the bias of the Weissman-Hill lnVaR-estimator with the obtention of estimates closer to the target value lnVaR_q, for $q = 1/n$. Note however that such a value of p is beyond the scope of Theorem 1. Such a reduction is particularly high for values of ρ close to zero, even when we work with models again out of the scope of Theorem 1, like the log-gamma and the log-Pareto. This is surely due to the high bias of the Weissman-Hill lnVaR-estimators for this type of models with $\rho = 0$.

5 Concluding Remarks

- The patterns of the estimators' sample paths are always of the same type, in the sense that for all k the lnVaR-estimator, $\ln Q_{H_p}^{(q)}$ decreases as p increases.
- It is clear that Weissman-Hill lnVaR-estimation leads to a strong over-estimation of the EVI and the MOP provides a more adequate lnVaR-estimation, being even able to beat the MVRB lnVaR-estimators in a large variety of situations.
- The results obtained lead us to strongly advise the use of the log-quantile estimator $\ln Q_p$ for an adequate value of p, provided by a bootstrap algorithm of the type devised for an EVI-estimation in Gomes et al. [27, 28].

Acknowledgments Research partially supported by National Funds through FCT—Fundação para a Ciência e a Tecnologia, projects PEst-OE /MAT /UI0006 /2011, 2014 (CEAUL) and EXTREMA, PTDC /MAT /101736 /2008.

References

1. Beirlant, J., Caeiro, F., Gomes, M.I.: An overview and open research topics in statistics of univariate extremes. Revstat **10**(1), 1–31 (2012)
2. Bingham, N., Goldie, C.M., Teugels, J.L.: Regular Variation. Cambridge University Press, Cambridge (1987)

3. Brilhante, M.F., Gomes, M.I., Pestana, D.: A simple generalisation of the Hill estimator. Comput. Stat. Data Anal. **57**(1), 518–535 (2013)
4. Brilhante, F., Gomes, M.I., Pestana, D.: The mean-of-order p extreme value index estimator revisited. In: Pacheco, A., Oliveira, R., Santos, R. (eds.) New Advances in Statistical Modeling and Application. Springer, Berlin (2014). (in press)
5. Caeiro, C., Gomes, M.I.: A semi-parametric estimator of a shape second order parameter. Notas e Comunicações CEAUL 07/2012 (2012)
6. Caeiro, F., Gomes, M.I.: Bias reduction in the estimation of a shape second-order parameter of a heavy right tail model. Preprint, CMA 22-2012 (2012)
7. Caeiro, F., Gomes, M.I., Pestana, D.: Direct reduction of bias of the classical Hill estimator. Revstat **3**(2), 113–136 (2005)
8. Caeiro, F., Gomes, M.I.: Minimum-variance reduced-bias tail index and high quantile estimation. Revstat **6**(1), 1–20 (2008)
9. Caeiro, F., Gomes, M.I.: Semi-parametric second-order reduced-bias high quantile estimation. Test **18**(2), 392–413 (2009)
10. Ciuperca, G., Mercadier, C.: Semi-parametric estimation for heavy tailed distributions. Extremes **13**(1), 55–87 (2010)
11. de Haan, L.: Slow variation and characterization of domains of attraction. In: Tiago de Oliveira, J. (ed.) Statistical Extremes and Applications, pp. 31–48. D. Reidel, Dordrecht (1984)
12. de Haan, L., Peng, L.: Comparison of tail index estimators. Statistica Neerlandica **52**, 60–70 (1998)
13. de Haan, L., Rootzén, H.: On the estimation of high quantiles. J. Stat. Plan. Inference **35**, 1–13 (1993)
14. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction. Springer Science+Business Media, LLC, New York (2006)
15. Ferreira, A., de Haan, L., Peng, L.: On optimising the estimation of high quantiles of a probability distribution. Statistics **37**(5), 401–434 (2003)
16. Fraga Alves, M.I., Gomes, M.I., de Haan, L.: A new class of semi-parametric estimators of the second order parameter. Portugaliae Mathematica **60**(2), 194–213 (2003)
17. Geluk, J., de Haan, L.: Regular Variation, Extensions and Tauberian Theorems. CWI Tract 40, Center for Mathematics and Computer Science. Amsterdam, Netherlands (1987)
18. Gnedenko, B.V.: Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. **44**, 423–453 (1943)
19. Goegebeur, Y., Beirlant, J., de Wet, T.: Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. Revstat **6**(1), 51–69 (2008)
20. Goegebeur, Y., Beirlant, J., de Wet, T.: Kernel estimators for the second order parameter in extreme value statistics. J. Stat. Plan. Inference **140**(9), 2632–2654 (2010)
21. Gomes, M.I., Martins, M.J.: "Asymptotically unbiased" estimators of the tail index based on external estimation of the second order parameter. Extremes **5**(1), 5–31 (2002)
22. Gomes, M.I., Oliveira, O.: The bootstrap methodology in statistical extremes—choice of the optimal sample fraction. Extremes **4**(4), 331–358 (2001)
23. Gomes, M.I., Figueiredo, F.: Bias reduction in risk modelling: semi-parametric quantile estimation. Test **15**(2), 375–396 (2003)
24. Gomes, M.I., Pestana, D.: A sturdy reduced bias extreme quantile (VaR) estimator. J. Am. Stat. Assoc. **102**(477), 280–292 (2007)
25. Gomes, M.I., Canto e Castro, L., Fraga Alves, M.I., Pestana, D.: Statistics of extremes for iid data and breakthroughs in the estimation of the extreme value index: Laurens de Haan leading contributions. Extremes **11**(1), 3–34 (2008)
26. Gomes, M.I., Miranda, C., Pereira, H., Pestana, D.: Tail index and second order parameters' semi-parametric estimation based on the log-excesses. J. Stat. Comput. Simul. **80**(6), 653–666 (2010)
27. Gomes, M.I., Figueiredo, F., Neves, M.M.: Adaptive estimation of heavy right tails: the bootstrap methodology in action. Extremes **15**, 463–489 (2012)

28. Gomes, M.I., Martins, M.J., Neves, M.M.: Generalised jackknife-based estimators for univariate extreme-value modelling. Commun. Stat.—Theory Methods **42**, 1227–1245 (2013)
29. Hall, P., Welsh, A.H.: Adaptive estimates of parameters of regular variation. Ann. Stat. **13**, 331–341 (1985)
30. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
31. Matthys, G., Beirlant, J.: Estimating the extreme value index and high quantiles with exponential regression models. Statistica Sinica **13**, 853–880 (2003)
32. Matthys, G., Delafosse, M., Guillou, A., Beirlant, J.: Estimating catastrophic quantile levels for heavy-tailed distributions. Insur.: Math. Econ. **34**, 517–537 (2004)
33. Reiss, R.-D., Thomas, M.: Statistical Analysis of Extreme Values, With Application to Insurance, Finance, Hydrology and Other Fields, 3rd edn. Birkhäuser Verlag, Boston (2007)
34. Weissman, I.: Estimation of parameters and large quantiles based on the *k* largest observations. J. Am. Stat. Assoc. **73**, 812–815 (1978)

Adaptive Choice and Resampling Techniques in Extremal Index Estimation

Dora Prata Gomes and M. Manuela Neves

Abstract This work deals with the application of resampling techniques together with the adaptive choice of a 'tuning' parameter, the block size, b, to be used in the bootstrap estimation of the extremal index, that is a key parameter in extreme value theory in a dependent setup. Its estimation has been considered by many authors but some challenges still remain. One of these is the choice of the number of upper order statistics to be considered in the semiparametric estimation. Block-bootstrap and Jackknife-After-Bootstrap are two computational procedures applied here for improving the behavior of the extremal index estimators through an adaptive choice of the block size for the resampling procedure. A few results of a simulation study will be presented.

Keywords Adaptive choice · Block size · External index · Exterme value theory · Resampling techniques

1 Introduction and Preliminaries

When natural calamities of great magnitude happen, we are concerned with the occurrence and the frequency of such events because of their human and economic effects. There are a large variety of fields of application such as, e.g., environment, finance, internet traffic, reliability and survival analysis, where values of interest are the extreme values.

The classical assumption in Extreme Value Theory (EVT) is that we have a set of independent and identically distributed (i.i.d.) random variables (r.v.'s), $X_1, \ldots, X_n$, from an unknown distribution function (d.f.) F and we are concerned with the

D. Prata Gomes (✉)
CMA and FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: dsrp@fct.unl.pt

M.M. Neves
CEAUL and ISA, Universidade de Lisboa, Tapada da Ajuda, 1349-017 Lisbon, Portugal
e-mail: manela@isa.ulisboa.pt

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_24

limit behavior of either $M_n \equiv X_{n:n} = \max(X_1, \dots, X_n)$ or $m_n \equiv X_{1:n} = \min(X_1, \dots, X_n)$, as $n \to \infty$.

For the case of the maximum value, whenever it is possible to normalize M_n so that we get a non-degenerate limit as $n \to \infty$, the resulting limit is of the Extreme Value (EV) type d.f.,

$$EV_\gamma(x) := \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}), & 1+\gamma x > 0 \text{ if } \quad \gamma \neq 0 \\ \exp(-\exp(-x)), & x \in \mathbf{R} \qquad \text{if} \quad \gamma = 0. \end{cases} \tag{1}$$

In d.f. (1) the parameter γ is called *the extreme value index*, whose estimation is of primordial importance. Other parameters of extreme events can appear, such as a high quantile, the probability of exceedance or the return period of a high quantile, that need to be suitably estimated on the basis of an available sample.

The estimation is usually performed under a semiparametric approach based on probabilistic asymptotic results on the tail of the unknown distribution. Several difficulties arise when we intend to obtain reliable semiparametric estimates of the parameters of extreme events. These estimates are usually calculated on the basis of the largest k order statistics in the sample, and they are strongly dependent on k.

For overcoming this difficulty Jackknife and Bootstrap procedures are known as giving more stable estimates around the target value, for i.i.d. sequences. However, in many practical situations, the i.i.d assumption is not always valid. For example, for the amount of rain in a given location on consecutive days, it is obvious that the chance of rain after a rainy day is higher than the chance of rain after a dry day.

Whenever independence is no longer valid, some important dependent sequences have been studied and the limit distributions of their order statistics under some dependence structures are known. Stationary sequences are examples of such sequences and are realistic for many real problems.

Suppose we have n observations from a stationary process $\{Y_n\}_{n\geq 1}$ with marginal distribution function F. For large n and a high level u_n, we have

$$F_n(u_n) = P[max(Y_1, \dots, Y_n) \leq u_n] \approx F^{n\theta}(u_n),$$

where $\theta \in [0, 1]$ is a constant for the process, known as the *extremal index*. This concept, even appearing in papers from Newell [24], Loynes [22], O'Brien [25], was only well characterized by Leadbetter [19].

As dependence in stationary sequences can assume several forms, some conditions have to be imposed. The first condition, known as the $D(u_n)$ dependence condition, Leadbetter et al. [21], ensures that any two extreme events can become approximately independent as n increases when separated by a relatively short interval of length $l_n = o(n)$. Hence, $D(u_n)$ limits the long-range dependence between such events.

Provided that a stationary sequence $\{Y_n\}_{n\geq 1}$ has limited long-range dependence at extreme levels, the maxima of this sequence follow the same distributional limit law as the associated independent sequence, $\{X_n\}$, but with other values for the parameters. Actually, see Leadbetter et al. [21], if $\{Y_n\}_{n\geq 1}$ is a stationary sequence with marginal distribution F, $\{X_n\}_{n\geq 1}$ an i.i.d. sequence of r.v.'s with the same distribution F, $M_n := \max(Y_1, \ldots, Y_n)$ and $\widetilde{M_n} := \max(X_1, \ldots, X_n)$, under the $D(u_n)$ condition, with $u_n = a_n x + b_n$, $Pr\left\{(\widetilde{M_n} - b_n)/a_n \le x\right\} \underset{n\to\infty}{\longrightarrow} G_1(x)$ as $n \longrightarrow \infty$, for normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$, if and only if, $Pr\{(M_n - b_n)/a_n \le x\} \underset{n\to\infty}{\longrightarrow} G_2(x)$ where $G_2(x) = G_1^\theta(x)$, for a constant θ such that $0 < \theta \le 1$.

So, given that $G_1(\cdot) \equiv EV_\gamma(\cdot)$, the limit law $G_2(\cdot) \equiv EV_\gamma^\theta(\cdot)$ is an extreme value d.f. with location, scale and shape parameters $(\mu_\theta, \sigma_\theta, \gamma_\theta)$ given by

$$\mu_\theta = \mu - \sigma\frac{1-\theta^\gamma}{\gamma}, \quad \sigma_\theta = \sigma\theta^\gamma \quad \text{and} \quad \gamma_\theta = \gamma,$$

where (μ, σ, γ) are the location, scale and shape parameters, respectively, of the limit law of the i.i.d sequence.

The quantity θ is the aforementioned extremal index. This parameter, apart from being of interest in its own right, is crucial for determining the limiting distribution of extreme values from the stationary process.

The extremal index θ, $0 \le \theta \le 1$ is directly related to the clustering of exceedances: $\theta = 1$ for i.i.d. sequences and $\theta \to 0$ whenever dependence increases. The case $\theta = 0$ appears in pathological situations. For ‘almost all cases’ of interest, we have $\theta > 0$.

For illustration of the behavior of a stationary process for some values of θ let us consider the following example:

Example 1 Moving Maximum Process (MM process). Let $\{Z_n\}_{n\ge0}$ be a sequence of i.i.d. variables from the model $F(z) = \exp(-z^{-1})$, with $z \ge 0$ and for $a \ge 0$ define

$$Y_0 = Z_0, \quad Y_j = (a+1)^{-1}\max\left\{aZ_{j-1}, Z_j\right\}, \quad j = 1, 2\ldots$$

$\Pr\{M_n \le u_n\} = \Pr\{\widetilde{M}_n \le u_n\}^\theta$ as $n \to \infty$ where $\theta = 1/(a+1)$ lies in the interval $[1/2, 1]$, Davison [3].

Figure 1 shows a partial realization of variables following the above process with $a = 0.2;\ 0.4;\ 1$, which corresponds to $\theta = 0.833;\ 0.714;\ 0.5$, respectively. The

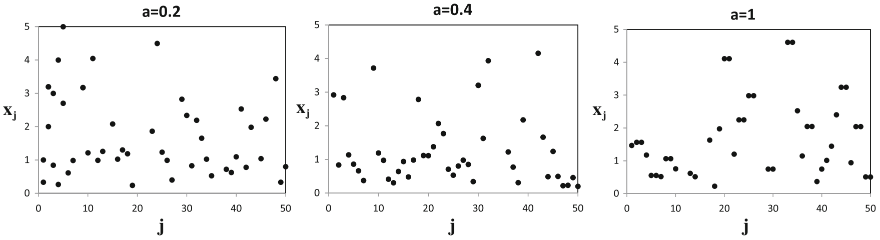


Fig. 1 Values from the moving maximum process with $a = 0.2$ ($\theta = 0.833$), $a = 0.4$ ($\theta = 0.714$) and $a = 1$ ($\theta = 0.5$). As $a(\theta)$ increases (decreases) we notice some clusters appearing

maxima show increasing clustering as $a \to 1$ which corresponds to decreasing values of θ.

In next section some of the classical estimators of θ are referred to. Their asymptotic properties are pointed out. However despite nice asymptotic properties, for finite samples the estimates strongly depend on the upper level u_n.

2 Extremal Index Estimation

Classical estimators of θ have been developed based on characterizations of θ given by Leadbetter [19], O'Brien [26].

We consider the most common interpretation of θ, as being the reciprocal of the 'mean time of duration of extreme events' which is directly related to the exceedances of high levels, Hsing et al. [14] and Leadbetter and Nandagopalan [20]

$$\theta = \frac{1}{\text{limiting mean size of clusters}}.$$

Identifying clusters by the occurrence of downcrossings or upcrossings, we can write

$$\theta = \lim_{n\to\infty} Pr[X_2 \leq u_n | X_1 > u_n] = \lim_{n\to\infty} Pr[X_1 \leq u_n | X_2 > u_n]. \tag{2}$$

The classical up-crossing estimator (UC-estimator), $\widehat{\Theta}^{UC}$ Gomes [6–8] Nandagopalan [23] is a naive estimator that can be derived directly as an empirical counterpart of (2),

$$\widehat{\Theta}^{UC} \equiv \widehat{\Theta}^{UC}(u_n) := \frac{\sum_{i=1}^{n-1} I\,(X_i \leq u_n < X_{i+1})}{\sum_{i=1}^{n} I\,(X_i > u_n)}, \tag{3}$$

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 $\tau \equiv \tau_n$ fixed, the underlying d.f. F verifies

$$F(u_n) = 1 - \tau/n + o(1/n), \quad n \to \infty \text{ and } \tau/n \to 0.$$

Additional estimators of the extremal index can be defined by different forms of identifying clusters. Let us refer to two very popular and well studied estimators: *the block estimator* and *the runs estimator*, Hsing [12, 13]. The *blocks estimator*, is derived by dividing the data into approximately k_n blocks of length r_n, where $n \approx k_n \times r_n$, i.e., considering $k_n = [n/r_n]$. Each block is treated as one cluster and the number of blocks in which there is at least one exceedance of the threshold u_n is counted. The *block estimator*, $\widehat{\Theta}_n^B(u_n)$, is then defined as

$$\widehat{\Theta}_n^B(u_n) := \frac{\sum_{i=1}^{k_n} I\left(\max\left(X_{(i-1)r_n+1}, \ldots, X_{ir_n}\right) > u_n\right)}{\sum_{i=1}^{n} I\left(X_i > u_n\right)}.$$

If we assume that a cluster consists of a run of observations between two exceedances, then the “runs estimator” is defined as:

$$\widehat{\Theta}_n^R(u_n) := \frac{\sum_{i=1}^{n} I\left(X_i > u_n, \max\left(X_{i+1}, \ldots, X_{i+r_n-1}\right) \le u_n\right)}{\sum_{i=1}^{n} I\left(X_i > u_n\right)}.$$

Under mild conditions, $\lim_{n\to\infty} \theta_n^B = \lim_{n\to\infty} \theta_n^R = \theta$. Other properties of these estimators have been well studied by Smith and Weissman [30] and Weissman and Novak [31].

In this paper our attention will be focused on the UC-estimator and given the sample $\mathbf{X}_n := (X_1, \ldots, X_n)$ and the associated ascending order statistics, $X_{1:n} \le \cdots \le X_{n:n}$, we shall consider the level u_n as a deterministic level $u \in [X_{n-k:n}, X_{n-k+1:n}[$.

The $UC-$estimator, in (3) can now be written as a function of k, the number of top order statistics above the chosen threshold,

$$\widehat{\Theta}^{UC} \equiv \widehat{\Theta}^{UC}(k) := \frac{1}{k}\sum_{i=1}^{n-1} I\left(X_i \le X_{n-k:n} < X_{i+1}\right).$$

For many dependent structures, the bias of $\widehat{\Theta}^{UC}(k)$ has two dominant components of orders k/n and $1/k$ (see Gomes et al. [9]),

$$Bias[\widehat{\Theta}^{UC}(k)] = \varphi_1(\theta)\left(\frac{k}{n}\right) + \varphi_2(\theta)\left(\frac{1}{k}\right) + o\left(\frac{k}{n}\right) + o\left(\frac{1}{k}\right), \tag{4}$$

whenever $n \to \infty$ and $k \equiv k(n) \to \infty$, $k = o(n)$.

The Generalized Jackknife methodology has the properties of estimating the bias and the variance of any estimator, leading to the development of estimators with bias and mean squared error often smaller than those of an initial set of estimators.

Using the information obtained from (4) and based on the estimator $\widehat{\Theta}^{UC}$ computed at the three levels, k, $[k/2]+1$ and $[k/4]+1$, where [x] denotes, as usual, the integer part of x, Gomes et al. [9] derived a reduced-bias estimator for θ, the Generalized Jackknife estimator of order 2, $\widehat{\Theta}^{GJ}$, defined as

$$\widehat{\Theta}^{GJ} \equiv \widehat{\Theta}^{GJ}(k) := 5\widehat{\Theta}^{UC}([k/2]+1) - 2\left(\widehat{\Theta}^{UC}([k/4]+1) + \widehat{\Theta}^{UC}(k)\right).$$

This is an asymptotically unbiased estimator of θ, in the sense that it can remove the two dominant components of bias referred to in (4). Under certain conditions, estimators $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ are consistent and asymptotically normal if $\theta < 1$, see Gomes et al. [9] and Nandagopalan [23].

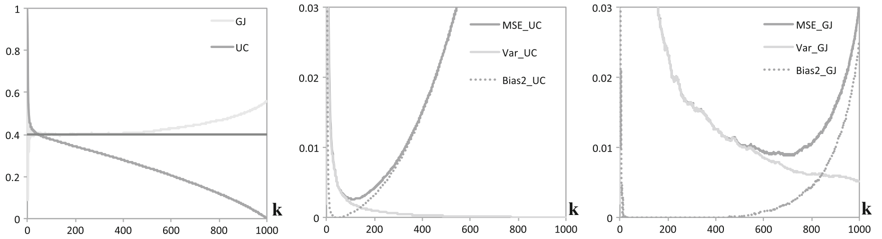


Fig. 2 Simulated mean values, *MSE*, variance and *Bias*^{2} of the *UC*-estimator and Generalized Jackknife estimator (from *left* to *right*) for the ARMAX process with $\theta = 0.4$ and a sample size $n = 1000$

For illustrating the properties of the $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ estimators, let us consider the following max-autoregressive process:

Example 2 Max-Autoregressive Process (ARMAX process). Let $\{Z_i\}_{i\geq 1}$ be a sequence of independent, unit-Fréchet distributed random variables. For $0 < \theta \leq 1$, let

$$Y_1 = Z_1 \quad Y_i = \max\{(1-\theta)Y_{i-1}, \theta Z_i\} \quad i = 2, \ldots$$

For $u_n = ny$, $0 < y < \infty$, $\Pr\left\{M_n \leq u_n\right\} \to \exp\left(-\theta/y\right)$, as $n \to \infty$. The extremal index of the sequence is equal to θ, Beirlant et al. [1].

Some remarks on Fig. 2.

- The $\widehat{\Theta}_n^{UC}$ estimator, shows a very strong bias. The bias is the dominant component of the *MSE*, see Fig. 2 (middle).
- $MSE(\widehat{\Theta}_n^{UC})$ is very sharp, which reveals a need for a very accurate way of choosing k in order to obtain a reliable estimate of θ.
- The Generalized-Jackknife estimator, $\widehat{\Theta}_n^{GJ}$, shows a more stable simulated mean value, near the target value of the parameter but at expenses of a very high variance, which does not enable it to outperform the original estimator, regarding *MSE* at optimal levels.
- $MSE(\widehat{\Theta}_n^{GJ})$ is not so sharp as $MSE(\widehat{\Theta}_n^{UC})$, suggesting then less dependence on the value u_n, for obtaining the estimate of θ.

Recently the use of adequate bootstrap procedures has resulted in improving the behavior of the estimators for a finite sample. Let us refer to Prata Gomes and Neves [27], Prata Gomes and Neves [28], Gomes et al. [10] for some results on the use of resampling procedures in extreme value estimation.

However the choice of the level k still remains an interesting research topic. Regarding the compromise between bias and variance given by the mean squared error, *MSE*, some authors have shown that, in extreme value theory, resampling methodologies have been performing quite well for estimating the optimal number of order statistics to be used in the estimation of parameters of rare events, see Gomes et al. [9, 10].

3 Resampling Techniques for Dependent Case

In their classical form, as first proposed by Efron [4], bootstrap methods are designed for use in samples collected under an independent set-up.

In context of dependent data, the situation is more complicated. Singh [29] mentioned the inadequacy of the classical bootstrap under dependence.

Several attempts have been made to extend the bootstrap method to the dependent case. A breakthrough was achieved when resampling of single observations was replaced by block resampling.

Bootstrap methods using different blocks have been proposed to attempt to reproduce difference aspects of the dependence structure of the observed data in the resampled data: Moving Block Bootstrap (MBB), Non-Overlapping Block Bootstrap (NBB), Circular Block Bootstrap (CB) and Stationary Bootstrap (SB) are the most well known.

The accuracy of block bootstrap estimators, critically depends on the block length that must be supplied by the user.

As a simple illustration, let us examine Fig. 3 with the values from a sample of size $n = 100$ generated from the ARMAX process with $\theta = 0.1$ and the values obtained by bootstrapping the original sample, through the classical procedure and by block resampling using several block sizes.

The orders of magnitude of the "optimal" block sizes are known in some inference problems [2, 11, 15, 16, 18] as $b \sim Cn^{1/k}$, with $k = 3, 4$ or 5, values to be used for the estimation of bias, variance, or one-sided distribution function/two-sided distribution function, respectively. This result, of practical and theoretical interest, will be used here as the basis for choosing the "optimal" block length for resampling. Two main approaches can be pointed out: the cross validation method proposed by Hall et al. [11] and the plug-in method based on Lahiri et al. [18].

Extremal index estimators usually have a high bias; so much so that in most cases the bias is the main component of the *MSE*. There is then a need for bias reduction. Based on Lahiri et al. [18], a nonparametric plug-in (NPPI) method for selecting the "optimal" block length in order to reduce the bias, will be considered. This method employs nonparametric resampling procedures to estimate the relevant constants in the leading term of the "optimal" block length and, hence, does not require the knowledge and/or derivation of explicit analytical expressions for the constants.

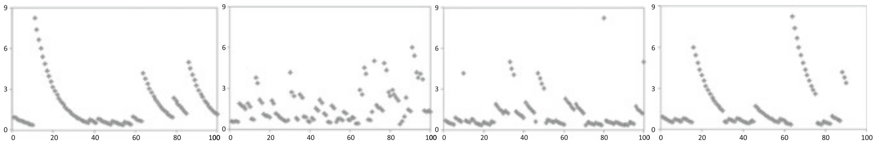


Fig. 3 Values from a sample of size $n = 100$ generated from the ARMAX process with $\theta = 0.1$ and resampled equal size samples considering the classical i.i.d. resample, blocks of size 2, 5 and 15, from *left* to *right*

3.1 Main Steps of NPPI Method

Given the sample $(X_1, X_2, \ldots, X_n)$ from an unknown model F, let us consider $\widehat{\Theta}_n$, any estimator of θ and let $\widehat{\Theta}_n^*(b)$ be the corresponding bootstrap estimator based on blocks of size b. Let us denote by $\phi_n \equiv Bias\big(\widehat{\Theta}_n\big) = E\big(\widehat{\Theta}_n\big) - \theta$, the bias of $\widehat{\Theta}_n$, so $\widehat{\phi}_n^*(b) \equiv \widehat{Bias}(b) = E_*\big(\widehat{\Theta}_n^*(b)\big) - \widehat{\Theta}_n$ is the corresponding block bootstrap estimator of $Bias\big(\widehat{\Theta}_n\big)$ based on blocks of size b.

Under suitable regularity conditions, the variance and the bias of a block bootstrap estimator admit expansions of the form (see Lahiri [16]),

$$n^{2a} Var\left(\widehat{\phi}_n^*(b)\right) = C_1 n^{-1} b^r + o\left(n^{-1} b^r\right) \quad \text{as} \quad n \to \infty \tag{5}$$

$$n^a Bias\left(\widehat{\phi}_n^*(b)\right) = C_2 b^{-1} + o\left(b^{-1}\right) \quad \text{as} \quad n \to \infty \tag{6}$$

over a suitable set of possible block lengths $b \in \{2, \ldots, n\}$, where $C_1 \in (0, \infty)$ and $C_2 \in \mathbf{R}$ are population parameters, $r \geq 1$ is an integer, and $a \in [0, \infty[$ is a known constant. For $\phi_n \equiv Bias$, Hall et al. [11] consider that (5) and (6) hold with $r = 1$ and $a = 1$.

It is known Hall et al. [11] that the variance of a block bootstrap estimator is an increasing function of the block length b while its bias is a decreasing function of b. From (5) and (6) an expansion for $MSE(\widehat{\phi}_n^*(b))$ is obtained and leads to the asymptotic MSE-optimal block length, $b^0 \equiv b_n^0$ (see Hall et al. [11] and Lahiri [16]):

$$b_n^0 = \left(\frac{2C_2^2}{C_1}\right)^{1/(r+2)} n^{1/(r+2)}(1 + o(1)). \tag{7}$$

Estimation of C_1 and C_2, under the NPPI method, is done considering the leading part of (5) and (6),

$$C_1 \sim n b^{-r} n^{2a} Var\big(\widehat{\phi}_n^*(b)\big) \quad \text{and} \quad C_2 \sim b n^a Bias\big(\widehat{\phi}_n^*(b)\big).$$

This suggests the use of consistent estimators of $Var\big(\widehat{\phi}_n^*(b)\big)$ and $Bias\big(\widehat{\phi}_n^*(b)\big)$ and define estimators of the parameters C_1 and C_2 as

$$\widehat{C}_1 = n b^{-r} n^{2a} \widehat{Var}_n \quad \text{and} \quad \widehat{C}_2 = b n^a \widehat{Bias}_n, \tag{8}$$

where $\widehat{Var}_n \equiv \widehat{Var}_n(b_1)$ and $\widehat{Bias}_n \equiv \widehat{Bias}_n(b_1)$ are consistent estimators of the variance and the bias, respectively, of the block bootstrap estimator $\widehat{\phi}_n^*(b)$ based on some suitable initial block length b_1.

Following the suggestion of Lahiri et al. [18] of using the Jackknife-After-Bootstrap (JAB) method of Efron [5] and Lahiri [17] for estimating $Var\big(\widehat{\phi}_{kn}^*(b)\big)$, Prata Gomes and Neves [27] built a computational procedure based on the estimation of the variance and the bias of the Block Bootstrap estimator in order to estimate the "optimal" block length, in the sense of minimizing the mean square error of the estimator of the Bias of extremal index estimators.

JAB estimator of the variance of $\widehat{\phi}_n^*(b)$ is defined as:

$$\widehat{VAR}_{JAB}\Big(\widehat{\phi}_n^*(b)\Big) = \frac{m}{(n_b - m)M} \sum_{i=1}^{M} \Big(\widetilde{\phi}_{kn}^{(i)*}(b) - \widehat{\phi}_{kn}^*(b)\Big)^2 \qquad (9)$$

where $n_b = n - b + 1$ is the number of overlapping blocks of length b, contained in $(X_1, \ldots, X_n)$ and $\widetilde{\phi}_n^i(b) = m^{-1}\big(\ell\widehat{\phi}_n(b) - (\ell - m)\widehat{\phi}_n^i(b)\big)$ is the ith block-deleted jackknife *pseudo-value* of $\widehat{\phi}_n^*(b)$, $i = 1, \ldots, M$, $b = c_1 n^{1/5}$, with $c_1 = 1$ and $m = c_2 n^{1/3} b^{2/3}$ with $c_2 = 1$.

A consistent estimator of $Bias(\widehat{\phi}_n^*(b))$, Lahiri et al. [18], is given by

$$\widehat{Bias}_n \equiv \widehat{Bias}_n(b) = 2\big(\widehat{\phi}_n^*(b) - \widehat{\phi}_n^*(2b)\big). \qquad (10)$$

The NPPI estimator $\widehat{b}_n^0$ of the "optimal" block length b_n^0 is then obtained from (7) and (8) as

$$\widehat{b}_n^0 = \left(\frac{2\widehat{C}_2^2}{\widehat{C}_1}\right)^{1/(r+2)} n^{1/(r+2)}(1 + o(1)). \qquad (11)$$

4 A Few Results from a Simulation Study

An extensive simulation study is being carried out by the authors and some results have already been shown in Prata Gomes and Neves [28]. Here one illustration of that study is given, using the MM and ARMAX processes defined previously.

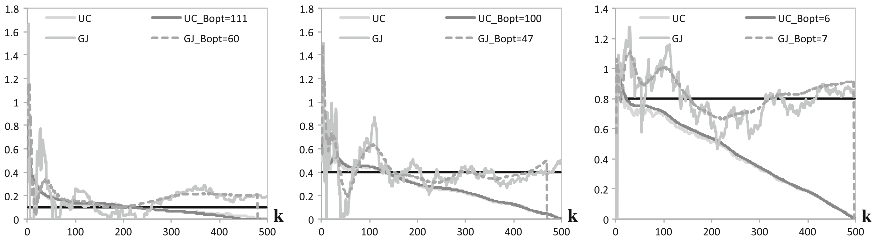


Fig. 4 One sample path of the estimators $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ for a sample of size $n = 500$ from the ARMAX process and block bootstrap estimates using "optimal" block length for $\theta = 0.1$, $\theta = 0.4$ and $\theta = 0.8$

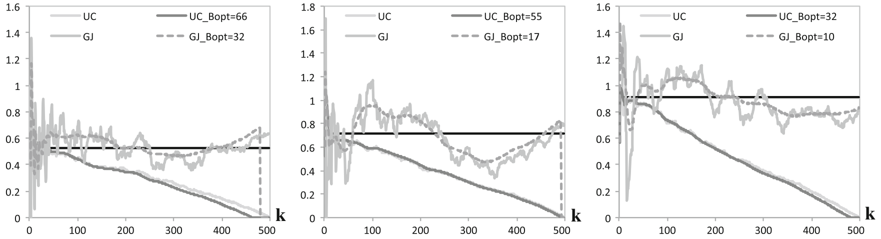


Fig. 5 One sample path of the estimators $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ for a sample of size $n = 500$ from the moving Maximum process and block bootstrap estimates using "optimal" block length for $\theta = 0.5263$, $\theta = 0.7143$ and $\theta = 0.9091$

A simple path of the estimators $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ is generated. The optimal block length is obtained for each sample and block bootstrap estimates, using the optimal block length, are then obtained for each estimator.

The simulation procedure is developed in the following steps:

- Generate a random sample of size $n = 500$ (in our study obtained from the ARMAX process with $\theta = (0.1, 0.4, 0.8)$ and from the MM process with $a = (0.1, 0.4, 0.9)$, what corresponds to $\theta = (0.9091, 0.7143, 0.5263)$;
- Define an initial block size, $b_1 = C_3 n^{1/5}$, see Sect. 3. when $\widehat{Var_n}$ and $\widehat{Bias_n}$ were defined, and the JAB blocking parameter, $m = C_4 n^{1/3} {b_1}^{2/3}$. Lahiri et al. [18] pointed out that with $C_3 = 1$ for the initial block length b_1 gives the best result for different functionals of interest, and the value of C_4 for calculating m is $C_4 = 1$ for the bias and variance functionals.
- The NPPI method was applied for estimating the "optimal" block length for resampling, which depends on the value of k. Now the value of block size with the highest frequency was adopted as the "optimal" block length.
- Finally block bootstrap estimates for the $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ were obtained.

Figures 4 and 5 show sample paths for $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ estimates, and the corresponding block bootstrap estimates, where the block size is also indicated.

5 Some Overall Conclusions

A general method for estimating the "optimal" block length for resampling in the situation of dependence was presented. This was used in a simulation study for estimating the extremal index. Two estimators of that parameter were used and bootstrap versions of those estimators based on a block resampling scheme were considered. Estimates from the Generalized Jackknife estimator revealed promising results, showing a more stable path.

Monte Carlo simulations allow us to analyze the behavior of our procedures, given that the true values of the parameters are known. In real case studies it is not obvious how to choose the threshold appropriately. Without an adequate procedure for choosing the level k to be used in the estimation, it will thus be difficult to justify any particular estimate for the extremal index. This is a topic out of the scope of this paper, but that constitutes work already in progress.

We have presented a procedure that allows us not only to obtain more stable estimators but also enables the development of reduced bias estimators. To improve the procedure used is the next step.

Acknowledgments Research partially supported by National Funds through FCT – Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2014 (CEAUL) and PEst-OE/MAT/UI0297/2014 (CMA/UNL) and UID/MAT/00297/2013 (Centro de Matemática e Aplicações).

References

1. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., Ferro, C.: Statistics of Extremes: Theory and Applications. Wiley, New York (2004)
2. Bühlmann, P., Künsch, H.: Block length selection in the bootstrap for time series. J. Comput. Stat. Data Anal. **31**, 295–310 (1999)
3. Davison, A.: Statistics of Extremes. Courses 2011–2012. École Polytechnique Fédérale de Lausanne EPFL (2011)
4. Efron, B.: Bootstrap methods: another look at the jackknife. Ann. Stat. **7**, 1–26 (1979)
5. Efron, B.: Jackknife-after-boostrap standard errors and influence functions (with discussions). J. Roy. Stat. Ser. B **54**, 83–111 (1992)
6. Gomes, M.I.: Statistical inference in an extremal Markovian model. COMPSTAT 257–262 (1990)
7. Gomes, M.I.: Modelos extremais em esquemas de dependência. I Congresso Ibero-Americano de Esdadistica e Investigacion Operativa, 209–220 (1992)
8. Gomes, M.I.: On the estimation of parameters of rare events in environmental time series. Statistics for the Environment, Barnett et al. (eds.), 226–241 (1993)
9. Gomes, M.I., Hall, A., Miranda, C.: Subsampling techniques and the Jackknife methodology in the estimation of the extremal index. J. Comput. Stat. Data Anal. **52**(4), 2022–2041 (2008)
10. Gomes, M.I., Figueiredo, F., Neves, M.M.: Adaptive estimation of heavy right tails: resampling-based methods in action. Extremes **15**, 463–489 (2012)
11. Hall, P., Horowitz, J.L., Jing, B.-Y.: On blocking rules for the bootstrap with dependent data. Biometrika **50**, 561–574 (1995)
12. Hsing, T.: Estimating the parameters of rare events. Stoch. Proc. Appl. **37**, 117–139 (1991)
13. Hsing, T.: Extremal index estimation for a weakly dependent stationary sequence. Ann. Stat. **21**, 2043–2071 (1993)
14. 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)
15. Künsch, H.: The jackknife and the bootstrap for general stationary observations. Ann. Math. **17**, 1217–1241 (1989)
16. Lahiri, S.: Theoretical comparisons of block bootstrap methods. Ann. Stat. **27**, 386–404 (1999)
17. Lahiri, S.: On the jackknife after bootstrap method for dependent data and its consistency properties. Econom. Theory **18**, 79–98 (2002)
18. Lahiri, S., Furukawa, K., Lee, Y.-D.: Nonparametric plug-in method for selecting the optimal block lengths. Stat. Methodol. **4**, 292–321 (2007)
19. Leadbetter, M.R.: Extremes and local dependence in stationary sequences. Z. Wahrsch. Verw. Gebiete **65**(2), 291–306 (1983)

20. 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)
21. Leadbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Series. Springer, New York (1983)
22. Loynes, R.M.: Extremes values in uniformly mixing stationary stochastic processes. Ann. Math. Stat. **36**(3), 993–999 (1965)
23. Nandagopalan, S.: Multivariate extremes and estimation of the extremal index. Ph.D Thesis, University of North Carolina, Chapel Hill (1990)
24. Newell, G.: Asymptotic extremes for m-dependent random variables. Ann. Math. Stat. **35**, 1322–1325 (1964)
25. O'Brien, G.: The maximum term of uniformly mixing stationary processes. Z. Wahrsch. Verw. Gebiete **30**, 57–63 (1974)
26. O'Brien, G.: Extreme values for stationary and Markov sequences. Ann. Probab. **15**(1), 281–289 (1987)
27. Prata Gomes, D., Neves, M.M.: Resampling methodologies and the estimation of parameters of rare events. In: Proceedings of the AIP Conference on Numerical Analysis and Applied Mathematics. **1389**, (ICNAAM 2011). 1475–1478 (2011)
28. Prata Gomes, D., Neves, M.M.: Bootstrap and other resampling methodologies in statistics of extremes. Accepted (2013)
29. Singh, K.: On the asymptotic accuracy of the Efron's bootstrap. Ann. Stat. **9**, 345–362 (1981)
30. Smith, R., Weissman, I.: Estimating the extremal index. J. R. Stat. Soc. B **56**, 515–528 (1994)
31. Weissman, I., Novak, S.: On blocks and runs estimators of the extremal index. J. Stat. Plan. Inference **66**, 281–288 (1998)

Some Estimation Techniques in Reliability and Survival Analysis Based on Record-Breaking Data

Inmaculada Barranco-Chamorro and Sneh Gulati

Abstract In this paper we review some of the classical and Bayesian results on statistical inference from records that can be used in reliability and survival analysis. We focus on some important lifetime models, giving special attention to heavy-tailed distributions in order to consider applications of record-breaking data to the study of extreme events. Results on the estimation of the number of observations needed to attain a given number of records are also studied in depth. Numerical illustrations and results on the estimation of cost functions are included as well. This chapter can serve as a guide for people interested in making inferences in the fields of reliability and survival analysis when only record values are available.

Keywords Record values · Heavy tailed distributions · Classical inference · Bayesian inference · Sample-size estimation

1 Introduction

Human beings are fascinated with records. We devour literature on record-breaking events and hold our breath for the next one; who will beat Paul Biedermann's 200 m freestyle world record of 1:42, what will be next high reached by the DOW Jones index; will there be a hurricane that will top Hurricane Katrina's record 25 billion dollar loss; just to name a few examples. It is not just the general public that is fascinated with records, it turns out that the statisticians also love records and record-breaking data as is evidenced by the plethora of papers and books on the subject. They want to develop the mathematical theory that defines record-breaking data

I. Barranco-Chamorro (✉)
Departamento de Estadistica e I.O. Facultad de Matematicas,
Universidad de Sevilla, Avd. Reina Mercedes s/n, 41012 Sevilla, Spain
e-mail: chamorro@us.es

S. Gulati
Department of Mathematics and Statistics, Florida International University,
Miami, FL 33199, USA
e-mail: gulati@fiu.edu

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_25

and use it to predict the next record. From a statistical perspective, record-breaking data routinely arise in reliability and survival analysis where units are observed sequentially and only successive maxima (or minima) are recorded. The resulting data not only lead to considerable measurement savings, but can often serve equally well in making the required inference in certain situations. For example, in reliability and survival studies, the practitioner is often interested in estimating a guarantee value; essentially an upper or a lower quantile. Clearly when the observed data consist of successive maxima (or minima) an upper (or a lower) quantile can be easily estimated (see [11] for details.) Moreover, as seen in [2, 5], and references therein, records are also potentially useful in diverse fields such as insurance theory, industrial stress-testing, climatology and geosciences, etc.

Mathematically record-breaking data are defined as follows:

Let $Y_1, Y_2, \ldots$ be independent identically distributed (iid) random variables (rv's) with common cumulative distribution function (cdf) F. The cdf F is assumed to be absolutely continuous with probability density function (pdf) f. In addition, we assume that the rv's Y_j are observed sequentially, i.e., Y_j is observed at time j.

An observation Y_j is called an upper record value (or simply a record) if its value exceeds that of all previous observations, i.e. Y_j is an upper record if $Y_j > Y_i$ for all $i < j$. Lower records can be defined similarly. Details and general properties about records can be found in [2, 5] or [20].

The following variables are associated with record-breaking data:

The record time sequence:

$$\begin{aligned} &L(1) = 1 \\ &L(j) = \min\left\{i \ : \ i > L(j-1), \ \ Y_i > Y_{L(j-1)}\right\}, \quad \text{for } j \geq 2. \end{aligned}$$

The record value sequence:

$$X_j \ = \ Y_{L(j)}, \qquad j \geq 1.$$

Note that X_1 is taken as the reference value or trivial record. The rest are nontrivial records.

The record counting process:

$$N^{(n)} \ : \ \text{“number of records among } Y_1, \ldots, Y_n\text{”} \tag{1}$$

Since the Y_i's are iid absolutely continuous rv's, $\{L(j)\}$ and $N^{(n)}$ are distribution free, i.e., the parent cdf F will not affect distributions of these variables. The distribution of the record values X_j is obviously affected by F.

Record-Breaking data were first studied in a stochastic setting by [8] who showed that the record times (and the inter-record times) have an infinite expectation. Since then the number of authors who have studied and continue to study record-breaking data both from a stochastic and inferential perspective has mushroomed. Inferential procedures from records have spanned a wide range of topics: parameter estimation, hypothesis testing, lifetime and hazard function estimation and prediction [11].

The purpose of this paper is to review some results on inference from records that can be used in risk and reliability analysis especially in the context of estimation and prediction of extreme events. We focus our attention on the study of records from heavy tailed distributions since such distributions are often used to describe insured losses and serve as a model for extreme events. From an insurer's perspective, it is critical to know when the next catastrophic event will strike and what its magnitude will be so that they can obtain the necessary reinsurance coverage. Thus estimation and prediction of these extremes is especially important from a statistical viewpoint. Equally important for an insurance company is the ability to estimate the original sample size based on the number of observed records. Given the highest recorded flood levels in a city, it is of interest to know many days of flooding did the city actually experience. Similarly, while it is important to know what the extreme hurricane losses were, it is just as important to know the number of times a company had to pay out due to hurricane losses as well as the actual amounts. The problem of sample size estimation is also important in geostatistics. For instance, [14] estimated the number of glacial advances from the number of surviving debris (called moraines). He dealt with this issue as a problem of estimation of the sample size from the number of records. In reliability and survival analysis, the problem of sample size estimation manifests itself in fatigue tests. These tests are mandatory to ensure the integrity of structures; however since they are essentially destructive, the underlying sample size must be carefully chosen in order to find a good balance between the accuracy of the estimation and the cost of the experiment.

The organization of the paper is as follows. We study the estimation and prediction of extreme events (in the context of record values) from a classical perspective in Sect. 2, while in Sect. 3 we study the same topics in a Bayesian framework. Finally, in Sect. 4, we review results on the estimation of the original sample size based on the number of observed records. Numerical examples to illustrate the applications of some of these results are presented at the end of every section.

2 Distribution Theory for Records from Heavy Tailed Distributions

As mentioned in the introduction, manufacturers and insurers need to hedge themselves against extreme events and thus be able to predict them accurately. Given past extreme events, we can think of the future extreme event as a future record-breaking event and use appropriate record-breaking theory for estimation and prediction. Hence in this section, we review some inferential results on record-breaking data from some of the heavy tailed distributions used to model extreme events.

2.1 Pareto and Related Distributions

The Pareto distribution has extensive applications in hydrology, income distributions and the insurance claims. Arnold and Press [4] stated that this model is useful for approximating data that arise from distributions with fat tails [12]. Since Pareto-type distributions are the most commonly used class of distributions to describe extreme events, we will use this class as our focal point.

Definition 1 A rv Y follows a generalized Pareto distribution (GPD) if its cdf is given by

$$F_{\theta,\sigma,\beta}(y) = \begin{cases} 1 - \left[1 + \beta\left(\frac{y-\theta}{\sigma}\right)\right]^{-1/\beta} & \text{for } \beta \neq 0 \\ 1 - \exp\left(-\frac{y-\theta}{\sigma}\right) & \text{for } \beta = 0 \end{cases} \tag{2}$$

This model has three parameters: location θ, $\theta \in I\!R$; scale σ, $\sigma > 0$; and shape β, $\beta \in I\!R$. For $\beta \geq 0$ the support for y is given by $y \geq \theta$, whereas for $\beta < 0$ the support is $\theta \leq y \leq \theta - \frac{\sigma}{\beta}$.

A rv with cdf given in (2) is denoted by $Y \sim GPD(\theta, \sigma, \beta)$. (Note that the GPD is equivalent to the two-parameter exponential distribution for $\beta = 0$, while for the case when both $\beta = 0$ and $\theta = 0$ it reduces to the exponential distribution. Finally, for $\beta > 0$ and $\theta = \sigma/\beta$, the GPD is equivalent to the classical Pareto distribution.)

2.2 Results for the Generalized Pareto Distribution

Consider the first k upper records: $X_1, \ldots, X_k$ from the GPD defined in (2). Let Z_i denote the standardized records, that is $Z_i = (X_i - \theta)/\sigma$. (Note then that $Z_i \sim GPD(0, 1, \beta)$.) The single moments of the m^{th} standardized record value Z_m, $E[Z_m^r]$, denoted by $\mu_m^{(r)}$, are given by (from [27]; originally in [6])

$$\mu_m^{(r)} = \frac{1}{\beta^r} \sum_{i=0}^{r} \frac{C_i^r (-1)^{r-i}}{(1-\beta i)^m}, \quad r = 1, 2, \ldots, \tag{3}$$

with $C_i^r = \binom{r}{i}$.

The double moments of the record values Z_m and Z_n with $m < n$, $E[Z_m^r Z_n^s]$, denoted by $\mu_{m,n}^{(r,s)}$, are given by

$$\mu_{m,n}^{(r,s)} = \frac{1}{\beta^{r+s}} \sum_{j=0}^{s} \sum_{i=0}^{r} \frac{C_j^s C_i^r (-1)^{s+r-i-j}}{(1-\beta j)^{n-m}(1-\beta j - \beta i)^m}, \quad r, s = 1, 2, \ldots \quad m < n \tag{4}$$

Using the generalized least-squares approach and the moments given above the Best Linear Unbiased Estimators (BLUEs) of θ and σ, denoted by θ^* and σ^*, are given by (see [5, 7] for details):

$$\theta^* = \left\{ \frac{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{1}'\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}) - (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^2} \right\} \mathbf{X} = \sum_{i=1}^{k} a_i X_i \tag{5}$$

$$\sigma^* = \left\{ \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1} - \mathbf{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\mathbf{1}'\boldsymbol{\Sigma}^{-1}}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}) - (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^2} \right\} \mathbf{X} = \sum_{i=1}^{k} b_i X_i \tag{6}$$

where $\mu_i = E[Z_i]$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$, $\sigma_{i,j} = \text{Cov}[Z_i, Z_j]$, $\boldsymbol{\Sigma} = ((\sigma_{i,j}))$ with $1 \le i, j \le k$; $\mathbf{X} = (X_1, \ldots, X_k)'$ and $\mathbf{1} = (1, \ldots, 1)'$.

Their variances and covariances are

$$\text{Var}[\theta^*] = \sigma^2 \left\{ \frac{\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}) - (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^2} \right\} = \sigma^2 V_1 \tag{7}$$

$$\text{Var}[\sigma^*] = \sigma^2 \left\{ \frac{\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}) - (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^2} \right\} = \sigma^2 V_2 \tag{8}$$

$$\text{Cov}[\theta^*, \sigma^*] = \sigma^2 \left\{ \frac{-\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1}}{(\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})(\mathbf{1}'\boldsymbol{\Sigma}^{-1}\mathbf{1}) - (\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\mathbf{1})^2} \right\} = \sigma^2 V_3. \tag{9}$$

The BLUEs are used to propose pivots and confidence intervals for the location and scale parameters. The pivotal quantities are

$$R_1 = \frac{\theta^* - \theta}{\sigma\sqrt{V_1}}, \quad R_2 = \frac{\sigma^* - \sigma}{\sigma\sqrt{V_2}}, \quad \text{and} \quad R_3 = \frac{\theta^* - \theta}{\sigma^*\sqrt{V_1}}. \tag{10}$$

R_1 and R_3 are used to make inferences about θ when σ is known and unknown, respectively, while R_2 is used to make inferences about σ. Percentage points of these pivotal quantities were computed in [27] via Edgeworth approximations and Monte Carlo simulations.

Finally, the Best Linear Unbiased Predicted value (BLUP) of the next record, X^*_{k+1}, can be written as a linear function of BLUEs of θ and σ based on the first k records (θ^*_k and σ^*_k) as

$$X^*_{k+1} = \theta^*_k + \sigma^*_k \mu_{k+1}, \tag{11}$$

and prediction intervals for the next record value X_{k+1} can be given by using the pivotal quantity

$$T_{k+1} = \frac{X_{k+1} - X_k}{\sigma^*_k}. \tag{12}$$

As before percentage points of the above pivotal quantity are usually computed via simulations.

At this point, it is important to note that the methodology followed in [27] for the GPD is general, and it can be applied to other similar settings where we want to estimate the location and scale parameters of a distribution as well as predict the next record based on the first k records. Once again, we refer the reader to [7] or [5] for the relevant general methodology.

2.3 Results for Some Other Distributions

In this subsection, we summarize some results related to record values for two other heavy tailed distributions. Results for additional distributions are not presented since the general methodology presented in Sect. 2.2 applies to most of these distributions.

Generalized Exponential Distribution

Definition 2 A rv Y follows a Generalized Exponential Distribution (GED) if its cdf is given by

$$F_{\theta,\sigma,\beta}(y) = \left\{1 - e^{-(y-\theta)/\sigma}\right\}^{\beta}, \quad y > \theta,\ \theta \in I\!R,\ \sigma > 0,\ \beta > 0. \tag{13}$$

This distribution was introduced in [13] as an alternative to the gamma and Weibull models. Note that above distribution is characterized by a location, scale and a shape parameter. Using lower records and the general methodology outlined in the previous section, [22] developed inferential results for this distribution. These include expressions for the standardized moments of lower records, the BLUEs of θ and σ, and predictions of future records from a classical point of view.

Modified Weibull Distribution

Definition 3 A rv Y follows a Modified Weibull Distribution (MWD) if its pdf is given by

$$f_{a,b,\lambda}(y) = a(b + \lambda y)y^{b-1}e^{\lambda y}\exp(-ay^{b}e^{\lambda y}), \quad y > 0,\ a > 0,\ b \geq 0,\ \lambda > 0. \tag{14}$$

This distribution was introduced in [15] as a new lifetime model, it can also be used to model tail behavior, and records from this distribution were studied by [26]. Note that the MWD includes the Weibull distribution as a special case ($\lambda = 0$) and the Type I extreme value distribution ($b = 0$). Although, in this model, we have a shape parameter b and two scale parameters, (a, λ), the generalized least-squares approach can be applied again. Sultan [26] derived the single and the product moments for upper record values from the MWD, new recurrence relations between them, the BLUEs of the underlying parameters and some characterizations.

2.4 Numerical Illustrations

Next we include a numerical illustration given in [5]. They used the data given by [23] on 1h mean concentrations of sulfur dioxide (in pphm) form Long Beach, California for the years 1956–1974 and extracted upper records for the month of October. These values are given by: 26, 27, 40 and 41. Chan [10] considered log-record values of the above data (given by 3.258, 3.296, 3.689 and 3.714) and assumed that they had come from a Type I extreme value distribution (min) with parameters θ and σ (cdf given by $F(y) = 1 - exp(-e^{\frac{y-\theta}{\sigma}})$.) For the extreme value distribution, the expressions (5), (6) and (11) can be simplified to (see [5] for details):

$$\theta^* = \frac{\alpha_k}{k}\sum_{i=1}^{k-1} X_i + (1-\alpha_k)X_k \tag{15}$$

$$\sigma^* = X_k - \frac{1}{k-1}\sum_{i=1}^{k} X_i \tag{16}$$

and

$$X^*_{k+1} = X_k + \frac{1}{k(k+1)}\sum_{i=1}^{k-1}(X_k - X_i) \tag{17}$$

where $\alpha_k = -\gamma + \sum_{i=1}^{k} \frac{1}{i}$ and γ is the Euler's constant.

Using (15) and (16), then the BLUE's of θ and σ are calculated to be $\theta^* = 3.338$ and $\sigma^* = 0.300$. Moreover, using (17), an estimate of the next log-record value is given by 3.788166. Arnold et al. [5] also computed a conditional 90 % interval for the next upper log-record to be (3.714, 3.9737). Note that one can exponentiate this interval to obtain the corresponding 90 % interval for the next upper record in terms of the original data.

Another application consisting of real record values for the Weibull distribution can be seen in [25].

3 Bayesian Prediction of Future Records

While the results summarized in Sect. 2 have involved classical theory, estimation and prediction problems in the context of records can be satisfactorily solved in the Bayesian framework (see [3]). This is due to the fact that from a sequence of n iid rv's, we expect to have only a few records, so additional prior information is usually welcome and useful for inferential purposes. In this sense, Bayesian inferential methods can provide better results than classical ones. The general methodology applicable to all families of distributions, and detailed in [3], is as follows:

Suppose that the data consist of the first upper record values $X_1, X_2, \ldots, X_k$ from a continuous distribution $F(x|\theta)$, with $\theta \in \Theta \subseteq I\!R^p$ where Θ denotes the underlying parameter space. The likelihood function of the record values is given in [5]

$$L(\theta, \mathbf{x}) = f(x_k|\theta) \prod_{i=1}^{k-1} \frac{f(x_i|\theta)}{1 - F(x_i|\theta)}, \qquad \text{where } \mathbf{x} = (x_1, \ldots, x_k). \tag{18}$$

The prior information of the experimenter is expressed in terms of a proper conjugate prior distribution, $\pi(\theta)$, defined on the parameter space Θ. Multiplying $\pi(\theta)$ by the likelihood of the records, given in (18), updates the prior to posterior density $\pi^*(\theta|\mathbf{x})$. The posterior density can be used not only to compute the Bayesian estimates of the underlying parameters but also provides the following predictive density of the future sth record $y = x_s^*$, given the first k records as

$$f^*(y|\mathbf{x}) = \int_\Theta g_k(y|\theta)\ \pi^*(\theta|\mathbf{x}) d\theta \tag{19}$$

where $g_k(\cdot)$ is the conditional pdf of the sth record given that the kth record has been observed, with $s > k$ and is given by (see [5])

$$g_k(y|\theta) = g_k(y|\mathbf{x}, \theta) = \frac{\{H(y) - H(x_k)\}^{s-k-1}}{\Gamma(s-k)} \frac{f(y|\theta)}{1 - F(x_k|\theta)}, \qquad x_k < y \tag{20}$$

Here $H = -\ln(1 - F)$ and $\ln(\cdot)$ denotes the natural logarithm.

Finally, the Bayes point predictor of the sth future record under squared error loss is obtained as the expected value of (19).

AL-Hussaini and Ahmad [3] used this methodology to obtain prediction intervals for future records from the Burr, Pareto, and Weibull distributions. Bayesian estimators of the underlying parameters and their point predictors for these models were also considered in [1].

3.1 Applications to the Pareto Distribution

As mentioned earlier, the Pareto distribution is one of the most important classes of distributions when it comes to describing extreme events. Thus we now look at results on the prediction of future records from the Type I Pareto distribution.

Type I Pareto distribution

Prediction of future records for the Type I Pareto distribution was addressed not only in [3] using the methods detailed above, but also in [1, 16]. Madi and Raqab [16] considered prediction of future record values and record average for the Type I Pareto distribution only, whereas [1, 3] considered Bayesian estimation for the parameters of several life distributions, including the Pareto. Here we will present the results of [1] as applied to the Type I Pareto distribution. This is because, amongst the references stated above, these authors are the only ones who explicitly gave the expression for the point predictor of the future record value as well as a confidence interval for a future record.

We assume then that the data consist of k first upper record values $X_1, X_2, \ldots, X_k$ from the Type I Pareto distribution with cdf

$$F(x|\theta,\alpha) = 1 - \left(\frac{\theta}{x}\right)^{\alpha}, \qquad x \geq \theta,\ \theta > 0,\ \alpha > 0. \tag{21}$$

Using an appropriate conjugate prior, the posterior density of (θ, α) is computed as

$$\pi^*(\theta,\alpha|\mathbf{x}) = \frac{(b+1)\,\{I(x_k, M)\}^{k+a}}{\Gamma(k+a)\theta\alpha^{-k-a}} e^{-\alpha I(x_k,\theta)}, \qquad 0 < \alpha, \quad 0 < \theta < M \tag{22}$$

where a, b, c and d are positive real numbers (hyperparameters of the prior density), $I(u, v) = ln(u) + ln(c/v^{b+1})$ and $M = min\{x_1, d\}$.

Ahmadi and Doostparast [1] used the posterior density of (θ, α) to compute the following Bayes point predictor for the X_s, the sth future upper record ($s \geq k + 1$):

$$\hat{X}_{s(B)} = \frac{X_k E^{s+a-1}}{B(k+a, s-k)} \sum_{j=0}^{s-k-1} E^{-j} \int_0^1 \frac{\{E - \ln z\}^{j-s-a}}{z^2} dz \tag{23}$$

where $E = I(X_k, M)$.

Moreover, [1] showed that for X_s, $t(X_s) = I(x_k, M)/I(X_s, M)$ follows a Beta distribution with parameters $k+a$ and $s-k$, which is independent of X_s, and thus it is a pivotal quantity for X_s. It can therefore be used to construct a prediction interval for X_s which is given as

$$\left(X_k \exp\left\{I(X_k, M)\left(\frac{1}{b_{1-\gamma/2}} - 1\right)\right\},\ X_k \exp\left\{I(X_k, M)\left(\frac{1}{b_{\gamma/2}} - 1\right)\right\}\right) \tag{24}$$

where b_γ is the γth percentage point of the $Beta(k + a, s - k)$ distribution.

3.2 Numerical Illustration

In [1], seven records were simulated from a $P(2, 3)$ distribution. Their values were: 3.0889, 3.3358, 3.7241, 4.4956, 5.3649, 5.8284, 6.0071. They considered (22) with prior hyperparameters $a = 1, b = 2, c = 20, d = 15$. The Bayes estimates of α and β were $\widehat{\alpha}_B = 5.6930$ and $\widehat{\beta}_B = 2.8969$. By applying (24), a 95 % prediction interval for the 8th upper record in the sample is (6.0339; 13.6832).

We now turn to a review of some results on estimation of the complete sample size based on the number of records.

4 Methods for Estimating the Sample Size Based on the Number of Record-Breaking Data

The aim of this section is to review some methods for estimating the unknown sample size based on the record counting process, $N^{(n)}$, defined in (1). Methods in the literature for estimating n are given in [9, 17, 18], where estimators of the unknown sample size n are proposed based on the number of records, $N^{(n)} = k$. Practical applications of these techniques can be seen in [14, 24].

First, we review some general results about the record counting process $\left\{N^{(n)}\right\}_{n\geq 1}$. Details can be found in [5].

Lemma 1 *1. The probability mass function (pmf) of $N^{(n)}$ is*

$$P\left[N^{(n)} = k\right] = \frac{s(n,k)}{n!}, \qquad k \in \{1, \ldots, n\}, \tag{25}$$

where $s(n,k)$ denotes an (unsigned) Stirling number of the first kind.

2. For large n, [21] proposed the following approximation to (25)

$$P\left[N^{(n)} = k\right] = e^{-\lambda_n}\frac{\lambda_n^{k-1}}{(k-1)!} + O\left(\frac{1}{\ln(n)}\right) \tag{26}$$

with $\lambda_n = \ln(n) + \gamma - 1$ and γ is Euler's constant ($\gamma = 0.5772\ldots$).

3. $N^{(n)}$ can be written as

$$N^{(n)} = \sum_{j=1}^{n} I_j \tag{27}$$

where I_j are independent Bernoulli rv's with $I_j \sim Ber\left(\frac{1}{j}\right)$. Expression (27) can be used to derive properties of $N^{(n)}$. For instance, we have that the probability generating function (pgf) of $N^{(n)}$, $G_{N^{(n)}}(s)$, is

$$G_{N^{(n)}}(s) = \frac{1}{n!}\prod_{j=1}^{n}(j+s-1), \qquad s \in I\!R \tag{28}$$

and the following approximation to $E\left[N^{(n)}\right]$

$$E\left[N^{(n)}\right] \approx \ln(n) + \gamma \tag{29}$$

4. *For inferential purposes, we highlight that the support of $N^{(1)}$ is only one point, $N^{(1)} = 1$. So, by applying a result given in [18], we have that the family of pmf's of $N^{(n)}$ is complete.*

Now we turn our attention to methods in the literature for estimating n.

4.1 Methods for Estimating n

From the properties of $N^{(n)}$ given in Lemma 1, [17] proposed three methods for estimating n: an unbiased estimator, the maximum likelihood and the moments method estimator. They are based on the observed value of $N^{(n)}$ and are listed next.

4.1.1 Unbiased Estimation of *n*

An unbiased estimator of n, T_1, is given by

$$T_1 = 2^{N^{(n)}} - 1 \tag{30}$$

Properties of T_1

1. T_1 is the only unbiased estimator of n based on $N^{(n)}$. This fact follows from the fact that the family of pmf's of $N^{(n)}$ is complete.
2. The variance of T_1 is given by

$$\mathrm{Var}[T_1] = \frac{(n+3)(n+2)(n+1)}{6} - (n+1)^2. \tag{31}$$

4.1.2 Maximum Likelihood Estimation of *n*

Let k_0 denote the observed value of records and let $p_n(k) = P\left[N^{(n)} = k\right]$. The maximum likelihood estimator (MLE) is $T_2 = \hat{n}$ with $\hat{n}$ satisfying

$$p_{\hat{n}}(k_0) = \max_n p_n(k_0) \tag{32}$$

Properties of T_2

1. T_2 was characterized in [17] in terms of the mode of pmf of $N^{(n)}$, $Mo(N^{(n)})$. Explicitly

$$T_2 = \min\{n \;/\; Mo(N^{(n)}) = k_0\}, \quad k_0 \;\text{ being the observed value of } N^{(n)}. \tag{33}$$

2. Appropriate software should be used to calculate T_2 by using (33). If the calculation of (33) is cumbersome then the next approximation can be useful.
3. Cramer [9] derived bounds for T_2 and proposed the following approximation for T_2 when $N^{(n)} \geq 3$

$$n_+ = \left\lfloor \exp\left\{ \frac{N^{(n)}}{2} + \frac{3}{4} - \gamma + \sqrt{\frac{(2N^{(n)}-3)^2}{16} - \zeta(2) + \zeta(3)} \right\} \right\rfloor \tag{34}$$

   where $\lfloor x \rfloor$ denotes the integer part of x and $\zeta(\cdot)$ the Riemann zeta function. n_+ provides a good approximation to the MLE of n as can be seen in [9], Table 3.
4. Cramer [9] derived the following approximations for $E[T_2]$ and Var$[T_2]$

$$E[T_2] \in \left[G_{N^{(n)}}(e)e^{-2\gamma},\ G_{N^{(n)}}(e)e^{-\gamma} \right] \tag{35}$$

$$\text{Var}[T_2] \in \left[\frac{e^{-4\gamma}}{\Gamma(e^2)} n^{e^2-1},\ \frac{e^{-2\gamma}}{\Gamma(e^2)} n^{e^2-1} \right], \tag{36}$$

where $G_{N^{(n)}}(\cdot)$ was given in (28). For large n, $G_{N^{(n)}}(e)$ in (35) can be approximated by $\frac{n^{e-1}}{\Gamma(e)}$.

4.1.3 Method of Moments Estimator of *n*

From (29) a Method of Moments Estimator (MME) of n is given by $T_3 = e^{N^{(n)}-\gamma}$.

Properties of T_3

The mean and variance of T_3 are

$$E[T_3] = G_{N^{(n)}}(e)e^{-\gamma} \tag{37}$$

$$Var[T_2] = e^{-2\gamma} \left[G_{N^{(n)}}(e^2) - G_{N^{(n)}}(e)^2 \right] \tag{38}$$

with $G_{N^{(n)}}(\cdot)$ given in (28). In order to obtain an integer estimate of n, T_3 should be modified to T_3^*

$$T_3^* = nint(T_3) = nint\left(e^{N^{(n)}-\gamma}\right) \tag{39}$$

where $nint(\cdot)$ denotes the nearest integer function.

4.1.4 Additional Properties

It is illustrated in [9, 17] that for $k_0 \geq 3$ the following relationship holds for T_1, T_2, and T_3:

$$T_1 < T_2 \leq T_3 \tag{40}$$

Remark 1 For $k_0 = 1$, $T_1 = T_2 = 1$ and $T_3^* = 2$, while for $k_0 = 2$ we have that $T_1 = 3$, $T_2 = 2$ and $T_3^* = 4$.

Also we highlight that, while we have explicit expressions for T_1, T_3 and their characteristics, computational methods and/or approximations are needed to calculate T_2, its mean and variance.

Finally, note that T_2 and T_3 are biased estimators of n, since from (40)

$$n < E[T_2(N^{(n)})] \leq E[T_3(N^{(n)})].$$

The biases of T_2 and T_3 can be assessed by using (35) and (37), respectively.

4.1.5 Numerical Illustration

We use the following example from [5] to illustrate the practical use of the methods given in Sect. 4. A rock crushing machine has to be reset if, at any operation, the size of the rock being crushed is larger than any that has been crushed before. In this setting, let us suppose we just know the number of times that the machine has been reset, that is the number of record-breaking values. We want to estimate the total number of operations based on the number of times it has been reset.

Suppose that we observe that the machine has been reset three times, i.e. $k_0 = 3$. By applying (30), (33), and (39), we have the following estimates of n given in Table 1. Note that the data set for the sizes of the rocks up to the third time the machine had to be reset as given in [5] and consisted of the following n = 12 values: 9.3, 0.6, 24.4, 18.1, 6.6, 9.0, 14.3, 6.6, 13.0, 2.4, 5.6, 33.8. Thus, for this data set, the best estimator of the sample size turns out to be T_3^*.

We also want to point out here that for the above data set, Arnold et al. [5] assumed an underlying exponential distribution and using a vague prior, they computed a 95 % Bayesian prediction interval for the fourth record as (33.8, 91.7). Arnold et al. [5] also investigated the use of a one-parameter Rayleigh distribution (and a two-parameter Rayleigh) for the same data, to compute a 90 % prediction interval for the fourth record as (33.8, 49.6) (under the assumption of a two-parameter Rayleigh, the interval was computed as (33.8, 69.1).)

To conclude this section, we propose a result that can be used to estimate a function of the sample size.

Table 1 Estimates of n for $k_0 = 3$

$N^{(n)} = k_0$	T_1	T_2	T_3^*
3	7	8	11

4.2 Unbiased Estimation of Functions of the Sample Size

It is possible to apply the results given in [18] to $N^{(n)}$ to develop unbiased estimators of functions of sample size. These estimators can be useful in stress and fatigue tests where the practitioner is often interested in the cost of the experiment which is typically a function of the number n of items in the test.

Let $S(N^{(n)})$ denote the support of $N^{(n)}$ and $p_n(k)$ the pmf of $N^{(n)}$. Note that $S(N^{(n)}) = \{1, 2, \ldots, n\}$.

Lemma 2 *The family of pmf's of $N^{(n)}$ is recursive on $S(N^{(n)})$, (see [18]).*

This result means that the support and the pmf of $N^{(n)}$ satisfy certain recursive relations. Explicitly

1. The support of $S(N^{(n)})$ satisfies

$$S(N^{(1)}) = 1 \quad \text{and} \quad S(N^{(n+1)}) = S(N^{(n)}) \cup \{n+1\}$$

2. For $n \geq 1$, the pmf of $N^{(n)}$ satisfies the recursive relation

$$p_{n+1}(k) = q_{n+1} p_n(k) + (1 - q_{n+1}) p_n(k-1),$$

where $q_{n+1} = \frac{n}{n+1}$ and $p_1(1) = 1$.

Theorem 1 *Any function of sample size, $h(n)$, admits an unbiased estimator based on $N^{(n)}$, $g(N^{(n)})$. This is characterized by the following relationships*

$$g(1) = h(1) \tag{41}$$

$$g(n+1) = \left\{ (n+1)h(n+1) - nh(n) - \sum_{i=1}^{n} g(i) p_n(i-1) \right\} \frac{1}{p_n(n)} \tag{42}$$

Since $S(N^{(1)}) = \{1\}$, the family of pmf's of $N^{(n)}$ is complete. Therefore the unbiased estimator of $h(n)$ based on $N^{(n)}$ is uniquely determined. (See [18])

4.2.1 Application: Estimation of Cost Functions in Stress and Fatigue Testing

Engineers often test items to determine their operating life under a given level of stress, breaking points or safe usage limits. These kind of studies are also usual in

Table 2 Estimates of some cost functions for $N^{(n)} = 3$

$N^{(n)} = 3$	$TC(n) = n^2 + n + 1$	$MC(n) = 2n + 1$	$V(n) = \frac{n^2+n+1}{n}$
Estimates	39	15	8

material sciences where they are known as fatigue testing. Details and examples can be seen in [19]. In this setting, inference based on records can be of interest. Specifically, suppose that we are sampling from a continuous rv and we only know the number of records, but we want to estimate the cost of the experiment given as a function of the number of items in the test , n. Some cost functions used in Economics are

Total Cost denoted by $TC(x)$.
Marginal Cost defined as $MC(x) = \frac{d}{dx} TC(x)$.
Average Cost defined as $AV(x) = \frac{TC(x)}{x}$.

Particular cases of total cost functions of the sample size that can be of interest in reliability and survival analysis are

1. $TC(n) = a + bn + cn^2$, with $c > 0$. This cost function leads to increasing marginal costs.
2. $TC(n) = a + bn + cn^2 + dn^3$, with appropriate coefficients b, c, d, produces U-shaped marginal costs.

Clearly the above cost functions are non linear functions of n, and therefore results in Theorem 1 can be applied.

As illustration consider the total cost function $TC(n) = n^2 + n + 1, n \geq 1$. In this case $MC(n) = 2n + 1$ and $AV(n) = \frac{n^2+n+1}{n}$. If we observe 3 records, by applying (42), we have the estimates given in Table 2.

Acknowledgments The authors thank the referees for their constructive suggestions, which helped to improve the presentation of the paper. Barranco-Chamorro's research was supported by grant UJA2013/08/01.

References

1. Ahmadi, J., Doostparast, M.: Bayesian estimation and prediction for some life distributions based on record values. Stat. Pap. **47**, 373–392 (2006)
2. Ahsanullah, M.: Record Values. Theory and Applications. University Press of America, Lanham (2004)
3. AL-Hussaini, E.K., Ahmad, A.E.: On Bayesian interval prediction of future records. Test **12**(1), 79–99 (2003)
4. Arnold, B.C., Press, S.J.: Bayesian inference for Pareto populations. J. Econom. **21**, 287–306 (1983)
5. Arnold, B.C., Balakrishnan, N., Nagaraja, N.: Records. Wiley, New York (1998)

6. Balakrishnan, N., Ahsanullah, M.: Recurrence relations for single and product moments of record values from generalized Pareto distribution. Commun. Stat.: Theory Methods **23**(10), 2841–2852 (1994)
7. Balakrishnan, N., Cohen, A.C.: Order Statistics and Inference: Estimation Methods. Academic Press, San Diego (1991)
8. Chandler, K.N.: The distribution and frequency of record values. J. R. Stat. Soc. Ser. B **14**(2), 220–228 (1952)
9. Cramer, E.: Asymptotic estimators of the sample size in a record model. Stat. Pap. **41**(2), 159–171 (2000)
10. Chan, P.S.: Interval estimation of location and scale parameters based on record values. Stat. Probab. Lett. **37**(1), 49–58 (1998)
11. Gulati, S., Padget, W.J.: Parametric and Nonparametric Inference from Record-Breaking Data. Lecture Notes in Statistics, vol. 172. Springer, Berlin (2003)
12. Gulati, S., Shapiro, S.S.: Goodness of fit tests for the Pareto distribution. In: Vonta, F., Nikulin, M., Limnios, N., Huber, C. (eds.) Statistical Models and Methods for Biomedical and Technical Systems, pp. 263–277. Birkhauser, Boston (2008)
13. Gupta, R.D., Kundu, D.: Generalized exponential distribution. Aust. N. Z. J. Stat. **41**(2), 173–188 (1999)
14. Johnson, R.: Record values and surviving glacial moraines. Teach. Stat. **25**(3), 66–69 (2003)
15. Lai, C.D., Xie, M., Murthy, D.N.: A modified Weibull distribution. IEEE Trans. Reliab. **52**, 33–37 (2003)
16. Madi, M.T., Raqab, M.Z.: Bayesian prediction of temperatures using the Pareto model. Environmetrics **15**, 701–710 (2004)
17. Moreno-Rebollo, J.L., Barranco-Chamorro, I., López-Blázquez, F., Gómez-Gómez, T.: On the estimation of the unknown sample size from the number of the records. Stat. Probab. Lett. **31**, 7–12 (1996)
18. Moreno-Rebollo, J.L., López-Blázquez, F., Barranco-Chamorro, I., Pascual-Acosta, A.: Estimating the unknown sample size. J. Stat. Plan. Inference **83**, 311–318 (2000)
19. Nelson, W.B.: Accelerated Testing-Statistical Models, Test Plans, and Data Analysis. Wiley, New York (2004)
20. Nevzorov, V.B., Balakrishnan, N.: A record of records. In: Handbook of Statistics-16: Order Statistics: Theory and Methods, pp. 515–570. North-Holland, Amsterdam (1998)
21. Pfeifer, D.: Some remarks on Nevzorov's record model. Adv. Appl. Probab. **23**, 823–834 (1991)
22. Raqab, M.: Inferences for generalized exponential distribution based on record statistics. J. Stat. Plan. Inference **104**, 339–350 (2002)
23. Roberts, E.M.: Review of statistics of extreme values with application to air quality data. Part II. Applications J. Air Pollut. Control Assoc. **29**, 733–740 (1979)
24. Soh Fotsing, B.D., Anago, G.F., Fogue, M.: Statistical techniques of sample size estimating in fatigue tests. Int. J. Eng. Technol. **2**(6), 477–481 (2010)
25. Soliman, A.A., Abd Ellah, A.H., Sultan, K.S.: Comparison of estimates using record statistics from Weibull model: Bayesian and non-Bayesian approaches. Comput. Stat. Data Anal. **51**, 2065–2077 (2006)
26. Sultan, K.S.: Record values from the modified Weibull distribution and applications. Int. Math. Forum **2**(41), 2045–2054 (2007)
27. Sultan, K.S., Moshref, M.E.: Record values from generalized Pareto distribution and associated inference. Metrika **51**, 105–116 (2000)

Risk Scoring Models for Trade Credit in Small and Medium Enterprises

Manuel Terradez, Renatas Kizys, Angel A. Juan, Ana M. Debon and Bartosz Sawik

Abstract Trade credit refers to providing goods and services on a deferred payment basis. Commercial credit management is a matter of great importance for most small and medium enterprises (SMEs), since it represents a significant portion of their assets. Commercial lending involves assuming some credit risk due to exposure to default. Thus, the management of trade credit and payment delays is strongly related to the liquidation and bankruptcy of enterprises. In this paper we study the relationship between trade credit management and the level of risk in SMEs. Despite its relevance for most SMEs, this problem has not been sufficiently analyzed in the existing literature. After a brief review of existing literature, we use a large database of enterprises to analyze data and propose a multivariate decision-tree model which aims at explaining the level of risk as a function of several variables, both of financial and non-financial nature. Decision trees replace the equation in parametric regression models with a set of rules. This feature is an important aid for the decision process of risk experts, as it allows them to reduce time and then the economic cost of their decisions.

Keywords Trade credit · Scoring models · Small and medium enterprises · Multivariate regression · Decision trees

M. Terradez (✉) · A.M. Debon
Universitat Politecnica de Valencia, Valencia, Spain
e-mail: mterrade@eio.upv.es

A.M. Debon
e-mail: andeau@eio.upv.es

R. Kizys
Portsmouth University, Portsmouth, UK
e-mail: renatas.kizys@port.ac.uk

A.A. Juan
IN3 - Open University of Catalonia, Barcelona, Spain
e-mail: ajuanp@uoc.edu

B. Sawik
AGH University of Science and Technology, Krakow, Poland
e-mail: bsawik@zarz.agh.edu.pl

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_26

1 Introduction

The current financial crisis has renewed the interest in research and development of failure prediction models for all of the corporate and retail sectors [3]. The literature on the modeling of credit risk for large, listed companies is extensive, and it either uses historical accounting data to predict insolvency or models that rely on market information. However, market information is not available for small and medium enterprises (SMEs), which require risk management tools and methodologies specifically developed for them. Research on credit risk management for SMEs is relatively scarce. This research aims to partially fill this void by analyzing the risk of trade credit operations in SMEs. Trade credit (TC) involves supplying goods and services on a deferred payment basis; that is, giving the customer time to pay. Thus, TC is an 'implicit short-term loan from non-financial suppliers to their clients. It occupies a prominent place in the world of business and is one of the most important forms of credit available to businesses. Because TC represents such an important share of total assets or liabilities, managing it is critical for the businesses, especially for SMEs.

While the actual cost of institutional credit remains close to the nominal cost, the cost of TC can vary widely. In effect, if significant discounts for early payment are considered, TC can become an expensive way of borrowing. The cost of TC is reflected in both the level of credit (the amount purchased on credit) and the length of the credit period (the number of days taken before payment is made). We discuss here some basic concepts and models to predict default risk in SMEs based on TC indicators. We explore the use of several models, including both classical statistical and econometric models (e.g., logistic regression and multiple discriminant analysis) as well as data-mining techniques (e.g., decision trees, neural networks, nearest neighbor, etc.). As discussed in the abstract, our results suggest that decision trees have the best fit, where CHAID (Chi-squared Automatic Interaction Detector) provides better prediction of defaults than CART (Classification and Regression Tree). We also find that the most important predictor is the ratio of accounts payable over total liabilities, with larger values of this ratio implying a greater risk of default. Other important predictors are the ratio of accounts payable over accounts receivable and sales growth.

The remainder of this paper is organized as follows. Section 2 contains a literature review on the topic. Section 3 addresses measurement and estimation issues of default risk. Section 4 describes the data employed in this study. Section 5 provides an overview of our methodology. Finally, Sect. 6 offers some concluding remarks.

2 Literature Review

Early research into corporate failure prediction involved determining which accounting ratios best predict failure, primarily employing Multiple Discriminant Analysis (MDA) or Logit/Probit models. Usually, ratios are calculated a year before bankruptcy or default and thus these are static models. Altman [1] and Ohlson [20] pioneered models to predict failure using these financial ratios. Altman used MDA, which was echoed by Deakin [11] and Micha [19], inter alia. Ohlson introduced a logistic regression model, which has several advantages over MDA (see next section for a discussion on this matter) and a wealth of studies followed this direction [4, 5].

Credit risk models for private companies are limited by data availability. Market datum are not available for unlisted firms. Moreover, some of the datum required to calculate accounting ratios in studies of the failure of listed companies is not available for SMEs. Other studies using a variety of statistical techniques, have contributed to the knowledge of the insolvency indicators, both financial [2, 9] and non-financial [3, 14] that arise in SMEs. In particular, Fantazzini [12] propose a non-parametric survival approach with a random-forest model, but they also conclude that a simpler logit model outperforms the random-forest model in the out-of-sample validation.

As we discussed in the previous section, managing trade credit is critical for the businesses, especially for SMEs, as it represents an important share of total assets. Therefore, it is not surprising that recent research has focused on the links between the management of TC (and delays in payment) and the liquidation or bankruptcy of enterprises, or even the refinancing (or restructuring) of debt [10, 29]. Commercial lending involves credit risk due to exposure to default that can have negative effects on probability and liquidity [8]. According to [21], the proper management of the TC offered (as a supplier) is critical to the survival and success of business. These authors also conclude that most SMEs are not proactive in their management of credit, and that they do not employ risk models (according to them, about 83 % of SMEs do not classify their customers using risk categories).

As noticed by Boissay [7], credit-constrained firms facing liquidity problems from their customers are more likely to not pay their suppliers. However, because of the difficulty of obtaining data, the line of research that studies the relation between the management of TC and risk is not sufficiently developed.

3 Measuring Default Risk in SMEs

According to European standards, SMEs have less than 250 employees and sales figures under 50 million EUR (or total assets under 43 million EUR). As pointed out by Altman et al. [3], two of the main factors behind failed SMEs are insufficient capitalization and lack of planning. In the related literature it is common to find terms related to high levels of risk, such as: insolvency, bankruptcy, failure, default, etc. All of these terms are quite similar, albeit with small differences. In fact, they can be used interchangeably in a modeling framework, since they are usually transformed into a binary variable that takes on the value 1 if the event occurs and takes on the value 0 otherwise. In this paper we use the term default. Accordingly, we use the probability of default (PD) as a measure of risk. Notice, however, that failure and closure are different concepts: while failure generally implies closure, the inverse is not true a firm's closure may be due to other reasons. Several factors can affect a PD, such as the firm's leverage, profitability or cash flows. A scoring model specifies how to combine the different pieces of information in order to get an accurate assessment of the PD.

Assume we have have annual firm-level data on default factors and default behavior. The binary variable of default will take on value 1 if the firm eventually defaults in the year following the one observed for the factors, and zero otherwise. A score summarizes the information contained in factors $x_1, x_2, ..., x_k$ that affect the PD. Ideally, the scoring model should predict a high PD for those firms that eventually will default and a low PD for those that will not. Logistic regression models can be used to predict default because the response variable is binary and they yield a score between 0 and 1, which can be interpreted as the client's PD. The model coefficients signal the importance of each predictor in the explanation of the estimated probability of default. A score summarizes the information contained in factors that affect the PD, e.g.: $score = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$. The logistic function, $P = 1/(1 + exp(-score))$, is usually applied to link scores to PDs. The ratio $P/(1 - P)$ is called odds-ratio and facilitates the model interpretation. $Log[P/(1 - P)]$ is called $logit(P)$ and thus the associated models are called logit models. A natural way of estimating the coefficients of the model is throughout the maximum likelihood method, i.e., the coefficients are chosen such that the probability of observing the given default behavior is maximized.

Alternative nonlinear techniques that can be used to approach this problem include decision trees and neural networks, among others [6, 16]. A decision tree is a set of conditions organized in a hierarchical structure, so that the final decision can be determined by the fulfillment of the rules from the root of the tree to one of its end nodes. One of the great advantages of this technique is that the possible options from a given condition are exclusive, the analysis of a situation, which allows one to analyze a situation, follow the tree properly, get an action or take a decision.

Two of the main techniques for developing trees are CART (Classification and Regression Tree) and CHAID (Chi-squared Automatic Interaction Detector): CART performs binary partitions and assigns a mean and variance to each node, trying to select partitions that reduce the variance of the child nodes; CHAID performs non-binary partitions and uses a Chi Square test to determine the optimal partition.

Instead of the well-known R^2 statistic, suitable for linear models, in the case of nonlinear models we can report the $Pseudo - R^2$, which is also bounded by 0 and 1—with higher values indicating a better fit. However, in this work we use an alternative measure of fit which is frequently used in binary models: the ROC curve. A ROC curve is a technique for visualizing, organizing and selecting classifiers based on their performance. It has its origin in Signal Detection Theory [28] and has been widely accepted and commonly used in fields such as Psychology [18] and Medicine [30]. It has also been introduced in other fields that are more related with our work, such as Economics [27] and Data-mining [15]. ROC curves are particularly useful for comparing the classification power of different estimated models. Details of ROC curves are provided in [13]. The degree of predictability of the model is defined by the area under the ROC curve (AUC), which is constructed for all possible cutoff points to classify positive or negative events. Since the AUC is a portion of the area of the unit square, its value will always be between 0 and 1, where the random guessing procedure has an area of 0.5. As with the $Pseudo - R^2$, the greater is the AUC the better is the classifier.

Alternative measures of risk that can be used to approach the problem of trade credit risk are Value-at-Risk (VaR) and Conditional Value-at-Risk (CVaR). VaR and CVaR have been widely used in the field of financial engineering (e.g. [25, 26]). CVaR is used in conjunction with VaR and is applied for estimating the risk with non-symmetric cost/return distributions. Rockafellar and Uryasev [23, 24] introduced a new approach to select a set of investments with the reduced risk of high losses. The problem was solved by calculating VaR and minimizing CVaR simultaneously. For trade credit VaR can be defined as accepted threshold of risk by decision-maker, in that case CVaR would be worst case accepted level of TC risk. Optimization models with VAR and CVaR could be used to shape the distribution of TC risk in a favorable way for a decision maker.

4 Data Sources

The Iberian Balance sheet Analysis System (SABI) is a database that includes information on the balance sheets of more than 1.2 million Spanish and more than 0.4 million Portuguese companies. A random sample (extracted from SABI) with more than 5,000 active Spanish SMEs was used. We selected active SMEs offering accounting records for the previous year (e.g., 2011), so that we could extract data on at least one of the following variables: accounts receivable, or accounts payable. The SABI database does not provide information on companies default behavior, but it provides some risk measures. One of these measures is the scores from the Multi Objective Rating Evaluation (MORE), a proprietary scoring model. The MORE rating consists of 10 categories indicated by traditional symbols used by rating agencies: AAA to D, with CC being the 8th if we rank them from most creditworthy to less creditworthy. One of our goals is to develop a scoring model, alternative to MORE, based on predictors related to TC. Our scoring model should have a transparent and replicable methodology under the assumption that SMEs typically apply homogeneous risk rules to all their TC customers. In contrast to MORE, in this research we also propose a model that allows for customized TC rules that suit different customers and, thus, leads to reduced levels of default risk. In our model we use a binary dependent variable (response) which, based on the MORE score, classifies SMEs into two categories: risky and non-risky companies. The former comprise companies with standard CC or lower rating featuring a relatively high probability of default. The non-risky companies comprise all the remaining SMEs in the sample with a relatively low probability of default.

As for the independent variables (predictors) most of them are variables related to TC. Specifically, we used: (a) DAR: "Days accounts receivable (debtors)"; (b) DAP: "Days accounts payable (creditors)"; (c) AR_Assets: "Ratio (Accounts receivable/Total assets)"; (d) AP_Liab: "Ratio (Accounts payable/Total liabilities)";

Table 1 Descriptive statistics of the considered variables

Variable	N	Min	Max	Mean	St. Dev
Descriptive statistics					
Employees	5094	1	244	9.20	18.673
Age	5093	3.0056	87.7889	15.038873	8.9831294
Log_AR_Assets	5094	−15.26	0.00	−2.2883	1.90025
Log_AR_Liab	5094	−12.56	4.05	−1.7992	1.52436
Log_AP_AR	5048	−10.04	13.56	0.5515	1.88502
Log_SalesGrowth	5094	−9.34	6.24	−0.0826	62420
Log_ARGrowth	5094	−9.71	10.74	−0.0681	1.14850
Log_APGrowth	5094	−8.34	7.05	−0.0333	0.81845
Log_DAP	5094	−2.66	13.20	4.8826	1.45649
Log_DAR	5094	−2.30	14.05	3.5273	2.01879
ValidN (listwise)	5047				

(e) AP_AR: "Ratio (Accounts payable/Accounts receivable)"; (f) APGrowth: "Ratio (Accounts payable [last year]/Accounts payable [previous year])"; and (g) ARGrowth: "Ratio (Accounts receivable [last year]/Accounts receivable [previous year])". We also used some other factors which, although not directly related to default on TC, help us account for the existing heterogeneity of SMEs. These factors are: number of employees, age (years in operation), activity sector, and sales growth in the last year. All the variables were obtained for the last accounting year, that is, 2011. Growth variables were derived comparing 2011 and 2010. Table 1 shows some descriptive statistics of the variables. Also, the associated correlations are given in Table 2.

5 Methodology and Results

All enterprises with less than three years of operation were excluded from the study, since their accounting ratios and business behavior are not consolidated enough and their inclusion could mislead the results of the analysis. In the first step, we performed a log transformation of our continuous independent variables featuring high concentration on low values but long positive tails. We also used dummy predictors to include one categorical variable; namely activity sector (agriculture, manufacturing, building, services). In a second step we estimated various models, including both classical statistical and econometric models (logistic regression and multiple discriminant analysis) as well as data-mining techniques (CART and CHAID decision trees, neural networks, and nearest neighbor). Our results suggest that decision trees show the best fit. An appealing feature of decision trees is that they are easy to implement and interpret. Neural networks provided equivalent results in our case, but they are

Table 2 Correlations between pairs of variables

Pearson correlations

	Employees	Age	Log_AR_ Assets	Log_AP_Liab	Log_AP_AR	Log_ SalesGrowth	Log_ ARGrowth	Log_ APGrowth	Log_DAP	Log_DAR
Employees	1	0.242[a]	0.113[a]	0.059[a]	−0.077[a]	0.048[a]	0.020	0.005	0.001	0.096[a]
Age	0.242[a]	1	−0.001	−0.094[a]	−0.086[a]	−0.017	−0.014	−0.047[a]	−0.014	0.086[a]
Log_AR_ Assets	0.113[a]	−0.001	1	0.338[a]	−0.676[a]	0.041[a]	0.299[a]	0.037[a]	−0.003	0.512[a]
Log_AP_ Liab	0.059[a]	−0.094[a]	0.338[a]	1	0.411[a]	0.054[a]	0.028[b]	0.258[a]	0.309[a]	0.027
Log_AP_ AR	−0.077[a]	−0.086[a]	−0.676[a]	0.411[a]	1	0.003	−0.272[a]	0.167[a]	0.280[a]	−0.580[a]
Log_ SalesGrowth	0.048[a]	−0.017	0.041[a]	0.054[a]	0.003	1	0.155[a]	0.223[a]	−0.065[a]	−0.157[a]
Log_ ARGrowth	0.020	−0.014	0.299[a]	0.028[b]	−0.272[a]	0.155[a]	1	0.236[a]	0.015	0.217[a]
Log_ APGrowth	0.005	−0.047[a]	0.037[a]	0.258[a]	0.167[a]	0.223[a]	0.236[a]	1	0.243[a]	0.020
Log_ DAP	0.001	−0.014	−0.003	0.309[a]	0.280[a]	−0.065[a]	0.015	0.243[a]	1	0.156[a]
Log_ DAR	0.096[a]	0.086[a]	0.512[a]	0.027	−0.580[a]	−0.157[a]	0.217[a]	0.020	0.156[a]	1

[a] Correlation is significant at the 0.01 level (2-tailed)
[b] Correlation is significant at the 0.05 level (2-tailed)

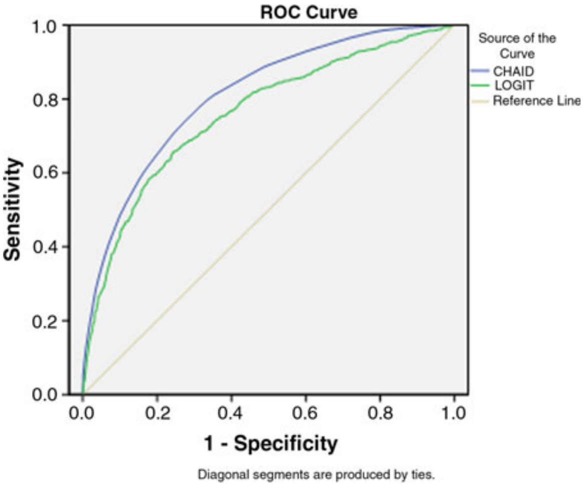
much more difficult to interpret so we preferred to go with the decision trees. Results obtained with CHAID and CART decision trees are very similar, but CHAID seems to provide slightly better predictions for the target category (“default”). Therefore, our preferred model is a decision tree based on the CHAID technique. Our model has a maximum depth of 5 levels, a minimum node size of 20 individuals, the Pearson Chi-Square statistic is used to decide the joining and division of nodes, and there are 6 intervals for the continuous predictors. The tree has 51 terminal nodes, which can be considered too many and could lead to an over-fitting problem. In order to minimize this problem, we carry out a 10-fold Cross Validation to validate the results.

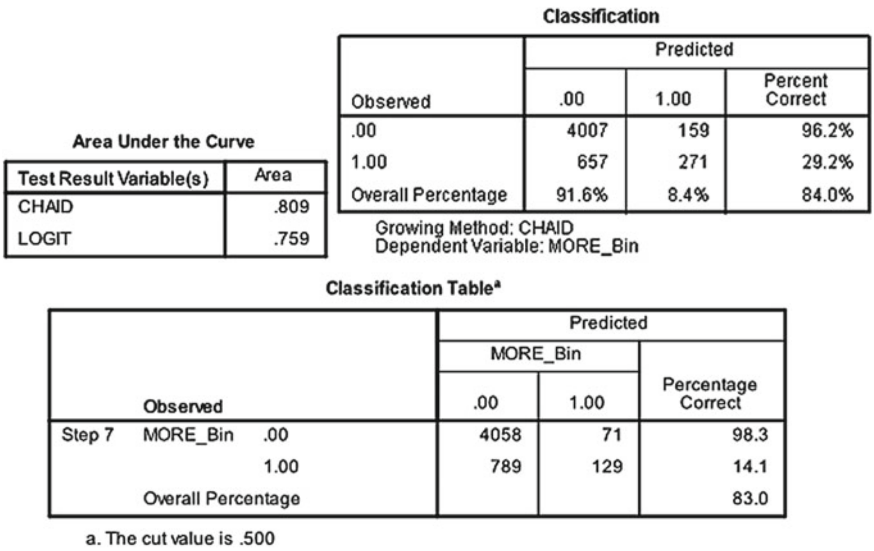
Figures 1 and 2 suggest that our CHAID model shows a fairly acceptable performance: AUC = 0.808 with respect to MORE, and 84 % of success in predicting the right category. In fact, it shows a reasonable success in predicting both categories (96.2 % of non-defaults and 29.2 % of defaults were successfully predicted). Furthermore, it shows a better goodness of fit than the logit model.

Figure 3 shows the first two branches (13 nodes) of the final tree due to space limitations we do not depict the whole tree here.

The most important predictor is the “accounts payable/total liabilities” ratio. Larger values of this ratio imply a greater risk of default, with thresholds located at 0.36 and 0.59 (after transforming the model back to the levels). Other important variables are the “ratio of accounts payable over accounts receivable” and “sales growth” (SG). Our analysis of nodes 1 and 3 and their child nodes indicates that when AP_Liab is high or low, the second important variable to look at is SG. As expected, the lower the value of SG the riskier is the company. However, when AP_Liab is medium (node 2), the second important variable is AP_AR, with larger values relating to riskier firms, too. While all of the initial variables are included in the final model, it result that the variables “accounts payable growth” and “activity sector” have less influence on the firm’s score than the other ones.

Fig. 1 ROC curves for the CHAID and logit models





Area Under the Curve

Test Result Variable(s)	Area
CHAID	.809
LOGIT	.759

Classification

Observed	Predicted .00	Predicted 1.00	Predicted Percent Correct
.00	4007	159	96.2%
1.00	657	271	29.2%
Overall Percentage	91.6%	8.4%	84.0%

Growing Method: CHAID
Dependent Variable: MORE_Bin

Classification Table[a]

Observed			Predicted MORE_Bin .00	Predicted MORE_Bin 1.00	Predicted Percentage Correct
Step 7	MORE_Bin	.00	4058	71	98.3
		1.00	789	129	14.1
	Overall Percentage				83.0

a. The cut value is .500

Fig. 2 Numerical performance of the CHAID and logit models

In order to test the performance of our model, we used a reference model: the classical logit model proposed by Pozuelo et al. [22]. This logit model offers results which are fairly close to those obtained with MORE (AUC = 0.92 with respect to MORE). The logit model takes into account the main financial dimensions, such as: solvency, liquidity, profitability, leverage, etc. This logit model provides a good benchmark because it uses the same database (SABI) and the same population (Spanish SMEs) as in our study. An advantage of this model over the Altman model is that financial ratios used by Altman are not very common in Spanish balance sheets. Accordingly, some studies [17] emphasize the limited applicability of the Altman model on Spanish firms.

6 Conclusions and Future Work

Our research, based on a credit scoring model, provides evidence that trade credit is a good proxy of risk for Spanish SMEs. This means that firms can reduce their risk by managing TC properly, which implies adjusting the ratio of accounts payable over total liabilities as a first step. Our research also makes a significant contribution to the relatively scarce literature on the application of decision trees to credit risk analysis. The decision tree is distinguished by several aspects that provide better practical results than the parametric models. The results obtained using tree methods for classification or regression can be summarized in a series of (usually few) logical

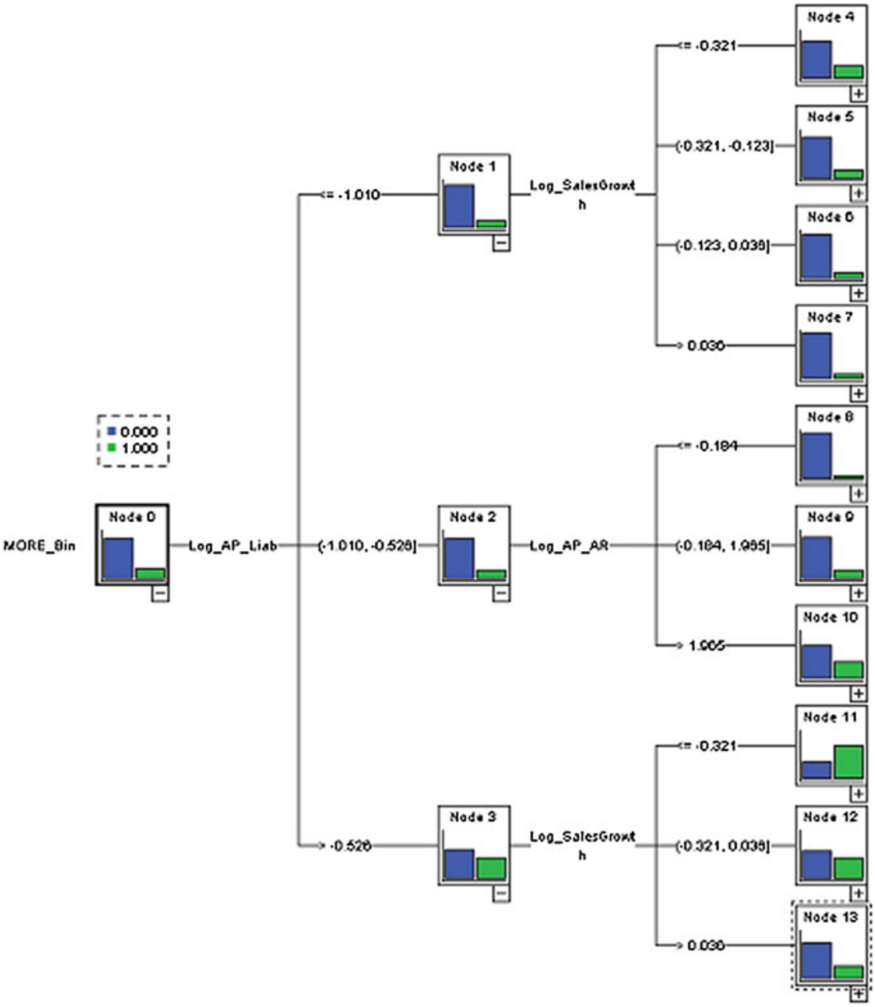


Fig. 3 First two branches of our CHAID tree model

if-then conditions (tree nodes). This makes it easy to understand and interpret the model.

This paper aims at being a first step towards a project where the main goal is to help companies reduce their global risk by customizing TC rules to different customers. In future work, we plan to combine the scoring model developed here with metaheuristic algorithms in order to support TC risk decision-making in SMEs. We also plan to develop new optimization models with VaR and CVaR as risk measures for TC. These kind of models will provide a decision maker with a tool for evaluating the relationship between expected and worst-case TC risk level.

Acknowledgments This work has been partially supported by the NCN grant (6459/B/T02/2011/40) and AGH grant (11.11.200.274).

References

1. Altman, E.I.: Financial ratios, discriminant analysis and prediction of corporate bank ruptcy. J. Financ. **23**, 589–609 (1968)
2. Altman, E.I., Sabato, G.: Modeling credit risk for SMEs: evidence from the US market. ABACUS **43**, 332–357 (2007)
3. Altman, E.I., Sabato, G., Wilson, N.: The value of non-financial information in SME risk management. J. Credit Risk **6**, 1–33 (2010)
4. Aziz, A., Emanuel, D.C., Lawson, G.H.: Bankruptcy prediction: an investigation of cash flow based models. J. Manag. Stud. **25**, 419–437 (1988)
5. Becchetti, L., Sierra, J.: Bankruptcy risk and productive efficiency in manufacturing firms. J. Bank. Financ. **27**, 2099–2120 (2002)
6. Berry, M.J.A., Linoff, G.: Data Mining Techniques. Wiley, New York (1997)
7. Boissay, F., Gropp, R.: Payment defaults and interfirm liquidity provision. Rev. Financ. 1–42, (2013), doi:10.1093/rof/rfs045
8. Cheng, N., Pike, R.: The trade credit decision: evidence of UK firms. Manag. Decis. Econ. **24**, 419–438 (2003)
9. Correa, A., Acosta, M., Gonzalez, A.L.: La insolvencia empresarial: un anlisis emprico para la pyme. Revista de Contabilidad **6**, 47–79 (2003)
10. Cunat, V.: Trade credit: suppliers as debt collectors and insurance providers. Rev. Financ. Stud. **20**, 491–527 (2007)
11. Deakin, E.B.: A discriminant analysis of predictors of business failure. J. Account. Res. **10**, 167–179 (1972)
12. Fantazzini, D., Figini, S.: Random survival forest models for SME credit risk measurement. Methodol. Comput. Appl. Probab. **11**, 29–45 (2009)
13. Fawcett, T.: An introduction to ROC analysis. Pattern Recognit. Lett. **27**, 861–874 (2006)
14. Grunert, J., Norden, L., Weber, M.: The role of non-financial factors in internal credit ratings. J. Bank. Financ. **29**, 509–531 (2004)
15. Hastie, T, Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Berlin (2001)
16. Hernandez, J., Ramirez, M.J., Ferri, C.: Introducción a la minería de datos. Pearson Prentice Hall (2004)
17. Lizarraga, F.: Modelos de prevision del fracaso empresarial: funciona entre nuestras empresas el modelo de Altman de 1968? Revista de Contabilidad **1**, 137–164 (1998)
18. Metz, C.E., Kronman, H.B.: Statistical significance tests for binormal ROC curves. J. Math. Psychol. **22**, 218–243 (1980)
19. Micha, B.: Analysis of business failures in France. J. Bank. Financ. **8**, 281–291 (1984)
20. Ohlson, J.: Financial ratios and the probabilistic prediction of bankruptcy. J. Account. Res. **18**, 109–131 (1980)
21. Poutziouris, P., Michaelas, N., Soufani, K.: Financial management of Trade Credits in SMEs. Working paper. Concordia University. http://www.efmaefm.org/efma2005/papers/241-soufani_paper.pdf
22. Pozuelo, J., Labatut, G., Veres, E.: Análisis descriptivo de los procesos de fracaso empresarial en microempresas mediante técnicas multivariantes. Revista Europea de Direccin y Economa de la Empresa, **19**, 47–66 (2010)
23. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. J. Risk **2**(3), 21–41 (2000)

24. Rockafellar, R.T., Uryasev, S.: Conditional value-at-risk for general loss distributions. J. Bank. Financ. **26**, 1443–1471 (2002)
25. Sarykalin S., Serraino G., Uryasev S.: Value-at-risk vs. conditional value-at-risk in risk management and optimization. In: Chen, Z-L., Raghavan, S., Gray, P., (Eds.) Tutorials in Operations Research, INFORMS Annual Meeting, Washington DC, USA, October 12–15 (2008)
26. Sawik, B.: Downside risk approach for multi-objective portfolio optimization. In: Klatte, D., Lthi, H.-J., Schmedders, K. (eds.) Operations Research Proceedings 2011, Operations Research Proceedings, pp. 191–196. Springer, Heidelberg (2012)
27. Sobehart, J.R., Keenan, S.C.: A practical review and test of default prediction models. RMA J. **84**, 54–59 (2001)
28. Swets, J.A.: Signal Detection Theory and ROC Analysis in Psychology and Diagnostics. Collected Papers Lawrence Erlbaum Associates (1996)
29. Wilner, B.: The exploitation of relationships in financial distress: the case of trade credit. J. Financ. **55**, 153–178 (2000)
30. Zweig, M.H., Campbell, G.: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. **39**, 561–577 (1993)

Signatures of Systems with Non-exchangeable Lifetimes: Some Implications in the Analysis of Financial Risk

Roy Cerqueti and Fabio Spizzichino

Abstract We review the basic aspects of the concept of signature for a coherent system. The cases of exchangeability and non-exchangeability are compared in view of possible applications to the analysis of financial risk. The case of a special class of basket option is finally analyzed.

Keywords Coherent systems · Signature · Excheangeability · Option theory

1 Introduction

The concept of *signature*, introduced by [12], is a simple and useful tool for the analysis of a reliability system. Since its first inception, the relevance of this concept when dealing with "coherent systems" (see [2]) became evident. For a wide review of this topic we address the reader to the references cited in the bibliography and, in particular, to [3, 5, 6, 13].

One basic problem in system reliability lies in the analysis of the relationship between the reliability of a system and that of each of its single components. The concept of signature produces, in a sense, a change of perspective and focuses on the (random) number of components' failures that lead the system to its own failure.

Initially, signature has been employed under the condition of components with independent and identically distributed lifetimes. Such a concept, in fact, is specially relevant in that case, where a large part of the casualty in the system's lifetime is induced by the casuality in the temporal order in which the different components fail. Systems with i.i.d. components, on the other hand, do not always fit with

R. Cerqueti (✉)
Department of Economics and Law, University of Macerata,
Via Crescimbeni 20, 62100 Macerata, Italy
e-mail: roy.cerqueti@unimc.it

F. Spizzichino
Department of Mathematics, Sapienza University of Rome,
Rome, Italy
e-mail: fabio.spizzichino@uniroma1.it

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_27

real-world applications. It has then been noticed that the definition of signature can be extended, in a completely natural way, to the case of exchangeable components (see [9, 10]). This extension is particularly important, because it allows us to consider components' lifetimes that are conditionally i.i.d. rather than just i.i.d.

As a further generalization, more recent studies dealt with the concept of signature even in cases of non-exchangeability (see e.g. [8, 11, 15]). The case of non-exchangeability leads to two different concepts of signature: the first one is only related to the structure of the system, while the other one is related to both the structure of the system and to the joint distribution of the components lifetimes. The former is concerned with the symmetry properties of the system [15], while the latter can play a role in the computation and approximation of the system reliability in particular (see e.g. [8, 11, 15]).

To the best of our knowledge, the concept of signature has been employed so far exclusively in the field of reliability systems. The case of non-exchangeable components provides however a realistic representation of a wider family of systems and networks appearing in the applied sciences. The possibility of extending this concept to non-exchangeable cases, open then the path to applications to other fields. In particular we think that signature can play a useful role in the field of Economics and financial risk even if this path remains unexplored.

As a preliminary task in the direction of filling this gap, it is important to understand the differences, as far as properties and meaning of signatures are concerned, between the two cases of exchangeability and non-exchangeability.

In this note we deal with some aspects of this issue and point out some relevant implications. Some of such implications will be also demonstrated by considerations of financial character.

More precisely, the remaining part of the paper is organized as follows. In Sect. 2 we briefly recall the concepts of signatures and present preliminaries and notation. In Sect. 3 we discuss the main differences between the cases of exchangeable and non-exchangeable lifetimes. A discussion about some related aspects from the point of view of financial applications, will be presented in the Sect. 4, with a specific focus on basket options.

2 Preliminaries, Notation, and Definitions of Signatures

We consider a *reliability system* S formed by *n components* $C_1, \ldots, C_n$. Given $j = 1, \ldots, n$ and a time $t > 0$, the *status of the* j*th component* at time t is a binary variable $Y_j(t)$ defined by

$$Y_j(t) = \begin{cases} 1 & \text{if } \mathrm{C_j} \text{ is working at time t} \\ 0 & \text{if } \mathrm{C_j} \text{ is down at time t.} \end{cases}$$

Each component is assumed to be working at time $t = 0$, and hence $Y_j(0) = 1$, for each $j = 1, \ldots, n.$ The *status of the system* can analogously be defined by letting

$$Y_S(t) = \begin{cases} 1 & \text{if S is working at time t} \\ 0 & \text{if S is down at time t,} \end{cases}$$

Fix $t \geq 0$. One assumes that $Y_S(t)$ is exclusively determined by $Y_1(t), \ldots, Y_n(t)$ and defines the *structure function of the system* as the function $\varphi_S : \{0, 1\}^n \to \{0, 1\}$ such that:

$$Y_S(t) = \varphi_S\left(Y_1\left(t\right), \ldots, Y_n\left(t\right)\right).$$

φ_S is usually assumed to be *coherent*, i.e. the following conditions are satisfied:

- $\varphi_S(0, \ldots, 0) = 0$, $\varphi_S(1, \ldots, 1) = 1$;
- φ_S is non-decreasing with respect to its components;
- each component of S is *relevant*

Now, denote by $\mathscr{G}$ the set of the *path vectors* of the system, i.e.

$$\mathscr{G} = \{\mathbf{y} \in \{0, 1\}^n | \varphi_S\left(\mathbf{y}\right) = 1\}.$$

Trivially $\mathbf{y} = (1, \ldots, 1)$ is a path vector and thus $Y_S(0) = 1$.

The *lifetime* of S and that of the jth component are respectively given by X_S and X_j, where

$$X_S = \inf\{t \geq 0 \,|\, Y_S(t) = 0\} = \inf\{t \geq 0 \,|\, (Y_1(t), \ldots, Y_n(t)) \notin \mathscr{G}\},$$

and

$$X_j = \inf\{t \geq 0 \,|\, Y_j(t) = 0\}.$$

Furthermore, $R_S(t)$ denotes the *reliability function* of the system at time t, namely:

$$R_S(t) \equiv P\{X_S > t\}, \qquad \forall t \geq 0. \tag{1}$$

The term $R_S(t)$ depends both on the structure function φ_S and on the joint distributions of the components' lifetimes. As we are going to discuss in the following, the concept of signature provides an insight about the structure of such dependence.

We first recall the formal definitions of *structure signature* and *probability signature.* For this purpose we need the following further assumptions and notation.

First of all, it is convenient to imagine that each component continues to work until its own failure, even if the system has already failed, so that all the lifetimes $X_1, \ldots, X_n$ are well defined and can be eventually observed. We assume furthermore that the joint distribution of the elements of the vector $\mathbf{X} = (X_1, \ldots, X_n)$ is such that

$$P\{X_1 \neq \cdots \neq X_n\} = 1. \tag{2}$$

By considering the order statistics $X_{(1)}, \dots, X_{(n)}$ of the vector $\mathbf{X}$ we thus have:

$$P\{X_S = X_{(k)}\}, \tag{3}$$

for one and only one $k = 1, \dots, n$.

Before continuing, let us remark that the failures of the subsequent components give rise to the progressive observation of a permutation of $\{1, \dots, n\}$. All the $n!$ possible permutations can be observed and each permutation describes a possible temporal order in which the different components fail.

Consider the events E_k, defined by

$$E_k \equiv \{X_S = X_{(k)}\}, \qquad k = 1, \dots, n. \tag{4}$$

$E_1, \dots, E_n$ form then a partition of the sample space, i.e. one and only one of them will be observed.

Denote by $\mathscr{P}$ the set of all the permutations of $\{1, \dots, n\}$ and consider the random vector $(J_1, \dots, J_n)$ defined by:

$$J_k = i \text{ when } X_{(k)} = X_i, \qquad \forall k = 1, \dots, n, \tag{5}$$

i.e. J_k indicates the identity of the component which fails in correspondence of the kth observed failure. We also set

$$A_k \equiv \{(j_1, \dots, j_n) \in \mathscr{P} \mid J_1 = j_1, \dots, J_n = j_n \Rightarrow X_S = X_{(k)}\}.$$

While the events $E_1, \dots, E_n$ form a partition of the sample space, the sets $A_1, \dots, A_n$ form a partition of the set $\mathscr{P}$ and we have

$$\sum_{k=1}^{n} |A_k| = n!.$$

As to the logical relation between these two partitions, we can write

$$\{(J_1, \dots, J_n) \in A_k\} = \{X_S = X_{(k)}\} \equiv E_k.$$

One basic remark is that $(A_1, \dots, A_n)$ is determined by the structure function φ_S.

We can now recall the definitions of two different notions of signatures

Definition 1 • The *structure signature* of S is $\mathbf{p} \equiv (p_1, \dots, p_n)$, where

$$p_k = \frac{|A_k|}{n!}, \qquad k = 1, \dots, n.$$

- The *probability signature* of S is $\hat{\mathbf{p}} \equiv (\hat{p}_1, \ldots, \hat{p}_n)$, where

$$\hat{p}_k = P(E_k), \qquad k = 1, \ldots, n.$$

See [8, 10–12, 15]. Concerning the events $\{E_k\}_{k=1,\ldots,n}$ we can write, by recalling (4),

$$X_S = \sum_{k=1}^{n} X_{(k)} \mathbf{1}_{E_k}, \tag{6}$$

and, by applying the law of total probabilities and by (1), (4) and (6), we can conclude:

$$R_S(t) \equiv \sum_{k=1}^{n} P\left(E_k\right) \cdot P\{X_{(k)} > t | E_k\}. \tag{7}$$

In view of the definitions above, the decomposition (7) can be rewritten in terms of the probability signature:

$$R_S(t) \equiv \sum_{k=1}^{n} \hat{p}_k \cdot P\{X_{(k)} > t | E_k\}. \tag{8}$$

3 Two Different Scenarios and Different Roles of Signatures

Concerning the two concepts of probability signature and of structure signature we observe different properties depending on the type of joint distribution that is assessed for the lifetimes of the components. As mentioned above, the scenario obtained under the condition of exchangeability is fairly special and it is rather different from the one that emerges in the non-exchangeable cases. Even the relations existing between the two concepts and their roles in applied problems are generally different in the two cases. These differences will be briefly outlined in this section, where the cases of exchangeability and non-exchangeability will be treated separately. The condition (2) is however assumed in any case, since it is necessary for the definitions of signatures to be meaningful. More arguments on this topic can be found in the cited references; some potentially useful examples are discussed in [16].

Before starting our discussion here, it is useful to pay attention to a couple of simple remarks. First, we notice that, from a purely mathematical viewpoint, both the vectors $\mathbf{p}$ and $\hat{\mathbf{p}}$ can be seen as probability distributions over the space $\{1, \ldots, n\}$. The probability signature $\hat{\mathbf{p}}$, in particular, can be seen as the probability distribution of the random variable M, defined as follows: M is the number of the observed component failures up to the failure of the system. A very special class of coherent systems is relevant in the reliability field and in a signature-based analysis, in particular. This is the class of systems of the type $k : n$ (for $k = 1, \ldots, n$). A system $k : n$ is one which is able to work as long as at least k of its components are working, namely it fails at the instant of the $(n-k+1)$th components' failure. In particular, a *parallel system* is a system $1:n$ and a *series system* is a $n:n$ system. In the case of a $k:n$ system we have $P(M=n-k+1)=1$ and both $\mathbf{p}$, $\hat{\mathbf{p}}$ are degenerate probability distributions, with $p_{n-k+1}=\widehat{p}_{n-k+1}=1$. Notice also that, in these cases, the structure of the system is perfectly symmetric. In other words all the components in the system contribute in a same way in maintaining the system in its working state.

3.1 The Exchangeable Case

We first consider the case where the components' lifetimes $X_1,\ldots,X_n$ are exchangeable random variables. Namely the joint distribution of $\mathbf{X}$ is invariant with respect to permutations of the variables. As an immediate consequence of this assumption, the random permutation $(J_1,\ldots,J_n)$, defined in (5), is distributed uniformly over $\mathscr{P}$, i.e.:

$$P\{(J_1,\ldots,J_n)\in B\}=\frac{|B|}{n!},\qquad \forall\, B\subseteq\mathscr{P}.$$

This entails the following simple result (see e.g. the discussion in [15]).

Proposition 1 *1. For $k=1,\ldots,n$, one has*

$$\hat{p}_k=p_k;$$

2. the events $\big(X_{(k)}>t\big)$ and E_k are independent;
3. the reliability function of the system is:

$$R_S(t)=\sum_{k=1}^{n}p_k P\{X_{(k)}>t\}. \tag{9}$$

We notice that item 3. is an immediate consequence of 1. and 2. and of the total probability formula (8). Moreover, items 1. and 2 are immediate consequences of the assumption that all the permutations $(j_1,\ldots,j_n)$ are equally probable as values for $(J_1,\ldots,J_n)$. We recall that each permutation describes a different temporal order in which the different components fail.

The statements in Proposition 1 are relevant from an applied point of view. From 1. we see that, in the exchangeable case, structure signature and reliability signature collapse into one and the same concept. Thus the probability distribution of the random variable M only depends on the structure of the system and it is not influenced by the joint probability law of the lifetimes $X_1,\ldots,X_n$. This lack of interaction is confirmed by item 2.

Let us now examine item 3. in details. It is clear that $R_S(t)$ generally depends on the pair $(\varphi_S,F_{\mathbf{X}})$ where φ_S is the structure of the system and $F_{\mathbf{X}}$ denotes the joint probability distribution function of the lifetimes $X_1, \ldots, X_n$. Such dependence may turn out to be rather complex, in some cases. The special form (9) of the formula of total probabilities (8) has then the following interpretation: when $X_1, \ldots, X_n$ are exchangeable (9) shows that $R_S(t)$ depends on φ_S only through the system signature $\mathbf{p}$ (which is only a function of φ_S and is it is not influenced by $F_{\mathbf{X}}$). On the other hand, $R_S(t)$ is influenced by $F_{\mathbf{X}}$ only through the vector of the marginal distributions of the order statistics $X_{(1)}, \ldots, X_{(n)}$.

These facts entail the following implications:

1. Consider two coherent systems S' and S'' formed with different sets of components $C'_1, \ldots, C'_n$ and $C''_1, \ldots, C''_n$, respectively, and such that they share the same structure functions, i.e. $\varphi'_S \equiv \varphi''_S$. Then, as long as the vectors of the components' lifetimes are exchangeable, S' and S'' share the same (probability and structure) signature, even if the joint distributions are different.
2. Think of a coherent system S, all the components of which play similar roles as to the system's capability to work. In such a case, we are allowed to interchange the respective positions of any two components in the system. This situation is met, for instance, in a network where all the components are just transmission nodes, possibly with different capacities but similar in nature. For such a system S, consider a permutation $\pi \in \mathscr{P}$ and denote by $\mathbf{S}_\pi$ the system obtained by permuting the components through π. Then the reliability functions $R_S(t)$ and $R_{S_\pi}(t)$ coincide, for any t.

3.2 The Non-exchangeable Case

In this subsection we consider the case when $X_1, \ldots, X_n$ are not exchangeable, so that we cannot rely anymore on Proposition 1. As a first consequence, the structure signature and the reliability signature do not necessarily coincide. We can still consider the structure signature $\mathbf{p}$ which, by definition is a combinatorial invariant, only determined by the structure φ_S. But this vector does not carry complete information about the probabilities $\widehat{p}_1, \ldots, \widehat{p}_n$. Actually, the vector $\hat{\mathbf{p}}$ is influenced also by the choice of the joint distribution function $F_{\mathbf{X}}$. Moreover, the formula (8) cannot be reduced to (9). Generally both the vectors $\hat{\mathbf{p}}$ and $\big(P\{X_{(1)} > t|E_1\}, \ldots, P\{X_{(n)} > t|E_n\}\big)$, whose scalar products produce $R_S(t)$, depend on both the data φ_S, $F_{\mathbf{X}}$.

It is now interesting to briefly point out the different roles of $\hat{\mathbf{p}}$ and $\mathbf{p}$ in reliability problems.

$\hat{\mathbf{p}}$ can be applied in different ways. It can be used in particular for defining the *projected system*, which provides in a sense the best approximation of the original system [8, 11]. Furthermore it could be used for extending to the non-exchangeable case comparisons, between two systems, that have been developed for i.i.d components and that are based on the structural signature. See also below.

For the purpose of analyzing the possible role of $\mathbf{p}$, it is again convenient to consider a coherent system S whose components have similar roles, so that interchanging the respective positions of any two components makes sense. For these cases we would like to investigate what happens if we permute, according to some permutation $\pi \in \mathscr{P}$, the positions of the components.

Fix a permutation π and recall that S_π denotes the new system, obtained by applying π on the components of S. The structure function of S_π is just given by

$$\varphi_\pi(\mathbf{y}) = \varphi(\mathbf{y}_\pi) . \tag{10}$$

Since the reliability function depends on the probability signature and the latter depends on the joint distribution function of the lifetimes then, generally, $R_{S_\pi}(t) \neq R_S(t)$, for $t > 0$. We denote by $R^*(t)$ the *symmetrized* reliability function defined as follows:

$$R^*(t) = \frac{1}{n!} \sum_{\pi \in \mathscr{P}} R_\pi(t) . \tag{11}$$

Notice that we implicitly identified $R_{(1,\ldots,n)}$ with R_S, where $\{1, \ldots, n\}$ is the identical permutation.

It is also useful to adopt the notation $R_S^{(F)}(t)$, in order to stress the dependence of the reliability function on the joint law F of the components lifetimes $X_1, \ldots, X_n$. Furthermore we denote by F_π the joint law of the permuted vector $\mathbf{X}_\pi$. One can see that

$$R_\pi^{(F)}(t) = R_{\{1,\ldots,n\}}^{(F_\pi)}(t).$$

Denote now by Π a random permutation of $\{1, \ldots, n\}$, uniformly distributed over $\mathscr{P}$, and set

$$\left(X_1^*, \ldots, X_n^*\right) = \left(X_{\Pi_1}, \ldots, X_{\Pi_n}\right).$$

Finally we denote by F^* the joint distribution function of $\left(X_1^*, \ldots, X_n^*\right)$ and by $R_S^{(\mathbf{F}^*)}(t)$ the reliability function of the system S when the lifetimes of its components are $\left(X_1^*, \ldots, X_n^*\right)$. The random vector $\left(X_1^*, \ldots, X_n^*\right)$ is exchangeable and it is such that the vectors of the order statistics $\left(X_{(1)}^*, \ldots, X_{(n)}^*\right)$ and $\left(X_{(1)}, \ldots, X_{(n)}\right)$ share the same joint law. All these properties and positions lead us to the following result.

Proposition 2 *All the systems S_π, for $\pi \in \mathscr{P}$, share the same structure signature* $\mathbf{p}$. *Furthermore*

$$R_S^{(\mathbf{F}^*)}(t) = \sum_{k=1}^{n} p_k P\{X_{(k)} > t\}, \tag{12}$$

$$R^*(t) = R_S^{(\mathbf{F}^*)}(t) . \tag{13}$$

See [15] for details. Thus we see that $R^*(t)$ can be interpreted as the reliability function of a fictitious system (the *average system*), having the same structure of S and same components of S; but such that the components are distributed at random among the different positions. As shown by (12) $R^*(t)$ is, typically, more easily computed than $R(t)$. Even if its meaning is fictitious it can still be of interest. Consider in this respect the function

$$\mu(t) = |R_S^{(\mathbf{X})}(t) - R^*(t)|. \tag{14}$$

We expect that $R_S^{(\mathbf{X})}(t) - R^*(t) \geq 0$ when the system is correctly designed. The function $\mu(t)$ expresses a sort of distance between the reliability function and the *symmetrized* reliability function $R^*(t)$ for the system S. It is related to the amount of asymmetry of the system S: the larger the symmetry level of the structure function φ_S, the smaller the difference in the left-hand side of (14). On the other hand, it can be argued that the more the structure signature is a concentrated probability distribution, the smaller is the asymmetry of the system. Recall in this respect that, as we had noticed above, degenerate signatures, in particular, correspond to the completely symmetric structures of the type $k : n$. We then see that the structure signature has a double role: it allows us to compute $R^*(t)$ by means of (12) and provides us with some information about the error that arises, in the computation of the reliability function, when we approximate $R(t)$ by $R^*(t)$, namely when we replace the "true" distribution of $X_1, \ldots, X_n$ with the exchangeable distribution which gives rise to the same joint distribution for the order statistics.

Let S', S'' be two systems with the same number of components and let $\mathbf{p}'$, $\mathbf{p}''$ be their structural signatures respectively. As already mentioned $\mathbf{p}'$, $\mathbf{p}''$ also permit one to compare S', S'' in the following sense: different types of stochastic orderings between the probability distributions $\mathbf{p}'$, $\mathbf{p}''$ imply corresponding stochastic orderings between the reliability functions of S', S'', when a vector of the same i.i.d. components is installed in the two systems (see [6]). This can be a good way to compare the two systems, even for cases when the components are not exchangeable. Furthermore one can conjecture that results similar to those in [6] could be extended to non-exchangeable case, in terms of $\widehat{\mathbf{p}}'$, $\widehat{\mathbf{p}}''$.

4 A Special Class of Basket Options and Implications of Non-exchangeability

In this section we focus attention on financial applications and, more precisely, on the risk associated to the so-called *basket options*. On one hand we point out that the topic of signature can be of some interest also in this field. On the other hand we further discuss, just from an economic viewpoint, the implications related with the difference between exchangeability and non-exchangeability, as far as signature is concerned.

Basket options constitute one of the most popular and traded structured products, and belong to the wide family of *exotic options* (see [17]). The success of this financial product lies in low prices, in the management of the risk profile through an appropriate selection of correlated assets in the basket and in the reduction of the transaction costs. The payoff of this product is linked to the performance of a collection (basket) of assets. On such a basis, the option may be of various typologies in nature. We will consider here a particular model of basket options, where the basket is composed of a set of n assets, formed with a subset of r "important" assets and a set of s "standard" assets, $n = r + s$. For all the assets, irrespectively of whether they are important or not, a lower barrier is considered which should not be crossed until the maturity time of the option (see e.g. [1, 4, 7]).

We can think of an important asset as one for which a very big amount of stocks is traded on the market. We can then expect that its volatility is smaller than that of the assets with less stocks and this may reflect in a lower riskiness.

Let $T > 0$ be the expiration time (or time to maturity) for the option and $\alpha > 0$ be the common barrier for all the assets in the basket. Furthermore, for $t \geq 0$ and $j = 1, \ldots, n$, let $\Lambda_j(t)$ be the stochastic process describing the evolution of the return of the jth asset. We consider then the n-dimensional vector of (random) failure times $\mathbf{X} = (X_1, \ldots, X_n)$ such that:

$$X_j = \inf\{t > 0 \mid \Lambda_j(t) \leq \alpha\}. \tag{15}$$

X_j will be then interpreted hereafter as the lifetime of the jth asset and it can be also convenient to set

$$Y_j(t) = \begin{cases} 1 & \text{if } X_j > t, \\ 0 & \text{otherwise.} \end{cases} .$$

A basket option will be viewed as a coherent system S whose n components $C_1, \ldots, C_n$ are the assets in the basket. Once the financial structure of the option has been fixed, one defines the failure time of the option a random variable X_S, suitably defined as a function of $X_1, \ldots, X_n$.

At the expiration time T the holder of the option obtains a return $Ret_T > 0$, under the condition

$$X_S > T.$$

For $t \geq 0$, the reliability function of the option at time t is

$$R_S(t) \equiv P\{X_S > t\}.$$

Generally, the price of a financial product is clearly related with its risk level. For our basket option, an appropriate measure of riskiness is the value $R_S(T)$, which then plays a relevant financial role.

In order to exactly define the very nature of the options that we consider or, in other words, to describe the structure function of the system, we in particular focus attention on financial models defined in terms of a nonincreasing function

$$\rho : \{1, \ldots, n\} \to \{0, 1, \ldots, r + 1\},$$

satisfied the condition with a meaning described as follows: the option has a fatal default at the first time in which the failures of k assets are observed, with k such that at least $\rho(k)$ failures are due to the more important assets. It is natural to assume that the function $\rho(k)$ is nonincreasing. A few more precise details about its definition are however in order.

The condition $\rho(k) = 0$ obviously means that the failure of k standard assets is enough to determine the default of the option. The position $\rho(k) = r + 1$ means that k is so small that the failure of k assets cannot produce the option's default, even in the case when all the failed assets are important ones. The minimum number of failures able to determine the default is the minimum value of k that satisfies the condition $\rho(k) \leq k$. The maximum possible number of failures that can be conceptually observed up to the default coincides with the minimum value of k such that $\rho(k) = 0$.

Let us now proceed to formally define the option's default time X_S.

Set

$$N_k \equiv \sum_{j=1}^{r} \left(1 - Y_j\left(X_{(k)}\right)\right).$$

N_k then denotes the number of assets that have already failed at the moment of the kth overall failure. We let

$$X_S = X_{(k)}$$

if and only if

$$N_k \geq \rho(k), \;\; N_h < \rho(h),$$

for $h = 1, \ldots, k - 1$.

In other words, the family of the path vectors of the systems is defined by

$$\left\{ \mathbf{y} \in \{0,1\}^n \mid r - \sum_{j=1}^{r} y_j < \rho\left(n - \sum_{j=1}^{n} y_j\right)\right\}. \qquad (16)$$

We notice that such a system manifests the following structure of partial symmetry: all the important assets share a common role and also all the standard assets share a common role of their own. In a sense this structure could be seen as a natural generalization of the famous k-out-of-n models. To designate our models, we may use the term $(n - \rho(k))$-*out-of-n systems*.

Remark 1 In the field of basket options, a further generalization could be sometimes more realistic: one may admit that the above numbers $\rho(k)$ are replaced by numbers $\rho(k; J)$ also depending on the subsets $J \subset \{1, \ldots, s\}$ of standard assets that failed up to the time $X_{(k)}$. The assumption that $\rho(k)$ is a non-decreasing function of k, should be replaced by a new condition involving also the monotonicity with respect to J. Models of this type are also related to the concept of system with weighted components, analyzed in [14].

To the best of our knowledge, coherent systems of the type $(n - \rho(k))$-out-of-n have not been considered so far from the point of view of a signature analysis. The following result shows the form of their structure signature $\mathbf{p} = (p_1, \ldots, p_n)$. Denote by I_ρ the set $\{k \in \{1, \ldots, n\} \,|\, 0 < \rho(k) \leq k\}$.

Proposition 3 *(a) Let $k \in I_\rho$. Then:*

$$p_k = \sum_{j=0}^{\rho(k-1)-1} \frac{\binom{r}{j}\binom{n-r}{k-j-1}}{\binom{n}{k-1}} - \sum_{j=0}^{\rho(k)-1} \frac{\binom{r}{j}\binom{n-r}{k-j}}{\binom{n}{k}}; \qquad (17)$$

(b) $p_k = 0$ if $\rho(k) = r + 1$;
(c) $p_k = 0$ if $\rho(k) = \rho(k-1) = 0$;
(d) Let k be such that $\rho(k) = 0$, $\rho(k-1) > 0$. Then

$$p_k = 1 - \sum_{h \neq k} p_h.$$

Proof (a) First, we recall that the structure signature of a system coincides with the probability signature, where the latter is computed under the assumptions that the components are i.i.d. Thus we need to compute the probabilities

$$P\left(X_S = X_{(k)}\right), \qquad k = 1, \ldots, n,$$

under the assumption that the assets' lifetimes $X_1, \ldots, X_n$ are i.i.d.
For $k \in I_\rho$, we consider the quantity $\overline{P}_k := \sum_{h=k+1}^{n} p_h$, so that

$$\overline{P}_k = P\left(X_S > X_{(k)}\right) = P\left(N_k < \rho(k)\right) = \sum_{j=0}^{\rho(k)-1} P\left(N_k = j\right).$$

Then

$$p_k = \overline{P}_{k-1} - \overline{P}_k = \sum_{j=0}^{\rho(k-1)-1} P\left(N_{k-1} = j\right) - \sum_{j=0}^{\rho(k)-1} P\left(N_k = j\right).$$

In view of the assumption that the assets' lifetimes $X_1, \ldots, X_n$ are i.i.d., the terms $P\left(N_k = r\right)$ are given by hypergeometric probabilities. More precisely:

$$P\left(N_k = j\right) = \frac{\binom{r}{j}\binom{n-r}{k-j}}{\binom{n}{k}}.$$

(b) The condition $\rho(k) = r + 1$ means that the observation of k failures cannot cause the default of the option. Thus $p_k = 0$.
(c) If $\rho(k) = 0$ then k failures cause the default of the option, if the latter had not defaulted before. Thus the probability of a default at $X_{(k)}$ is null when $\rho(k-1) = 0$.
(d) It trivially follows from (a), (b) and (c), since $\sum_{k=1}^{n} p_k = 1$.

As already discussed in Sect. 2, the signature analysis of a system is strongly influenced by the conditions of exchangeability or non-exchangeability among the components.

In the present context, exchangeability of $X_1, \ldots, X_n$ is reflected by a symmetry condition among the behavior of the assets' returns $\Lambda_1, \ldots, \Lambda_n$ and the following statement can in particular be made: at any fixed time t, the probability that $h < n$ returns are above the threshold α, while the remaining $n - h$ returns are below α, is independent on the specific selection of the h assets.

We are in the non-exchangeability case when such a statement is no longer true. In this respect, non-exchangeability can be viewed as a condition of "heterogeneity" among the assets of the basket. Specifically, in analyzing the joint behavior of the assets at the expiration date T, the identity of any single asset matters. This is actually a typical circumstance in the above setting. Exchangeability is then only an extreme and idealized condition, for us.

Let us briefly mention some relevant implications of non-exchangeability on the signature analysis.

As a first remark, we can say that the special structure $(n - \rho(k))$-out-of-n is just appropriate for the financial model of heterogeneity, where the assets can be of only two "types".

We can moreover recall that the probability signature is different from the structure signature detailed in (17). When the important assets are more reliable than the other ones, the probability signature is stochastically larger than the structure signature. This circumstance would guarantee that the "projected system" is less risky than the "average system", where the "projected system" provides a better approximation of the reliability of the system (of the option, in our case) than the "average system" [11].

A further remark concerns the effect of some possible piece of new information about the market. Suppose that short after time 0, an event A is observed that modifies the evaluation of the future performance of the assets (such as e.g. the failure of an important asset, outside the basket). This has a double effect on the terms in the r.h.s. of Eq. (8). Not only the factors $P\{X_{(k)} > t | E_k\}$ change into $P\{X_{(k)} > t | E_k \cap A\}$, but also the weights $\hat{p}_k$ are influenced by the replacement of joint distribution (prior to A) with a different one (posterior to A), when at least one of the two is not exchangeable. This circumstance may have the following relevant consequence. On the basis of a same set of assets, consider two different options O_1 and O_2, characterized by different and non-comparable functions $\rho_1(k)$ and $\rho_2(k)$. Compare then O_1 and O_2 in terms of their levels of riskiness and then in terms of their price: it can happen that the ordering between O_1 and O_2 posterior to A is the opposite of the ordering if the comparison had been made prior to A. In view of the validity of the formula (9), this situation cannot manifest when the prior and posterior joint distributions are both exchangeable.

The condition of non-exchangeability is even more intrinsic to the nature of the option, when we consider the models mentioned in Remark 1. In such models a character of heterogeneity is present and it makes sense to compare two different options obtained by different arrangements in the system of a same set of assets. The problem then arises of determining the most efficient permutation. It is useful to recall in this respect that the structure signature and probability signature are of help in such an analysis. The fact that probability signature can be influenced by arrival of new information can be an interesting issue for further research.

Acknowledgments The authors are grateful to the participants in the International Conference on Risk Analysis (ICRA5) Conference in Tomar, Portugal (May-June, 2013). FS has been partially supported by Project “Aspetti di dipendenza stocastica e rischio in modelli probabilistici”, Univ. La Sapienza, 2014.

References

1. Alexander, C.: Market Models. Wiley, West Sussex (2001)
2. Barlow, R.E., Proschan, F.: Statistical Theory of Reliability and Life Testing, To Begin With, Silver Spring, Maryland (1981)
3. Boland, P.J., Samaniego, F.J.: The signature of a coherent system and its applications in reliability. In: Mazzucchi, T., Singpurwalla, N., Soyer, R. (eds.) Mathematical Reliability: An Expository Perspective, pp. 3–30. Kluwer Academic Publishers, Boston (2004)
4. Cerqueti, R., Rotundo, G.: Options with underlying asset driven by a fractional Brownian motion: crossing barriers estimates. New Math. Nat. Comput. **6**(1), 109–118 (2010)
5. Gertsbakh, I.B., Shpungin, Y.: Models of Network Reliability. CRC Press, Boca Raton (2010)
6. Kochar, S., Mukerjee, H., Samaniego, F.J.: The signature of a coherent system and its application to comparisons among systems. Nav. Res. Logist. **46**(5), 507–523 (1999)
7. Kou, S.G., Wang, H.: Option pricing under a double exponential jump diffusion model. Manag. Sci. **50**(9), 1178–1192 (2004)
8. Marichal, J.-L., Mathonet, P.: Extensions of system signatures to dependent lifetimes: explicit expressions and interpretations. J. Multivar. Anal. **102**(5), 931–936 (2011)
9. Navarro, J., Rychlik, T.: Reliability and expectation bounds for coherent systems with exchangeable components. J. Multivar. Anal. **98**, 102–113 (2007)
10. Navarro, J., Balakrishnan, N., Bhattacharya, D., Samaniego, F.: On the application and extension of system signatures in engineering reliability. Nav. Res. Logist. **55**, 313–327 (2008)
11. Navarro, J., Spizzichino, F., Balakrishnan, N.: Applications of average and projected systems to the study of coherent systems. J. Multivar. Anal. **101**(6), 1471–1482 (2010)
12. Samaniego, F.J.: On closure of the IFR class under formation of coherent systems. IEEE Trans. Reliab. **R34**, 60–72 (1985)
13. Samaniego, F.J.: System Signatures and Their Applications in Engineering Reliability. International Series in Operations Research and Management Science, vol. 110. Springer, New York (2007)
14. Samaniego, F.J., Shaked, M.: Systems with weighted components. Stat. Probab. Lett. **78**(6), 815–823 (2008)

15. Spizzichino, F.: The role of signature and symmetrization for systems with non-exchangeable components. In: Advances in Mathematical Modeling for Reliability, pp. 138–148. IOS, Amsterdam (2008)
16. Spizzichino, F., Navarro, J.: Signatures and symmetry properties of coherent systems. In: Recent Advances in System Reliability. Springer Series in Reliability Engineering, pp. 33–48. Springer, New York (2012)
17. Zhang, P.G.: Exotic Options. World Scientific Publishing Co., Singapore (1997)

Detecting an IO/AO Outlier in a Set of Time Series

Vassiliki Karioti

Abstract The analysis of time series is an important area of statistics and it is necessary to understand the nature of outliers, in order to use appropriate methods to detect, or accommodate them. An interesting aspect is the case of detecting an outlier (Type IO or Type AO) in a set of autoregressive time series at the same time point. For example, consider a phenomenon in neighbouring regions. Measurements of the phenomenon in each region is a time series, so a set of series is creating. It is possible, an external factor affecting all regions, to cause unusual values, and then an outlier is appeared in each series of the set at the same time point. Tests for an innovative outlier affecting every member of a set of autoregressive time series at the same time point are developed. In one model, the outliers are represented as independent random effects; likelihood ratio tests are derived for this case and simulated critical values are tabulated. In a second model, assuming that the size of the outlier is the same in each series, a standard regression framework can be used and correlations between the series are introduced. In the case of additive outlier, the outliers are represented only as independent random effects.

Keywords Time series · Innovative outliers · Additive outliers · Autoregressive models

1 Introduction

Sets of time series are modeled in various ways in the large statistical literature on longitudinal data, but the detection of outliers does not seem to have been investigated in detail in any of the approaches. The detection of unusual values can play an important role in risk assessment as well as in many other areas. For example, consider a phenomenon that is measured in each of several neighboring regions, thus creating

V. Karioti (✉)
Department of Accounting, Technological Educational Institution of Patras,
Patras, Greece
e-mail: vaskar@teipat.gr; vaskar@teiwest.gr

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_28

a set of series. An external factor could affect all regions, creating unusual values, so that an outlier appears in each series of the set at the same time point.

In this paper, we can consider the case of detecting outliers, when the data consist of a set of time series. Models of classical time series form are considered, in particular AR(p). We construct tests for outliers based on likelihood ratio and investigate their performance in detail for the AR(1) case.

There are two basic types of outliers in time series:

1. Additive Outlier (AO) and
2. Innovative Outlier (IO).

The first type is an outlier that affects a single observation. The second one acts as an addition to the noise term at a particular series point.

This paper is organized as follows. In Sect. 2, we present our basic models of time series in case that an outlier of type IO appears in each series of the set at the same time point. In Sect. 3, we present the time series models in case that an outlier of type AO appears in each series of the set at the same time point. In Sect. 4, we construct tests for IO outliers based on likelihood ratio and investigate their performance in detail for the AR(1) case. A simulation study for determining the critical values of the test statistic for a random IO and the powers of the likelihood tests are presented in Sect. 5. One typical example, of the profits of ten businesses for ten years, will be used as an illustration in Sect. 6 of this paper. Finally, the conclusions are represented in Sect. 7.

2 Models for Innovative Outlier

We construct models for the innovative outlier (IO) introduced by [2], in which the value of the “innovation” or noise is extreme. This affects not only the particular observation at which it occurs but also subsequent observations.

Two models are considered. In one model, the outliers are represented as independent random effects. In a second model, assuming that the size of the outlier is the same in each series, a standard regression framework can be used and correlations between the series are introduced.

For an IO of random size, we take the stationary AR(p) model for a single time series with IO in the form used by [2, 4], and extend it to a set of series. For a set of $i = 1, \dots, m$ time series Y_{it}, where the length of the ith series is the n_i, the autoregressive parameters $\{a_r\}$ are the same in each series, and every series is affected by an outlier at the same time q, then

$$Y_{it} = \sum_{r=1}^{p} a_r Y_{i,t-r} + \eta_{it},$$

$$Y_{iq} = \sum_{r=1}^{p} a_r Y_{i,q-r} + \Delta_i + \eta_{iq}.$$

The outlier in the ith series is $\Delta_i \sim \mathcal{N}(\Delta, \sigma_\delta^2)$ and is independent of the innovation terms n_{it}, which are i.i.d. $\mathcal{N}(0, \sigma^2)$ for all i and t. Thus, writing $u_{it} = Y_{it} - \sum_{r=1}^{p} a_r Y_{i,t-r}$, we have $u_{it} \sim \mathcal{N}(0, \sigma^2)$ for $t = 1, \ldots, n_i$, $t \neq q$, $u_{iq} \sim \mathcal{N}(\Delta, \sigma^2 + \sigma_\delta^2)$ for $i = 1, \ldots, m$. We will examine conditional likelihoods given the first p terms of the series, so we assume that $q > p$. We will further assume that $n_i > q$ for all i so that the outlier does in fact appear in every series.

To construct a test for the existence of the IO in our models, we use the two-stage maximum likelihood principle, [1]. Let T_q be a likelihood ratio test statistic for an outlier at the specific time point q. Then the test statistic for an outlier at unknown time is $T^* = \max(T_q)$, or $\min(T_q)$ as appropriate. To implement this, we must first write the likelihoods under the null and alternative hypotheses. For the case of random IO, we do this by looking at the distribution of the terms u_{it} defined above. The likelihood for each series (conditional on its first p terms) is the product of $n_i - p$ terms of the form $\mathcal{N}(0, \sigma^2)$ under H_0 (no outlier). Similarly, under H_1 (outlier at known time q) the conditional likelihood of each series is the product of n_i-p-1 terms of the form $\mathcal{N}(0, \sigma^2)$ and one of the form $\mathcal{N}(\Delta, \sigma^2 + \sigma_\delta^2)$.

Thus, the maximized log–likelihood (without the constant terms) under null and alternative hypothesis respectively are given by:

$$y_{it} = a y_{i,t-1} + n_{i,t-1}.$$

So, $T = -[N - m(p+1)] \ln \tilde{\sigma}^2 + (N - mp) \ln \hat{\sigma}^2 - m \ln \tilde{\tau}^2$ where

$$\sigma^2 = \tfrac{1}{N-mp} \sum_i \sum_t \hat{u}_{it}^2,\ \hat{u}_{it} = y_{it} - \sum_{i=1}^{p} \hat{a}_r y_{i,t-r}$$

(the estimators of the parameters under H_0),

$$\tilde{\Delta} = \tfrac{1}{m} \sum_i \tilde{u}_{iq},$$

$$\tilde{\tau}^2 = \tfrac{1}{m} \sum_i (\tilde{u}_{iq} - \tilde{\Delta})^2, \text{ where } \tau^2 = \sigma^2 + \sigma_\delta^2,$$

$$\tilde{\sigma}^2 = \tfrac{1}{N-m(p+1)} \sum_i \sum_{t \neq q} \tilde{u}_{it}^2,$$

$$\tilde{a} = \frac{\tilde{\sigma}^{-2} \sum_i \sum_{t \neq q} y_{it} y_{i,t-1} + \tilde{\tau}^2 \sum_i (y_{iq} - \tilde{\Delta}) y_{i,q-1}}{\tilde{\sigma}^{-2} \sum_i \sum_{t \neq q} y_{it}^2 + \tilde{\tau}^2 \sum_i y_{i,q-1}^2}$$

(estimators of the parameters underH_1).

For an IO of fixed size, we suppose that $\Delta_i = \Delta$. Writing the model in regression format, the model for each series is given by the following equation:

$$\begin{aligned}
\underset{\sim}{y_i} &= X_i\underset{\sim}{\beta} + \underset{\sim}{\varepsilon_i}, \quad \text{where} \\
\underset{\sim}{y_i} &= (y_{i,p+1}, \ldots, y_{in})', \\
\underset{\sim}{\varepsilon_i} &= (n_{p+1}, \ldots, n_p)', \\
\underset{\sim}{\beta} &= (a_1, \ldots, a_p)', \\
X_i, &\quad \text{is an } (n-p) \times p\text{matrix containing lagged values of } y.
\end{aligned}$$

Assuming that all m series have the same length n, the above model can be written as:

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}.$$

The vectors $\underset{\sim}{y}$, $\underset{\sim}{\varepsilon}$ and the matrix X are formed by stacking $\underset{\sim}{y_i}$, $\underset{\sim}{\varepsilon_i}$ and X_i respectively. The error structure $\underset{\sim}{\varepsilon}$ is $V = E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \Sigma \otimes I$, where Σ is the $m \times m$ covariance between the values of the innovations n_{it} occurring at the same time t in the different series.

In all models, values at different times are always assumed to be independent, within as well as between series. Under H_1, an IO outlier of size Δ is occurring at time q. The matrices X_i acquire an extra column $(0, \ldots, 0, 1, 0, \ldots, 0)$ where the solitary non-zero element occurs in the position corresponding to time q, and the vector of coefficients acquires an extra element Δ.

We consider three different models:

1. Heteroscedasticity. $n_{it} \sim \mathcal{N}(0, \sigma_i^2)$ for each i, $\Sigma = \text{diag}(\sigma_{12}, \ldots, \sigma_{m2})$, as the series are considered independent.
2. Σ is unrestricted. In this model unspecified correlations between the series is considered.
3. All the series are equally correlated. $\Sigma = \sigma^2(1-\rho)\mathbb{I} + \rho\mathbb{J}$, where ρ is the common correlation, $\mathbb{I}$ the identity matrix and $\mathbb{J}$ the matrix whose elements are all unity.

To fit these models we draw on standard theory for the estimation of systems of equations.

3 Models for Additive Outlier

On the other hand, the Additive Outlier (AO) acts like an error of observation occurring at that time only and, since it does not enter into the structure of the series, does not effect subsequent observations.

The distribution of u_{it} is conditional to the outlier Δ_i, so the distribution is as follow:

$$\begin{aligned} &\mathscr{N}(0,\sigma^2) \quad \text{for} \quad p < t < q \quad \text{and for} \quad p+q < t \leq n_i, \\ &\mathscr{N}(\Delta_i,\sigma^2) \quad \text{for} \quad t = q, \\ &\mathscr{N}(-a_{t-q}\Delta_i,\sigma^2) \quad \text{for} \quad q < t \leq p+q. \end{aligned}$$

We consider the case of the random AO as we did for the random IO, as the case of fixed AO is more complicated as there are non linear constraints of β.

The likelihood under the null hypothesis is the same as the null in the case of random IO. But, under the alternative,

$$l_1 = -\tfrac{\ln\sigma^2}{2}\sum_{i=1}^{m}(n_i - p) - \tfrac{\ln 2\pi}{2}\sum_{i=1}^{m}(n_i - p) - \tfrac{m}{2}\ln\theta^2 - \tfrac{m}{2}\ln\Gamma - \\ \tfrac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{t=p+1}^{n_i} U_{ti}^2(a) + \tfrac{1}{2\Gamma}\sum_{i=1}^{m}\left(\tfrac{B_i}{\sigma^2} - \tfrac{\delta}{\theta^2}\right)^2 - \tfrac{m\delta^2}{2\theta^2},$$

where

$$\begin{aligned} \Gamma &= \tfrac{A}{\sigma^2} + \tfrac{1}{\theta^2}, \\ A &= \sum_{t=q}^{p+q} a_{t-q}^2 = \sum_{j=0}^{p} a_j^2, \quad a_0 = -1, \\ B_i &= \sum_{t=q}^{p+q} a_{t-q} U_{ti} = \sum_{j=0}^{p} a_j U_{(j+q)i}, \\ U_{ti} &= X_{ti} - \sum_{j=1}^{p} a_j X_{(t-j)i}, \\ \Delta &\sim \mathscr{N}(\delta,\theta^2). \end{aligned}$$

It is obvious that the above equations are complicated and a simulation is necessary in order to find out the estimators of the parameters.

4 Testing

Even if the distribution of a test statistic T_q, for an outlier at a specific point is Known, the distribution of T^* is usually unknown, because of the correlations between the various T_q. A common solution is to use Bonferroni adjustments, [1, 3]. Therefore, $a/(n-p)\%$ critical values in the tail of the distribution of T_q provide a conservative a % level test for T^*. It is often the case that the degree of conservatism is very small in testing for a single outlier, so that the critical values are very good approximations, [3].

In the present problem, the distribution of T_q is unknown. However, being based on likelihood ratios, T_q, can be tested approximately against chi-squared critical values, if the standard asymptotic result holds. Consequently, a test of T^* at the nominal 5 % level could be carried out by comparing the test statistic $T = -2(\hat{l}_0 - \hat{l}_1)$ to the upper $5/(n-p)\%$ point of the chi-squared distribution.

In fact, the theory does not apply to the random effects model, because the null hypothesis includes the restriction $\sigma_\delta^2 = 0$, which falls on the boundary of the parameter space. However, it does apply to equally sized outliers. The relevant chi-squared distribution has one degree of freedom (since H_0 differs from H_1 by the constraint $\Delta = 0$).

5 Simulation Study

The main aim of our simulation study is to verify if the critical values provided by the above approximation work well enough, and to provide simulated percentage points. Consequently, we examined data simulated under the null hypothesis. We restricted our attention to models of order $p = 1$, and also assumed that every time series in a set had the same length n. In all cases, we generated series of length $n + 50$, then discarded the first 50 observations and retained the last n for analysis.

For the random IO, using the pseudorandom standard normal distribution RNNOR from the IMSL library $n + 50$ standard normal variates are generated for each series. The observations $y_i t$ were then obtained using the equation $y_{it} = ay_{i,t-1} + n_{i,t-1}$ (where the subscript on the single autoregressive parameter a_1 has been dropped).

The following Table 1, presents simulated critical values for testing for IO random outlier. They appear to be independent of the autoregressive parameter a. Critical values are higher for larger n because the statistic is the maximum of $n - p$ values, but they are also expected to be smaller for larger m because the statistic at any particular time represents an analysis across the m series.

Moreover, in order to examine the performances of the test statistic in the presence of an IO affecting every series, further simulations were carried out. Data were generated as described above, with the addition of the outlier at a selected time point. The test works well (results not shown). Power appears to be independent of the autoregressive parameter.

6 Example

For illustration, the above procedure, for random IO, has been applied to the profits of ten businesses for ten years and particular from 2002 until 2011. The data are plotted in Fig. 1, which suggests that at the year 2008 may be an outlier in each series of the set.

Table 1 Simulated critical values for IO random outlier

No. of series m	No. of series n	a	1 %	5 %
10	25	0.1	17.143	13.714
		0.5	17.272	13.686
		0.75	17.564	13.676
		0.9	17.530	13.872
	50	0.1	18.958	15.218
		0.5	18.505	15.154
		0.75	18.968	15.203
		0.9	18.751	15.227
20	25	0.1	16.637	13.049
		0.5	16.501	13.014
		0.75	16.288	13.014
		0.9	16.174	12.857
	50	0.1	17.762	14.33
		0.5	17.759	14.449
		0.75	17.666	14.348
		0.9	17.604	14.322

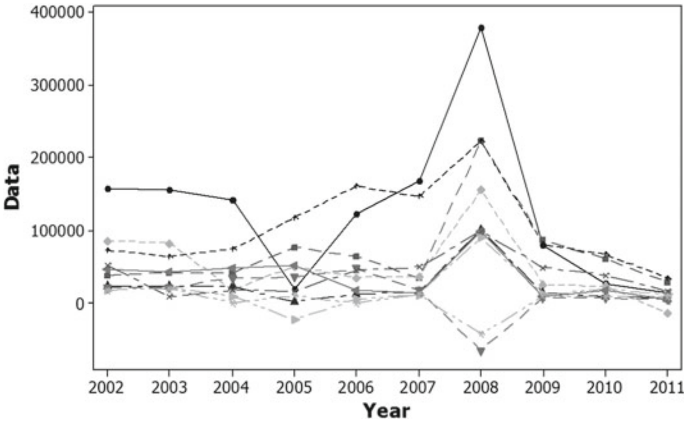


Fig. 1 Time series plot of the profits of the ten businesses

We apply the likelihood test with m = 10 and n = 10 for each series. The two–stage likelihood statistic is 61.2556 at position 7 that means at the year 2008, and comparing this with the tabulated critical values, we conclude that there is an outlier. Probably that year, an external factor had been affected all businesses and caused unusual values.

7 Conclusions

In the present paper, we return to the question of outlying values, but under the assumption that every series in the set is affected at the same time. Our results show that the analysis for an IO is quite simple, particularly if the size of the outlier can be assumed to be the same in every series. In that case, chi-squared significance levels can be employed with our likelihood ratio tests. In the case of outliers of random size, simulated significance levels are preferred.

Important extensions that should be investigated include extending the random IO model to allow correlations between series, allowing unequal autoregressive coefficients between series, and other structures for the covariance matrix Σ beyond the equicorrelation case. For example, when the time series in the set arise from different geographical areas, the correlation between two series might incorporate a function based on the geographical distance between them, or be non-zero only for neighboring areas.

Although, AO outliers may also be investigated, they are not so likely applicable to a set of time series. It is straightforward to produce a likelihood ratio test for random AO, but the theory for an AO of fixed size is more complicated.

References

1. Barnett, V., Lewis, T.: Outliers in Statistical Data, 3rd edn. Wiley, Chichester (1994)
2. Fox, A.J.: Outliers in time series. J. R. Stat. Soc. B **34**, 350–363 (1972)
3. Hawkins, D.M.: Identification of Outliers. Chapman & Hall, London (1980)
4. Muirhead, C.R.: Distinguishing outlier types in time series. J. R. Stat. Soc. B **48**, 39–47 (1986)

Response Surface Methodology: A Review of Applications to Risk Assessment

Teresa A. Oliveira, Conceição Leal and Amílcar Oliveira

Abstract Risk Analysis has assumed a crucial relevance over the past few years, particularly in dynamical systems with increasing complexity. Thanks to recent technological advances, the use of simulation techniques to estimate models has become the norm rather than the exception. These simulated models are used to predict the behavior of a system, to compute the probability of occurrence of a specific event and to predict the consequence of the said event. Uncertainty associated with the simulation, either in model parameters or in experimental data, requires its quantification as a prerequisite in probabilistic risk assessment. The computational costs of numerical simulations are often very high, thus the use of metamodels arises as a pressing necessity. Response Surface Methodology is known to be a suitable tool, both for the estimation of metamodels for the behaviors of systems and risk assessment, and for the quantification of uncertainty. A review of applications and of various aspects on the use of Response Surface Methodology in Risk Assessment Systems will be presented.

Keywords Monte Carlo method · Risk analysis · RSM · Sensitivity analysis · Uncertainty

1 Introduction

Risk Analysis is the process of systematically identifying and assessing potential risks and uncertainties that occur in a system and then find a viable strategy to more efficiently control these risks. It involves the likelihood of occurrence and the

T.A. Oliveira (✉) · A. Oliveira
CEAUL, Lisbon, Portugal
e-mail: Teresa.Oliveira@uab.pt

A. Oliveira · T.A. Oliveira · C. Leal
Open University, Lisbon, Portugal
e-mail: amilcar.oliveira@uab.pt

C. Leal
e-mail: conceicao.leal2010@gmail.com

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_29

magnitude of consequences of a specified hazard realization. It is a topic with great impact on modern society, whether in the context of research or within the applications' area, since it is the analytical process providing information on undesirable events which may pose a potential danger. The different perspectives by which risk is addressed in several scientific areas, the multitude of applications and different social connotations assigned to it, make it difficult to attain objectification, assessment, management and risk communication, and make the boundaries that separate these aspects ambiguous.

Risk Assessment will be approached as a scientific process whose methodology can be qualitative, quantitative, or semi-quantitative, if it combines the above two forms of analysis. In qualitative risk assessments, the results are expressed in a descriptive way, while in quantitative processes, risk is quantified by combining the probability or frequency of occurrence of an imminent danger to the magnitude of the result of this occurrence [40]. Risk assessment methodology and the way it quantifies error varies according to the application areas. However, the ultimate goal is always risk characterization, in order to provide data for decision making.

The work fields for application of risk analysis are extremely wide, as evidenced by the extensive documentation on it. It varies from the application to project management or industrial mega-projects, to different Engineering fields, from environmental and ecological protection to possible natural disasters or those resulting from human error, from public health to the financial system, from transmission of information to terrorism or sabotage. The complexity of most systems, the impossibility to use real systems, the lack of data arising from such failure or the high costs of obtaining it, make the use of simulation an almost mandatory option in many situations. This tool allows estimation of models to predict behaviors of systems, in particular those concerning the identification of hazards, to estimate the probability of occurrence of a particular event and the consequences from that occurrence. The form of uncertainty present in the simulation, either in the model parameters or in the data used, or in the form of the model itself, show that uncertainty quantification is a prerequisite in probabilistic risk assessment.

The deterministic method of risk assessment relies on the assumption that the events are completely predetermined and the evaluation takes only some values into account, for example: the extreme values, the mean value, the 95th percentile and the optimum value. This approach has several drawbacks, as it uses only few values with the same weight, which is not realistic. Also the interdependence between the input values and the different impact they have on the output values is not considered, provoking the oversimplification of the model and the resulting reduction of its accuracy.

In probabilistic risk assessment, uncertainty is considered and the risk is characterized by a probability distribution, whose model is used to create/simulate different risk scenarios. The numerical simulation often involves high computational costs, in such way that it requires the use of metamodels. Response Surface Methodology is a suitable tool for the metamodels estimation, both in the case of systems' behaviors and risk assessment and in the quantification of uncertainties, thus revealing itself as a good alternative to Monte Carlo simulations. The next section presents a review of various aspects on the use of Response Surface Methodology in risk assessment.

2 Response Surface Methodoloy and Risk

The identification and characterization of hazards, the identification of patterns of exposure, the identification and analysis of main risk factors and events that may affect a system in what concerns the impact, the likelihood of occurrence and the propagation of uncertainties, are aspects to take into account in risk analysis. In all these matters the simulation and modeling have a fundamental role, since one seeks to evaluate different scenarios, to anticipate actions, to prevent, to mitigate and, if possible, to eliminate situations that could cause damage. It is through the implementation of these actions that the Response Surface Methodology (RSM) [5, 6, 29] plays an important role. The application of this methodology provides models that allow the characterization and/or optimization of a system or its components, or simple metamodels that replace complex numerical simulation models, and can be used within a framework of computationally intensive uncertainty analysis, making it a tool that must be taken into account in risk analysis.

Although this methodology has applications in increasingly diverse areas, it is in industry and especially in engineering projects that the widest range of applications has visible impact. RSM is an extremely useful tool for planning products and processes, for modeling and for optimizing systems. The quest of quality improvement and innovation in products and industrial processes at the lowest possible costs has inspired the need to improve statistical tools and seek new approaches, thus RSM has followed this trend. Many authors are important references as they address the methodology in several publications emphasizing its relevance in the referred areas of industry and particularly in engineering, for example Douglas Montgomery, Raymond Myers, George Box and his co-authors.

Response Surface Methodology consists of building a function that simulates the behavior of the real model in the space of the input variables. This methodology is based on the assumption that the answer η to a product, a process or a system is a function of a set of variables $x_1, x_2, \ldots, x_k$ and that this function can be approximated by a function f in such way that $\eta = f(x_1, x_2, \ldots, x_k) + \varepsilon$. The form of the true function f is unknown and ε is an error component that represents the variation sources other than those referred to in f. In the traditional way of application of the methodology, it is assumed that ε is normally distributed with mean value zero and constant variance, but other forms of implementation have arised which are free of these assumptions. The function f is estimated with a set of experimental or simulated points. One can introduce in the model controlled variables (factors), or include random variables that represent the system's uncertainties—stochastic response surface. To replace the real function, different mathematical models can be used, namely the expansion in Taylor polynomials and polynomial chaos, whose parameters need to be estimated since they are unknown.

This methodology has been used successfully in the treatment of risk, in areas such as radioactive waste disposal [19], environmental aspects [23, 24, 47], geological aspects [30, 32, 33, 38], structural problems of reliability [3, 7, 11, 13, 16, 17, 22, 28, 35–37, 39, 42–45], etc. The stochastic form of methodology with expansion in polynomial chaos is widely used. Also in terms of risk analysis the importance of RSM in industry and in particular to engineering projects becames crucial. These projects often involve very complex systems with several risks and effectively managing the balance between productivity and security is a challenge in many industries which operate critical engineering systems. This complexity leads to complex computational models, underscoring the need for accurate studies and thus involving high associated computational cost. RSM plays a key role in the simulation and analysis of these systems.

The most widely known quantitative definition of risk is that the risk of an event is the product of the probability of the occurrence of the event by the magnitude of its consequences (potential loss). In this approach, the product of the response probability model for the response of the model of consequences, in each scenario, provides a probabilistic measure of the risk of the event. A measure of the overall risk is obtained by adding a measure of the risk of each individual event in the system.

The risk curve represents the variation of the magnitudes of the consequences of the event based on the estimated probabilities for the occurrence thereof. Part of the difficulty of assessing the risk lies in the estimation of its components: the likelihood of an adverse event and the potential loss arising from the occurrence of such event. The two components of risk assessment are estimated using numerical simulation models or metamodels. In any case, the modelling of physical systems is complicated by the existence of several sources of uncertainty. However, despite the difficulty in incorporating uncertainties in the modeling process, they should be considered, since they allow the evaluation of the accuracy of the risk estimation.

The simulation model of risk may include variables controlled by the investigator but must include random variables that represent the uncertainty of the system, so it can assess its relevance in the system and its spread in the response, see Fig. 1. For example, [33] consider an integrative approach of response surface in which the simulation model of underground CO_2 storage includes both types of variables. However, in many approaches only the uncertain variables are included in the risk assessment model.

2.1 Sensitivity Analysis

The number of input variables in the model determines the computational cost of the simulation process of a probabilistic scenario (set of events that can occur in a system, planned or proposed from real data). Since it is necessary to simulate different scenarios to obtain estimations for the components of risk quantification, the lower the number of variables in the model, the lower will be the computational cost. The sensitivity analysis consists in assessing the uncertainty of each variable involved in the system and the variability of the phenomenon, allowing the identification of the distributions where uncertainty has a greater impact on the model response.

Response Surface Methodology can be used in the sensitivity analysis, especially in stochastic processes. The use of a metamodel which is simpler than numerical simulation model allows the reduction of computational cost, besides allowing the identification of possible interactions between variables. Reference [2] present RSM as an efficient tool in the sensitivity analysis. Reference [23] use the Response Surface for the sensitivity analysis in a study of the impact of the transfer of radionuclides to man after the release of gas from a nuclear installation. Reference [41] use a methodology based on RSM for the sensitivity analysis in a hydrologic model.

2.2 Uncertainty Analysis

Uncertainty is the lack of knowledge about the true value of a variable, the lack of knowledge about the model that best describes a system of interest or about which of several alternative probability distribution functions must represent an amount of interest [14]. Uncertainty may be associated to various system elements such as measurements in the input data, values of the parameters and the model structure and even to algorithms for obtaining the model and the human behavior. Thus, it is common to consider three components in uncertainty: structural uncertainty regarding the ignorance about the true model, the uncertainty in the parameters, introduced with the

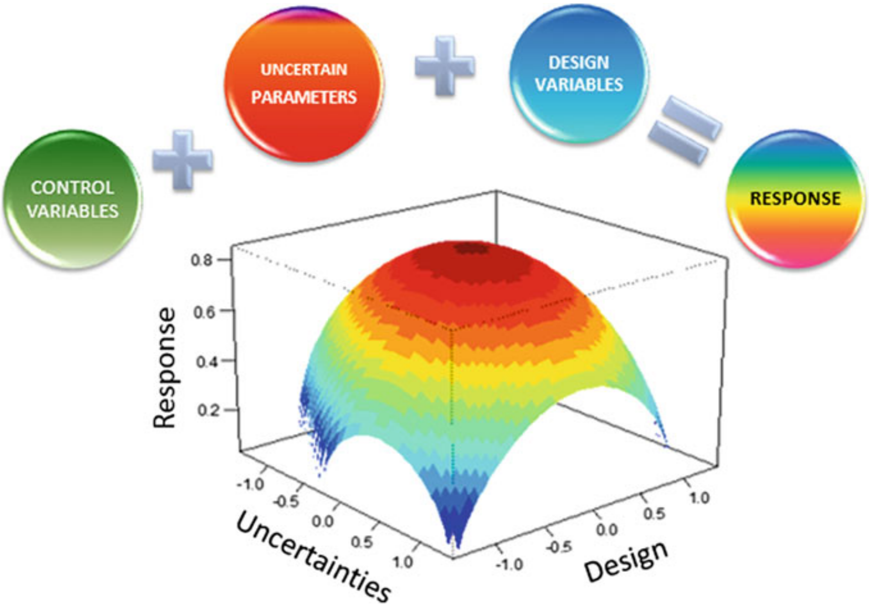


Fig. 1 Illustration of risk simulation model

need to use estimates for their values, and stochastic uncertainty resulting from the possibility of the variation of the parameters or of other important system amounts.

The uncertainty in the model can be addressed in two perspectives: one that assumes that the model for risk assessment or of its components does not vary, and another in which they vary according to the time or space in which the risk is assessed. In the first case, the significance of model uncertainty is in knowing whether and when the model can be applied to produce reasonable results or when it will fail. Validation is the best way to assess the model uncertainty [20, 21]. To implement the validation process several techniques can be applied. Examples include the comparison of values predicted by the model with numerous data sets obtained independently and under the same conditions as those underlying the risk assessment, cross-validation, the Bootstrap methodology or Jackknife method [10]. In the case where the risk evaluation model varies in time or space, it is possible to use several approaches to quantify the impact of uncertainty in the model. One can assess the consequences of this variation through the different simulation models, compare different values of the input variables in different models and use different Bayesian approaches to analyze the model uncertainties [4, 8, 9].

The stochastic uncertainty analysis—in the model parameters and the input variables—is the found in literature of risk analysis and the one that has aroused the most interest from the point of view of science or decision-making, given the fact that these sources of uncertainty will have an impact on the response risk assessment model. The propagation of uncertainty in the response of the model is of major importance in risk analysis, since decision-making is influenced by the risk estimate obtained from the model response. The uncertainty analysis allows checking out the confidence level in model estimates, identifying the main sources of uncertainty and quantifying the degree of confidence in the data and the existing model.

Several quantitative methodologies have been developed to analyze the propagation of stochastic uncertainties, and probability theory, along with statistics, provide the main concepts of its implementation, given the need to estimate the parameters and the quantification of randomness. These methodologies vary with the complexity of the system and with the model that is used to assess the risk. We will focus on some methods based on sampling by Monte Carlo simulation or Latin Hypercube and Response Surface Methods.

Stochastic uncertainty analysis comprises three main stages: (1) the characterization of uncertainty in model parameters or input variables based on their probability density functions (PDF) or cumulative distribution function (CDF), (2) the spread of these functions by model equations to obtain the PDF or CDF functions of the variable(s) response, and (3) the management of uncertainty.

The characterization of uncertainty in model parameters or in the input variables is based on their respective PDF functions. However, these functions are usually unknown and must be estimated using experimental or simulated data or assumptions on it must be taken into account.

Characterization of the uncertainty in the response variable is provided by the probability distribution of the responses of the model. Since this is unknown, an estimate can be obtained by numerical simulation of a high number of input data samples to use in the model, thus to obtain a high number of responses. Monte Carlo sampling or the Latin Hypercube are the most used methods to obtain samples of the input values.

For each input parameter with associated uncertainty or variability, the application of Monte Carlo method requires that a probability distribution (or frequency distribution) and the uncertainty limits for each parameter be given. The method consists of generating repeated independent pseudo-random values of uncertain input variables, from the known distribution (assumed or estimated) and within the limits of the imposed restrictions, followed by the application of the model using these values to generate a set of responses that are analyzed statistically in order to obtain an empirical probability distribution from the responses.

As an alternative to the traditional sampling method of Monte Carlo, the Latin Hypercube Design can be used to select the samples of the input values in a relatively simple way and without losing generality in applications. In addition, with this method one can obtain samples that reflect the shape of the function of density from which the sample is generated more accurately. This allows to obtain an estimate of the probability distribution that, in general, is better or equal to that obtained with the Monte Carlo [19].

To produce an accurate estimate of the probability distribution function, it is necessary to simulate a very large number of scenarios. Since the method described is computationally intensive, its use may be impractical because of the high computational costs, in case of a too complex system or whenever complex models are involved.

In the next section we will explore a methodology that solves some of these problems, since generally it converges more quickly to the solution.

3 The Stochastic Response Surface Methodology—Expansion into Polynomial Chaos

The use of Monte Carlo method or Latin Hypercube to study the propagation of uncertainty and to estimate the probability distribution of the response may have, as mentioned above, very high computational costs. For this reason, it is necessary to rely on methodologies that converge more quickly to the solution.

Stochastic Response Surface Methodology (SRSM) allows the generation of a reduced response model, computationally less demanding and statistically equivalent to the complete numerical model. For the estimation of its coefficients only the results of a limited number of simulations of the complete model are needed. Two case studies are presented by [25]. The basic idea of the methodology is to represent the response of a model to changes in variables, using a response surface defined with an orthogonal polynomial basis with respect to a probability measure on the space of parameters. SRSM relies on the assumption that the random variables, whose probability density functions are square integrable, can be approximated by the expansion in stochastic series of random variables or their direct transformation [1].

In the classic version of the methodology, a vector of random variables $\xi = (\xi_i), i = 1, \ldots, n$, is selected, under $N(0, 1)$ distribution, representing uncertain variables of a model in such way that $x_i = h(\xi_i)$. This selection made, response variables are represented as a function of the same vector of random variables: $Y = f(c, \xi)$, with c being a vector of coefficients to estimate. Estimates of model coefficients are obtained through the response of the system model to the various achievements of ξ. The coefficients c_i quantify the dependence of response Y on the input vector ξ, for each realization of x.

The form of the function f is the result of the polynomials chaos expansion (Ψ_i polynomials which form a base of orthogonal polynomials to a given probability measure) and is expressed by:

$$Y = f(c, \xi) = c_0\Psi_0 + \sum_{i1=1}^{\infty} c_{i1}\Psi_1(\xi_{i1}) + \sum_{i1=1}^{\infty}\sum_{i2=1}^{i_1} c_{i1i2}\Psi_2(\xi_{i1,\xi_{i2}}) + \cdots$$

In the case of the classical approach, the measure is Gaussian and the polynomials are the Hermite polynomials, see [15, 46]. Reference [48–51] showed that it is possible to obtain a better approximation of the response variables using non-Gaussian expansions in polynomials chaos. In this case, the Hermite polynomials are replaced by orthogonal polynomials with respect to the probability measure of input variables [48]. This approach was designated as the generalized polynomial chaos expansion. [12] presented conditions on the probability measures involving the mean square convergence of the generalized polynomials chaos expansion.

Reference [30] proposed a new generalization of the methodology, called arbitrary polynomial chaos expansion or data-driven chaos expansion. In this new approach, the probability distributions and the probability measures are arbitrary. Statistical moments are the only source of information that is propagated in the stochastic model. Probability distributions may be discrete, continuous, or continuous discretized, may be specified through an analytical way (PDF or through CFD), numerically using a histogram or by using raw data. In this approach, all distributions are admissible for the input variables of a given model, as long as they have a finite number of moments in common. Thus, in the case of considering a truncated polynomial, only a finite number of moments needs to be known, with no need for complete knowledge of the probability density function or even its existence, which frees the researcher from the need of assumptions that may not always be supported by existing data. According to the literature, this expansion converges exponentially and faster than the classical expansion.

The estimation of model parameters depends on model complexity [24]. In case the model is invertible, the parameters can be obtained directly from the input random variables $(\xi_i)_{i=1}^{n}$. If the model equations are mathematically manipulated, in spite of nonlinearities, then the model coefficients can be obtained afterwards, by an appropriate norm minimization of residuals, replacing the input random variables by the respective transformations in terms of Gaussian variables $N(0, 1)$ (Galerkin method). When the model equations are difficult to manipulate the coefficients can be estimated by the collocation points methods. Each set of points chosen such that the model estimates are accurate at these points, gives a set of N linear equations whose solution allows us to obtain the N parameters.

Reference [24] present some methods for parameter estimation, all based on the collocation points methods: Probabilistic Collocation Method, Efficient Collocation Method and Regression Based Method and these authors discuss advantages and disadvantages for each method.

The expansion in polynomials chaos is a simple but powerful tool for stochastic modeling. Probability density functions, probability distribution functions or other statistics of interest can be estimated and quickly evaluated via Monte Carlo simulation, once the evaluation of a polynomial function is faster than the original equations model evaluation.

In the case of risk analysis, to use arbitrary expansion one can directly consider a set of large-sized data or probability density function of maximum or relative minimum entropy, since, in this case, the relevant moments of the polynomial chaos expansion are compatible with those of the input variables. The bootstrap resampling method may be used to obtain more precise estimates of the moments from a reduced set of data available, providing a more accurate estimation of the risk assessment model. Reference [34] propose such an application on calibration models to history matching for CO_2 storage in underground reservoirs.

4 Applications and Computational Resources

Response Surface Methodology plays a key role on the generation of fast models, or metamodels (proxy models), replacing the simulator in complex processes which requires many simulations. The applications are varied and many of them concern Stochastic Response Surface Methodology, a specially suited approach for the quantification of uncertainty.

Besides the examples already mentioned, some other applications deserve a special reference, like simulations taking place in underground stocking of CO_2, see [27, 38] as an example of classical methodology, [31–34] using approaches with polynomial chaos expansion; risks associated with natural or human threats, see [22]; and seismic vulnerability of structures and buildings, see [39].

Reference [26] use the methodology to assess the potential of flooding resulting from a tropical cyclone, and [18] use the probabilistic risk assessment methodology in the probabilistic assessment of the risk in an accident with a nuclear reactor. Reference [24] apply the methodology to two case studies: one for the analysis of uncertainty concerning the carcinogenic effects of the perchloroethylene in humans and the other directed to assess the concentrations of environmental pollutants and of emission sources.

Table 1 Some R package

Package	Description
rsm	Provides functions to generate response-surface designs, fit first-and second-order response-surface models, make surface plots, obtain the path of steepest ascent, and do canonical analysis
propagate	Propagation of uncertainty using higher-order Taylor expansion and Monte Carlo simulation
FME	Provides functions to help in fitting models to data, to perform Monte Carlo, sensitivity and identifiability analysis. It is intended to work with models written as a set of differential equations that are solved either by an integration routine from package *deSolve*, or a steady-state solver from package *rootSolve*
lhs	Provides a number of methods for creating and augmenting Latin Hypercube Samples

The implementation of Response Surface Methodology in its classical form, as far as the optimization and the response surface exploration is concerned, is available, for example, in the commercial software Design-Expert, Optimus or SAS. Free software R has a specific package for the implementation of the methodology in its classical form, *rsm*, and some packages with tools that enable the implementation of more current forms of this methodology, namely those concerning the generation of designs, different from the classical ones, the implementation of Monte Carlo (cf. Table 1). However, there is no record of any package that enables the implementation of the stochastic form of the methodology. There are some free tools that assist the implementation of the SRSM, particularly those provided by the Community Portal for Automatic Differentiation and by The DAKOTA Project.

5 Considerations and Conclusion

In general, response surface methods are well developed and find applications in many fields. The application of RSM in association with other techniques such as neural networks, computer simulation or genetic algorithms can be found in many applications ranging from Industries, Physical and Chemical Sciences, Engineering, Biological and Clinics Sciences, Food Sciences, Social Sciences, Agriculture, Aeronautics to other (countless) areas. Response surface approximations serve as surrogate models for the full mathematical model that can be used to quickly interrogate regions of the input space that were not sampled. Risk Assessment is a field where the RSM can be used, namely for modeling consequences from a event and uncertainties in the model, where SRSM may prove itself as a very useful tool.

It's a fact that there are some research projects in which the sharing experiences become valuable for the tools and assistance they provide in the implementation of the SRSM, with particular emphasis on uncertainty analysis. Nevertheless, it would be interesting to go further and explore, in R Project, the construction of computational tools enabling implementation of the stochastic approach of RSM, since the project is of free access. This would be specially interesting in the particular case of polynomial chaos expansion.

Acknowledgments The authors are very grateful to the referees for their useful comments and suggestions. This research was partially sponsored by national founds through the Fundação Nacional para a Ciência e Tecnologia, Portugal - FCT under the projects: (PEst-OE/MAT/UI0006/2011 and PEst-OE/MAT/UI0006/2014).

References

1. Balakrishnan, S., Roy, A., Ierapetritou, M.G., Flach, G.P., Georgopoulos, P.G.: Uncertainty reduction and characterization for complex environmental fate and transport models: an empirical Bayesian framework incorporating the stochastic response surface method. Water Resour. Res. **39**(12), 1350 (2003)
2. Bauer, K.W., Parnell, G.S., Meyers, D.A.: Response surface methodology as a sensitivity analysis tool in decision analysis. J. Multi-Criteria Decis. Anal. **8**(3), 162–180 (1999)
3. Baysal, R.E., Nelson, B.L., Staum, J.: Response surface methodology for simulating hedging and trading strategies. In: Simulation Conference, WSC, Winter, December 2008, pp. 629–637. IEEE (2008)
4. Bouda, M., Rousseau, A.N., Konan, B., Gagnon, P., Gumiere, S.J.: Bayesian uncertainty analysis of the distributed hydrological model HYDROTEL. J. Hydrol. Eng. **17**(9), 1021–1032 (2011)
5. Box, G.E., Draper, N.R.: Empirical Model-Building and Response Surfaces. Wiley, New York (1987)
6. Box, G.E., Wilson, K.B.: On the experimental attainment of optimum conditions. J. R. Stat. Soc. Ser. B (Methodological) **13**(1), 1–45 (1951)
7. Bucher, C.G., Bourgund, U.: A fast and efficient response surface approach for structural reliability problems. Struct. Saf. **7**(1), 57–66 (1990)
8. Cheung, S.H., Oliver, T.A., Prudencio, E.E., Prudhomme, S., Moser, R.D.: Bayesian uncertainty analysis with applications to turbulence modeling. Reliab. Eng. Syst. Saf. **96**(9), 1137–1149 (2011)
9. Der Kiureghian, A.: Bayesian analysis of model uncertainty in structural reliability. In: Reliability and Optimization of Structural Systems'90, pp. 211–221. Springer, Berlin (1991)
10. Efron, B.: The Jackknife, The Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia (1982)
11. El-Masri, H.A., Reardon, K.F., Yang, R.S.: Integrated approaches for the analysis of toxicologic interactions of chemical mixtures. CRC Crit. Rev. Toxicol. **27**(2), 175–197 (1997)
12. Ernst, O.G., Mugler, A., Starkloff, H.J., Ullmann, E.: On the convergence of generalized polynomial chaos expansions. ESAIM: Math. Model. Numer. Anal. **46**(02), 317–339 (2012)
13. Feraille, M., Marrel, A.: Prediction under uncertainty on a mature field. Oil Gas Sci. Technol.–Rev. IFP Energ. Nouv. **67**(2), 193–206 (2012)
14. Frey, H.C., Mokhtari, A., Zheng, J.: Recommended practice regarding selection, application, and interpretation of sensitivity analysis methods applied to food safety process risk models. US Department of Agriculture. http://www.ce.ncsu.edu/risk/Phase3Final.pdf (2004)
15. Ghanem, R.G., Spanos, P.D.: Stochastic Finite Elements: A Spectral Approach. Springer, New York (1991)
16. Groten, J.P., Feron, V.J., Sühnel, J.: Toxicology of simple and complex mixtures. Trends Pharmacol. Sci. **22**(6), 316–322 (2001)

17. Gupta, S., Manohar, C.S.: An improved response surface method for the determination of failure probability and importance measures. Struct. Saf. **26**(2), 123–139 (2004)
18. Ha, T., Garland, W.J.: Loss of coolant accident (LOCA) analysis for mcmaster nuclear reactor through probabilistic risk assessment (PRA). In: Proceedings of 27th Annual Conference of the Canadian Nuclear Society Toronto, Ontario, Canada, 11–14 June 2006
19. Helton, J.C.: Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. Reliab. Eng. Syst. Saf. **42**(2), 327–367 (1993)
20. Hoffman, F.O., Miller, C.W., Ng, Y.C.: Uncertainties in radioecological assessment models (No. IAEA-SR-84/4; CONF-831032-1). Oak Ridge National Laboratory, TN (USA); Lawrence Livermore National Laboratory, CA, USA (1983)
21. Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, 3rd edn. Wiley, New York (2013)
22. Iervolino, I., Fabbrocino, G., Manfredi, G.: Fragility of standard industrial structures by a response surface based method. J. Earthq. Eng. **8**(06), 927–945 (2004)
23. Iooss, B., Van Dorpe, F., Devictor, N.: Response surfaces and sensitivity analyses for an environmental model of dose calculations. Reliab. Eng. Syst. Saf. **91**(10), 1241–1251 (2006)
24. Isukapalli, S.S., Georgopoulos, P.G.: Computational Methods for Sensitivity and Uncertainty Analysis for Environmental and Biological Models. Environmental and Occupational Health Sciences Institute, New Jersey (2001)
25. Isukapalli, S.S., Roy, A., Georgopoulos, P.G.: Stochastic response surface methods (SRSMs) for uncertainty propagation: application to environmental and biological systems. Risk Anal. **18**(3), 351–363 (1998)
26. Kennedy, A.B., Westerink, J.J., Smith, J.M., Hope, M.E., Hartman, M., Taflanidis, A.A., Dawson, C.: Tropical cyclone inundation potential on the Hawaiian Islands of Oahu and Kauai. Ocean Model. **52**, 54–68 (2012)
27. Kleijnen, J.P., van Ham, G., Rotmans, J.: Techniques for sensitivity analysis of simulation models: a case study of the CO_2 greenhouse effect. Simulation **58**(6), 410–417 (1992)
28. Liel, A.B., Haselton, C.B., Deierlein, G.G., Baker, J.W.: Incorporating modeling uncertainties in the assessment of seismic collapse risk of buildings. Struct. Saf. **31**(2), 197–211 (2009)
29. Myers, R.H., Montgomery, D.C., Anderson-Cook, C.M.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley, New York (2009)
30. Oladyshkin, S., Nowak, W.: Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. Reliab. Eng. Syst. Saf. **106**, 179–190 (2012)
31. Oladyshkin, S., Class, H., Helmig, R., Nowak, W.: Highly efficient tool for probabilistic risk assessment of CCS joint with injection design. Comput. Geosci. **13**, 451–467 (2009)
32. Oladyshkin, S., Class, H., Helmig, R., Nowak, W.: An integrative approach to robust design and probabilistic risk assessment for CO_2 storage in geological formations. Comput. Geosci. **15**(3), 565–577 (2011)
33. Oladyshkin, S., Class, H., Helmig, R., Nowak, W.: A concept for data-driven uncertainty quantification and its application to carbon dioxide storage in geological formations. Adv. Water Resour. **34**(11), 1508–1518 (2011)
34. Oladyshkin, S., Class, H., Nowak, W.: Bayesian updating via bootstrap filtering combined with data-driven polynomial chaos expansions: methodology and application to history matching for carbon dioxide storage in geological formations. Comput. Geosci. **17**, 1–17 (2013)
35. Patel, T., Telesca, D., George, S., Nel, A.: Toxicity profiling of engineered nanomaterials via multivariate dose response surface modeling (2011)
36. Que, J.: Response surface modelling of Monte-Carlo fire data. Doctoral dissertation, Victoria University (2003)
37. Risso, F., Schiozer, D.: Risk analysis of petroleum fields using Latin hypercube, Monte carol and derivative tree techniques. J. Pet. Gas Explor. **1**(1), 014–021 (2011)
38. Rohmer, J., Bouc, O.: A response surface methodology to address uncertainties in cap rock failure assessment for CO_2 geological storage in deep aquifers. Int. J. Greenh. Gas Control **4**(2), 198–208 (2010)

39. Rossetto, T., Elnashai, A.: A new analytical procedure for the derivation of displacement-based vulnerability curves for populations of RC structures. Eng. Struct. **27**(3), 397–409 (2005)
40. Royal Society: Risk: Analysis, Perception and Management. Report of a Royal Society Study Group, London, The Royal Society, pp. 89–134 (1992)
41. Song, X., Zhan, C., Xia, J., Kong, F.: An efficient global sensitivity analysis approach for distributed hydrological model. J. Geogr. Sci. **22**(2), 209–222 (2012)
42. Steffen, O.K.H., Contreras, L.F., Terbrugge, P.J., Venter, J.: A risk evaluation approach for pit slope design. 42nd US rock mechanics symposium and 2nd U.S.-Canada Rock Mechanics Symposium, held in San Francisco, 29 June–2 July 2008
43. Taflanidis, A.A., Kennedy, A.B., Westerink, J.J., Smith, J., Cheung, K.F., Hope, M., Tanaka, S.: Probabilistic hurricane surge risk estimation through high fidelity numerical simulation and response surface approximations. ASCE April 2011
44. Tanase, F.N.: Seismic performance assessment using response surface methodology. Constr.: J. Civ. Eng. Res. **2**, 13 (2012)
45. Wang, X., Song Z.: Reliability analysis of evacuation B improved response surface method. In: 2nd International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT-2012). Published by Atlantis Press, Paris, France (2012)
46. Wiener, N.: The homogeneous chaos. Am. J. Math. **60**(4), 897–936 (1938)
47. Wilde, M.L., Kümmerer, K., Martins, A.F.: Multivariate optimization of analytical methodology and a first attempt to an environmental risk assessment of β*-blockers* in hospital wastewater. J. Braz. Chem. Soc. **23**(9), 1732–1740 (2012)
48. Xiu, D., Karniadakis, G.E.: Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. Comput. Methods Appl. Mech. Eng. **191**(43), 4927–4948 (2002)
49. Xiu, D., Karniadakis, G.E.: The Wiener-Askey polynomial chaos for stochastic differential equations. SIAM J. Sci. Comput. **24**(2), 619–644 (2002)
50. Xiu, D., Karniadakis, G.E.: Modeling uncertainty in flow simulations via generalized polynomial chaos. J. Comput. Phys. **187**(1), 137–167 (2003)
51. Xiu, D., Karniadakis, G.E.: A new stochastic approach to transient heat conduction modeling with uncertainty. Int. J. Heat Mass Transf. **46**(24), 4681–4693 (2003)

FF-Type Multivariate Models in an Enforced Regression Paradigm

Jerzy K. Filus and Lidia Z. Filus

Abstract We consider the stochastic dependence of a given random variable Y on a set of its explanatory variables. Using our earlier *method of parameter dependence* we obtain a description of this dependence in the form of a conditional probability distribution of Y, given any realization of the explanatory variables. We obtain a wide class of conditional distributions, including most of the important non-Gaussian cases, in an explicit, tractable, analytical form which, basically, is not known in the current literature. This fact automatically prompts one to extend the existing regression models, which usually are given in form of conditional expectations, to models based on the corresponding conditional probability distributions, given the same values of the data. The latter models, obviously, contain more statistical information and thus are expected to give better predictions. We also included some, related to the conditional, multivariate probability densities.

Keywords Regression · Conditional probability distributions as enforced (extended) regression · Multivariate probability distributions

1 Introduction

We apply the method of parameter dependence [1–4], used for the construction of a variety of multivariate probability distributions, to analyze closer the (***weak***) dependence of a given random variable Y from a set of explanatory random variables, say, $X_1, \ldots, X_k$ [5].

J.K. Filus (✉)
Department of Mathematics and Computer Science,
Oakton Community College, Des Plaines, IL 60016, USA
e-mail: jfilus@oakton.edu

L.Z. Filus
Department of Mathematics, Northeastern Illinois University,
Chicago, IL 60625, USA
e-mail: L-Filus@neiu.edu

C.P. Kitsos et al. (eds.), *Theory and Practice of Risk Assessment*,
Springer Proceedings in Mathematics & Statistics 136,
DOI 10.1007/978-3-319-18029-8_30

High generality of the method of parameter dependence opens the possibility to present it now as a base for settings of the theory that, in a sense, extends and modifies the theory of classical regression and the underlying models.

Recall that, perhaps up to the presence, the vast majority of (stochastic) models and methods, employed in the classical regression theory and its applications, were essentially related to the conditional expectation

$$E[Y|x_1, \dots, x_k] \tag{1}$$

of a random variable (or a vector) of the main interest Y, given realizations $x_1, \dots, x_k$ of a set of explanatory random variables $X_1, \dots, X_k$. This kind of stochastic model for the statistical theory of regression is limited with respect to an amount of statistical information and thus the precision of the corresponding predictions is limited as well.

Obviously, a better (i.e., more precise) model compared to (1) would be the stochastic model in the form of a whole conditional probability distribution (not just its expected value) of the same random variable Y, given the same occurrences of the explanatory values $x_1, \dots, x_k$.

A practitioner statistician would certainly be *more happy* having the possibility to enrich the typical regression model (1) by the following conditional cdf:

$$P(Y \le y) = G_k(y|x_1, \dots, x_k), \tag{2}$$

or, whenever it exists, the corresponding conditional pdf:

$$g_k(y|x_1, \dots, x_k). \tag{3}$$

As model (1) is *part* of the (2) or (3) model, (2) as well as its theory, is the, proposed here, natural extension of this part of the classical regression theory which deals with (1)—like models. Of course, this possibility was well-known fact for a long time. But using the stochastic dependence models in the form (2) and (3) was (possibly, only, with exception of the conditionals associated with the classical multivariate normal pdfs) seen as *not realistic*. The reason for this shortcoming is explained as follows. In almost all non-Gaussian situations there were no well defined and handy enough analytical forms of the conditional distributions (2) that would explicitly and meaningfully reflect the stochastic dependences of the random variables Y from the explanatory ones $X_1, \dots, X_k$. Discovering the pattern (of the parameter dependence) [1, 2] for an explicit description of the stochastic dependences and, associated with it, new possibilities for construction of the conditional distributions (2) on a *mass-scale* gave us hope to enrich the statistical regression theory in a new way.

The main goal for all that is an expected (in many practical cases) improvement in accuracy of predictions for the phenomena so far modeled by regular (conditional expectation) regression methods.

At this point let us recall the known fact that some (to a measure similar) results were obtained and the corresponding theory exists in the literature. This theory, known as the *quantile regression methods* was first introduced by Koenker and Bassett in 1978 [6], and is continued by some authors up to now. Similar to ours, this approach is aimed to complement the classical linear regression analysis, mainly by replacing the typical investigations centered on the conditional expectation $E[Y|x_1, \dots, x_k]$ by more general considerations on the conditional quantiles directly associated with the (whole) conditional probability distribution. The main advantage of the extension is the possibility to obtain more statistical information than by using traditional methods of regression. The essential difference between those methods and our present approach mainly relies on the parametric character of the, here presented, extended regression models. In practice this means that our models are capable of *embracing* (in one analytic formula) the probabilities $P(Y \leq y|x_1, \dots, x_k)$ for all the y values *at the same time* (for any y, this probability as a function of the arguments $x_1, \dots, x_k$ we call the *enforced regression function of the second kind*) not just for *each* quantile y separately, which, practically, reduces the investigations to few values of y, only. As for the multivariate normal case exception, it should be admitted that in the extended regression theory presented here, both the common restrictions on the normality and the linearity of the regression function, can be relaxed in a natural way.

The general pattern for our stochastic dependence description is obtained by means of some method of conditioning. This method relies on a kind of randomization of the originally constant parameter(s) (the baseline, i.e., *no explanatory random variables influence case*) of pdf $g_0(y)$ of the random variable Y. Namely, we consider these parameters as continuous (and in general nonlinear) function of the independent explanatory random variables $X_1, \dots, X_k$, each of them having some known probability distribution. For example, Y may be considered as the life time of an object and $X_1, \dots, X_k$ may be extra stresses put on this object.

This procedure yields the determination of a set of classes of conditional pdfs $g_k(y|x_1, \dots, x_k)$ or cdfs $G_k(y|x_1, \dots, x_k)$ of the random variable Y, given the realizations of the independent (explanatory) random variables $X_1, \dots, X_k$. Our purpose is to provide a method for explicit determination of any conditional probability $P(a \leq Y \leq b|x_1, \dots, x_k)$ for all: $-\infty \leq a < b \leq \infty$, (also assuming that $P(Y = \pm\infty) = 0$) in a concise analytical form as a function of the realizations $x_1, \dots, x_k$ of the explanatory random variables $X_1, \dots, X_k$. In particular, in Sect. 5.2 we consider conditional survival functions $P(Y \geq y|x_1, \dots, x_k)$ with Y interpreted as the life time of an object and $X_1, \dots, X_k$ as stresses that the object endures.

The above probabilities that the object would survive at least time period of length y (for any y) is an analytically given continuous function, say $r(x_1, \dots, x_k)$ of the stresses realizations $x_1, \dots, x_k$. This function is also considered (see Sect. 5.2) as the *enforced regression of the second kind*, and this is the main stochastic model we construct in this paper. In Sect. 5.1 we precede this by the notion of *enforced regression of the first kind* which also covers, in a new way, conditional expectations $E[Y|x_1, \dots, x_k]$ when the baseline probability distribution of Y is neither normal nor exponential. When Y is normally distributed, the enforced regression of the first kind becomes identical with the ordinary (not necessarily linear) regression function. Both first and second kind of the enforced regression are based on the parameter dependence (also called *weak dependence*) between the random variable Y and the set of random variables $X_1, \dots, X_k$. We explain this kind of stochastic dependence in Sect. 2. In Sects. 3 and 4 we consider the, defined in Sect. 2, conditional probability distributions in the wider framework of joint multivariate distributions. The main results concerning the enforced regression are placed in Sect. 5.

2 Parametric Description of the Stochastic Dependences

2.1 The Probability Distribution of the Life-time

We are interested in the probability distribution of a random variable Y which, in particular, may correspond to the (random) life-time of an object. In case of **actuary** applications, Y may be considered as the residual life-time of a client whose age at the moment of registration is t. Assume that in the (regular situation) the residual life-time Y has a given known probability density function $f(y; \theta)$, where θ is its scalar or vector parameter. In some significant cases the situation complicates because of the presence of extra (disturbing facts) taking place when potential clients were subjected to special stresses that could affect their residual life-time. In some cases the amount of a given stress can be measured by a (possibly random) quantity X_i. For example, X_i can be the time the stress was endured, multiplied by its intensity, for example as the amount of nicotine or alcohol consumed in an average day. Thus, the *disturbing factors* having an impact on the residual life-time Y, (its probability distribution) may be considered as either a single random variable X or as a set of such (independent) random variables $X_1, X_2, \dots, X_k, k = 2, 3, \dots$

2.2 General Stochastic Mechanism

We aim to find an efficient way to describe, analytically, the general stochastic mechanism by which a random stress X *influences* the random life-time Y. In the general version (of the *extended regression theory*) the question can be formulated as: *how can one analytically express the influence of any random variable X on another random variable Y probability distribution, when the two are not independent and with no given explicit algebraic relation (transformation) between them.*

The stochastic (indirect) influence of a value x, corresponding to the random event (observation) $X = x$, on possible values y of the random variables Y (the stochastic impact of x on y) is understood as the impact of x on the probability (or probability density or hazard rate) of occurring y, when $X = x$ happens. This, in general, means that different values x, x^* of X may bring different probabilities (or densities) of the given random event $Y = y$. Speaking more generally, in the considered setting, the random event $X = x$, *influences* the probabilities $P(Y \leq y)$, for any fixed real value y.

In other words, we are seeking for the conditional probability distributions $P(Y \leq y|X = x)$, or the corresponding conditional probability densities, hazard rates, etc. In the framework we have chosen, the phrase: *Size of an incentive (or, in particular, stress) x changes the probability density $f(y;\theta)$ of the life-time Y* or (equivalently), the *x changes Ys hazard (failure) rate $\lambda(y;\theta)$ through a change of the parameter* θ will be understood as: *the stress X changes, (proportionally to its magnitude x) the numerical value of the parameter(s) θ.* In yet other words *a change in θ is proportional to a magnitude x of the stress X*, given the random event $X = x$. To describe the (deterministic) relationship between θ and x one can impose, as new values of θs, a *hypothetical* continuous (not necessarily linear) function of x: $\theta = \theta(x)$. For example, we may let $\theta \rightarrow \theta(x) = \theta \cdot (1 + ax + bx^2)$ and then, assuming such a model, statistically estimate the (new) parameters a and b, where the factor θ on the right hand side of the above equality may be known from *previous* (baseline) estimation procedures. We obtain the conditional density $g_2(y|x)$ of Y, given $X = x$, as defined by

$$g_2(y|x) = f(y; \theta(x)). \tag{4}$$

(Here notice that to define any new object we usually need to prove its existence and uniqueness. In the above case however, the proof is straightforward, simply given *by indication*. The existence and uniqueness of $g_2(y|x)$ above (as well as in all other similar cases present in this work) directly follows from the existence and uniqueness of any $f(y;\theta)$, which is a known fact, and from the fact that the function $\theta(x)$ is chosen to be a known. Obviously, two different continuous functions $\theta(x)$ will produce two different objects $f(y;\theta(x))$ for the same original density $f()$. The only requirement that *a given function $\theta(x)$ fits to $g_2(y|x)$* is: *the range of that function is included in the range of the values of the parameter θ* (often the set of all positive reals) which is satisfied. These recognitions complete the required proof).

The last key formula yields directly the extended regression theory formulation (see Sect. 5). Equivalently, one can define the conditional hazard (failure) rate by transforming the original hazard rate $\lambda_2(y,\theta)$ of Y to the conditional one

$$\lambda_2(y|x) = \lambda_2(y, \theta(x)). \tag{5}$$

It is examined closer, below.

3 Further Development

3.1 Bivariate Normal Case

Having defined the conditional pdf $g_2(y|x)$ and the marginal pdf $g_1(x)$ of the stress X, we automatically obtain the joint pdf $h(x, y)$ of the random variables X and Y

simply as the arithmetic product

$$h(x, y) = g_2(y|x)g_1(x) \tag{6}$$

(when the random vector (X, Y) is meaningful in the investigations).

Anyway this, apparently, is the method (also called the *parameter dependence*) for constructing bivariate and multivariate probability densities. Being a general method of construction it can be considered as an extension of the paradigm on which the construction of classical bivariate and multivariate Gaussian densities is based [1, 2]. The same applies to the associated Gaussian conditional densities. So, the method of parameter dependence is just an extension of the method of how the classical multivariate normal were constructed. (The same statement may apply to the, considered in Sect. 5, enforced regression.) We explain this closer in below.

Realize that, as any bivariate pdf, the classical bivariate normal $g(x, y)$ can be uniquely represented as the product

$$f_2(y|x)f_1(x) = g(x, y), \tag{7}$$

where here $f_1(x)$ is any univariate (here the marginal) normal $N(\mu_1, \sigma_1)$ pdf of the marginal X. $f_2(y|x)$ is the conditional pdf of Y, given the random event $(X = x)$ realizes, obviously with the probability density $f_1(x)$. Notice that as the marginal normal (*not yet influenced* baseline) pdf $f_2(y)$ of Y is

$$f_2(y) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{\frac{-(y-\mu_2)^2}{2\sigma_2^2}} \tag{8}$$

the corresponding conditional pdf $f_2(y|x)$ of Y (*influenced* by the event '$X = x$') is

$$f_2(y|x) = \frac{1}{s_2\sqrt{2\pi}} e^{\frac{-(y-\mu_2-a(x-\mu_1))^2}{2s_2^2}}, \tag{9}$$

where $a = \rho\frac{\sigma_2}{\sigma_1}$ and $s_2^2 = \sigma_2^2(1 - \rho^2)$, ρ being the linear correlation coefficient. It is quite clear now that for the bivariate normal pdfs device the transition $f_2(y) \to f_2(y|x)$ may be regarded as the result of the, here considered, *stochastic action* of the random variable X on Y, resulting in the new random variable Y^* (for simplicity, in our notation will be $Y^* = Y$) or just as an *action* of the random variable X (arbitrary) realization x on the (density of) y, $(x \to f_2(y))$.

As in the case considered above, this action (impact) affects the parameter μ_2 of the marginal $N(\mu_2, \sigma_2)$ of Y in such a way that it is transformed from μ_2 to $\mu_2^* = \mu_2 + a(x - \mu_1)$. The latter value μ_2^* of the parameter clearly *became* a continuous function of the argument x, $\mu_2(x)$.

3.2 FF-Normal (Pseudonormal) Models

It should now be obvious that the above *old* classical normal bivariate model obeys exactly the same pattern of influence $X \to Y$ as we outline in this paper. The (only) difference between the two approaches is that the (parameter) function $\mu_2(x) = \mu_2 + a(x - \mu_1)$, used in the normal pdf case, is only linear, and is applied only for this particular parameter μ_2 of this particular conditional distribution. However, in general, there is no need for such restrictions (to the normal cases only), even if sometimes an extension of that normal pattern results with some (usually modest) prices. We may enrich the above linear function $\mu_2 + a(x - \mu_1)$ by adding a (*correcting*) quadratic term to obtain other parameter (function)

$$\mu_2^*(x) = \mu_2 + a(x - \mu_1) + A(x - \mu_1)^2, \tag{10}$$

and obtain another, corresponding, conditional, but still normal, pdf $f_2^*(y|x)$ of Y, given the value x of X. The anticipated *gain* is, in some cases, a better accuracy in the sense of a (new) model's fit to the given data.

Note that as $A \to 0$ the new (pseudonormal) model approaches the original classical normal. There is also no reason to avoid further extensions of the quadratic polynomial parameter function (3) to polynomials of higher degrees, whenever it may improve the models accuracy. But instead of polynomials we may apply any other suitable continuous parameter function $r(x - \mu_1)$, such as exponential, logarithmic, trigonometric, etc. in $(x - \mu_1)$.

Now the *action* $x \to Y$ parallels transforming the parameter μ_2 of Y in the way $\mu_2 \to \mu_2 + r(x - \mu_1)$, which results in a new (still normal) conditional pdf of $Y|x$:

$$f_2(y|x) = \frac{1}{s_2\sqrt{2\pi}} e^{\frac{-(y-\mu_2-r(x-\mu_1))^2}{2s_2^2}} \tag{11}$$

Also the parameter σ_2 of Y can be affected by (a stress) x, so that $\sigma_2 \to (q(x-\mu_1))\sigma_2$, where $\sigma_2 = \frac{s_2}{\sqrt{(1-\rho_2)}}$, and $q(x-\mu_1)$ is any (proper) non-negative continuous function of the $(x - \mu_1)$, while ρ is the linear correlation coefficient of the original bivariate normal version of the now extended model.

Now we can formulate general

Definition 1 Definition of a pseudonormal extension of the bivariate normal pdf as given by the product

$$g(x, y) = f_2^{**}(y|x) f_1(x), \tag{12}$$

where the marginal pdf $f_1(x)$ of X is the, as before, ordinary $N(\mu_1, \sigma_1)$ pdf, while the other factor in (12), the (general) conditional pdf, is given as

$$f_2^{**}(y|x) = [((q(x - \mu_1))\sigma_2\sqrt{2\pi})]^{-1} e^{\frac{-(y-\mu_2-r(x-\mu_1))^2}{2(q(x-\mu_1))^2\sigma_2^2}} \tag{13}$$

Particular Examples of the bivariate pseudonormal pdfs easily follow formulas (10) and (11).

For earlier results on, what we called 'pseudonormals', see for instance [7, 8] and the books [9, 10].

Here, however, we propose to **rename** the notion of *pseudonormal* to *FF-normal* or *FF-Gaussian*. In parallel, we change the names of similar notions such as *pseudoexponential*, *pseudoWeibullian*, *pseudogamma* (distributions) to *FF-exponential*, *FF-Weibullian*, *FF-gamma* ..., respectively, and we will use them from now on throughout.

In order to construct more stochastic dependence $X \rightarrow Y$ models (where X is considered as an explanatory variable for Y), realize that there is no reason to reduce the considerations to the $normal \rightarrow FF - normal$ extension pattern. We may, instead, apply the above *method of parameter dependence* to almost any other parameter dependent probability distribution (density) of a Y. Moreover, there is neither theoretical necessity nor practical need for assuming that the explanatory variable X (stress) and the variable of interest Y belong to the same class of probability distributions. Actually, each of them separately can be chosen from any reasonable class of cdfs. In particular, the pdf $f_1(x)$ of X present in formula (12) need not necessarily be normal. It can be, for example, the gamma, but in such cases the model given by (12) and (13) is not FF-normal any more (still, however, may be considered as an *extension of the FF-normal*, say *semi FF-normal*). That freedom (and relative easiness) in the models construction procedure yields a remarkable generality, and therefore it brings a *promise* to be applied in many other than the, so far considered, real-life problems.

3.3 Non FF-Normal Models

Example 1 Consider the following *FF - exponential* case. Suppose the residual life-time Y, in absence of a given stress X, has the exponential pdf

$$f(y;\theta) = \theta^{-1} e^{\frac{-x_1}{\theta}} \tag{14}$$

Consider a non-zero stress $X = x$, affecting the life-time Y, so turning it into the random variable $Y|x$. Denote the involved stochastic mechanism by $(x \rightarrow Y) \rightarrow (Y|X = x)$, or as *densities transformation* $f(y) \rightarrow f(y|x)$. As a result of that, the parameter θ of density (14) transforms into a parameter $\theta^* = \theta\varphi(x)$, that is continuously dependent on an amount x of the stress X. Mathematically, that results in the following determination of the conditional density of $Y|X = x$, by formula: $f(y|x) = f(y; \theta\varphi(x))$, where $\varphi(x)$ is assumed to be any (properly chosen, for a given practical situation) continuous function of x, when random event $X = x$ happens. One such a function with nice analytical properties (easy calculations), we propose, is

$$\varphi(x) = 1 + Ax^r \tag{15}$$

[or, in higher dimension cases $\varphi_k(x_1, \ldots, x_k) = 1 + A_1x_1^r + A_2x_2^r + \cdots + A_kx_k^r$] where r, A (or r, $A_1, \ldots, A_k$) are positive real parameters with specially important cases, for $r = 1$ or $r = 2$. The parameters are to be estimated and verified by statistical methods, provided that the given above function $\varphi(x)$ is a properly chosen (sub)model. (So basically, most of the statistical work to be done retains its *parametric character*). Under the assumption (15) the formula (14) takes the form:

$$f(y|x) = (\theta(1 + Ax^r))^{-1} e^{\frac{-y}{\theta(1+Ax^r)}} \tag{16}$$

If the marginal pdf $g(x)$ of stress X has an exponential pdf., then the product $g(x)f^*(y|x) = g(x, y)$ is regarded as the bivariate FF-exponential pdf of the random vector (X, Y). However, it is not the only possibility. The marginal pdf $g(x)$ of X can belong to any class of pdfs or cdfs, including discrete type. For example, the stress X can be Gaussian. We then obtain *semi FF-exponential* bivariate distribution once we extend the class of the considered models. Now again, admit that there may be more than one stress that affects the life-time Y.

4 Several Covariates Case

In practice we rather should consider a set of few (the most *influential* factors) stresses $X_1, \ldots, X_k$, that all together have an essential impact on, say, a client residual life-time: $Y|x_1, \ldots, x_k$. The stresses $X_1, \ldots, X_k$ must be either stochastically independent or their joint pdf or cdf, should be known. Consider, now, closer the case of *multiple stresses* say, $X_1, \ldots, X_k$, so we need to determine the conditional pdfs $f(y|x_1, \ldots, x_k)$ of the (life-time) Y, given the stresses $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$. But the transition from a single stress to the multiple is, actually, easy and also may be considered as the extension of the multivariate Gaussian distribution paradigm. Therefore, we first refer to this classical concept. In that, the conditional pdf $f(x_{k+1}|x_1, \ldots, x_k)$ of X_{k+1}, k = 1, 2, . . ., given values $x_1, \ldots, x_k$ (with the random vector $(X_1, \ldots, X_k)$ distributed according to the k-variate normal distribution with, in particular, k = 2) is defined as:

$$f_{k+1}(x_{k+1}|x_1, \ldots, x_k) = \frac{1}{s_{k+1}\sqrt{2\pi}} e^{\frac{-(x_{k+1}-\mu_{k+1}-a_1(x_1-\mu_1)-a_2(x_2-\mu_2)-\cdots-a_k(x_k-\mu_k))^2}{2s_{k+1}^2}} \tag{17}$$

Thus, we can achieve our goal (the construction), ones we extend the class of linear functions: $a_1(x_1 - \mu_1) + a_2(x_2 - \mu_2) + +a_k(x_k - \mu_k)$ by a wider class of continuous (nonlinear, in particular quadratic) parameter functions, say: $R_k((x_1 - \mu_1), (x_2 - \mu_2), \ldots, (x_k - \mu_k))$.

The general idea is to *replace* the parameter μ_{k+1} of the *original* normal pdf of X_{k+1} by a linear (for the normal) or nonlinear (for the FF-normal) parameter function

$R_k(\ldots)$ i.e., the procedure obeys the scheme:
$\mu_{k+1} \rightarrow R_k((x_1 - \mu_1), (x_2 - \mu_2), \ldots, (x_k - \mu_k))$.

This scheme corresponds to the considered stochastic impact of the random quantities (the stresses) $X_1, \ldots, X_k$ on the random variable (the life-time) $X_{k+1} = Y$. It is, we have described the stochastic relation: $(X_1, \ldots, X_k) \rightarrow Y$.

Recall that, the *linear case* corresponds to an ordinary normal, while *nonlinear* to the FF-normal. The constructions that use the *method of parameter dependence* turn out to be *universal* to a quite large extend.

Remark 1 In case of any parameter- dependent distribution function of a random variable Y, the parameter, say θ can be considered (after being replaced) as a *parameter-function* $\theta(X_1, \ldots, X_k)$ of k random arguments. The parameter in the form $\theta(X_1, , X_k)$ is, of course, a random variable. Realize, however, that this kind of the parameters randomization is not just (the simple, and very well known) compounding concept. In the compounding concept (for k = 1) we have always $\theta(X_1) = X_1$ so that identity function is the only parameter function applied. Thus, the compounding may only be considered as the very special case of the parameter dependence pattern.

5 The Enforced Regression as the Conditional Probability

There are two basic ways the conditional densities defined in this work may be applied to the concept of *enforced regression*. Firstly, remaining closer to the classical concept of the regression we may, in a new way, enrich that classical models by using the conditional expectation to adopt it to non-Gaussian situations. For example if the r.v. of the main interest Y has either Weibull, Gompertz or gamma distribution its rather unlikely that its conditional expectation would be a linear function the explanatory rvs. $X_1, \ldots, X_k$, whatever they represent. Notice that the baseline expectations are functions of the distribution parameters and the parameters will then become functions (linear, in particular) of $X_1, \ldots, X_k$. Such approach is (according to our best knowledge) not known in current literature. Secondly, we may apply our conditionings to determine, in one general formula, all the conditional probabilities: $P(a \leq Y \leq b | x_1, \ldots, x_k)$ for all: $-\infty \leq a < b \leq \infty$ (and $P(Y = \pm\infty) = 0$) so, in particular, all the conditional quantiles of Y. These conditional probabilities as general functions of realizations $x_1, \ldots, x_k$ of the random variables $X_1, \ldots, X_k$ may be considered as a new version of the stochastic models alternative to the well known models of classical regression. It is expected that they will contain significantly more information than the classical regression functions.

5.1 Conditional Expectations for the Parameter Dependence Pattern

As an example, consider the following three nonGaussian probability distributions: gamma, Weibullian and Gompertz of a random variable, here commonly denoted by Y. We use the following parameterizations of their densities:

1. For the gamma:

$$f(y;\beta,\theta) = y^{\beta-1}e^{\frac{-y}{\theta}}/\theta^{\beta}\Gamma(\beta) \tag{18}$$

2. For the Weibullian:

$$f(y;\beta,\theta) = \theta\beta y^{\beta-1}e^{-\theta y^{\beta}} \tag{19}$$

3. For the Gompertz:

$$f(y;\beta,\theta) = \beta\theta e^{\beta y+\theta}e^{-\theta e^{\beta y}} \tag{20}$$

As in above text, we assume that each time the parameters β and θ continuously depend on realizations $x_1,\ldots,x_k$ of explanatory random variables $X_1,\ldots,X_k$. This assumption turns the above three densities into the conditional densities of Y, given $x_1,\ldots,x_k$ while the class of the densities is invariant, so remains the gamma or Weibullian or Gompertz, respectively, in y. In parallel, the (original) expected values of the three densities turn into the following conditional expectations, given realizations of the $X_1,\ldots,X_k$. According to that rule we define the following *enforced regression functions of first kind*, say $R(x_1,\ldots,x_k)$:

1*. For the gamma:

$$R(x_1,\ldots,x_k) = E[Y|x_1,\ldots,x_k] = \beta(x_1,\ldots,x_k)\theta(x_1,\ldots,x_k). \tag{21}$$

2*. For the Weibull:

$$R(x_1,\ldots,x_k) = E[Y|x_1,\ldots,x_k] = \theta(x_1,\ldots,x_k)\Gamma\left(1+\frac{1}{\beta(x_1,\ldots,x_k)}\right). \tag{22}$$

3*. For the Gompertz:

$$R(x_1,\ldots,x_k) = E[Y|x_1,\ldots,x_k] = \left(\frac{1}{\beta(x_1,\ldots,x_k)}\right)e^{\theta(x_1,\ldots,x_k)}Ei(-\theta(x_1,\ldots,x_k)) \tag{23}$$

where $Ei()$ is known to be $Ei(z) = \int\limits_{-z}^{\infty} v^{-1}e^{-v}dv$.

In each of the above cases, a choice of the parameter functions $\beta(x_1, \dots, x_k)$, $\theta(x_1, \dots, x_k)$ theoretically is arbitrary but in applications would be dictated by the fit to data criterion. More specifically, for cases 1* and 2* it is reasonable to chose

$$\beta(x_1, \dots, x_k) = \beta(1 + \varphi(x_1, \dots, x_k)) \text{ and } \theta(x_1, \dots, x_k) = \theta(1 + \mu(x_1, \dots, x_k))$$

where the values β and θ of the baseline expectations parameters on the right hand site of above equalities (understood there as the *old parameters*) could be assumed to be known.

Here, for example, both the functions $\varphi(x_1, \dots, x_k)$ and $\mu(x_1, \dots, x_k)$ can be chosen linear or quadratic forms in $x_1, \dots, x_k$ or exponent or logarithm of such forms. Obviously, other choices are permitted too. In case of Gompertz conditional expectation 3*, requirement that $\theta(x_1, \dots, x_k)$ is an independent of $x_1, \dots, x_k$ constant, say θ^* would be reasonable for sake of the simplicity.

5.2 Conditional Survival Probabilities

The conditional expectations, as considered above, still bring much less information about a modeled reality than the conditional probabilities of events that one may be most interested with. For example, in a medical trial doctors may be interested in a probability that a patient will survive the next five years time period. More precisely, suppose that a given fixed age person is diagnosed with a given kind of cancer. If the person is not a smoker and will be given a specific treatment the probability she will survive at nearest five years is known to be p_0. For *the same* person who has been smoking tobacco it is less than p_0 but depends on amount x of smoking. This amount can be measured, for example, as arithmetic product of length of smoking time period and an intensity of smoking (number of milligrams of nicotine per day per one kilo of weigh). Our expectation is to find the probability p of, say, five years survival as a continuous function of the smoking level x: $p = p(x)$ with $p(0) = p_0$.

Besides of smoking, the patient could endure other stresses such as drinking alcohol, excessive consuming sugar, time spending in prison etc. Since (especially when **actuary** problems are involved) more often than with a single patient the doctors researchers have to do with a whole population of them, we propose to consider the stresses as random variables $X_1, X_2, \dots, X_k$. Any (measured) realizations $x_1, x_2, \dots, x_k$ of these random variables are assumed to uniquely determine the conditional probability of five years survival, given these (stresses) realizations: $P(Y \geq 5 | x_1, x_2, \dots, x_k)$ which is a continuous function of that arguments $x_1, x_2, \dots, x_k$.

As it was mentioned, similar theory under the name *quantile regression methods* is known in literature and was initiated in [6], where the probabilities $P(Y \leq y_0 | x_1, x_2, \dots, x_k)$ were investigated for a fixed y_0 by nonparametric methods.

Unlike, what we present here is the parametric approach (assuming that the considered classes of the parameter functions $\beta(x_1, \dots, x_k)$ and $\theta(x_1, \dots, x_k)$ are parametric such as the linear, the power, the exponential etc., whose the parameters are to be estimated) where we aim to find **analytic formulas** valid for any (variable) value **y** instead of a single constant value y_0, only.

Basically, in what follows, we limit ourselves to the conditional *survival probabilities*: $P(Y \geq y|x_1, x_2, \ldots, x_k) = r(y; x_1, x_2, \ldots, x_k)$. These probabilities as functions of y and of the realizations $x_1, x_2, \ldots, x_k$ we called the *enforced regression functions of second kind.*

For the three probability distributions considered in Sect. 5.1 one obtains the following three formulas as survival functions of the arguments $y, x_1, x_2, \ldots, x_k$:

1.** For the gamma:

$$\begin{aligned}&P(Y \geq y|x_1, x_2, \ldots, x_k)\\&\quad = 1 - \gamma[\beta(x_1, x_2, \ldots, x_k), y/\theta(x_1, x_2, \ldots, x_k)]/\Gamma[\beta(x_1, x_2, \ldots, x_k)],\end{aligned} \tag{24}$$

where $\gamma[s, x] = \int\limits_0^x t^{s-1}e^{-t}dt$ is the lower incomplete gamma function and meaning of the gamma density's parameters $\beta(), \theta()$ is given by (18).

2.** For the Weibull:

$$P(Y \geq y|x_1, x_2, \ldots, x_k) = e^{-\theta(x_1,x_2,\ldots,x_k)y^{\beta(x_1,x_2,\ldots,x_k)}}, \tag{25}$$

where the meaning of the Weibull density's parameters $\beta(), \theta()$ is given by (19).

3.** For the Gompertz:

$$P(Y \geq y|x_1, x_2, \ldots, x_k) = e^{-\theta(x_1,x_2,\ldots,x_k)e^{\beta(x_1,x_2,\ldots,x_k)y}-1}. \tag{26}$$

Here, the meaning of the parameters is given by (20).

Recall, that the Gompertz distribution is commonly applied in modeling of residual life time of adults for actuary and demographic purposes while the Weibull model is typically of use in reliability of technical devices investigations. Also the gamma is pretty often used in reliability. Notice, that for Gaussian and exponential distributions all the formulas from Sect. 5 become almost trivial so we omitted them.

With exception of 1* the above *enforced regression functions* as the functions of $x_1, x_2, \ldots, x_k$ may, in general, be pretty complicated. This, of course, brings limitations on (parametric) classes of the corresponding parameter functions $\beta(x_1, x_2, \ldots, x_k)$ and $\theta(x_1, x_2, \ldots, x_k)$. Between others, number k of the explanatory variables should not be too large (say, no more than 4 or 5). To simplify underlying statistical investigations they should be proceeded by some numerical analysis. Depending on how important is an expected gain of the above stochastic models precision (of the conditional probabilities predictions) we may afford for less or more complexity of the chosen parameter functions. Also more complex models will require more data (higher sample sizes). To apply the above defined *enforced* *regression* successfully many of numerical and statistical problems could be involved requiring proper solutions. This part of work is, however, beyond the scope of this paper.

References

1. Filus, J.K., Filus, L.Z.: A method for multivariate probability distributions construction via parameter dependence. Commun. Stat. Theory Methods **42**(4,15), 716–721 (2013)
2. Filus, J.K., Filus, L.Z.: Multivariate pseudodistributions as natural extension of the multivariate normal density pattern theory. In: American Institute of Physics Conference Proceedings 1479, Numerical Analysis and Applied Mathematics, vol. 1479, pp. 1417–1420 (2012)
3. Filus, J.K., Filus, L.Z.: On some new classes of multivariate probability distributions. Pak. J. Stat. **22**, 21–42 (2006)
4. Filus, J.K., Filus, L.Z.: On new multivariate probability distributions and stochastic processes with systems reliability and maintenance applications. Methodol. Comput. Appl. Probab. J. **9**, 426–446 (2007)
5. Filus, J.K., Filus, L.Z.: Weak stochastic dependence in biomedical applications. In: American Institute of Physics Conference Proceedings 1281, Numerical Analysis and Applied Mathematics, vol. III, pp. 1873–1876 (2010)
6. Koenker, R., Bassett, G.W.: Regression quantiles. Econometrica **46**, 33–50 (1978)
7. Filus, J.K., Filus, L.Z.: A class of generalized multivariate normal densities. Pak. J. Stat. **16**, 11–32 (2000)
8. Filus, J.K., Filus, L.Z., Arnold, B.C.: Families of multivariate distributions involving triangular transformations. Commun. Stat. Theory Methods **39**(1), 107–116 (2010)
9. Filus, J.K., Filus, L.Z.: Some alternative approaches to system reliability modeling. In: Pham, H. (ed.) Recent Advances in Reliability and Quality Design. Springer Series in Reliability Engineering, pp. 101–135 (2008)
10. Kotz, S., Balakrishnan, N., Johnson, N.L.: Continuous Multivariate Distributions, vol. 1, 2nd edn. Wiley, New York (2000)