Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

Paulo Eduardo Oliveira
Maria da Graça Temido
Carla Henriques
Maurizio Vichi *Editors*

Recent Developments in Modeling and Applications in Statistics

Springer

Studies in Theoretical and Applied Statistics
Selected Papers of the Statistical Societies

For further volumes:
http://www.springer.com/series/10104

Series Editors

Spanish Society of Statistics and Operations Research (SEIO)

Société Française de Statistique (SFdS)

Società Italiana di Statistica (SIS)

Sociedade Portuguesa de Estatística (SPE)

Paulo Eduardo Oliveira • Maria da Graça Temido
Carla Henriques • Maurizio Vichi
Editors

Recent Developments in Modeling and Applications in Statistics

Springer

Editors
Paulo Eduardo Oliveira
Maria da Graça Temido
CMUC, Mathematics
University of Coimbra
Coimbra
Portugal

Carla Henriques
CMUC, Escola Superior de Tecnologia
e Gestão
Instituto Politécnico de Viseu
Viseu
Portugal

Maurizio Vichi
Dipartimento di Scienze Statistiche
Sapienza Università di Roma
Rome
Italy

ISSN 2194-7767 ISSN 2194-7775 (electronic)
ISBN 978-3-642-32418-5 ISBN 978-3-642-32419-2 (eBook)
DOI 10.1007/978-3-642-32419-2
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012950594

© Springer-Verlag Berlin Heidelberg 2013
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Foreword

Dear reader, on behalf of the four Scientific Statistical Societies: SEIO, Sociedad de Estadística e Investigación Operativa (Spanish Statistical Society and Operation Research); SFdS, Société Française de Statistique (French Statistical Society); SIS, Società Italiana di Statistica (Italian Statistical Society); SPE, Sociedade Portuguesa de Estatística (Portuguese Statistical Society), we inform you that this is a new book series of Springer entitled: "Studies in Theoretical and Applied Statistics," with two lines of books published in the series: "Advanced Studies"; " Selected Papers of the Statistical Societies."

The first line of books offers constant up-to-date information on the most recent developments and methods in the fields of theoretical statistics, applied statistics, and demography. Books in this series are solicited in constant cooperation among statistical societies and need to show a high-level authorship formed by a team preferably from different groups to integrate different research points of view.

The second line of books proposes a fully peer reviewed selection of papers on specific relevant topics organized by editors, also in occasion of conferences, to show their research directions and developments in important topics, quickly and informally, but with a high quality. The explicit aim is to summarize and communicate current knowledge in an accessible way. This line of books will not include proceedings of conferences and wishes to become a premier communication medium in the scientific statistical community by obtaining the impact factor, as it is the case of other book series such as "Lecture Notes in Mathematics."

The volumes of selected papers of the statistical societies will cover a broad scope of theoretical, methodological as well as application-oriented articles, surveys and discussions. A major purpose is to show the intimate interplay between various, seemingly unrelated domains and to foster the cooperation among scientists in different fields by offering well-based and innovative solutions to urgent problems of practice.

On behalf of the founding statistical societies I wish to thank Springer, Heidelberg and in particular Dr. Martina Bihn for the help and constant cooperation in the organization of this new and innovative book series.

Rome, Italy
June 2012

Maurizio Vichi

Preface

Statistics has been turning into a main tool in almost every field of activity, either purely theoretical or as an essential part of scientific applied work, supporting conclusions and giving insight into new possibilities of usage of methods and results, thus being an essential tool helping in *looking for faceless facts*. The construction of models describing populations or phenomena including random behavior uses a wide range of methods. The recollection of data is at the basis of this modeling and verification of assumptions, thus sampling methodologies appear naturally as an essential object of interest in statistical applications. The modeling process must then be conducted using the suitable techniques giving the researcher the means to search for hidden facts and behaviors. This modeling may be addressed by fitting well predefined shapes and distributions to the data or leaving the data to unveil its intrinsic properties by using the usually more computational demanding, nonparametric methods. If these imply the usage of larger data sets, the parametric description redirects the work toward estimation and hypothesis testing, which can help detecting inconvenient initial choices for the modeling effort.

Reflecting the above-mentioned problems, this volume is organized into six chapters, each one dedicated different approaches to the description and modeling process. The first chapter will include a few contributions on sampling and modeling concerned with direct description of the population under study, possibly using different statistically based methodologies. The second chapter includes contributions concerned with estimation problems, deferring to later chapters a few papers concerned with estimation problems in some more particular frameworks or with testing hypothesis. The final chapter includes contributions of more general flavor.

This volume contains peer reviewed selected contributions presented at the XVIII Annual Congress of the Portuguese Statistical Society that took place at S. Pedro do Sul from September 29 to October 2, 2010, bringing together not only both theoretical and applied statisticians mainly working in Portugal but also a number of Brazilian and Spanish specialists, a few invited well-known specialists and a minicourse opening the conference. The main emphasis on the call for contributions, selection of plenary talks, and minicourse was put on estimation with a particular interest on nonparametric methods and also on the usage of sampling procedures and methodologies, due to the approach of the Portuguese Census. The XVIII Annual Congress of the Portuguese Statistical Society was attended by more than 170 participants, including those who participated in a special session organized

by the Italian Statistical Society as a consequence of a cooperation between several national societies, which also produced the interest on publishing this volume. These participants presented 81 contributed sessions, scattered by 29 sessions and 42 poster contributions.

The editorial committee wishes to express its deep gratitude to everyone who contributed their work to the present volume. The editorial committee is largely indebted to the anonymous referees who kindly helped us on the evaluation process that preceded the final preparation of this volume.

Viseu, Portugal — Carla Henriques
Coimbra, Portugal — Paulo Eduardo Oliveira
Rome, Italy — Maria da Graça Temido
June 2012 — Maurizio Vichi

Contents

Contributors

M. Isabel Fraga Alves Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal, isabel.alves@fc.ul.pt

Nelson Antunes CEMAT and FCT of University of Algarve, Portugal, nantunes@ualg.pt

Fernando Bação ISEGI, New University of Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal, bacao@isegi.unl.pt

Carlos A. Braumann Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal, braumann@uevora.pt

M. F. Brilhante Universidade dos Açores (DM) and CEAUL, Campus de Ponta Delgada, Apartado 1422, 9501-801 Ponta Delgada, Portugal, fbrilhante@uac.pt

Nuno M. Brites Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal, d7488@alunos.uevora.pt

Frederico Caeiro CMA and Faculdade de Ciências e Tecnologia da UNL, 2829–516 Caparica, Portugal, fac@fct.unl.pt

Pedro Campos LIAAD – INESC Porto, L.A and Faculdade de Economia da Universidade do Porto, LIAAD: R. Ceuta, 118, 4050, Porto, Portugal, pcampos@fep.up.pt

Helena Cardoso ICBAS University of Porto and HGSA Hospital Center of Porto, Portugal, helenacardoso@icbas.up.pt

Alda Carvalho ISEL and CEMAPRE, Rua Conselheiro Emídio Navarro, 1, 1959-007, Lisboa, Portugal, acarvalho@adm.isel.pt

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

Carlos A. Coelho Departamento de Matemática and Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre 2829-516, Caparica, Portugal, cmac@fct.unl.pt

Álvaro Cunha CEC and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal, acunha@fe.up.pt

José Carlos Dias BRU-UNIDE and ISCTE-IUL Business School, Av. Prof. Aníbal Bettencourt, 1600-189 Lisboa, Portugal, jose.carlos.dias@iscte.pt

Sandra Dias Departamento de Matemática, Ed. de Ciências Florestais, Apartado 1013, 5001-801 Vila Real, Portugal, sdias@utad.pt

Patrícia A. Filipe Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal, pasf@uevora.pt

João Figueiredo Faculdade de Economia do Porto, R. Roberto Frias, 4200 Porto, Portugal, jfabelha@gmail.com

M. Ivette Gomes CEAUL and FCUL, Campo Grande, 1749-016, Lisboa, Portugal, ivette.gomes@fc.ul.pt

Esmeralda Gonçalves CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal, esmerald@mat.uc.pt

Gonçalo Jacinto DMAT/CIMA-UE and ECT of University of Évora, Portugal, gjcj@uevora.pt

Manuela Larguinho Department of Mathematics, ISCAC, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal, mlarguinho@iscac.pt

Anthony Ledford AHL Research, Man Research Laboratory, Eagle House, Walton Well Road, Oxford, OX2 6ED, United Kingdom, aledford@ahl.com

Joana Leite Institute of Accounting and Administration of Coimbra, Quinta Agrícola, Bencanta, 3040-316, Coimbra, Portugal, jleite@iscac.pt

Victor Lobo ISEGI, New University of Lisbon, and CINAV - Portuguese Navy Research Center, and Escola Naval - Alfeite, 2810-001 Almada, Portugal, vlobo@isegi.unl.pt

Enno Mammen Department of Economics, University of Mannheim, D-68131 Mannheim, Germany, emammen@rumms.uni-mannheim.de

Catarina Marques Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Av. das Forças Armadas, 1649-026 Lisbon, Portugal, catarina.marques@iscte.pt

Filipe J. Marques Departamento de Matemática and Centro de Matemática e Aplicações, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre 2829-516, Caparica, Portugal, fjm@fct.unl.pt

Paula Marques Instituto Superior Dom Dinis, Av. 1.º Maio, 2430, Marinha Grande, Portugal, paulafartaria@gmail.com

João Paulo Martins School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and Applications of University of Lisbon, Portugal, jpmartins@estg.ipleiria.pt

Nazaré Mendes-Lopes CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal, nazare@mat.uc.pt

Rute Meneses U.Fernando Pessoa of Porto, Portugal, rmeneses@ufp.edu.pt

João Menezes Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Av. das Forças Armadas, 1649-026 Lisbon, Portugal, joao.menezes@iscte.pt

Jorge Mendes CEAUL, FCUL, University of Lisbon, and ISEGI, New University of Lisbon and Campus de Campolide, 1070-312 Lisboa, Portugal, jmm@isegi.unl.pt

Miguel Mendes FEUP and CMUP, Universidade do Porto, Portugal, mmendes@fc.up.pt

Denisa Mendonça ICBAS University of Porto and ISPUP, Institute of Public Health, University of Porto, Portugal, dvmendon@icbas.up.pt

Manuel Cabral Morais CEMAT and Mathematics Department, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal, mjmorais@math.ist.utl.pt

Isabella Morlini Department of Economics, University of Modena & Reggio Emilia, Italy, isabella.morlini@unimore.it

Paula Milheiro–Oliveira CMUP and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal, poliv@fe.up.pt

António Pacheco CEMAT and Mathematics Department, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal, apacheco@math.ist.utl.pt

Luísa Pedro IPL of Lisbon, Portugal, luisa.pedro@estesl.ipl.pt

Dinis Pestana Faculty of Sciences of Lisbon, University of Lisbon, CEAUL — Center of Statistics and Applications of University of Lisbon, Portugal, dinis.pestana@fc.ul.pt

Alexandra Pinto Laboratory of Biomathematics, Faculty of Medicine of Lisbon, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal, apinto@fm.ul.pt

Alexandra Ramos FEP and CMUP, Universidade do Porto, Rua Dr Roberto Frias, 4200-464-Porto, Portugal, aramos@fep.up.pt

Patrícia Ferreira Ramos CEMAT, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal, patriciaferreira@ist.utl.pt

Elizabeth Reis Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Av. das Forças Armadas, 1649-026 Lisbon, Portugal, ear@iscte.pt

José Luís Pais Ribeiro FPCE, University of Porto and UIPES, Portugal, jlpr@fpce.up.pt

Lígia Henriques-Rodrigues CEAUL and Instituto Politécnico de Tomar, Portugal, Ligia.Rodrigues@aim.estt.ipt.pt

Teresa Rodrigues Laboratory of Biomathematics, Faculty of Medicine of Lisbon, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal, teresarodrigues@fm.ul.pt

Carlos J. Roquete Universidade de Évora, Instituto de Ciências Agrárias e Ambientais Mediterrânicas, Núcleo da Mitra, Apartado 94, 7002-774 Évora, Portugal, croquete@uevora.pt

Christoph Rothe Toulouse School of Economics, 21 Allée de Brienne, F-31000 Toulouse, France, rothe@cict.fr

Maria de Fátima Salgueiro Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal, fatima.salgueiro@iscte.pt

Carlos Santos ISEC, Alameda das Linhas de Torres, 179, 1750-142, Lisboa, Portugal, carlos.santos@isec.universitas.pt

Paulo Araújo Santos Departamento de Informática e Métodos Quantitativos, Escola Superior de Gestão e Tecnologia, Instituto Politécnico de Santarém, Complexo Andaluz, Apartado 295, 2001-904 Santarém, Portugal, paulo.santos@esg.ipsantarem.pt

Rui Santos School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and Applications of University of Lisbon, Portugal, rui.santos@ipleiria.pt

Melanie Schienle School of Business and Economics, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany, melanie.schienle@wiwi.hu-berlin.de

Wolfgang Schmid Department of Statistics, European University Viadrina, Grosse Scharrnstrasse 59,15230 Frankfurt (Oder), Germany, schmid@euv-frankfurt-o.de

Patrícia Serra Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Lisboa, Portugal, patricia.serra@iscte.pt

António Martins da Silva ICBAS University of Porto and HGSA Hospital Center of Porto, Portugal, ams@icbas.up.pt

Isabel Silva U.Fernando Pessoa of Porto, Portugal, isabels@ufp.pt

Maria da Graça Temido CMUC and DMUC, University of Coimbra, 3001-454 Coimbra, Portugal, mgtm@mat.uc.pt

Paula C. R. Vicente Instituto Universitário de Lisboa (ISCTE-IUL), BRU-IUL, Portugal, pvicente@netcabo.pt

Pedro Vieira University of Trás–os–Montes e Alto Douro, Quinta de Prados, Vila Real, Portugal and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal, pmvieira@fe.up.pt

Estela Vilhena Polytechnic Institute of Cávado and Ave and ICBAS University of Porto, Portugal, evilhena@ipca.pt

Part I

Sampling and Modeling

Using Latent Variables in Model Based Clustering: An E-Government Application

Isabella Morlini

Abstract

Besides continuous variables, binary indicators on ICT (Information and Communication Technologies) infrastructures and utilities are usually collected in order to evaluate the quality of a public company and to define the policy priorities. In this chapter, we confront the problem of clustering public organizations with model-based clustering, and we assume each observed binary indicator to be generated from a latent continuous variable. The estimates of the scores of these variables allow us to use a fully Gaussian mixture model for classification.

1 Introduction

In order to handle mixed continuous and binary variables for classification purposes, in this work we assume each observed categorical variable to be generated from a latent continuous variable. For estimating the scores of these latent variables, we use the method proposed in Morlini [3]. In economics, the latent variables may be interpreted as utility functions. The assumption is that the responses (e.g., the presence or the absence of a public or private service) are determined by the crossing of certain thresholds in these functions. The advantages of using the scores of the latent variables in place of the original categories and then specifying a full Gaussian model are threefold. (1) Classification is possible also for a large number of variables, while most of the models currently used for variables with mixed scale types [2, 5] are feasible only with few categorical indicators. (2) Many forms of restrictions can be imposed on the variances and the covariances, to obtain parsimony. (3) Local dependencies can be specified not only for couples of

I. Morlini (✉)
Department of Economics, University of Modena & Reggio Emilia, Italy
e-mail: isabella.morlini@unimore.it

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_1, © Springer-Verlag Berlin Heidelberg 2013

continuous variables, but also for couples of (original) categorical variables and for a continuous and a categorical variable.

As shown by Vermunt and Magidson [5], the possibility to include local dependencies among the indicators may prevent the possibility of ending with a solution that has too many clusters since, often, a simpler solution with less groups may be obtained by including some direct effects between the indicators. Moreover, relaxing the local independence assumption may yield a better classification of objects since omitting a significant bivariate dependency from a latent class cluster model leads to too high weights of the indicators in the classification.

In this work, we propose an e-government application with the data collected for the UNDERSTAND project (European Regions UNDER way towards STANDard indicators for benchmarking information society) of the Emilia–Romagna Region (Italy). The data consist of a set of categorical indicators and continuous variables on ICT, comparable at European level. The chapter is organized as follows. In Sect. 2, we summarize the method used for estimating the latent variables scores. In Sect. 3 we briefly describe the data set and we report results on the application.

2 Estimation of the Latent Variables Scores

Let consider a general set up in which the values of p binary attributes and q quantitative variables are collected for n objects and let $g = p + q$. We indicate with x_k $(k = 1, \ldots, p)$ the p binary attributes and with y_j $(j = 1, \ldots, q)$ the q quantitative variables. For each object i $(i = 1, \ldots, n)$, the p-dimensional vector $\mathbf{x}_i = [x_{i1} \ldots x_{ip}]$ contains the values of the binary attributes and the q-dimensional vector $\mathbf{y}_i = [y_{i1} \ldots y_{iq}]$ contains the values of the quantitative variables. We suppose that the binary values are generated from latent continuous variables ξ_k $(k = 1, \ldots, p)$, and we obtain a new $(n \times g)$ matrix of quantitative variables by estimating the score ξ_{ik} for each object i $(i = 1, \ldots, n)$ and each latent variable k $(k = 1, \ldots, p)$. The score ξ_{ik} is associated with the observed categorical value x_{ik} as follows: $x_{ik} = 1$ if $\xi_{ik} \geq T_k$ and $x_{ik} = 0$ if $\xi_{ik} < T_k$, where T_k is the threshold, obtained from the data, for the k-th latent variable. The method includes the following consecutive steps:

1. Estimate the threshold T_k of each latent variable and the tetrachoric correlation coefficient r_{kl} between each pair $\{k, l\}$ of latent variables.
2. Perform a principal component analysis on the matrix of the tetrachoric correlations and obtain the eigenvectors and the eigenvalues.
3. Estimate the score of each principal component for each object, given the eigenvectors and the eigenvalues.
4. Estimate the score of each latent variable for each object, given the scores of the principal components.

We construct a contingency table for each pair of variables x_l and x_k $(l, k = 1, \ldots, p)$, with the following cell frequencies: The estimated value for the threshold generating the variable x_l is the value T_l satisfying $\Phi(T_l) = (a_{kl} + c_{kl})/n$. For

	$x_l = 0$	$x_l = 1$	Tot.
$x_k = 0$	a_{kl}	b_{kl}	$a_{kl} + b_{kl}$
$x_k = 1$	c_{kl}	d_{kl}	$c_{kl} + d_{kl}$
Tot.	$a_{kl} + c_{kl}$	$b_{kl} + d_{kl}$	n

variable x_k, it is the value T_k satisfying $\Phi(T_k) = (a_{kl} + b_{kl})/n$, where Φ is the standard normal cumulative distribution function. We then estimate the matrix of tetrachoric correlations $\mathbf{R} = (r_{kl})$ $(k, l = 1, \ldots, p)$ conditional on the thresholds T_l and T_k, via maximum likelihood. The tetrachoric correlation, introduced by Pearson [4], is the correlation coefficient r_{kl} that satisfies

$$\frac{d_{kl}}{n} = \int_{T_l}^{\infty} \int_{T_k}^{\infty} \phi(\xi_k, \xi_l, r_{kl})\, \mathrm{d}\xi_k \mathrm{d}\xi_l, \tag{1}$$

where $\phi(\xi_k, \xi_l, r_{kl})$ is the bivariate normal density function:

$$\phi(\xi_k, \xi_l, r_{kl}) = \frac{1}{2\pi\sqrt{1 - r_{kl}^2}} \exp\left[-\frac{1}{2(1 - r_{kl}^2)}(\xi_k^2 - 2r_{kl}\xi_k\xi_l + \xi_l^2)\right]. \tag{2}$$

The solution may be found iteratively or by using one of the analytic formula proposed in the seminal work of Pearson [4]. Since the thresholds and the tetrachoric correlation coefficient are identifiable if no frequency in the contingency table is equal to zero, we replace the zero by one half. We perform a principal component analysis on the matrix $\mathbf{R}$ and consider the following model:

$$t_{ih} = \alpha_{h1}\xi_{i1} + \alpha_{h2}\xi_{i2} + \ldots + \alpha_{hk}\xi_{ik} + \ldots + \alpha_{hp}\xi_{ip}, \tag{3}$$

where t_{ih} $(h = 1, \ldots, p, \quad i = 1, \ldots, n)$ is the score of the h-th principal component t_h for object i, α_{hk} $(k = 1, \ldots, p)$ are the loadings, with $\sum_{h=1}^{p} \alpha_{hk}^2 = 1$, and ξ_{ik} is the score for object i relative to the k-th latent variable. $\mathbf{t} \sim N(\mathbf{0}, \Lambda)$ where Λ is a diagonal matrix with elements $\lambda_h^2 = \sum_{k=1}^{p} \alpha_{hk}^2$, since the principal components are orthogonal. The variance of each component t_h and the coefficients α_{hk} $(h = 1, \ldots, p, \quad k = 1, \ldots, p)$ are estimated through the eigenvalues and the eigenvectors, respectively, of the matrix $\mathbf{R}$, without making any assumption about the distribution of the latent variables ξ_k. Given these values, we estimate the score of the principal components by expected a posteriori (EAP) estimates. This analysis does not require previous smoothing if the matrix is not positive definite. However, for the identifiability of the score estimates, all eigenvalues must be positive and a smoothing procedure is required if the matrix is positively semi-definite but not definite. We use the procedure implemented in Matlab, which adds a regularization term to the matrix. Different regularization terms lead to slightly different solutions. The EAP estimator of the h-th principal component score is the mean of the posterior distribution of t_h, which is expressed by:

$$E(t_h|\mathbf{x}_i,\mathbf{w}) = \int t_h f(t_h|\mathbf{x}_i,\mathbf{w})\mathrm{d}t_h = \frac{\int t_h f(\mathbf{x}_i|t_h,\mathbf{w})\phi(t_h|\mathbf{w})\mathrm{d}t_h}{\int f(\mathbf{x}_i|t_h,\mathbf{w})\phi(t_h|\mathbf{w})\mathrm{d}t_h}, \qquad (4)$$

where $f(\cdot)$ indicates the probability density function, $\mathbf{w}$ is the vector of known parameters (the thresholds and the eigenvectors, estimated geometrically by the principal component analysis on $\mathbf{R}$), and ϕ is the Gaussian distribution. In the following equations, for parsimony, $\mathbf{w}$ will be omitted. For every object i ($i = 1,\ldots,n$), the probability of the k-th binary attribute to be equal to 1, given the h-th principal component score, can be formalized as follows:

$$P(x_{ik}=1|t_h) = \frac{1}{\sigma_{hk}\sqrt{2\pi}} \int_{T_k}^{\infty} \mathrm{e}^{-\frac{(t_{ih}-\alpha_{hk}\xi_k)^2}{2\sigma_{hk}^2}} \mathrm{d}\xi_k. \qquad (5)$$

where $\sigma_{hk}^2 = \lambda_h^2 - \alpha_{hk}^2 = \sum_{l\neq k}\alpha_{hl}^2$. Introducing the change in the variable:

$$P(x_{ik}=1|t_h) = \frac{1}{\alpha_{hk}\sqrt{2\pi}} \int_{-\infty}^{\frac{t_{ih}-\alpha_{hk}T_k}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}} \mathrm{d}z, \qquad \text{when} \qquad \alpha_{hk} > 0, \quad (6)$$

$$P(x_{ik}=1|t_h) = \frac{1}{-\alpha_{hk}\sqrt{2\pi}} \int_{\frac{t_{ih}-\alpha_{hk}T_k}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}} \mathrm{d}z, \qquad \text{when} \qquad \alpha_{hk} < 0. \quad (7)$$

Letting $z_{hk} = (t_{ih} - \alpha_{hk}T_k)/\sigma_{hk}$, we may define the following quantities:

$$\begin{aligned} F_{hk}(t_h) &= |\alpha_{hk}|^{-1}\Phi(z_{hk}), && \text{when} \quad \alpha_{hk} > 0 \quad \text{and} \quad x_{ik} = 1 \\ &\quad \text{or} \quad \alpha_{hk} < 0 && \text{and} \quad x_{ik} = 0, \\ F_{hk}(t_h) &= |\alpha_{hk}|^{-1}[1-\Phi(z_{hk})], && \text{when} \quad \alpha_{hk} < 0 \quad \text{and} \quad x_{ik} = 1 \\ &\quad \text{or} \quad \alpha_{hk} > 0 && \text{and} \quad x_{ik} = 0, \end{aligned}$$

where Φ is the standard normal cumulative function. Assuming the independence of the binary attributes x_k ($k = 1,\ldots,p$) conditionally on each component t_h, we obtain $f(\mathbf{x}_i|t_h) = \prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1-F_{hk}(t_h)]^{1-x_{ik}}$. This assumption may be thought of as rather unrealistic, since at least one latent variable generating a binary attribute is dependent from the other latent variables. In fact, formally this is a weak point of our procedure, which allows for simple and fast computation. Considering S quadrature points, we estimate the scores as follows:

$$\tilde{t}_{ih} = \sum_{s=1}^{S} t_{sh}^q \frac{\phi(t_{sh})\prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1-F_{hk}(t_h)]^{1-x_{ik}}}{\sum_{s=1}^{S}\phi(t_{sh})\prod_{k=1}^{p} F_{hk}(t_h)^{x_{ik}}[1-F_{hk}(t_h)]^{1-x_{ik}}}, \qquad (8)$$

where t_{sh}^q are equally spaced points in $[-z_h,\, z_h]$ with $\Phi(-z_h/\lambda_h) = 0.001$, and $\phi(t_{sh}^q)$ are the density functions of these points in the $N(0,\lambda_h^2)$ curve times the interval size.

Given the estimates $\tilde{t}_{ih}$, the EAP estimates $\tilde{\xi}_{ik}$ of the latent variables may be reached through analogous steps. The EAP estimator of the k-th variable score is the mean of the posterior distribution of ξ_k, which is expressed by $E(\xi_k|x_k) = \int \xi_k f(\xi_k|x_k)\mathrm{d}\xi_k$. Let $\xi_k^+ = \xi_k$ if $\xi_k \geq T_k$ and $\xi_k^- = \xi_k$ if $\xi_k < T_k$. Then:

$$
\begin{aligned}
f(\xi_k|x_{ik}=1,\tilde{t}_{ih}) &= \frac{1}{\alpha_{hk}\sqrt{2\pi}} \int_{-\infty}^{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^+}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, && \text{if } \alpha_{hk} > 0,\\
f(\xi_k|x_{ik}=1,\tilde{t}_{ih}) &= \frac{1}{|\alpha_{hk}|\sqrt{2\pi}} \int_{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^+}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, && \text{if } \alpha_{hk} < 0,\\
f(\xi_k|x_{ik}=0,\tilde{t}_{ih}) &= \frac{1}{|\alpha_{hk}|\sqrt{2\pi}} \int_{-\infty}^{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^-}{\sigma_{hk}}} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, && \text{if } \alpha_{hk} < 0,\\
f(\xi_k|x_{ik}=0,\tilde{t}_{ih}) &= \frac{1}{\alpha_{hk}\sqrt{2\pi}} \int_{\frac{\tilde{t}_{ih}-\alpha_{hk}\xi_k^-}{\sigma_{hk}}}^{\infty} \mathrm{e}^{\frac{-z^2}{2}}\mathrm{d}z, && \text{if } \alpha_{hk} > 0.
\end{aligned}
\tag{9}
$$

Let $z_{hk}^+ = \frac{\tilde{t}_{ih}-a_{hk}\xi_{ik}^+}{\sigma_{hk}}$, $z_{hk}^- = \frac{\tilde{t}_{ih}-a_{hk}\xi_{ik}^-}{\sigma_{hk}}$,

$$
\begin{aligned}
F_{hk}^+(\xi_k) &= (\alpha_{hk})^{-1}\Phi(z_{hk}^+), && \text{when } \alpha_{hk} > 0,\\
F_{hk}^+(\xi_k) &= |\alpha_{hk}|^{-1}(1-\Phi(z_{hk}^+)), && \text{when } \alpha_{hk} < 0,\\
F_{hk}^-(\xi_k) &= |\alpha_{hk}|^{-1}\Phi(z_{hk}^-), && \text{when } \alpha_{hk} < 0,\\
F_{hk}^-(\xi_k) &= (\alpha_{hk})^{-1}(1-\Phi(z_{hk}^-)), && \text{when } \alpha_{hk} > 0.
\end{aligned}
\tag{10}
$$

Then $f(\xi_k|x_k) = \sum_{h=1}^{p} F_{hk}^+(\xi_k)^{x_{ik}} F_{hk}^-(\xi_k)^{1-x_{ik}} \times \phi(\tilde{t}_{ih})$. Considering S quadrature points, we estimate the scores as follows:

$$
\tilde{\xi}_{ik} = \sum_{s=1}^{S} \xi_{sk}^q \phi(\xi_{sk}) \sum_{h=1}^{p} (F_{hk}^+(\xi_s)^{x_{ik}} F_{hk}^-(\xi_s)^{1-x_{ik}} \times \phi(\tilde{t}_{ih})), \tag{11}
$$

where ξ_{sk}^q are equally spaced points in $[-z_k, T_k]$ when $x_{ik} = 0$, in $[T_k, z_k]$ when $x_{ik} = 1$, with $\Phi(-z_k) = 0.001$, $\phi(\xi_{sk}^q)$ being the density functions of these points in the $N(0,1)$ curve times the interval size.

3 An E-Government Application

In this section, we present a cluster analysis of the Emilia–Romagna municipalities, based on a set of back office and front office indicators. The indicators aim at establishing to what extent e-government is working within the region. The data

have been collected using an online questionnaire, a printed questionnaire sent by post, filled in face to face or obtained over the phone. The aim of this study is to obtain an insight into how municipalities are affected by ICTs and ICT-enabled developments. ICTs have opened up new possibilities for municipalities to overcome traditional disadvantages deriving from remoteness and distance. But instead of increasing the quality of service everywhere, they have been shown to exacerbate disparities. This is due to the difference in the speed and intensity of the adoption of ICTs, and also to the degree that these technological innovations are utilized. Regional investment in infrastructure related to the Information Society have increased over the past years, and regional decision makers are increasingly committed to the development of ICT in society. As a consequence, policy-makers need to be able to identify areas in which public investments and political support are most likely to be successful.

In order to cluster the 268 municipalities and identify the number of areas with a different ICT development level, we consider 20 binary features and 3 continuous variables. The binary indicators indicate the presence of the following online facilities: x_1: online resolutions of the public administration; x_2: call for bids; x_3: e-procurement platform; x_4: service delivery information; x_5: informative e-mails; x_6: telephone and e-mail index; x_7: web site organization for life events; x_8: web site organization for personalization; x_9: web site organization for subjects and/or offices; x_{10}: online questionnaires or forum related to the municipality activities; x_{11}: possibility to enter into the home page with call centers or sms or wap; x_{12}: SUAP; x_{13}: dynamic map; x_{14}: information on the government body; x_{15}: e-mail of the elected representative leadership; x_{16}: information on the possibility to access the restricted area; x_{17}: service chart; x_{18}: interactive site map; x_{19}: pages written in a foreign language; x_{20}: quality approved by W3C Markup Validator. The continuous variables are: y_{21}: percentage of employees with a digital signature; y_{22}: percentage of employees dedicated to ICT; y_{23}: percentage of employees that have received ICT training.

We estimate models from 2 to 5 groups with the Latent Gold package [6], considering a data set with all continuous variables. In this data set, the categorical values $x_1, \ldots, x_{20}$ are substituted with the latent variables scores $y_1, \ldots, y_{20}$ and all variables are treated as Normal in the mixture models. The first model for each number of classes assumes local independence. The other specifications are obtained by subsequently adding the direct relationship between couples of variables, on the basis of the Latent Gold's bivariate residuals information. The bivariate residuals computed by the package indicate how similar the estimated and the observed bivariate associations are. These residuals can be interpreted as lower bound estimates for the improvement in fit in the likelihood when the corresponding local independence constraints are relaxed and, in each model, is added the local dependency with the highest Latent Gold's bivariate residual in the previous model. For assessing and comparing the models, we use the BIC criterion. Table 1 reports the BIC values and the number of parameters. Variables from 1 to 20 are the latent variables scores, and variables y_{21}, y_{22}, and y_{23} are the continuous variables in

Table 1 BIC values and number of estimated parameters (par.)

Model	Description	2 clusters		3 clusters		4 clusters		5 clusters	
		BIC	par.	BIC	par.	BIC	par.	BIC	par.
1	Local independence	26591	70	25629	94	25230	118	25291	142
2	Model 1 + $\sigma_{y_{22}y_{19}}$	25336	72	24023	97	24239	122	23918	147
3	Model 2 + $\sigma_{y_{21}y_{18}}$	23710	74	22739	100	22519	126	22548	152
4	Model 3 + $\sigma_{y_{19}y_{14}}$	24994	76	22894	103	22361	130	21870	157
5	Model 4 + $\sigma_{y_{23}y_{20}}$	23394	78	22656	106	22790	134	22580	162
6	Model 5 + $\sigma_{y_{15}x_6}$	23229	80	22828	109	22383	138	21837	167
7	Model 6 + $\sigma_{y_7y_1}$	23582	82	22325	112	22478	142	21555	172
8	Model 7 + $\sigma_{y_9y_2}$	23066	84	22654	115	22951	146	21925	177
9	Model 8 + $\sigma_{y_{14}y_2}$	22996	86	22073	118	22135	150	22005	182
10	Model 9 + $\sigma_{y_8y_7}$	23187	88	21602	121	21974	154	21472	197
11	Model 10 + $\sigma_{y_{16}y_4}$	22849	90	22315	124	22240	158	21497	202

the original data set. σ_{kj} is the covariance between variables k and j. For each number of groups, the models with more local dependencies have the lowest BIC values. The accuracy of fit in all situations is improved with inclusions of direct relationships between variables: the local independence model always performs worst. The fact that working with more local dependencies may yield a simpler final model with less clusters is evident: model 11 with 2 clusters performs better than model 4 with 3 clusters and model 8 with 4 clusters. Table 2 reports the relative frequencies of category 1 in the binary variables $(x_1, \ldots, x_{20})$ and the mean values of the three continuous variables (y_{21}, y_{22}, y_{23}) in each group, in the 4-clusters partitions with all considered bivariate dependencies (model 11). The classification is in agreement with the criterion of segment addressability suggested by Chaturvedi et al. [1] and related to the degree according to which a clustering solution can be explained by variables that can be controlled by policy makers. Indeed, in group 3, the most densely populated municipalities are clustered, with the most efficient public nets that allow both the distribution of nearly all of the interactive services by the Public Administration and the development of other telecommunication services for citizens (call centers, sms, ...). This cluster is homogeneous with respect to the presence of online information and facilities like: resolutions of the public administration (x_1), call for bids (x_2), informative e-mails (x_5), telephone and e-mail index (x_6), web site organization for life events (x_7), web site organization for subjects and/or offices (x_9), information on the governing body (x_{14}), e-mail of the elected representative leadership (x_{15}), interactive site map (x_{18}), and pages written in a foreign language (x_{19}). This group has a small percentage of employees with a digital signature but higher percentages of employees dedicated to ICT support and employees that have received ICT training. In group 2, the smallest municipalities are clustered. These are usually mountain communities, not in tourist areas. These units are the least technologically advanced and are in areas where it is

Table 2 Cluster means for models 11 with 4 groups

Cluster	1	2	3	4	tot
x_1	0.49	0.00	0.73	0.36	0.44
x_2	0.88	0.34	1.00	0.82	0.82
x_3	0.01	0.00	0.27	0.00	0.03
x_4	0.46	0.06	0.93	0.55	0.44
x_5	0.78	0.66	1.00	1.00	0.78
x_6	0.78	0.22	0.87	0.55	0.71
x_7	0.09	0.00	0.67	0.00	0.10
x_8	0.12	0.00	0.47	0.18	0.13
x_9	0.93	0.31	1.00	0.73	0.85
x_{10}	0.06	0.03	0.33	0.00	0.07
x_{11}	0.05	0.00	0.40	0.18	0.07
x_{12}	0.17	0.00	0.80	0.09	0.18
x_{13}	0.52	0.16	0.93	0.55	0.50
x_{14}	1.00	0.00	1.00	0.91	0.87
x_{15}	0.72	0.13	0.93	0.45	0.65
x_{16}	0.31	0.00	0.87	0.36	0.31
x_{17}	0.02	0.00	0.27	0.00	0.03
x_{18}	0.12	0.00	0.93	0.00	0.15
x_{19}	0.00	0.00	1.00	1.00	0.10
x_{20}	0.14	0.00	0.47	0.00	0.13
y_{21}	6.28	3.72	3.13	10.3	6.05
y_{22}	0.95	0.84	1.89	0.79	1.00
y_{23}	8.44	2.77	22.7	10.7	8.71
n_c	210	32	15	11	268
pop	8706	2702	105060	15924	13678

Last column ("tot") reports the means in the sample. Last row ("pop") reports the average population of municipalities belonging to the cluster

practically impossible to build an optic fibers net. Due to this technological barrier, the online services offered are only the basic ones. The percentage of employees devoted to ICT support is comparable with the values in the other groups: this aspect denotes that the absence of front office services is due to the absence of a broad band internet connection and of communication infrastructures as opposed to local political will. In group 4, tourist places are clustered. These municipalities are characterized by the presence of online services that are more useful for tourists rather than citizens. Indeed, this cluster is perfectly homogeneous with respect to x_5 (presence of informative e-mails) and x_{19} (presence of pages written in a foreign language). The percentage of employees with a digital signature is much higher than in the other groups. Cluster 1 is the largest one and, obviously, the least homogeneous. It groups municipalities equipped with a public net able to support most interactive services that have not achieved "excellence" and may improve the opportunities for citizens.

References

1. Chaturvedi, A.D., Green, P.E., Carrol, J.D.: K-modes clustering. J. Classification **18**, 35–55 (2001)
2. Lawrence, C.J., Krzanowski, W.J.: Mixture separation for mixed-mode data. Stat. Comput. **6**, 85–92 (1996)
3. Morlini, I.: A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. ADAC **6**(1), 5–28 (2012)
4. Pearson, K.: Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. Phil. Trans. Roy. Soc. Lond. Ser. A **195**, 1–47 (1900)
5. Vermunt, J.K., Magidson, J.: Latent class cluster analysis. In: Hagenaars, J.A., McCutcheon, A.L. (eds.) Applied Latent Class Analysis, pp. 89–106. Cambridge University Press, Cambridge (2002)
6. Vermunt, J.K., Magidson, J.: Technical Guide for Latent GOLD 4.0: Basic and Advanced. Statistical Innovations Inc., Belmon, MA (2005)

Cross-Lagged Structural Equation Models vs Latent Growth Curve Models: A Study of Material Deprivation in Portugal with ICOR

Paula C.R. Vicente and Maria de Fátima Salgueiro

Abstract

Structural equation models with lagged and cross-lagged effects and latent growth curve models are proposed in this chapter to demonstrate the use of two different, but complementary, approaches to modeling material deprivation over time. Portuguese data (ICOR) from the European Statistics on Income and Living Conditions (EU-SILC) are used. The four waves for which the data are available, 2004–2007, are considered.

1 Introduction

The survey on living conditions and household income (ICOR) is a panel that was implemented with the objective of ensuring the Portuguese participation in the European database called European Statistics on Income and Living conditions (EU-SILC). This database was created with the aim of obtaining indicators on income, living conditions, and social exclusion, comparable between the various European Union countries. The Portuguese participation is assured by the Instituto Nacional de Estatística (INE) since 2004, on an annual basis. ICOR has the particularity of being a rotating panel.

In recent studies the level and the quality of life of families have been assessed through a concept of material deprivation, measured by indicators such as the ability to access a set of basic needs, possession of durables, housing conditions, and even environmental conditions of the place where families live (see [4]).

In this chapter we use data from the ICOR from 2004 to 2007 to study the concept of material deprivation in two dimensions: (1) the financial capacity to meet basic needs; and (2) the possession of durables. These data are used to longitudinally

P.C.R. Vicente (✉) · M. de Fátima Salgueiro
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Portugal
e-mail: pvicente@netcabo.pt; fatima.salgueiro@iscte.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_2, © Springer-Verlag Berlin Heidelberg 2013

Table 1 ICOR: number of households in the rotational panel, from 2004 to 2007

Panel: Year selected	Survey year 2004	2005	2006	2007
2004	1251			
2004	1242	1129		
2004	1252	1158	1097	
2004	1240	1144	1102	1017
2005		1184	1052	970
2005			1116	999
2005				1324
Total households	4985	4615	4367	4310

model material deprivation. A cross-lagged structural equation model and a latent growth curve model are proposed. The statistical software Mplus 5 is used to estimate all models presented.

2 The Data

The ICOR is a rotating panel, with a dynamic rotation of one-fourth of the sample in each year. Thus, no household or individual will remain in the sample for more than four consecutive years. The adopted scheme minimizes the inquiring efforts of the participants in the longitudinal study. The collected longitudinal data have been made available for researchers for the period 2004–2007, both at the household level and at the individuals level (see [5]).

In the current study, and in order to measure material deprivation, two dimensions were considered: financial capacity and possession of durables. The scores for each dimension are obtained as sums of items in the ICOR questionnaire. The financial capacity dimension includes five items: (1) can afford 1 week vacation out of home per year; (2) can afford a meal with meat or fish every 2 days; (3) can afford to keep home adequately warm; (4) can afford regular expenses with no difficulty; and (5) can afford unexpected expenses without resorting to loan. Being computed as a sum of the items the household can afford, the financial capacity variable takes values from 0 to 5. The possession of durables considers the following items: (1) tv set; (2) telephone; (3) washing machine; and (4) personal vehicle, and takes values from 0 to 4.

The sample under analysis includes 991 households with valid answers to all questions about financial capacity and possession of durables. In 2004, 20.7% of the households can afford the regular expenses with no difficulty, 49.6% can afford keeping the house warm, 94.2% can afford a meal with meat or fish every 2 days, 31.8% can afford 1 week vacation, and 77.3% can afford unexpected expenses without resorting to loan. In terms of the possession of durables, 65.9%, 91.7%,

98.7%, and 89.8% of the households mentioned the possession of personal vehicle, washing machine, tv set, and telephone, respectively.

3 Methodology

This study makes use of two complementary methodologies for modeling longitudinal data: structural equation models and latent growth curve models. A brief overview of each statistical technique is now given.

3.1 Cross-Lagged Structural Equation Models

The structural component of a structural equation model (SEM) specifies the set of dependencies among the constructs of interest in the model (see Bollen [1]). A SEM with lagged and cross-lagged effects appears as an application of the concepts of autoregressive models to SEMs with longitudinal data. Let us consider two variables of interest, y and w, measured at different time points. Each variable at time t is explained by its value at time $t-1$ (lagged effect) and by the value of the other variable at time $t-1$ (cross-lagged effect)—see Finkel [3].

Figure 1 displays the path diagram of a cross-lagged structural equation model with two processes (for y and w). Variables y_1, y_2, y_3, y_4 correspond to the four repeated measures of one process (e.g., the financial capacity dimension). Variables w_1, w_2, w_3, w_4 correspond to the repeated measures of the other process (possession of durables), at each of the four time points. The model can be represented by the following system of equations, with $t = 2, 3, 4$ and $i = 1, 2, \ldots, N$:

$$y_{it} = \beta_{y_t y_{t-1}} y_{it-1} + \beta_{y_t w_{t-1}} w_{it-1} + \epsilon_{y_{it}}, \tag{1}$$

$$w_{it} = \beta_{w_t y_{t-1}} y_{it-1} + \beta_{w_t w_{t-1}} w_{it-1} + \epsilon_{w_{it}}, \tag{2}$$

where $\beta_{y_t w_{t-1}}$ and $\beta_{w_t y_{t-1}}$ represent the cross-lagged effects and $\beta_{y_t y_{t-1}}$ and $\beta_{w_t w_{t-1}}$ represent the lagged effects. The residual terms in each process (ϵ_{y_t} and ϵ_{w_t}) have mean zero, are not autocorrelated, and are uncorrelated with the right-hand side variables. Since the two processes are modeled simultaneously, at each time point, ϵ_{y_t} and ϵ_{w_t} are allowed to be correlated. These covariances are denoted by $\psi_{w_t y_t}$ and are parameters in the model (and so are the variances $\psi_{w_t w_t}$ and $\psi_{y_t y_t}$). The residual variables are assumed to follow a multivariate normal distribution. For the purpose of model estimation the variance/covariance structure of the y_t and w_t implied by the specified model is equated to the observed variance/covariance structure among the four repeated measures of the two dimensions of interest.

The lagged and cross-lagged effects give information about the dependence structure of the variables of interest, at different time points. However, these relationships only reflect changes for all individuals considered, not being able

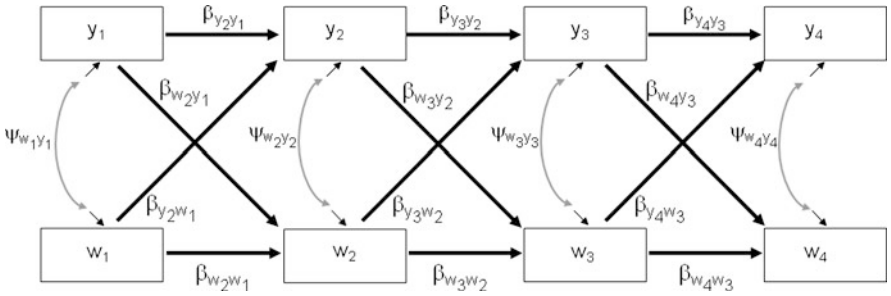

Fig. 1 Path diagram of a cross-lagged SEM (for y and w) and four time points

to access any information about individual changes. To account for individual differences in processes of change, latent growth curve models can be used (see [6]).

3.2 Latent Growth Curve Models

Latent growth curve models (LGCMs) are a useful statistical procedure for the analysis of longitudinal data. Through the estimation of a latent trajectory (linear or not), these models allow: (1) to estimate the average initial level and rate of change of all individuals; (2) to estimate the variability between individuals at the initial level and in the rate of change; and (3) to explain the heterogeneity observed in the trajectories, using conditional models. The idea underlying this type of longitudinal modeling is that the true trajectory is not observed, but latent, being estimated from the structure of means and correlations among the repeated measurements of the observed variables. LGCMs were developed within the framework of structural equation models, thus sharing many of its advantages (see [2, 7]).

Sometimes it is of interest to model joint trajectories for the repeated measures of two (or more) variables and multivariate LGCMs are required. Figure 2 displays the path diagram of a bivariate LGCM to simultaneously model the growth trajectories of y and w. The latent variables α^y and β^y represent the random intercept and the random slope for process y and are measured by the observed repeated measures of y. More generally, for the process y the value for individual $i = 1, \ldots, N$ at time $t = 1, \ldots, T$ is given by

$$y_{it} = \alpha_i^y + \lambda_t \beta_i^y + \epsilon_{it}^y, \tag{3}$$

and for the process w the value for individual i at time t is given by

$$w_{it} = \alpha_i^w + \lambda_t \beta_i^w + \epsilon_{it}^w, \tag{4}$$

where $\lambda_t = t - 1$ in the case of a linear growth trajectory (therefore factor loadings associated with the random slope are constrained to 0, 1, 2, and 3 in Fig. 2). The variances of the ϵ_{it} in each process are parameters in the model.

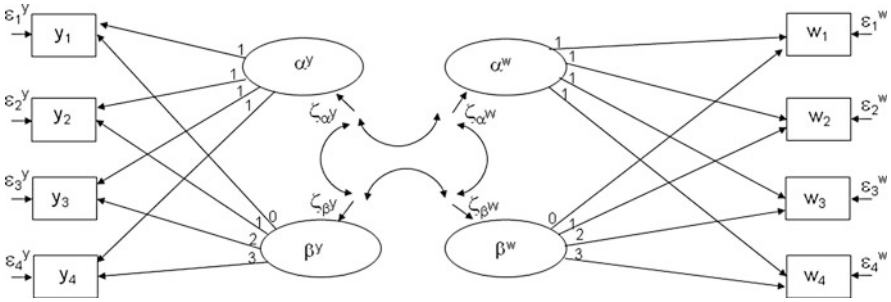

Fig. 2 Path diagram of a bivariate LGCM,where y_1, y_2, y_3, y_4 correspond to the four time point measures for one process and w_1, w_2, w_3, w_4 represent the four repeated measures for the other process

The parameters α_i^y and α_i^w, the random intercept for individual i in process y, or w, are given by (5). The parameters β_i^y and β_i^w, the random slope in process y, or w, for individual i, are given by (6),

$$\alpha_i = \mu_\alpha + \zeta_{\alpha_i} \tag{5}$$

$$\beta_i = \mu_\beta + \zeta_{\beta_i}, \tag{6}$$

where μ_α and μ_β are the mean of the intercept and the mean of the slope across all cases (in each process, y or w). The ζ_{α_i} and ζ_{β_i} are disturbances, with mean zero and uncorrelated with ϵ_{it} in each process, y or w, and represent the between individuals variability around the global mean. The covariance structure of the random intercepts and slopes in the two processes has variances $\Psi_{\alpha\alpha} = \mathrm{Var}(\zeta_\alpha)$, $\Psi_{\beta\beta} = \mathrm{Var}(\zeta_\beta)$ and covariance $\Psi_{\alpha\beta} = \mathrm{Cov}(\zeta_\alpha, \zeta_\beta)$. These covariances are represented by the double-headed curved arrows in Fig. 2.

The parameters of special interest in the model are the means (μ_α and μ_β) and variances ($\Psi_{\alpha\alpha}$ and $\Psi_{\beta\beta}$) of the random effects in each process, and the residual covariance structure between the random effects in the two processes ($\Psi_{\alpha\beta}$).

The equations presented so far assume that observed variables are continuous and normally distributed. If observed variables are ordinal, as is the case in this chapter, an auxiliary threshold model linking the ordinal variable y_{it} with an underlying continuous latent variable y_{it}^* is defined as:

$$y_{it} = c, \qquad \text{when} \qquad \tau_{c-1} < y_{it}^* \leq \tau_c \tag{7}$$

where $c = 1, 2, \ldots, C$ is the total number of ordered categories; τ_{c-1} and τ_c are the lower and upper thresholds for category c with $\tau_0 = -\infty$ and $\tau_c = +\infty$. The $c - 1$ threshold values are ordered from lowest to highest. The LGCM with ordinal manifest variables is then defined as

$$y_{it}^* = \alpha_i + \lambda_t \beta_i + \epsilon_{it} \tag{8}$$

by replacing y_{it} by y_{it}^{*} in (3). Similar reasoning applies for process w, and all remaining model definitions follow.

Mplus 5 was used to estimate all models presented in this chapter. The robust weighted least squares estimator (WLSMV) implemented in Mplus was used to deal with models with ordinal indicators.

In order to assess goodness of fit, the following measures are commonly used in the literature: the Tucker–Lewis Index (TLI), the Comparative Fit Index (CFI), and Weighted Root Mean Square Residual (WRMR). Recommended fit values are $\mathrm{TLI} > 0.95$, $\mathrm{CFI} > 0.95$, and $\mathrm{WRMR} < 0.9$.

4 Results

Two different models are considered: a cross-lagged structural equation model and a bivariate latent growth curve model. Results for the proposed cross-lagged SEM (Table 1) show that the model fits the data well: $\mathrm{CFI} = 0.992$, $\mathrm{TLI} = 0.992$, and $\mathrm{WRMR} = 0.950$. Table 2 displays the estimates (and t values in parenthesis) for the lagged and cross-lagged regression coefficients in a standardized solution (in which the variances of the variables are set to one so that the magnitude of the effects can be compared). Most of the lagged and cross-lagged effects are significant, leading to the conclusion that the financial capacity and the possession of durables at each time point are influenced by itself and by the other variable, in the preceding time point. One should point out that the lagged effects are always larger than the cross-lagged effects. The only nonsignificant effect is the effect of the financial capacity (2005) on the possession of durables (2006).

Results for the proposed bivariate LGCM (Fig. 2) also indicate an adequate model-data fit: $\mathrm{CFI} = 0.991$, $\mathrm{TLI} = 0.999$ and $\mathrm{WRMR} = 0.568$. Table 3 displays the estimates (and t-values in parenthesis) for the means, variances, and covariances of the random effect parameters. From Table 3 it is possible to conclude that in both processes (financial capacity and possession of durables) the mean of the slope is nonsignificant, suggesting that the average material deprivation does not change from 2004 to 2007. The estimated variances for the intercept and for the slope factors are both significant, leading to the conclusion that households vary, both regarding their 2004 material deprivation and their growth trajectories over time. The covariance between the intercepts of the two processes is significant, suggesting that the initial levels of financial capacity and possession of durables are associated. The covariance between the intercept of financial capacity and the slope of financial capacity is also significant. Its negative sign suggests that higher levels of financial capacity in 2004 are associated with lower mean growth trajectories from 2004 to 2007.

Table 2 Estimates (t-values), in a standardized solution, for the lagged and cross-lagged regression coefficients in the cross-lagged SEM

	Financial capacity05 (y_2)	Possession of durables05 (w_2)
Financial capacity04 (y_1)	**0.841** (55.980)	**0.146** (7.517)
Possession of durables04 (w_1)	**0.081** (3.901)	**0.816** (71.352)
	Financial capacity06 (y_3)	Possession of durables06 (w_3)
Financial capacity05 (y_2)	**0.815** (44.192)	0.026 (2.480)
Possession of durables05 (w_2)	**0.090** (3.476)	**0.956** (147.190)
	Financial capacity07 (y_4)	Possession of durables07 (w_4)
Financial capacity06 (y_3)	**0.769** (40.339)	**0.062** (3.416)
Possession of durables06 (w_3)	**0.135** (4.980)	**0.939** (86.219)

The boldface denotes a significant estimate, at the 5% level

Table 3 Estimates (t-values) for the random parameters of the LGCM

	Financial capacity	Possession of durables
Mean of the slope	−0.012 (−1.367)	0 (−0.025)
Variance of the intercept	**0.902** (61.733)	**0.968** (205.156)
Variance of the slope	**0.029** (7.905)	**0.005** (4.124)
Cov. between InterceptFC InterceptPD	**0.550** (22.521)	
Cov. between SlopeFC SlopePD	−0.004 (−1.282)	
Cov. between InterceptFC SlopeFC	**−0.057** (−6.834)	
Cov. between InterceptPD SlopePD	−0.008 (−1.107)	

The boldface denotes a significant estimate, at the 5% level

5 Discussion

Studies on poverty are often based on household disposable income. However, it can be argued that a multidimensional approach should be followed, and more than one indicator should be used to measure poverty and material deprivation. This chapter proposes measuring the financial capacity and the possession of durables (two dimensions of material deprivation) using several indicators.

A cross-lagged SEM and a LGCM are proposed to investigate and answer the following four research questions:

- How do the financial capacity and the possession of durables change globally, over the 4 years, considering all the households?
- Which evidence do we find for cross-lagged effects from financial capacity on possession of durables and from possession of durables on financial capacity, over time?
- Which evidence do we find for within and between individual growth variability over the four time points, both regarding financial capacity and possession of durables?

- How do average initial levels of financial capacity and possession of durables relate to the corresponding average growth rates, over the 4-year period under analysis?

The first two questions are answered by the proposed cross-lagged SEM. In fact, the results obtained with cross-lagged SEM allowed us to conclude that, at each time point, the household financial capacity is explained by their financial capacity and by their possession of durables in the previous moment in time. Similar conclusion can be drawn for the household possession of durables, with the exception that the effect of the 2005 financial capacity on the 2006 possession of durables is nonsignificant.

The bivariate LGCM proposed to jointly model the growth trajectories of the two dimensions of material deprivation allows us to answer the last two questions. The households average financial capacity and possession of durables does not change significantly over time. However, households vary both regarding their average material deprivation in 2004 and their growth of material deprivation from 2004 to 2007. Additionally, it is possible to conclude that higher initial levels of financial capacity are associated with lower mean growth trajectories in financial capacity from 2004 to 2007; the initial level of financial capacity is associated with the initial level of the possession of durables. Hence, the complementary nature of the two statistical approaches has been illustrated.

References

1. Bollen, K.: Structural Equations with Latent Variables. Wiley, New York (1989)
2. Bollen, K.A., Curran, P.J.: Latent Curve Models – A Structural Equation Perspective. Wiley, New Jersey (2006)
3. Finkel, S.: Causal Analysis with Panel Data. Sage, Thousand Oaks (1995)
4. Guio, A.-C.: What can be learned from material deprivation indicators in Belgium and in its regions. Institut Wallon de L' Evaluation de la Prospective et de la Statistique **901** (2009)
5. INE: Inquérito às Condições de Vida e Rendimento-ICOR. Documento Metodológico (2009)
6. McArdle, J.J.: Latent variable modeling of differences and changes with longitudinal data. Ann. Rev. Psychol. **60**, 577–605 (2009)
7. Salgueiro, M.F., Smith, P.W.F., Vieira, M.D.T.: A multi-process second-order latent growth curve model for subjective well-being. Quality & Quantity (2011). doi: 10.1007/s11135-011-9541-y

A Finite Mixture Approach to Uncover the Heterogeneity in the Relationship Between Visit Motivation and Activity Preferences

Catarina Marques, Elizabeth Reis, and João Menezes

Abstract

This research uses a model-based segmentation approach in partial least squares path modeling—the FInite MIXture Partial Least Squares (FIMIX-PLS) methodology—to capture the unobserved heterogeneity in the inner path model estimates of the influence of motivation to visit the Portuguese protected areas on visitors' preferences for different types of activities. Three distinctive visitor segments were identified. Results confirm the assumption of heterogeneity as the relationships differ across segments.

1 Introduction

Heterogeneity in the context of structural equation modeling (SEM) has typically been addressed by assuming that individuals can be assigned to segments a priori, for example, on the basis of demographic or psychographic variables (observed heterogeneity). However, if the substantive theory on the variables causing heterogeneity is not available, there is unobserved heterogeneity which can result in serious problems, e.g., biased parameter estimates thus leading to potentially flawed conclusions [7, 11, 16].

Different methods have been developed for the analysis of unobserved heterogeneity in SEM. Hahn et al. [4] extended the STructural Equation finite Mixture Model (STEMM) of Jedidi et al. [7, 8] to a Partial Least Squares (PLS) framework [9, 19] and named their model FIMIX-PLS. The data are supposed to be the result of the mixture of two or more populations in different proportions, with the aim

C. Marques (✉) · E. Reis · J. Menezes
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Av. das Forças Armadas, 1649-026 Lisbon, Portugal
e-mail: catarina.marques@iscte.pt; ear@iscte.pt; joao.menezes@iscte.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_3, © Springer-Verlag Berlin Heidelberg 2013

of estimating the (a posteriori) probability of each individual belonging to each of these subpopulations. The technique is based on the application of the Expectation-Maximization (EM) algorithm to the latent variable scores predicted in PLS; EM makes an estimation of the a posteriori probabilities used for the individuals' classification. In summary, the FIMIX-PLS approach combines the advantages of predicting path coefficients using PLS, with the maximum likelihood estimation of a finite mixture model.

The current research studies how the motivation to visit the Portuguese protected areas (PPA) influences visitors' preferences for different types of activities by accounting for unobserved heterogeneity. As this influence may differ for distinct groups of visitors, it is essential to identify visitor segments so that appropriate interpretations can be made that are of great importance to marketing decisions. Partial least squares path modeling (PLS-PM) is used to test the hypothesized relationships and model, and the finite-mixture PLS methodology is used to identify distinctive visitor segments. The use of PLS-PM is explained by its minimal demand on sample size and by its suitability to handle both reflective and formative measurement models [5]. As for the uncovered heterogeneity, FIMIX-PLS is regarded as the primary approach [17] and has become mandatory for evaluating PLS-PM results [5]. A conceptual model to explain how visit motivations influence visitors' preferences for particular nature or culture-based activities is proposed. Motivations to visit parks are based on the push and pull factors theory developed by Crompton [1].

2 The Influence of Visit Motivation on Activity Preferences

Crompton's [1] push and pull factors have been one of the most widely used theoretical framework to explain tourism motivations. The idea behind this two-dimensional approach lies in the assumption that individuals travel because they are pushed by their own internal forces and pulled or attracted by the external forces of the destination attributes, both forces being concurrently attractive in the choice of a particular destination [1,2,18]. Therefore, the push motivations explain the decision to travel, considered here as the motives to visit PPA, while the pull motivations justify the options when faced with the choice of travel destinations. In this study, the pull factors are considered simultaneously as the PPA attractions and activities.

Figure 1 displays a simplified diagram of the conceptual model used to test the proposed hypotheses. Marques [10] presents a detailed description of the model. Based on the literature of nature-based tourism, the proposed hypotheses are as follows:

H1.1: The motivations related to the enjoyment of nature positively influence the importance given to nature-based attractions (H1.1a) and observation and study attractions (H1.1b).

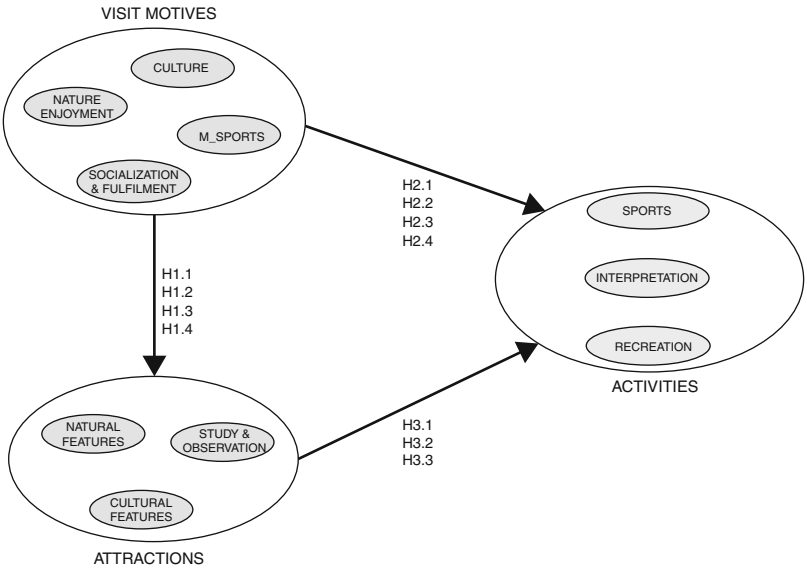

Fig. 1 Conceptual model and research hypotheses

H1.2: The motivation to visit PPA due of cultural related issues positively influences the importance of nature-based attractions (H1.2a) and local culture attractions (H1.2b).
H1.3: Motives related to sports positively influence the importance of nature-based attractions.
H1.4: Fulfilment and socialization positively influence the importance of nature-based (H1.4a), local culture (H1.4b), and study and observation (H1.4c) attractions.
H2.1: The motives related to the enjoyment of nature positively influence the preference for nature-based sports (H2.1a) and interpretation activities (H2.1b).
H2.2: Culture-related motives positively influence the preference for recreation activities in PPA.
H2.3: Sport-related motives positively influence the preference for nature-based sports.
H2.4: Fulfilment and socialization positively influence the preference for nature-based sports (H2.4a), interpretation (H2.4b), and recreation (H2.4c) activities.
H3.1: The importance of nature attractions positively influences the preference for interpretation activities.
H3.2: The importance of cultural attractions positively influences the preference for interpretation (H3.2a) and recreation (H3.2b) activities.
H3.3: The importance of study and observation attractions positively influences the preference for interpretation activities.

3 Model Specification

The PLS path model consists of two models, the outer and the inner models. The inner model depicts the relationships among the endogenous latent variables and the exogenous latent variables as posited by substantive theory. The outer model may be specified in two different ways: as a reflective or a formative mode. Traditionally, applications in marketing and business research are based on reflective measurement [3, which represents a factor model], where it is assumed the indicators reflect the variation in the latent variable, i.e., it is expected that changes in the construct will result in changes in all its indicators. In contrast, the formative models come into use when a combination of indicator variables underlies the latent variable. Under these models, changes in the indicators are expected to be reflected in changes in the latent variable (so the latter is an index). As mentioned, it is assumed that data result from a mixture of several populations combined in different proportions. Because the heterogeneity is concentrated in the inner model, each endogenous latent variable is distributed as a finite mixture of K conditional multivariate normal densities. The a posteriori probability of membership for each observation can be computed using Bayes' theorem, conditional on the estimates of the segment-specific parameters. Hahn et al. [4] and Ringle [12] present a detailed description of FIMIX-PLS algorithm.

4 Methodology

This research uses a survey approach based on a self-administered questionnaire. The target population includes inhabitants in Portugal aged 15 years or more. The survey was conducted between April and June 2005. There are 779 valid answers; 401 are considered visitors since they alleged to visit protected areas at least once per year. All variables (measures of visit motives, PPA attractions and activities) are measured on a 5-point Likert scale ranging from 1 for no importance to 5 for very important. The list of motives is in agreement with the literature on nature-based tourism motivation; The PPA attractions were identified by accessing PPA information and activity types were defined based on the orientations of the Portuguese Program of Nature Tourism. For these reasons, the outer models of PPA attractions and activities are assumed to be formative whereas the outer model of visit motive is assumed to be reflective.

The segmentation process follows a four-step approach and the statistical software SmartPLS 2.0 [13] was used. In the first step, the standard PLS-PM algorithm was run on all cases to test the hypotheses and model, i.e., to estimate the so-called overall model. The resulting scores for latent variables in the inner model were used as input for the FIMIX-PLS procedure in the second methodological step. As mentioned, the segment-specific heterogeneity is concentrated in the estimated relationships between latent variables. FIMIX-PLS captures this heterogeneity and calculates the a posteriori probability of membership for each observation, the

mixture proportions, and the segment-specific estimates for the inner relationships and for the regression variances of the latent endogenous variables. As the number of segments is not known a priori, it is necessary to repeat the FIMIX-PLS algorithm for a different number of segments to detect the optimal partition. The appropriate number of segments to be retained is decided based on the jointly consideration of AIC_3 (Corrected Akaike Information Criterion with a penalty factor of three) and CAIC (Consistent AIC) [16]. The Normed Entropy Statistic (EN) ensures that the segments are sufficiently distinct.

Step 3 consists of an ex post analysis; this analysis involves the identification of one or more external variables that best explain differences between the FIMIX-PLS segments. This approach not only helps the interpretability of these segments but also to overcome the limitation of FIMIX-PLS segmentation by focusing just on the inner model relationships [14]. The CHAID algorithm is applied by using the Classification Tree analysis of SPSS 17.0 with this purpose.

In the fourth step, data are classified by means of these explanatory variables, which serve as a new input for a group-specific estimation in the PLS-PM approach (a priori segmentation). The evaluation and interpretation of these group-specific PLS-PM outcomes require a PLS multigroup analysis to determine significant differences in both the inner and outer models.

5 Results

The results jointly obtained for the AIC_3 and CAIC statistics suggest a three-segment solution. The validity and reliability were assessed for the overall and the segment-specific models by applying different measures for both types of indicators, reflective, and formative. The results of the outer model show that the reflective items used to measure the constructs are deemed valid (discriminant validity, composite reliability, and convergent validity). All weight estimates are significant and the variance inflation factor computed for all formative indicators are much lower than the threshold value of 10 [5, 6]. Table 1 presents the values of R^2 and GoF measures for the overall and the segment-specific models. R^2 values for the endogenous latent variables in segment 1 are greater than in the overall model, indicating a real improvement in the explained variance for segment 1 model. In this segment, the activity type has relatively high values of R^2, indicating that visit motives and attractions have a strong influence on the importance of activities. Segment 3 shows the lowest R^2 values for the attraction latent variables; this means that the visit motives for these visitors have little influence on the importance given to PPA attraction (this segment-specific model also has the lowest GoF index). However, the values of the average R^2 are greater than those of the overall model.

Table 2 presents an overview of inner parameter estimates for the overall model and the three segment-specific models. The results of the overall model suggest that the main influences on activities are as follows: the nature-based sports are influenced by the motivation to do sports (path coefficient estimate of 0.3427); culture-related motives and attractions lead to valuing recreation (0.4316 and

Table 1 R^2 and GoF of the overall and the segment-specific models

Endogenous latent variables	R^2				Average R^2
	Overall	Segment 1	Segment 2	Segment 3	
Attractions					
NATURAL FEATURES	0.4167	0.6281	0.6485	0.2605	0.5060
CULTURAL FEATURES	0.3880	0.6325	0.5190	0.0961	0.4342
STUDY & OBSERVATION	0.4157	0.4858	0.6875	0.0658	0.3712
Activities					
SPORTS	0.1422	0.5284	0.1562	0.1437	0.3455
INTERPRETATION	0.1820	0.5777	0.1842	0.2504	0.4113
RECREATION	0.2010	0.5319	0.0538	0.1711	0.3423
GoF	0.4519	0.5990	0.5421	0.3420	
N	401	207	56	138	
%	100 %	52 %	14 %	34 %	

0.2312, respectively) and interpretation activities (0.3616); the motivation to enjoy nature also affects the interest in interpretation activities (0.2452).

Segments 1, 2, and 3 comprise 51.6 %, 14.0 %, and 34.4 % of the overall visitors group, respectively. Results from the multiple group comparison provide evidence of varying relationships among constructs for all segments. Kruskal–Wallis tests confirm that the segments are different in all inner relationships. Pairwise comparison tests (Dunnett's T3 for subsamples with unequal variances) are also computed for each inner relationship. No differences are shown between pairs of segments for only four relationships (referenced in Table 2 through the letters a, b, and c in the last column).

The main influences on activities can be summarized as follows: In segment 1, nature-based sports are influenced mainly by reasons of personal fulfilment (path coefficient estimate of 0.4814); recreation activities by the interest in cultural aspects of PPA (0.5328); and interpretation activities are affected by the desire to enjoy nature (0.4452) and the attraction of cultural features (0.3107). Study and observation also have a significant but negative effect on interpretation activities (-0.3562). In segment 2,[1] nature-based sports are only influenced by the motivation to do sports in PPA (0.5604), while interpretation activities are influenced both by the interest in cultural aspects of PPA (0.7050) and by being motivated by culture (through an indirect effect; total effect of 0.3867). Segment 3 is also different from the other two, because (1) the activities are influenced by the motives directly related to them (see hypotheses H2.1b, H2.2 and H2.3); (2) fulfilment and social motives have negative effects on all activities (although the impact on sports is not significant); and (3) attractions have only one significant effect on activities

[1] Some path coefficient estimates in segment 2 have high values but are nonsignificant. This situation may be the result of the highly skewed data, since the non-normal data inflate bootstrap standard errors and thus reduce statistical power [5].

Table 2 Standardized path coefficients for the overall and the segment-specific models

Hyp.	Inner paths	Overall model	Segment 1	Segment 2	Segment 3	KW statistics
H1.1a	NATURE ENJOYMENT → NATURAL FEATURES	0.4150*	0.5128*	0.5651*	0.5130*	32.55*[a]
H1.1b	NATURE ENJOYMENT → STUDY & OBSERVATION	0.4241*	0.5108*	0.4950*	0.2271	304.22*[b]
H1.2a	CULTURE → NATURAL FEATURES	0.1598*	0.1780*	−0.0409	−0.6154*	477.25*
H1.2b	CULTURE → CULTURAL FEATURES	0.4459*	0.6290*	0.5191*	−0.2126	429.39*
H1.3	M_SPORTS → NATURAL FEATURES	−0.0576	0.0832**	−0.1342	0.1201	283.91*
H1.4a	SOCIALIZATION & FULFILMENT → NATURAL FEATURES	0.1597**	0.1735*	0.4001**	−0.0497	401.54*
H1.4b	SOCIALIZATION & FULFILMENT → CULTURAL FEATURES	0.2016*	0.2245*	0.2165	−0.1233	193.64*
H1.4c	SOCIALIZATION & FULFILMENT → STUDY & OBSERVATION	0.2513*	0.2521*	0.3563**	−0.3211	272.41*
H2.1a	NATURE ENJOYMENT → SPORTS	−0.0707	0.1935*	−0.4103	−0.1243	385.20*
H2.1b	NATURE ENJOYMENT → INTERPRETATION	0.2452*	0.4452*	0.2520	0.3987*	107.80*
H2.2	CULTURE → RECREATION	0.4316*	0.1146	0.3965	0.5571*	277.96*
H2.3	M_SPORTS → SPORTS	0.3427*	0.1971*	0.5604*	0.5282*	369.17*[c]
H2.4a	SOCIALIZATION & FULFILMENT → SPORTS	0.1042	0.4814*	0.0088	−0.3121	377.20*
H2.4b	SOCIALIZATION & FULFILMENT → INTERPRETATION	−0.0270	0.2869*	0.1664	−0.3604*	367.53*
H2.4c	SOCIALIZATION & FULFILMENT → RECREATION	−0.2098*	0.1434**	−0.2764	−0.3888**	326.35*
H3.1	NATURAL FEATURES → INTERPRETATION	−0.1229	0.1035	−0.5078	0.0731	310.34*[a]
H3.2a	CULTURAL FEATURES → INTERPRETATION	0.3616*	0.3107*	0.7050*	−0.0932	503.64*
H3.2b	CULTURAL FEATURES → RECREATION	0.2312*	0.5328*	0.0694	−0.0291	397.17*
H3.3	STUDY & OBSERVATION → INTERPRETATION	−0.0064	−0.3562*	−0.4580	0.3120*	397.77*

Results of Dunnett's T3:
[a] No significant differences at 0.05 level between segments 1 and 2
[b] No significant differences at 0.05 level between segments 1 and 3
[c] No significant differences at 0.05 level between segments 2 and 3
*$p < 0.05$; **$p < 0.10$

(effect of study and observation on interpretation); this means that PPA attractions are not very important in explaining preferences for activities in this segment. Regarding the proposed hypotheses, it can be seen that (1) only two are not verified in the largest segment—segment 1; (2) in contrast, less than half are verified in the remaining segments; (3) only one relationship is simultaneously not significant in the overall and segment-specific models (H3.1: NATURAL FEATURES → INTERPRETATION; this means that interpretation activities are not influenced by the natural characteristics of PPA for all visitors).

6 Conclusion

This study uses the FIMIX-PLS approach to capture the unobserved heterogeneity that exists in inner model estimates of how visit motivation influences activity preferences. Overall, relationships differ across segments, with sign changes appearing in some paths; therefore, estimating a overall PLS model across all three segments might yield distorted results [11]. It is thus critical to identify different groups of visitors with distinct estimates in the inner model in order to shape effective marketing strategies [15].

The proposed conceptual model proves adequate to explain how visit motivations influence visitors' preferences for PPA activities for more than half of visitors (given the high R^2 values of endogenous latent variables and the number of hypotheses verified in segment 1). For these visitors, PPA attractions play an important possible mediating role in the relationship between visit motives and preferences for activities. Implications of the findings are that the Portuguese park management agency should target the segments differentially so as to provide activities for all visitors. Targeting the segments would also be facilitated by profiling them with visitor background variables.

References

1. Crompton, J.: Motivations for pleasure vacations. Ann. Tourism Res. **6**, 408–424 (1979)
2. Dann, G.: Tourist motivation: an appraisal. Ann. Tourism Res. **8**, 187–219 (1981)
3. Diamantopoulos, A.: Formative indicators: introduction to the special issue. J. Bus. Res. **61**, 1201–1202 (2008)
4. Hahn, C., Johnson, M.D., Herrmann, A., Huber, F.: Capturing customer heterogeneity using a finite mixture PLS approach. Schmalenbach Bus. Rev. **54**, 243–269 (2002)
5. Hair, J.F., Ringle, C.M., Sarstedt, M.: PLS-SEM: indeed a silver bullet. J. Market. Theor. Pract. **19**, 139–151 (2011)
6. Henseler, J., Ringle, C.M., Sinkowics, R.R.: The use of partial least squares path modeling in international marketing. Adv. Int. Market. **20**, 277–319 (2009)
7. Jedidi, K., Jagpal, H.S., DeSarbo, W.S.: Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. Market. Sci. **16**, 39–59 (1997a)
8. Jedidi, K., Jagpal, H.S., DeSarbo, W.S.: STEMM: a general finite mixture structural equation model. J. Classification **14**, 23–50 (1997b)

9. Lohmöller, J.B.: Latent Variable Path Modeling with Partial Least Squares. Physica-Verlag, Heidelberg (1989)
10. Marques, C.: On the heterogeneity and measurement error: an application to the visitors of portuguese protected areas. PhD Thesis, ISCTE Business School, Lisbon University Institute (2010)
11. Muthén, B.O.: Latent variable modeling in heterogeneous populations. Psychometrika **54**, 557–585 (1989)
12. Ringle, C.M.: Segmentation for path models and unobserved heterogeneity: The finite mixture partial least squares approach. Research Papers on Marketing and Retailing, vol 35, University of Hamburg (2006)
13. Ringle, C.M., Wende, S., Will, A.: SmartPLS 2.0 (M3) Beta. Available via http://www.smartpls.de (2005). Accessed 28 May 2009
14. Ringle, C.M., Wende, S., Will, A.: Finite mixture partial least squares approach: methodology and numerical examples. In: Esposito Vinzi, V., Chin, W.W., Henseler, J., Wang, H. (eds) Handbook of Partial Least Squares: Concepts, Methods and Applications in Marketing and Related Fields, pp. 195–218. Springer, Berlin (2010)
15. Sarstedt, M.: Accounting for unobserved heterogeneity in the analysis of antecedents and consequences of corporate reputation: an application of FIMIX-PLS. In: Proceedings of the 37th Annual Conference of the European Marketing Academy (EMAC), Brighton, England (2009). Available via http://www.ibl-unihh.de/emac2008%20download.pdf
16. Sarstedt, M., Becker, J.M., Ringle, C.M., Schwaiger, M.: uncovering and treating unobserving heterogeneity with FIMIXPLS: which model selection criterion provides an appropriate number of segments? Schmalenbach Bus. Rev. **63**, 34–62 (2011)
17. Sarstedt, M., Ringle, C.M.: Treating unobserved heterogeneity in PLS path modeling: A comparison of FIMIX-PLS with different data analysis strategies. J. Appl. Stat. **37**, 1299–1318 (2010)
18. Uysal, M., Jurowski, C.: Testing the push and pull factors. Ann. Tourism Res. **21**, 844–846 (1994)
19. Wold, H.O.: Soft modeling: The basic design and some extensions. In: Jöreskog, K.G., Wold, H.O. (eds.) Systems Under Indirect Observation, Part II, pp. 1–54. North Holland, Amsterdam (1982)

An Application of Structural Equation Modeling of Test Dispositional Optimism as Mediator or Moderator in Quality of Life in Patients with Chronic Disease

Estela Vilhena, José Luís Pais Ribeiro, Isabel Silva, Luísa Pedro, Rute Meneses, Helena Cardoso, António Martins da Silva, and Denisa Mendonça

Abstract

The aim of the present study was to test a hypothetical model to examine if dispositional optimism exerts a moderating or a mediating effect between personality traits and quality of life, in Portuguese patients with chronic diseases. A sample of 540 patients was recruited from central hospitals in various districts of Portugal. All patients completed self-reported questionnaires assessing socio-demographic and clinical variables, personality, dispositional optimism, and quality of life. Structural equation modeling (SEM) was used to analyze the moderating and mediating effects. Results suggest that dispositional optimism exerts a mediator rather than a moderator role between personality traits and

E. Vilhena (✉)
Polytechnic Institute of Cávado and Ave and ICBAS University of Porto, Portugal
e-mail: evilhena@ipca.pt

J.L.P. Ribeiro
FPCE University of Porto and UIPES, Portugal
e-mail: jlpr@fpce.up.pt

I. Silva
U.Fernando Pessoa of Porto, Portugal
e-mail: isabels@ufp.pt

L. Pedro
IPL of Lisbon, Portugal
e-mail: luisa.pedro@estesl.ipl.pt

R. Meneses
U.Fernando Pessoa of Porto, Portugal
e-mail: rmeneses@ufp.edu.pt

H. Cardoso · A. Martins da Silva · D. Mendonça
ICBAS University of Porto and HGSA Hospital Center of Porto, Portugal
e-mail: helenacardoso@icbas.up.pt; ams@icbas.up.pt; dvmendon@icbas.up.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_4, © Springer-Verlag Berlin Heidelberg 2013

quality of life, suggesting that "the expectation that good things will happen" contributes to a better general well-being and better mental functioning.

1 Introduction

Structural equation modeling (SEM) [2, 12, 14, 16, 18, 22] techniques are considered today to be a major component of applied multivariate statistical analysis for addressing complex scientific questions in a great variety of areas of research.

Quality of life (QoL) has become an important concept for health care. It is a construct composed of a number of factors that contribute to an individuals well-being and adjustment to chronic disease [13]. Patients with chronic diseases are required to live with the limitations imposed by their conditions. Some personality traits, such as neuroticism and extraversion, are factors that contribute to a poorer perception of health status and appear as determinants of QoL [5]. In studies that involve patients with chronic diseases, QoL is frequently studied, but its associations with personality traits and other variables, which contribute to the adjustment to illness, are not very well clarified. Optimism is an important variable that can affect the QoL [3, 8–10, 25], can promote a better emotional adjustment and physical health. It predicts positive attitudes and the tendency to plan for recovery, seek information, and reframe bad situations [25]. Data suggest the direct and indirect effects between optimism and QoL, and the importance of personality traits in this adjustment. But it is important to explore the nature of influence of optimism on QoL. Is it a domain specific rather than a generalized construct, with differential effects on QoL in different domains or contexts [25]? What type of effect does dispositional optimism exert between personality and QoL, is it mediational or moderating in nature? Several studies on optimism show its role as a moderator or mediator [3, 8–10, 25] but not in this context and without the application of SEM statistical methodology .

Using SEM, this study aimed to test two hypothetical models: (1) dispositional optimism exerts a moderating influence between personality traits (neuroticism and extraversion) and the psychosocial components of quality of life (general well-being and general mental health); (2) dispositional optimism exerts a mediating influence between these personality traits and the psychosocial components of quality of life, in Portuguese patients with chronic diseases.

2 Methods

2.1 Participants and Procedure

This is a cross-sectional study comprising a sequential sample of 540 volunteers, patients with chronic diseases who were approached by their physicians, in out-patient departments of the main hospitals. Participants were eligible to participate in the study if they met the following criteria: (1) diagnosis of epilepsy, diabetes,

multiple sclerosis, obesity, myasthenia gravis, and cancer; (2) age > 18 years at the time of the interview; (3) educational level higher than 6 years; (4) diagnosed at least 3 years prior to the study; (5) return to usual daily life with disease under control; (6) no psychiatric disturbances. Written informed consent was obtained from the institutional review board and from the subjects.

2.2 Measures

A self-report questionnaire was applied to evaluate neuroticism and extraversion, dispositional optimism, and the psychological components of quality of life. The data were collected by psychologists, after each consultation.

Socio-demographic and Clinical Variables
The demographic information included age (A), gender (G), and education (E). The medical information was obtained from medical record and includes time since diagnosis (TSD) and perception of severity of disease (PSD).

Neuroticism and Extraversion
A short version of the Revised NEO Personality Inventory (NEO-PI-R), the NEO Five-Factor Inventory (NEO-FFI), was used [4]. It is a measure, composed of 60 items, that assess the five domains of personality. In this study, we are only interested in Neuroticism (Neuro) and Extraversion (Extra), two of the five factors that include 4 of the 60 items.

Dispositional Optimism
Dispositional optimism was measured using a six-item scale, the Revised Life Orientation Test (LOT-R) developed to assess individual differences in generalized optimism versus pessimism. Three items reflect expectations of positive outcomes and three items reflect expectations of negative outcomes [20].

Quality of Life
This was measured with the SF-36 [26] questionnaire, a 36-item questionnaire, divided into the several dimensions and three components identified in the IQOLA project study [11]: physical health, mental health, and general well-being. In the present study we did not use the physical component. The general well-being component (GWB) includes the dimensions of general health (GH) and vitality (V): the general mental health component (GMH) includes role emotional (RE), social functioning (SF), and mental functioning (MF) dimensions.

2.3 Statistical Analysis

Data Analysis
Descriptive statistics were used to describe demographic and clinical characteristics. Pearson's correlation coefficient was used to analyze linear associations between

neuroticism, extraversion, dispositional optimism, and the psychosocial components of quality of life. Data were analyzed using the SPSS statistical package, version 17.

Structural Equation Modeling

SEM was used to test the two hypothetical models, using the traditional approach of moderation and mediation, described by Baron and Kenny [1].

SEM is a multivariate technique that allows questions that involve multiple regression analysis of factors. It can be viewed as a combination of factor analysis (measurement model) and regression or path analysis (structural model). It is a confirmatory rather than an exploratory technique, because it compares a hypothesized model's covariance matrix with that of observed data. The measurement model illustrates the relationships of the latent variables to the observed variable, whereas the structural model indicates the direct and the indirect effects of latent and observed variables on each other. Goodness of fit indices were used to test the adequacy of the models. The chi-squared test was used as an index of discrepancy between the original (sample) correlation matrix and the (population) correlation matrix estimated from the model. The value of χ^2, its p-value, and the number of degrees of freedom (df) were examined. A nonsignificant p-value ($p > 0.05$) and the ratio of $\frac{\chi^2}{df} < 3$ would represent a good model fit. Because the significance of chi-square tests is dependent on the number of subjects, other goodness-of-fit indices were also used: Comparative Fit Index ($\mathrm{CFI} > 0.9$) and Root Mean Square Error Approximation ($\mathrm{RMSEA} < 0.05$). CFI values derived from the comparison of the hypothesized model with the independent model. RMSEA values help to answer the question of how well the model would fit the population covariance matrix if it were available [2,12,14,16,18,22]. In SEM it is necessary to remember the problems of no unicity due to the presence of latent variables. Even if the parameters were perfectly identifiable, since the model gives a solution in terms of latent variables estimation by means of observed variables, there is not a unique solution for the latent variables of the model [23]. Analysis for constructing structural equation models was conducted with the EQS 6.1 program and were used Maximum Likelihood (ML) estimation with the Satorra–Bentler scaled chi-squared and adjustments to standard errors.

2.4 Results

Patient Characteristics

The 540 patients include: 67 with cancer, 103 with diabetes, 56 with epilepsy, 19 with myasthenia gravis, 101 with multiple sclerosis, and 194 with obesity. Of this group, 27.8 % were male, their mean age was 40.95 years ($sd = 11.4$), mean school 9.90 years ($sd = 4.7$), mean time since the diagnosis was 11.74 years ($sd = 9.36$), and mean severity of disease perception was 6.66 ($sd = 2.76$). Table 1 displays the mean scores and standard deviation of predictors, moderator/mediator, outcome variables, and a correlation matrix. Dispositional optimism, neuroticism,

Table 1 Descriptive statistics and correlations for dispositional optimism, personality traits, and QoL

	Mean	*sd*	1	2	3	4	5	6	7	8
1. Disp. Optimism	20.39	4.32	–	−0.52*	0.46*	0.35*	0.38*	0.27*	0.35*	0.50*
2. Neuroticism	25.94	7.66		–	−0.32*	−0.38*	−0.44*	−0.42*	−0.50*	−0.61*
3. Extraversion	29.77	6.45			–	0.30*	0.37*	0.21*	0.40*	0.33*
4. General health	48.24	20.36				–	0.50*	0.37*	0.42*	0.45*
5.Vitality	49.15	21.70					–	0.51*	0.55*	0.67*
6. Role emotion	65.26	40.70						–	0.53*	0.57*
7. Social functioning	73.02	25.91							–	0.62*
8. Mental functioning	59.21	22.66								–

*$p < 5\,\%$

and extraversion were moderately associated with scores on the psychosocial components of quality of life; only neuroticism shows an inverse relation with dispositional optimism and with the psychological components of QoL.

Structural Equation Modeling

Moderation Model

One of the hypotheses in this study is that neuroticism and extraversion could be associated with general well-being and mental health, and dispositional optimism would moderate this relation. Controlling for socio-demographic and clinical variables, the model fit statistics for the moderator model $\chi^2_{(50)} = 99.2623, p < 0.001$, $\text{CFI} = 0.97$, $\text{RMSEA} = 0.044$, $\frac{\chi^2}{df} = 1.98$) indicate a reasonable fit, but the effect of the interaction between personality traits and dispositional optimism was not statistically significant. These results reveal that the effect of neuroticism and extraversion on psychological components of quality of life is not moderated by dispositional optimism.

Mediation Model

The hypothesis that dispositional optimism mediates the relation between neuroticism/extraversion and QoL was evaluated by two models. According to Baron and Kenny criteria mediation [1], one model tests the significant effect of personality traits on QoL and the other model tests the role of dispositional optimism as a mediator between these variables. It means that the relationship between personality traits and quality of life is reduced, or is no longer significant in controlling for dispositional optimism. Both models are controlled for socio-demographic and clinical variables. The results of first model showed a statistically significant direct effect between personality traits (neuroticism and extraversion) and QoL (general well-being and mental health), $\chi^2_{(25)} = 53.847, p < 0.01$, $\text{CFI} = 0.98$,

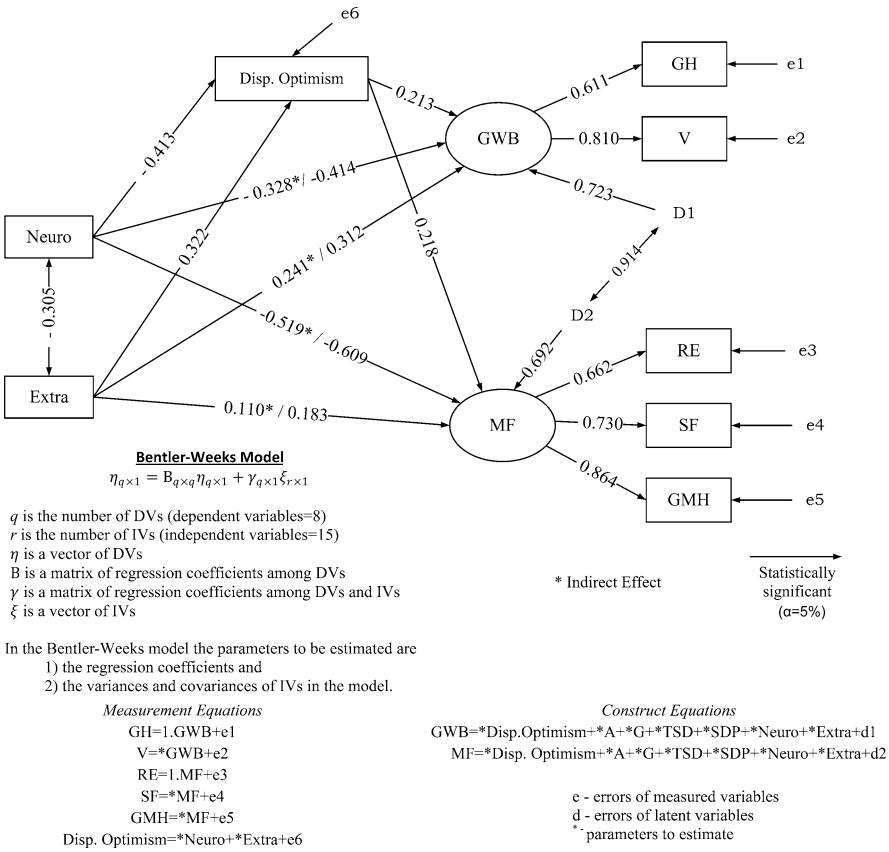

Fig. 1 SEM with standardized factor loadings and Bentler–Weeks Model [22]

RMSEA $= 0.048$, $\frac{\chi^2}{df} = 2.2$. The second (overall) model fits reasonably ($\chi^2_{(33)} = 76,645$, $p < 0.01$, CFI $= 0.98$, RMSEA $= 0.051$, $\frac{\chi^2}{df} = 2.3$) (Fig. 1). We found that patients with higher levels of neuroticism felt less satisfied with general well-being, general mental health, and they are less optimistic; extraverted patients tend to have a better general well-being, better general mental health, and tend to be more optimistic. After controlling for neuroticism/extraversion, we found that optimistic patients tended to be more pleased with general well-being and had a positive impact on general mental health. After controlling for dispositional optimism, it was revealed that the strength of the relation between personality traits and QoL was significantly reduced (evaluated by Sobel test [6, 15, 17]), demonstrating a partial mediation effect of dispositional optimism between neuroticism/extraversion and general well-being/general mental health. Results also reveal a statistically significant impact of perception of severity of disease ($b = -0.615$), showing that high levels imply poor general well-being.

3 Conclusions

Using SEM, the purpose of this study was to evaluate, in Portuguese patients with chronic diseases, the moderator or mediator effect of dispositional optimism between personality traits (neuroticism and extraversion) and psychological components of quality of life (general well-being and mental functioning).

Extensive research has documented links between personality traits and quality of life [10]. In chronic diseases, optimism has been shown to predict diverse and important benefits for individuals, affecting the adjustment of the quality of life [7, 10, 19, 25]. In recent years, this personality attribute has been the one most examined in the disease-adjustment relationship [21]. It influences several factors which help in the selection of coping strategies for adjustment to life.

The study results are consistent with those of other authors [25]. We found that neuroticism and extraversion influence dispositional optimism and, indirectly, affect the psychological components of quality of life. Dispositional optimism is more likely to play a mediating, rather a moderating, role in the association between personality traits and quality of life. An optimistic attitude, in other words, the expectation that good things will happen seems to promote a higher level of psychological well-being. This being so, interventions in patients should aim to achieve and maintain a level of optimism that would help to facilitate and improve the quality of life. The present study is accordance with the theory and research and makes it possible to clarify the relationship between optimism and quality of life, concluding that dispositional optimism is a mediator between neuroticism/extraversion and general well-being/mental health in Portuguese patients with chronic disease.

References

1. Baron, R.M., Kenny, D.A.: The moderator mediator variable distinction in social psychological-research-conceptual, strategic, and statistical considerations. J. Personal. Soc. Psychol. **51**(6), 1173–1182 (1986)
2. Byrne, B.M.: Structural Equation Modeling with EQS: Basic Concepts, Applications and Programming, 2nd edn. Lawrence Erlbaum Associates, London (2006)
3. Chang, E.: Does dispositional optimism moderate the relation between perceived stress and psychological well-being? a preliminary investigation. Personal. Individ. Differ. **25**, 233–240 (1990)
4. Costa, P.T., McCrae, R.R.: NEO-PI-R Professional Manual. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI). Psychological Assessment Resources (1992)
5. Dubayova, T., Nagyova, I., Havlikova, E., Rosenberger, J., Gdovinova, Z., Middel, B., et al.: Neuroticism and extraversion in association with quality of life in patients with Parkinson's disease. Qual. Life Res. **18**(1), 33–42 (2009)
6. Frazier, P.A., Tix, A.P., Barron, K.E.: Testing moderator and mediator effects in counseling psychology research. J. Counsel. Psychol. **51**(1), 115–134 (2004)
7. Grant, H., Higgins, E.T.: Optimism, promotion pride, and prevention pride as predictors of quality of life. Personal. Soc. Psychol. Bull. **29**(12), 1521–1532 (2003)

8. Gustavsson-Lilius, M., Julkunen, J., Hietanen, P.: Quality of life in cancer patients: The role of optimism, hopelessness, and partner support. Qual. Life Res. **16**(1), 75–87 (2007)
9. Hirsch, J.K., Wolford, K., LaLonde, S.M., Brunk, L., Morris, A.P.: Dispositional optimism as a moderator of the relationship between negative life events and suicide ideation and attempts. Cognit. Ther. Res. **31**(4), 533–546 (2007)
10. Ho, M.Y., Cheung, F.M., Cheung, S.F.: The role of meaning in life and optimism in promoting well-being. Pers. Indiv. Differ. **48**(5), 658–663 (2010)
11. Keller, S.D., Ware, J.E. Jr, Bentler, P.M., Aaronson, N.K., Alonso, J., Apolone, G., et al.: Use of structural equation modeling to test the construct validity of the SF-36 health survey in ten countries: results from the IQOLA project. J. Clin. Epidemiol. **51**(11), 1179–1188 (1998) doi: 10.1016/S0895-4356(98)00110-3
12. Kline, R.B.: Principles and Practice of Structural Equation Modeling, 3rd edn. The Guilford, New York (2011)
13. Krampe, H., Bartels, C., Victorson, D., Enders, C.K., Beaumont, J., Cella, D., et al.: The influence of personality factors on disease progression and health-related quality of life in people with ALS. Amyotroph. Lateral Scler. **9**(2), 99–107 (2008)
14. MacCallum, R.C., Austin, J.T.: Applications of structural equation modeling in psychological research. Ann. Rev. Psychol. **51**, 201–226 (2000)
15. Macckinnon, D.P., Fritz, M.S., Williams, J., Lockwood, C.M.: Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. Behav. Res. Methods **39**(3), 384–389 (2007)
16. Mulaik, S.: A Linear Causal Modeling with Structural Equations. Taylor & Francis Group, London (2009)
17. Preacher, K.J., Leonardelli, G.J.: Calculation for the Sobel Test: An interactive calculation tool for mediation tests Retrieved. Http://quantpsy.org/sobel/sobel.htm. Accessed 25 March 2010
18. Russell, D.W., Kahn, J.H., Spoth, R., Altmaier, E.M.: Analyzing data from experimental studies: A latent variable structural equation modeling approach. J. Counsel. Psychol. **45**(1), 18–29 (1998)
19. Scheier, M.F., Carver, C.S.: Effects of optimism on psychological and physical well-being – theoretical overview and empirical update. Cognit. Ther. Res. **16**(2), 201–228 (1992)
20. Scheier, M.F., Carver, C.S., Bridges, M.W.: Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem) – a reevaluation of the life orientation test. J. Personal. Soc. Psychol. **67**(6), 1063–1078 (1994)
21. Stanton, A.L., Revenson, T.A., Tennen, H.: Health psychology: Psychological adjustment to chronic disease. Ann. Rev. Psychol. **58**, 565–592 (2007)
22. Tabachni, B., Fidell, L.: Using Multivariate Statistics, 3rd edn. HarperCollins College Publishers, New York (1996)
23. Vittadini G.: Indeterminacy problems in the Lisrel Model. Multivariate Behav. Res. **24**(4), 397–414 (1989)
24. Ware, J.E., Snow, K.K., Kosinski, M., Gandek, B.: SF-36 Health Survey: Manual and Interpretation Guide. The Health Institute. New England Medical Center, Boston (1993)
25. Wong, W.S., Fielding, R.: Quality of life and pain in Chinese lung cancer patients: Is optimism a moderator or mediator? Qual. Life Res. **16**(1), 53–63 (2007)

Longitudinal Modeling of Job Satisfaction Using Portuguese Data from the European Community Household Panel

Maria de Fátima Salgueiro and Patrícia Serra

Abstract

This chapter proposes modeling job satisfaction over time using a Latent Growth Curve Model. Portuguese data from the European Community Household Panel are used. Job satisfaction is modeled longitudinally, considering a two factor structure: satisfaction with the extrinsic factors (material aspects of work) and satisfaction with the intrinsic factors (qualitative aspects of work). The characteristics of the complex survey design are taken into account in the estimation procedure. The methodological approach allows to capture differences between and within individuals in their job satisfaction trajectories.

1 Introduction

Latent Growth Curve Modeling (LGCM) has recently increased popularity as a statistical technique for longitudinal data analysis. LGCM allows the investigation of interindividual differences in change over time. Time is incorporated in the model by imposing some restrictions on the factor loadings in a latent variable model. Growth parameters are specified as latent variables (intercept and slope), and the shape of the growth trajectory, linear or quadratic, is defined by the number of latent variables included in the model.

One of the most important advantages of LGCM is that it is carried out within the Structural Equation Modeling (SEM) framework, sharing many of its strengths, but also some of its weaknesses. Although large sample sizes are required, LGCM allows to evaluate the adequacy of model fit by using model fit indices. LGCM also

M. de Fátima Salgueiro (✉) · P. Serra
Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit, Avenida das Forças Armadas, 1649-026, Lisboa, Portugal
e-mail: fatima.salgueiro@iscte.pt; patricia.serra@iscte.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_5, © Springer-Verlag Berlin Heidelberg 2013

has the ability to account for measurement error, by incorporating latent variables in the model, and the capacity to deal with missing data [4].

This chapter proposes using a LGCM with a linear growth trajectory to model job satisfaction over time. The concept of job satisfaction is considered multidimensional and, therefore, a two-factor structure is used to account for both the extrinsic and the intrinsic factors underlying job satisfaction [5]. The extrinsic factors are related to financial and other material rewards, such as pay, promotion, and security. The intrinsic factors are related to symbolic or qualitative aspects of work, for example, the relation with supervision and the work itself.

This chapter uses Portuguese data from the *European Community Household Panel* (ECHP). Salgueiro [6] used data from the ECHP to describe and explain growth trajectories of a latent life satisfaction factor, measured by four indicators, without taking into account the complex survey design. This chapter proposes a bidimensional LGCM to model job satisfaction growth trajectories and accounts for the complex structure of the survey. Information regarding job satisfaction available in the ECHP includes satisfaction with earnings, job security, type of work, working hours, working times, and working conditions/environment. The statistical software Mplus 5.0 is used to estimate the parameters of the proposed model. The structure of this chapter is as follows: Sect. 2 characterizes the sample under analysis and describes the complex survey design of ECHP. Section 3 provides a brief description of the proposed bivariate LGCM. Section 4 summarizes the main results of the statistical modeling undertaken, and Sect. 5 presents the discussion of the results and main conclusions.

2 The Data

The ECHP is a longitudinal survey, representative of the households in 12 European countries and conducted between 1994 (wave 1) and 2001 (wave 8) [2]. Information is collected both at the household level and at the individual level. The ECHP collects data on perceptions of job satisfaction and on demographic characteristics, including employment, income, health, and education. The current study uses the Portuguese data available for the period 1994–2001. The following questions are considered as indicators of job satisfaction: How satisfied are you with your present job in terms of earnings? How satisfied are you with your present job in terms of job security? How satisfied are you with your present job in terms of type of work? How satisfied are you with your present job in terms of working hours? How satisfied are you with your present job in terms of working times (day time, night time, shifts, etc.)? and How satisfied are you with your present job in terms of working conditions/environment? Respondents have been asked to rate their satisfaction level using a *Likert*-type scale from 1—not satisfied, to 6—totally satisfied.

The subsample of 2,502 respondents aged between 16 and 65 years old, employed throughout the period under analysis, working more than 15 h per week, with valid answers in all job satisfaction variables, and for whom longitudinal survey

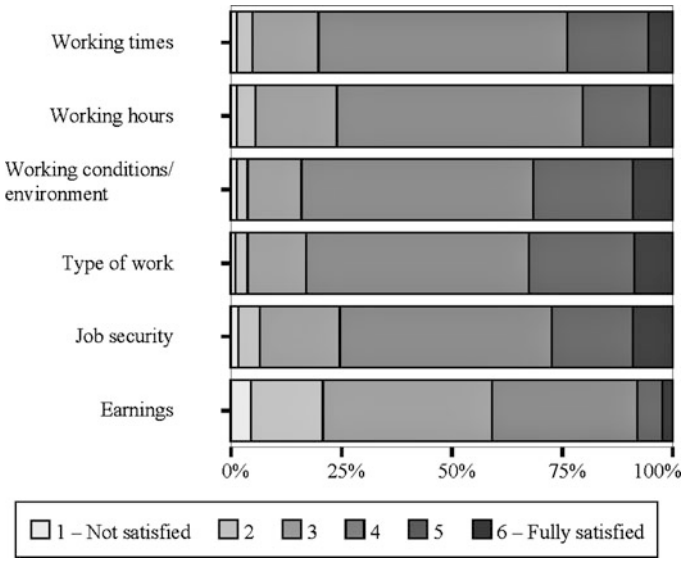

Fig. 1 Distribution of the job satisfaction variables in year 1994

weights were available, was considered. Figure 1 summarizes the distribution of the responses to the job satisfaction questions for the year 1994.

It is possible to conclude that the majority of the responses are at the level 4 of satisfaction for almost all items, except for satisfaction with earnings. In fact, the distribution of satisfaction with earnings is the most different: there is a higher percentage of responses in the lower levels of satisfaction (1, 2, and 3). The variables satisfaction with type of work and satisfaction with working conditions have the highest percentage of respondents with higher levels of satisfaction (4, 5, and 6).

Table 1 presents the means of the job satisfaction variables for the years 1994–2001. Satisfaction with earnings has the lowest average values for all waves, followed by satisfaction with working hours. For the whole period under analysis respondents are, on average, more satisfied with the type of work and with the working conditions/environment (these two items have the highest mean satisfaction levels). Some changes over time on the average levels of job satisfaction are also noticeable.

The average age of the respondents is 37 years old, and there are 1,595 males in the sample. As far as marital status is concerned, 72.5 % of the respondents are married. In terms of education, only 5.4 % have the recognized third level of education (ISCED 5–7), 10 % have the second stage of secondary level education (ISCED 3), and 82.4 % have less than the second stage of secondary level education (ISCED 0–2). Regarding the perceived health status, 73.3 % of the respondents consider it very good or just good, and only 26.7 % rate it as fair, bad, or very bad. In terms of the activity sector, 77.2 % of the respondents work on the private sector, while 22.8 % work on the public sector.

Table 1 Means of the job satisfaction variables from 1994 (wave 1) to 2001 (wave 8)

		1994	1995	1996	1997	1998	1999	2000	2001
Extrinsic	Earnings	3.264	3.219	3.197	3.215	3.329	3.359	3.398	3.434
Factors	Job security	4.037	4.087	4.044	4.115	4.142	4.169	4.170	4.188
Intrinsic	Type of work	4.194	4.237	4.218	4.213	4.244	4.223	4.263	4.262
Factors	Working hours	3.949	3.929	3.920	3.915	3.951	3.947	3.955	3.985
	Working times	4.036	4.104	4.056	4.027	4.069	4.041	4.075	4.084
	Working conditions	4.193	4.185	4.190	4.176	4.195	4.191	4.193	4.198

The ECHP adopts a complex multistage sampling scheme for collecting data. This scheme is defined once, for each household, at wave 1. In Portugal, the scheme has involved three different stages. First, the representativeness of the whole sample was defined a priori, concerning the seven Portuguese regions known as NUTS II. A first sample of 1,143 statistical sections[1] was then selected from those regions. In the third stage several households were selected from each statistical section, ensuring that each household had equal probability of being selected. The seven regions correspond to the seven strata of the stratification variable, and the 1,143 statistical sections are the primary sampling units (clusters). ECHP data also include longitudinal weights, computed for each individual that has participated in the panel since wave 1. Longitudinal weights are adjusted at each wave, to take account of dropouts and previous wave nonresponse.

3 Latent Growth Curve Models with Multiple Indicators

LGCM allows the investigation of interindividual differences in change over time. Unconditional LGCM describes the observed heterogeneity between individuals, and conditional LGCM aims at explaining it. If one wants to model the longitudinal trajectory of a latent factor measured by several observed indicators, an LGCM with multiple indicators (also known as a second-order LGCM) has to be considered.

Let y_{ijt} denote the repeated measures for individual i, at $t = 1, 2, \ldots, T$ consecutive time points of J ($j = 1, 2, \ldots, J$) observed ordinal variables. All manifest variables have the same number of categories and were measured at the same time points. Let y^*_{ijt} be the continuous latent variable determining the observed ordinal variable y_{ijt}. The connection between y_{ijt} and the underlying variable y^*_{ijt} is given by the auxiliary threshold equation:

$$y_{ijt} = c \text{ when } \tau_{c-1} < y^*_{ijt} \leq \tau_c,$$

[1]The statistical sections are contiguous geographic areas generally belonging to the same *freguesia*, although never exceeding the limits of the municipality to which they belong. In Portugal each statistical section has around 300 households.

where $c = 1, 2, \ldots, C$ is the total number of ordered categories of variable y_{ijt}; τ_{c-1} and τ_c are the lower and upper thresholds for category c with $\tau_0 = -\infty$ and $\tau_C = \infty$; and the $C-1$ finite threshold values are ordered from lowest to highest (for further details see [1]). Note that it is a common practice for LGCM with ordinal repeated measures to impose invariant thresholds for the repeated measures over time.

The value for indicator j for individual i at time t is given by:

$$y^*_{ijt} = \nu_{jt} + \Lambda_{jt}\eta_{it} + \upsilon_{ijt}, \tag{1}$$

where ν_{jt} is the intercept for indicator j at time t; Λ_{jt} is the factor loading for the indicator j at time t; η_{it} is the repeated latent variable for individual i and time t; and υ_{ijt} is the disturbance for the individual i at time t and indicator j (with mean zero and uncorrelated with the other elements of the equation). The disturbances are assumed to be normally distributed, with the structure of the variance matrix depending on which terms are allowed to be correlated.

The growth trajectory for the latent factor is given by:

$$\eta_{it} = \alpha_i + \lambda_t\beta_i + \zeta_{it}, \tag{2}$$

where α_i is the random intercept for individual i; β_i is the random slope for individual i; λ_t is the time trend variable; and ζ_{it} is the disturbance for individual i and time t, with mean zero, variance Ψ_t,, and uncorrelated with α_i, β_i, and λ_t. A usual coding convention for LGCM with linear growth is that $\lambda_t = 1 - t$ (setting $\lambda_1 = 0$ allows the intercept to reflet the mean of the trajectory at the initial time point, and the other values to incorporate the linear trajectory that was defined).

The random effects (intercept and slope) are given by:

$$\begin{aligned} \alpha_i &= \mu_\alpha + \zeta_{\alpha i}, \\ \beta_i &= \mu_\beta + \zeta_{\beta i}, \end{aligned} \tag{3}$$

where μ_α and μ_β are the mean of the intercept and the mean of the slope across all individuals. The model assumes that $\zeta_{\alpha i} \sim N(0, \Psi_\alpha)$ and $\zeta_{\beta i} \sim N(0, \Psi_\beta)$. The parameters of special interest in the model are the means and the variances of the random effects (intercept and slope) and the structure of the factor loadings.

Equations presented so far refer to modeling trajectories of a single process. According to Rose [5] job satisfaction is a bivariate concept because there are intrinsic and extrinsic factors, related among themselves, underlying job satisfaction. Confirmatory factor analysis (CFA) was used to validate a model with two correlated factors (at each time point): intrinsic job satisfaction (measured by four indicators) and extrinsic job satisfaction (with two indicators). Since there are two correlated dimensions of job satisfaction, their growth trajectories should be modeled simultaneously and a bivariate LGCM should be considered. Thus, two sets of Eqs. (2), (3) and (4) are required. Additional parameters of interest in the

model are the residual covariances among the random effects in the two processes (elements in $\Psi_{\alpha\beta}$).

ECHP data were collected under a complex survey design. The sampling literature recommends considering such design when performing model estimation. In Mplus, the estimator used for models with ordinal variables, for which data were collected using a complex survey design, is the Weighted Least Squares Means and Variance adjusted estimator (WLSMV). Such estimator belongs to a class of generalized least squares estimators and accounts for the sampling design features by considering weighted sample mean vectors and weighted sample covariance matrices. Standard errors are estimated by using a sandwich Huber-White type estimator, allowing for clustering and stratification in the sample. These methods produce consistent standard error estimates without making any distributional assumption and are also known as asymptotically distribution-free methods [3].

Measures of model fit used in this chapter include the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the chi-square test of model fit (χ^2) and corresponding degrees of freedom (df) [1].

4 Results for the Proposed LGCM

Figure 2 displays the path diagram of the unconditional bivariate second-order LGCM proposed in this chapter to describe the joint growth trajectories of the two dimensions of job satisfaction considered in this chapter: perceived satisfaction with intrinsic factors (JS Int t) and with extrinsic factors (JS Ext t). Recall JS Ext is measured by the ordinal variables earnings and job security. JS Int is measured by three ordinal variables: type of work; working hours; and times and conditions. The dashed rectangle in the top left corner of Fig. 2 illustrates the presence of the ordinal observed variable y_{jt}, and of the underlying continuous latent variable y^*_{jt} in the model. Because the two job satisfaction dimensions are correlated, their growth trajectories are modeled simultaneously—the induced correlation structure among intercepts and slopes in the two processes is depicted in Fig. 2.

Some methodological options were made: (1) the error terms of the repeated measures have been allowed to be correlated over time; (2) the scaling factors of the repeated measures have been fixed to 1 on the first wave and set free in the following waves; and (3) the thresholds for each repeated observed variable have been defined as time invariant [7].

The following goodness of fit measures were obtained: $\text{CFI} = 0.892$, $\text{TLI} = 0.957$, $\text{RMSEA} = 0.048$, and $\chi^2 = 200.066$ with estimated df = 31, suggesting that the model fits the data well. The estimates (and corresponding standard errors) obtained for the main parameters of interest are presented in Table 2. Note that because we are modeling the trajectories of latent factors, and given the convention adopted to include time in the model, the means of the intercepts are both zero, and no longer a parameter of interest in the model. The average growth rate trajectory is significant and positive for both processes (0.018 for extrinsic and 0.022 for

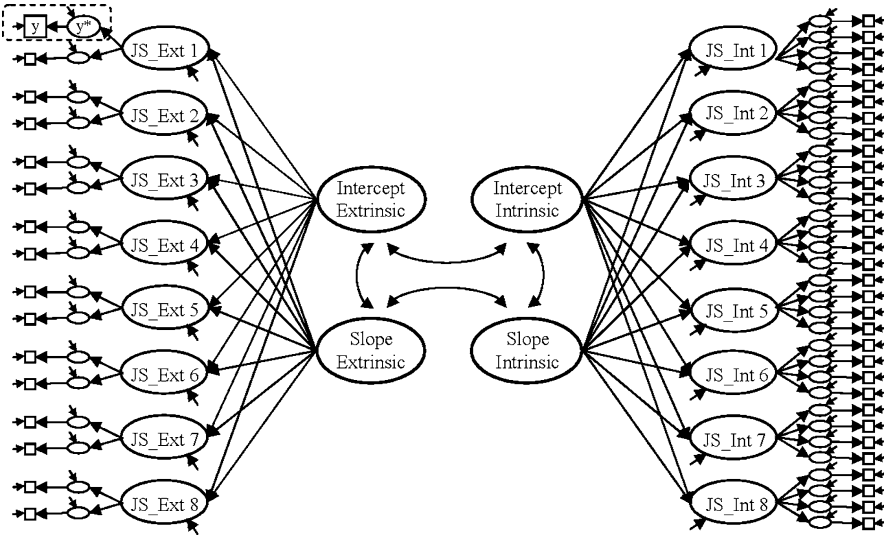

Fig. 2 The unconditional bivariate second-order LGCM proposed to describe joint growth trajectories of extrinsic and intrinsic job satisfaction

Table 2 Unconditional multiprocess second-order LGCM estimates

		Extrinsic	Intrinsic
Mean	Slope (μ_β)	0.018 (0.005)	0.022 (0.005)
Variance	Intercept (ψ_α)	0.161 (0.025)	0.309 (0.025)
	Slope (ψ_β)	0.003 (0.001)	0.008 (0.001)
Covariances	Intecept_Ext—Slope_Ext	−0.009 (0.003)	
	Intercept_Int—Slope_Int	−0.025 (0.004)	
	Intercept_Ext—Intercept_Int	0.139 (0.014)	
	Slope_Ext—Slope_Int	0.002 (0.000)	
	(parameters in $\Psi_{\alpha\beta}$)	0.002 (0.000)	

intrinsic JS), indicating that there is a significant increase in the mean extrinsic and intrinsic satisfaction levels from 1994 to 2001. Estimates for the variances of the intercept and of the slope are both significant. It is therefore possible to conclude that individuals vary significantly both in their mean levels of extrinsic and intrinsic satisfaction in 1994 and in their extrinsic and intrinsic satisfaction growth trajectories.

Regarding the estimates of the residual covariances between intercepts and slopes in both processes, it is possible to conclude that the mean levels of satisfaction in 1994 are significantly associated with the average satisfaction growth trajectories from 1994 to 2001. The estimates for both extrinsic and intrinsic factors are negative (-0.009 and -0.025, respectively) , suggesting that higher mean values in 1994 are associated with less pronounced average growth trajectories. The mean level of

satisfaction with extrinsic factors in 1994 is significantly and positively associated with the mean level of satisfaction with intrinsic factors in 1994 (estimated value of 0.139). Analogously, there is a significantly positive residual covariance between the slopes of both dimensions of satisfaction (estimated value of 0.002). As expected, the two processes are correlated and their growth trajectories should be jointly modeled.

5 Conclusion

This chapter has proposed modeling trajectories of two dimensions of job satisfaction using an unconditional bivariate second-order LGCM. The ordinal nature of the observed data has been considered. The complex survey design of the ECHP has been taken into account when estimating the proposed model, and a robust WLS estimator available in Mplus has been used.

Results suggest that there was a significant increase in the mean satisfaction levels, with both the intrinsic and the extrinsic job satisfaction dimensions, between 1994 and 2001. It is also possible to conclude that individuals differ significantly, both in their satisfaction levels in 1994 and in the way their satisfaction levels change over time.

The variability that was found between individuals has to be explained. A conditional LGCM should be considered, including in the model some covariates that could account for the variability found between individuals, namely socio-demographic characteristics. Thus, further work includes considering time-invariant and time-varying covariates in the proposed LGCM.

Acknowledgements This research was supported by the Fundação para a Ciência e a Tecnologia, grant PTDC/GES/72784/2006.

References

1. Bollen, K.A., Curran, P.J.: Latent Curve Models: A Strutural Equation Prespective. Wiley, New Jersey (2001)
2. EPUNet: ECHP User guide. http://epunet.essex.ac.uk/ECHP_USER_GUIDE-06-04-204b.pdf (2012). Accessed May 2012
3. Muthén, L., Muthén, B.O.: Mplus User's Guide, 5th edn. Muthén & Muthén, Los Angeles, CA (1998–2007)
4. Preacher, K.J., Wichman, A.L., MacCallum, R.C., Briggs, N.E.: Latent Growth Curve Modeling. Sage, Thousand Oaks (2008)
5. Rose, M.: Good deal, bad deal? Job satisfaction in occupations. Work Employ. Soc. **17**, 503–530 (2003)
6. Salgueiro, M.F.: Latent growth curve modelling of Portuguese life satisfactions trajectories. Int. J. Acad. Res. **2**, 188–194 (2010)
7. Serra, P.: Modelação longitudinal da satisfação laboral em Portugal. Master Thesis. Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal (2010)

Estimation of Underrepresented Strata in Preelection Polls: A Comparative Study

João Figueiredo and Pedro Campos

Abstract

In this work we aim at increasing the utility of a preelection poll, by improving the quality of the vote share estimates, both at macro and micro level. Three different methodologies are applied with that purpose: (1) polls aggregation, using existing auxiliary polling; (2) application of multilevel regression methods, using the multilevel structure of the data; and (3) methods of small area estimation, making use of auxiliary information through the application of the Empirical Best Linear Unbiased Prediction (EBLUP). These methods are applied to real data collected from a survey with the aim of estimating the vote share in the Portuguese legislative elections. When auxiliary information is required, we concluded that polls aggregations and EBLUP have to be applied with caution, since this information is extremely important for a good application of these models to the data set and to obtain good reliable forecasts. On the other hand, if auxiliary information is not available or if it is not of good quality, then multilevel regression can and should be seen as a safe alternative to obtain more precise estimates, either at the micro or macro level. Besides, this is the method which further improves the precision of the estimates. In the presence of good auxiliary information, EBLUP proved to be the method with greater proximity with real values.

J. Figueiredo
Faculdade de Economia do Porto, R. Roberto Frias, 4200 Porto, Portugal
e-mail: jfabelha@gmail.com

P. Campos (✉)
LIAAD – INESC Porto, L.A and Faculdade de Economia da Universidade do Porto,
LIAAD: R. Ceuta, 118, 4050, Porto, Portugal
e-mail: pcampos@fep.up.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_6, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

One of the dilemmas of preelection polls is the difficulty of estimating the proportion of voters in a specific political party, especially in situations where the number of sampling units is too small for assuring estimates with low sampling error. Assuming the population is divided into L strata, we propose three approaches for the problem of estimating, in a preelection poll, the proportion of voters, p, in a political party: aggregation of preelection polls, by applying a variation of the work of Erikson et al. [2]; multilevel regression methods, following the work of Gelman et al. [5, 6], Lax [8], Park et al. [11], Snijders and Bosker [15], and Leeuw and Meijer [9]; Small Area Estimation (SAE), using the estimator EBLUP (Empirical Best Linear Unbiased Predictor) at the area level [3, 14]).

The aim of this work consists in comparing the results from the application of these methods, which constitutes the originality of the paper, as these techniques have never been simultaneously compared and applied in the scope of preelection polls [4]. By applying these techniques, we seek not only to improve the quality of the estimates at the global level, but also to improve the quality of the estimates at other levels of disaggregation of data, including the district level, avoiding, ultimately, the need of developing a new preelection poll. We focused on the problem of estimating the proportion of votes in a given political party and implemented the three methodologies referred above. Aggregation of Polls and EBLUP are methods that have to be applied with caution as regards to the use of auxiliary information. The selection of inadequate information might lead to a deterioration of the results in forecasts provided by a survey carried out in isolation. By inadequate information we mean surveys that are made in different conditions and do not provide comparable information. Multilevel regression is a safe alternative to obtain more precise estimates, either at the micro or macro level when auxiliary information is not available or if it is of no good quality (e.g., when auxiliary information is not available at all or when it contains a small number of sampling units in some strata). Although all the models produce estimates which are closer to the real value than the existing poll, the method that obtains the closer results when compared to the results of the direct estimators from already existing polls is the EBLUP. Multilevel regression produces the estimates with greater quality.

The rest of this chapter is structured as follows: in Sect. 2 we present the three methods mentioned above. Data and methodology are introduced in Sect. 3. Sections 4, 5, and 6 contain respectively, results, conclusions, and future work.

2 Methods for Dealing with Underrepresented Strata

In Portugal, the election of the deputies to the Portuguese Parliament is made at the district level, here considered as the strata. Although it is often desirable to use a stratified sampling scheme to estimate the vote share in preelection polls,

Table 1 Sampling errors (B) of the preelectoral Poll of 2005, by district; the overall sample size $n = 997$

District	Aveiro	Beja	Braga	Bragança	Castelo Branco	Coimbra
Error of the poll	0.1187	0.25	0.1118	0.2425	0.2132	0.1507
District	Évora	Faro	Guarda	Leiria	Lisboa	Portalegre
Error of the poll	0.25	0.1601	0.2294	0.1525	0.0685	0.2887
Total sample	Porto	Santarém	Setúbal	Viana do Castelo	Vila Real	Viseu
0.0319	0.0765	0.1525	0.114	0.1826	0.2	0.1562

an underrepresentation of some sampling strata can lead to difficulty in obtaining accurate estimates. Information for this study is based on an existing poll of 2005, obtained by a stratified random sampling at the level of Portuguese mainland (corresponding to the whole country except the autonomous regions—Azores and Madeira), and includes 997 individuals who are selected randomly from people with fixed line telephones at home, and proportionally to the number of voters in the districts. Let p_{est} be the vote share at the population level. Therefore, to estimate p_{est} with a maximum sampling error of B and a degree of confidence $(1-\alpha)\times 100\,\%$ (α being usually 5 %), it involves an estimator that satisfies the condition:

$$P\left(|\hat{p}_{\text{est}} - p| \leq B\right) = 1-\alpha, \quad 0<\alpha<1 \tag{1}$$

which implies:

$$P\left(\hat{p}_{\text{est}} - B \leq p \leq \hat{p}_{\text{est}} + B\right) = 1-\alpha \tag{2}$$

Usually B is computed as $z\sqrt{V\hat{p}_{\text{est}}}$, where z is the two-tailed value of the Normal distribution corresponding to an area of $1-\alpha$. One can verify, by Table 1, that in our poll, at the district level, B varies between 0.0685 (district of Lisbon) and 0.2887 (district of Portalegre). Although it is not explicit in Table 1, we know that the smaller the samples of the districts, the larger are the errors of the estimates. The errors in some districts such as Beja, Bragança, Portalegre, and others are very high. Although there is not a consensual limit for the highest level B can assume, estimates based on errors higher than 0.20 are definitely not acceptable, due to small accuracy.

2.1 Aggregation of Preelection Polls

This method was first used by Erikson et al. [2], who aggregated 122 polls made by Columbia Broadcasting System News/New York Times (CBS/NYT). These polls were conducted from 1976 to 1988 in order to estimate the US public opinion. The basic idea of this method is to obtain more precise estimates by obtaining larger samples from the aggregation of the polls. However, if the opinion poll is not stable over time, applying this method will produce worse results than any single

poll in a specific time period. We can formalize this method as follows: consider a population with I strata, for which we have J independent samples for all segments or only for some. Let n_{ij} be the sample size for stratum i of the j^{th} poll. After the aggregation of the polls, the total sample size for stratum i is the sum of samples from each survey and is given by $n_i = \sum_{i=1}^{J} n_{ij}$. Let $\hat{p}_{ij}$ be an unbiased estimator of p_i obtained from the j^{th} poll. After the aggregation of the polls, the estimated proportion of elements in stratum i who vote in the political party A is given by:

$$\hat{p}_i = \frac{1}{n_i} \sum_{i=1}^{J} n_{ij} \hat{p}_{ij} \tag{3}$$

The aggregation of the total sample surveys is given by:

$$n = \sum_{i=1}^{I} n_i = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \tag{4}$$

The estimate of the proportion in the population who vote in the political party A is given by:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{I} n_i \hat{p}_i \tag{5}$$

2.2 Multilevel Regression

According to Gelman et al. [5], multilevel modeling can be viewed in two equivalent forms: (1) a generalization of linear regression, where the ordinate at the origin and the slopes may vary by group; (2) equivalently, we can see the multilevel modeling as a regression that includes a categorical explanatory variable representing the group where the observation belongs. Take the binary variable, y, representing the vote for a particular party, defined in the space $S = \{0, 1\}$, with 1 being *yes* and 0 being *no*. We can fit a logistic regression model for the probability of a *yes*, π_j, for respondents in the category j using a *logit* function: $logit\,(\pi_j) = X_j \beta$, where X is the matrix of covariates, X_j is the j^{th} column of X, and β is the vector of unknown regression parameters. The hierarchical model allows a partial aggregation of variable categories through the definition of a mixed effects model. The model can then be written in a standard form of a hierarchical logistic regression model, under a matrix form as $logit\,(p_i) = (X\beta)_i$, being p_i the parameter of the distribution of y_i, $y_i \approx Bernoulli\,(p_i)$ and β being the vector of the regression parameters, with $\beta \approx N\,(0, \Sigma_\beta)$. X is the matrix of covariates. We write the β vector as a collection of parameters $(\alpha, \gamma_1, \ldots, \gamma_m)$, where α is a subvector of disaggregated coefficients. Σ_β^{-1} is a diagonal matrix with 0 in each element of α, followed of σ_m, for each element of γ_m, and for each m. We use the notation p_i, for the probability corresponding to unit i, as distinguished from π_j, the probability corresponding to

the aggregate category j. Each γ_m, for $m = 1, \ldots, M$, is a subvector of coefficients γ_{km}, for each we adjust the hierarchical model: $\gamma_{km}^{ind} \approx N\left(0, \sigma_m^2\right)$, $k = 1, \ldots, K_m$, being K the number of coefficients for each γ_m. To obtain the parameters of the model, π_j, and the other parameters described above, Gelman et al. [5] uses the following Bayesian simulation: (1) perform Bayesian inference for the regression coefficients, β, and the hyperparameters σ_m through the data set; (2) for each of the J categories in the population, calculate: $p_j = logit^{-1}\left(X\beta\right)_j$; (3) calculate inferences for the parameters, summing up all the $N_j p_j$'s in each state.

Markov Chain Monte Carlo methods are used in WinBUGS [10] to simulate the models. Criteria to stop the simulation are based on the potential of reduction range factor, R, for each parameter, which is approximately the square root of the variance of the mixture of all Markov chains, $\hat{V}$ divided by the average within-chain variance. Gelman et al. [5] suggest taking $R < 1.1$ to indicate convergence. $\hat{V}$ and W are weighted by d. R is defined as follows:

$$R = \sqrt{\frac{(d+3)\hat{V}}{(d+1)W}} \tag{6}$$

After fitting the model $logit\left(p_i\right) = \left(X\beta\right)_i$ and obtaining the predictions for p_i, we can do a post-stratification to correct the sample of possible deviations from the population distribution, since the proportional sample strata size may differ from the real proportion in the population, which are known. For Gelman et al. [6], the purpose of post-stratification is the adjustment of the difference between sample and population. In the notation that follows, lowercase is used for variables that are observed only in the sample and uppercase is used for variables that are observed in the sample and are also known for the entire population. We assume we have an array of X indicators, whose joint distributions are known in the population, and a variable, y, whose distribution we are interested in estimating. We consider that the variables are discrete and list the categories of X as post-stratification cells, j with N_j elements in the population and n_j elements in the sample. With this notation, the total population is given by $N = \sum_{j=1}^{J} N_j$ and the sample size is given by $n = \sum_{j=1}^{J} n_j$. In the model of post-stratification, it is implicit that the data are collected by a simple random process in each of the J strata. The case of stratification (from which samples are taken by the sampling process of the poll) is nothing more than a case of post-stratification of the way it is formulated. It is assumed that the strata totals, N_j, are known for each category j. These categories include all cross-classifications of the predictors X. The population mean of any response, $\theta = E\left(Y\right)$, can be written as follows:

$$\theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} N_j} \tag{7}$$

where θ_j is the mean of the stratum j. Therefore, the post-stratification estimate $\hat{\theta}^{ps}$ is given by

$$\hat{\theta}^{ps} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j} \tag{8}$$

The simplest case used in this chapter is the post-stratification total, where the estimate for each stratum j is the mean estimate of the stratum, $\hat{\theta}_j = \overline{y}_j$, thus:

$$\hat{\theta}^{ps} = \frac{\sum_{j=1}^{J} N_j \overline{y}_j}{\sum_{j=1}^{J} N_j} \tag{9}$$

2.3 Empirical Best Linear Unbiased Predictor

SAE methods can be used in the cases where the sample size is too small for some geographical areas/domains to allow reliable estimates [1]. One of the most used classification of SAE methods is proposed by Rao [14], who suggests the following types of estimators: direct, or design-based, indirect, and model-based (synthetic and combined). In general, direct survey estimators are based on the domain-specific sample data. Indirect estimators are formed by information from other domains in order to borrow strength from related areas to increase the precision of the estimates. Synthetic estimators use information out of the domain in order to artificially increase the sample size, thus contributing to a gain in precision of the estimates. Finally, combined estimators use two components of estimation—a design-based and a model-based. Usually it is a combination between direct and synthetic estimators. Small area indirect estimators are often based on area level random effects models [12]. Under this class of models, when only aggregate specific covariates are available, the Best Linear Unbiased Predictor (BLUP) is obtained under the assumption of uncorrelated random area effects [3]. The empirical version of BLUP, the EBLUP, takes advantage of the between small area variation, especially when this is not large relative to the within small area variation. In this chapter we use EBLUP, a small area estimator of the combined type that estimates μ_d (unknown residuals at the population level) based on ϵ_d, (residuals obtained by fitting the model). There are essentially two different ways of applying the EBLUP, depending on the type of auxiliary information available: individual level (for which a link between direct and auxiliary information must exist at individual level) and area level (for which this link must exist only at the area/domain level). In this work we applied the EBLUP at area level, since the auxiliary information is aggregated to the domains (districts) for which we want to calculate the estimates. Let Y be the variable of interest and X the variable containing auxiliary information. Then, we have: $\hat{\overline{Y}}_i^{EBLUP} = \gamma \hat{\overline{Y}}_i^{Direct} (1-\gamma)\, \overline{X}^T \hat{\beta}$ where γ is a constant used to combine the direct and the synthetic estimators. To estimate the synthetic part of the model, $(1-\gamma)\, \overline{X}^T \hat{\beta}$, we have: $\hat{\overline{Y}}_i = \mu_i + \epsilon i$, where $\epsilon_i \approx N\left(0, \sigma^2\right)$. Being the true mean

of the area i given by: $\mu_i = \beta_0 + \beta_1 \overline{X}_i + Z$, where $\mu_i \approx N\left(0, \sigma_\mu^2\right)$. μ_i represents the random effects and Z reflects the structure of random effects. To estimate the variance of the model, Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods are used. We assume $\hat{\overline{Y}}_i \approx N\left(\mu; V\right)$. The estimates of small areas are given by: $\hat{\overline{Y}}_i = \hat{\beta} X_i + \hat{\mu}_i$. The variability of the estimator is given by the mean square error (MSE), calculated as follows: $MSE\left[\hat{\overline{Y}}_i\right] \approx G1 + G2 + 2G3$. $G1$ represents the uncertainty of the estimation of Y, $G2$ represents the uncertainty of the estimation of β, and $G3$ represents the uncertainty of the estimation of $\hat{\sigma}_\mu^2$.

3 Data and Methodology

To compare these techniques, we took the problem of estimating the proportion of votes in the Portuguese Socialist (PS) Party and applied the three methods using real data from a survey conducted by a survey company—IPOM (Instituto de Pesquisa de Opinião e Mercado)—in 2005, to estimate the PS vote share in the parliamentary elections stratified by district. Auxiliary data from two other surveys having the same goal (but at the district level) were used: the vote share at 1999 and 2002. For the application of multilevel regression we used R [13] and WinBUGS [10], with the code provided by Gelman [5], and adapted by the authors. Additional data, with the number of individuals older than 18 years, from Portuguese census [7] were used as an approximation of the numbers of voters, to produce a post-stratification of the results. The convergence of this model was obtained with 100 thousand iterations. The model is defined as:

$$
\begin{aligned}
&Pr\left(Y_i = 1\right) = \\
&\quad = logit^{-1}\left(\beta^0 + \beta^{sexo} \cdot sexo_i + \beta^{prof} \cdot prof_i + \beta^{sexo \cdot prof} \cdot prof_i\right. \\
&\qquad \left. + \alpha_{k[i]}^{idade} + \alpha_{l[i]}^{esc} + \alpha_{k[i],l[i]}^{idade \cdot esc} + \alpha_{j[i]}^{dis}\right), \qquad i = 1, 2, \ldots, 719, \\
&\alpha_{j[i]}^{dis} \approx N\left(\beta^{v \cdot prev}, \sigma_{distrito}^2\right), \\
&\alpha_k^{idade} \approx N\left(\mu^{idade}, \sigma_{idade}^2\right), \qquad k = 1, \ldots, 6, \\
&\alpha_l^{esc} \approx N\left(\mu^{esc}, \sigma_{esc}^2\right), \qquad l = 1, \ldots, 8, \\
&\alpha_{k,l}^{idade \cdot esc} \approx N\left(\mu_{idade \cdot esc}, \sigma_{idade \cdot esc}^2\right), \qquad k = 1, \ldots, 6,\ l = 1, \ldots, 8,
\end{aligned}
$$

where β^0 is the constant term for the ordinate at the origin; β^{sexo} and β^{prof} denote the coefficients for sex and profession, respectively; $sexo_i$ represents the sex of individual i, $prof_i$ is the indicator for the profession of the individual i, $\beta^{sexo \cdot prof}$ is the coefficient for the interaction between gender and profession, $\alpha_{k[i]}^{idade}$, $\alpha_{l[i]}^{esc}$, $\alpha_{j[i]}^{dis}$ and $\alpha_{k[i],l[i]}^{idade \cdot esc}$ are the coefficients for the origin of the coordinate variables age, school, district, and the interaction between age and schooling, respectively. $v.prev$

Table 2 Estimates for the year 2005 of the models computed by the three methods—best estimates, measured by the MAE are highlighted

District	Polls aggregation	Multilevel regression	EBLUP	Poll	Real
Aveiro	**0.4357**	0.3836	0.4169	0.3256	0.4109
Beja	0.2143	0.4054	0.4004	0.2143	0.5101
Braga	0.4697	0.4807	0.4382	0.4697	0.4542
Bragança	0.4000	0.4614	0.4279	0.4000	0.4207
Castelo Branco	0.8125	0.6273	0.4889	0.8125	0.5598
Coimbra	0.5769	0.5288	0.4541	0.5769	0.4541
Évora	0.6667	0.5616	0.4673	0.6667	0.4968
Faro	0.4444	0.4746	0.4344	0.4444	0.4933
Guarda	0.6667	0.5627	0.4673	0.6667	0.467
Leiria	0.2308	0.3493	0.4028	0.2308	0.3558
Lisboa	0.4351	0.437	0.4331	0.4351	0.4411
Portalegre	0.875	0.6102	0.4982	0.8750	0.5480
Porto	**0.4853**	0.4331	0.434	0.4412	0.4853
Santarém	0.5862	0.5352	0.4554	0.5862	0.4614
Setúbal	0.4737	0.4771	0.4388	0.4737	0.4371
Viana do Castelo	0.5000	0.5063	0.4427	0.5000	0.4197
Vila Real	0.4118	0.4647	0.4296	0.4118	0.4384
Viseu	0.6207	0.5601	0.4605	0.6207	0.404
Mainland prediction	0.4811	0.4637	0.4375	0.4655	0.4514
Mean absolute error	0.1163	0.0571	0.0306	0.1221	

is the mean of the results obtained by the party PS in the two previous elections. The variances of the variables district, age, education, and interaction between age and schooling were denoted, respectively, for $\sigma^2_{distrito}$, σ^2_{idade}, σ^2_{esc}, and $\sigma^2_{idade \cdot esc}$. To the parameters μ_{idade}, μ_{esc}, and $\mu_{idade \cdot esc}$ noninformative prior distributions are assigned. The application of EBLUP was implemented in R [13], and information of the results for the 1999 election was used as auxiliary information. Information of 2002 was not used here, since it is not correlated to the variable of interest. The model can be defined as: $\hat{y}_i = 0.368645 + 0.1479796 x_i$, where $\hat{y}_i$ is the model prediction of the proportion of votes in the PS in district i.

4 Results

Results are compared in two ways: either using the proximity of the estimate (measured by the mean absolute error—MAE—between the real values and the estimates, by districts, for each method) or taking into account the quality of the estimates (using a measure of relative precision—the CV, coefficient of variation). In terms of proximity, the EBLUP estimator produced the nearest forecasts in 11 of the 18 districts (cells with shadow in the corresponding column in Table 2). At mainland level, the method that obtains the best results is the multilevel regression (see

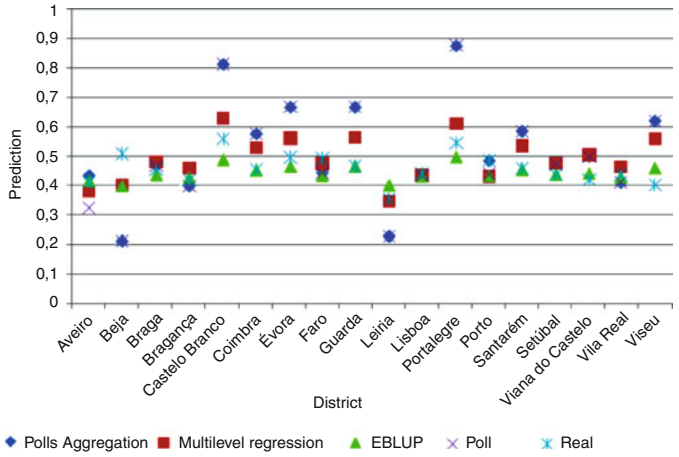

Fig. 1 Estimates of the three models, by district for the year 2005

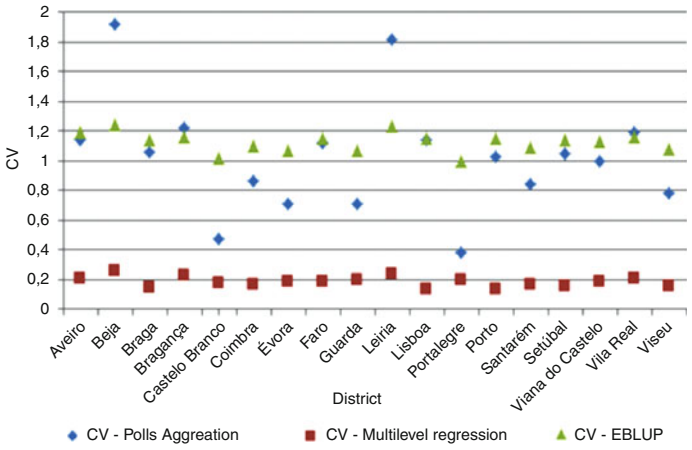

Fig. 2 Coefficients of variation (CVs) for the three models, by district, for the year 2005

corresponding column in Table 2). The aggregation of polls produces the nearest results in two districts, only. This method can be a valid alternative when applied with caution in relation to sampling methods of various surveys taken at different time steps. As we can see in Table 2, and in Fig. 1, the EBLUP prediction for Coimbra and the Polls Aggregation forecast for Porto are the same as the observed values (what happened just by chance) . To analyze the quality of the estimates, we plotted in Fig. 2 the coefficients of variation (CV). Multilevel regression is the method with lower CVs. When compared to the existing poll (column *Poll*), this method can improve the estimates of the survey both at the district level and at the mainland level (Table 3).

Table 3 Comparison of the coefficients of variation (CVs) for the three models and the real poll, by district

	Polls aggregation	Multilevel regression	EBLUP	Poll
Aveiro	1.14	0.21	1.20	1.46
Beja	1.99	0.26	1.25	1.99
Braga	1.07	0.15	1.14	1.07
Bragança	1.29	0.23	1.17	1.29
Castelo Branco	0.50	0.18	1.02	0.50
Coimbra	0.87	0.17	1.1	0.87
Évora	0.74	0.19	1.07	0.74
Faro	1.14	0.19	1.15	1.14
Guarda	0.74	0.20	1.07	0.74
Leiria	1.85	0.24	1.24	1.85
Lisboa	1.14	0.14	1.15	1.14
Portalegre	0.40	0.20	1.00	0.40
Porto	1.03	0.14	1.15	1.13
Santarém	0.85	0.17	1.10	0.85
Setúbal	1.06	0.16	1.14	1.06
Viana do Castelo	1.02	0.19	1.13	1.02
Vila Real	1.23	0.21	1.16	1.23
Viseu	0.80	0.16	1.08	0.80

5 Conclusions

Aggregation of Polls and EBLUP are methods that have to be applied with caution as regards to the use of auxiliary information in order to improve the quality of the estimates, measured by the coefficient of variation. The choice of appropriate information may result in substantial improvements in the quality of the estimates, while the choice of inappropriate information may lead to a worsening of results with respect to forecasts provided by a poll taken in isolation. As such, the choice of information should be made by the statistician, conscientious, reasoned, and attentive to any detail that might inspire a change in the tendency to vote. If auxiliary information is not available or if it is not of good quality, then multilevel regression can and should be seen as a safe alternative to obtain more precise estimates, either at the micro or macro level.

6 Future Work

As future work we intend to apply these techniques to future elections and other applications/surveys for other purposes, having dichotomous variables as well as other levels of aggregation, as, for example, a survey at district level with substrata at the subdistrict level (municipalities). We will also try to investigate possible

combinations of three methods. The use of resampling techniques in order to obtain convergence of the series is also to be developed in future applications. Finally, we aim at developing strategies to produce confidence intervals for these methods, so it is possible to analyze interval estimates instead of point estimates.

References

1. Coelho, P.: Estimadores combinados para pequenos domínios. Revista de Estatística **2**, 23–43 (1996)
2. Erikson, R.S., Wright, G.C., McIver, J.P.: Statehouse Democracy: Public Opinion and Policy in the American States. Cambridge University Press, Cambridge (1993)
3. Fay, R.E., Herriot, R.A.: Estimate of income for small places: an application of James-Stein procedures to census data. J. Am. Stat. Assoc. **74**, 269–277 (1979)
4. Figueiredo, J.: Estimação em Estratos sub-representados no contexto das Sondagens Eleitoriais – Uma comparação de métodos. Msc Thesis, Faculdade de Economia da Universidade do Porto (2010)
5. Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, Cambridge (2007)
6. Gelman, A., Little, T.C.: Poststratication into many categories using hierarchical logistic regression. Surv. Methodol. **23**, 127–135 (1997)
7. INE - Instituto Nacional de Estatítica: Censos 2001 - XIV Recenseamento Geral da População e IV Recenseamento Geral da Habitação, Resultados Definitivos, Lisboa (2002)
8. Lax, P.: How should we estimate public opinion in the states? Am. J. Polit. Sci. **53**, 107–121 (2009)
9. Leeuw, J., Meijer, E. (eds.): Handbook of Multilevel Analysis. Springer, New York (2008)
10. Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D.: WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. **10**, 325–337 (2000)
11. Park, D., Gelman, A., Bafumi, J.: Bayesian multilevel estimation with poststratification: state-level estimates from national polls. Polit. Anal. **12**, 375–385 (2004)
12. Pratesi, M., Salvati, N.: Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Stat. Methods Appl. **17**, 113–141 (2008)
13. R Development Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing (2010)
14. Rao, C.R.: Small Area Estimation. Wiley, New Jersey (2003)
15. Snijders, T., Bosker, R.: An Introduction to Basic and Advanced Multilevel Modeling. Sage, Beverly Hills (1999)

Medication and Polymedication in Portugal

Alexandra Pinto, Teresa Rodrigues, Jorge Mendes, Fernando Bação, and Victor Lobo

Abstract

Portugal is a country with a high per capita consumption of medical drugs. High levels of medication implies not only risk to the patient but also a strong burden to the National Health Service (Serviço Nacional de Saúde—SNS). Polymedication, according to many authors, is the consumption of at least five different drugs. Polymedication can have serious consequences to the patient, if it is not well controlled. A project named MAISCINCO was created in order to make the population aware of the risks of polymedication. Our study used the data from the 4th National (Portuguese) Health Survey to estimate the consumption of multiple drugs, in particular polymedication, and to identify groups with greater risk of being polymedicated. We also aimed to represent geographically, at a NUTS II level, the prevalence of the more frequent kinds of drug consumption. As expected from the literature we found that women and older people have greater probability of being polymedicated, and this trend was

A. Pinto (✉) · T. Rodrigues
Laboratory of Biomathematics, Faculty of Medicine of Lisbon, Av. Professor Egas Moniz, 1649-028 Lisboa, Portugal
e-mail: apinto@fm.ul.pt; teresarodrigues@fm.ul.pt

J. Mendes
CEAUL, FCUL, University of Lisbon, and ISEGI, New University of Lisbon and Campus de Campolide, 1070-312 Lisboa, Portugal
e-mail: jmm@isegi.unl.pt

F. Bação
ISEGI, New University of Lisbon, Campus de Campolide, 1070-312 Lisboa, Portugal
e-mail: bacao@isegi.unl.pt

V. Lobo
ISEGI, New University of Lisbon, and CINAV – Portuguese Navy Research Center, and Escola Naval – Alfeite, 2810-001 Almada, Portugal
e-mail: vlobo@isegi.unl.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_7, © Springer-Verlag Berlin Heidelberg 2013

observed and is similar in all regions. On the other hand, people in Madeira, Açores, and Algarve tend to use less medication.

1 Introduction

Medical drugs are used as therapy to relieve pain, improve functional capacity and quality of life, and prolong survival [10]. The number of medical drugs used by the National Health System has been increasing in Portugal. Between 2002 and 2008, consumption of drugs (such as tranquilizers, hypnotics, sedatives, and antidepressants) increased approximately 31.6 %, in all NUTS II[1] regions [2]. The high consumption of drugs and the aging of the population rise serious problems for the SNS and its sustainability in a near future [1]. The chronic use of medication (for more than 3 months) and polymedication may cause risks to the patient. There is no consensus regarding the number of prescribed drugs to consider the patient as polymedicated. However, it is usual to consider a minimum of five [4, 6, 10].

The main factors of polymedication [6, 7, 9] may be considered:

- Multiple pathologies (most common in old people)
- Multiple prescribers (without coordination among them)
- Auto medication (over-the-counter products and alternative medicine)
- Medicine advertising
- Socioeconomic aspects

Polymedication must be carefully thought about by prescribers to minimize drugs interactions (changes in activities of a drug when it is taken simultaneously with food or other drugs). It is estimated that the risk of adverse reactions is 6 % when two drugs are taken simultaneously. The risk increases to 50 % when five drugs are administered [4]. The increasing prevalence of polymedication, especially in an aging population, is becoming such a serious problem that Portugal established a counseling program, MAISCINCO, that aims to inform people about the dangers of taking too many drugs [9]. Previous studies have shown that polymedication and chronic use of drugs are more prevalent in women and elderly [5–7, 10]. A study held by the Faculty of Pharmacy from University of Lisbon revealed that the elderly take an average of seven drugs a day [9]. Polymedication itself leads to an increased risk for the patient and many hospitalizations as well. In the USA, approximately 28 % of admissions are due to polymedication, which could probably be considered the fifth cause of death if it were properly accounted for.

2 Objectives

The aim of our study is to estimate the prevalence of medical drug consumption with prescription or without it. Only drugs mentioned in National Health Survey (NHS) will be studied. The purpose of this study is threefold: to identify (1) which are the

[1] Standard Nomenclature of Territorial Units for Statistics purposes.

groups with higher consumption of drugs, (2) which are the drugs most consumed, and (3) what is the prevalence of polymedication both in Portugal in general, and in the seven regions of NUTS II. Moreover, the objective is to investigate if gender and age are in fact potential risk factors associated with polymedication.

3 Material and Methods

3.1 NHS

In our study we used data from the 2005/06 NHS, in particular, sections of individuals socio-demographic characterization and drug consumption, during the 2 weeks before the survey. This survey is the fourth in Portugal Continental and the first to cover the autonomous regions of Açores and Madeira. The NHS is an instrument that deals with health conditions of the population (individuals living in their usual residence during the period of data collection: February 2005 to February 2006). It is important to emphasize that all the individuals living in collective dwellings (such as prisons, hospitals, and hotels), representing less than 2 % of the Portuguese population, were excluded. The NHS sample is a stratified two-stage cluster sample drawn from the Statistics Portugal 2001 Master Sample, which was established from the 2001 Population and Housing Census results. The Statistics Portugal 2001 Master Sample uses stratification over NUTS III[2] in order to assure sufficient spread over the whole country. Each stratum was partitioned in areas (groups of Census sections) containing at least 240 usual residences. The first stage corresponds to the systematic selection of areas with probability proportional to number of usual residences. The sample comprises a total of 1,408 areas. The NHS sample introduces the second stage in the selection procedure. Indeed, in the second stage, the residences were drawn from the 1,408 areas of 2001 Master Sample. All the individuals from each selected residence were surveyed. The sample comprises a total of 19,950 residences (41,193 individuals) divided by the seven NUTS II: Norte (2,604), Centro (3,048), LVT (3,328), Alentejo (3,045), Algarve (3,220), Açores (2,304), and e Madeira (2,401).

3.2 Statistical Analysis

The statistical analysis concerns point estimates as well as their standard error between brackets immediately after the point estimate.The estimates are computed using a final weight for each individual. The final weight is obtained as follows:

- *Initial weight*: ratio between 2005 population estimates and the number of respondent individuals.

[2] 30 subregions of NUTS II.

- *Final weight*: calibration of the initial weight using the known marginal totals (gender and age groups), in order to get estimated totals similar to population totals in the following cells: males and females for 1–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85 and more years old.

The final weights were provided by INE (National Institute of Statistics of Portugal) in order to compute estimates based on the Horvitz–Thompson estimator [8]. Due to the complex sample design and to the calibration of weights, there is no closed form to compute the variance estimates. Accordingly, a Jackknife procedure was used to compute the variance of each estimate [11]. To do so, we were given replica weights in order to produce estimates for each replica and then compute the variance of replica estimates. A significance level of 5 % for hypotheses testing (based on asymptotic properties) is used. Chi-square tests, Adjusted Residuals, and Mann–Whitney tests for independent samples were conducted. Notwithstanding these tests are designed for data collected using a simple random sample design. We assume that the calibration of survey design weights reduces the distortion introduced by the two-stage cluster sample design and the final p-values can be viewed with fair confidence. A map was made with the results obtained, in order to illustrate drug consumption mainly at NUTS II level. Statistical analysis was performed using the statistics software IBM PASW 18, R Software, and Microsoft Excel 2007.

4 Results

4.1 Prescribed and Non-prescribed Drug Consumption

In the survey, 52.3 % (0.0042) of individuals had taken prescribed drugs (including contraceptive pills, ointments, creams, lotions, injections, and vaccines) and 9.4 % (0.0012) of individuals took medication without prescription (including vitamins and minerals), in the 2 weeks preceding the interview. The simultaneous consumption of prescribed and non-prescribed drugs occurred in 5.0 % (0.00072) of individuals. On the other hand, 43.2 % (0.004) of individuals had no medication at all. Among those who had used prescribed drugs, 9.6 % (0.042) had also consumed drugs without prescription. We found out that 62.7 % (0.0451) of prescribed consumption and 56.7 % (0.051) of non-prescribed consumption concerned women.

4.2 Medication and Polymedication

About 5.3 % (0.00072) of individuals are polymedicated with prescribed drugs. When both drugs, prescribed and not prescribed, are taken into account, the percentage of polymedicated individuals raises to 6.0 % (0.00001).

All results presented forward are related to the total of drug consumption either prescribed or not (Prescribed + Non-prescribed).

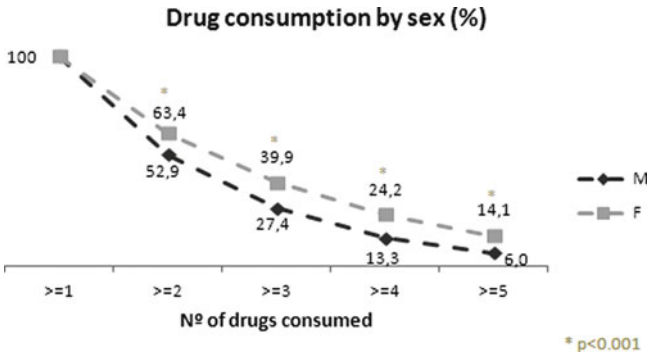

Fig. 1 Drug consumption by sex

4.3 Drug Consumption Profile by Sex

We analyzed the consumption by sex, considering classes of individuals that consumed at least 2, 3, 4, or 5 drugs. We found that consumption was significantly different (chi-square test; $p < 0.001$ for consumption ≥ 2, ≥ 3, ≥ 4, and ≥ 5 drugs). We also observed that women have higher consumption than men (Fig. 1). The difference between consumption in both genders peaked in three or more drugs, starting to decline from four or more.

4.4 Drug Consumption Profile by Age

The average consumption in individuals under 15 years old was 1.19 (0.021) drugs a day, rising to 3.14 (0.035), in people aged 65 and older. Our study confirms the strong trend of increasing use of drugs with age (Fig. 2). All age classes were significantly related with drug consumption (chi-square test; $p < 0.001$, for ≥ 2, ≥ 3, ≥ 4, and ≥ 5 drugs). Older individuals appeared associated with polymedication and younger individuals with lower consumption. About 53.2 % of polymedicated individuals are over 64 years old.

4.5 Drug Consumption Profile by Age and Sex

Age and gender appear as factors associated with drug use, as shown in Fig. 3. Consumption increases with age, with higher incidence in females. Significant differences were seen between gender, for all age groups (Mann–Whitney test; $p < 0.001$).

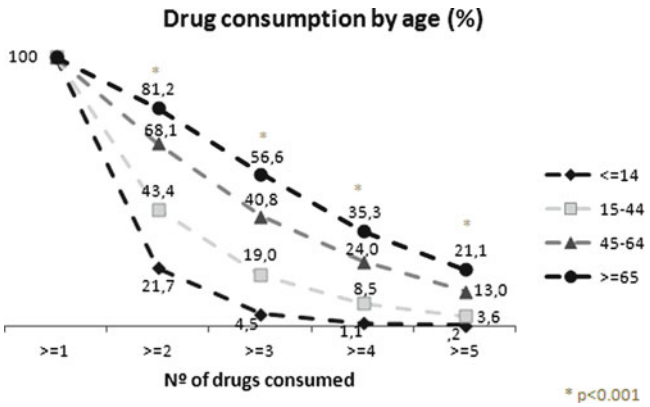

Fig. 2 Drug consumption by age

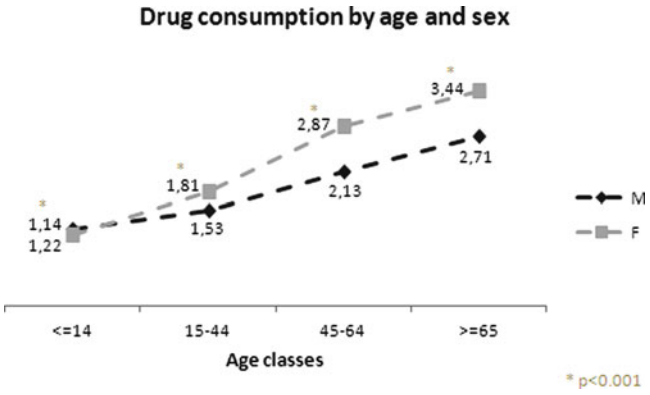

Fig. 3 Average drug consumption by sex in the four age classes

4.6 Consumption and Polymedication Profile by Region

Despite some differences, drug consumption showed the same trend in terms of age and gender distribution in all NUTS II regions. Significant statistical differences were seen in all levels of consumption (≥ 2, ≥ 3, ≥ 4, and ≥ 5) between each region (chi-square test; $p < 0.001$).

Lisboa e Vale do Tejo (LVT) are well known for having the highest consumption. On the other hand, Madeira is the region with the lowest intakes (Fig. 4).

When polymedication is analyzed among medicated individuals, Algarve and Açores follow Madeira. Norte, Centro, and Alentejo have an intermediate position with an identical polymedication profile between them (Fig. 4).

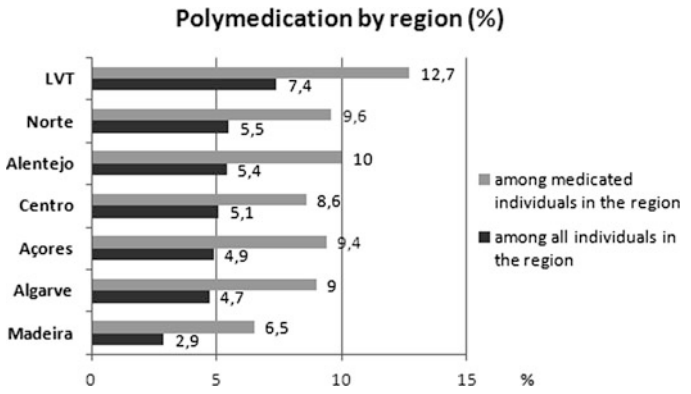

Fig. 4 Polymedication by NUTS II

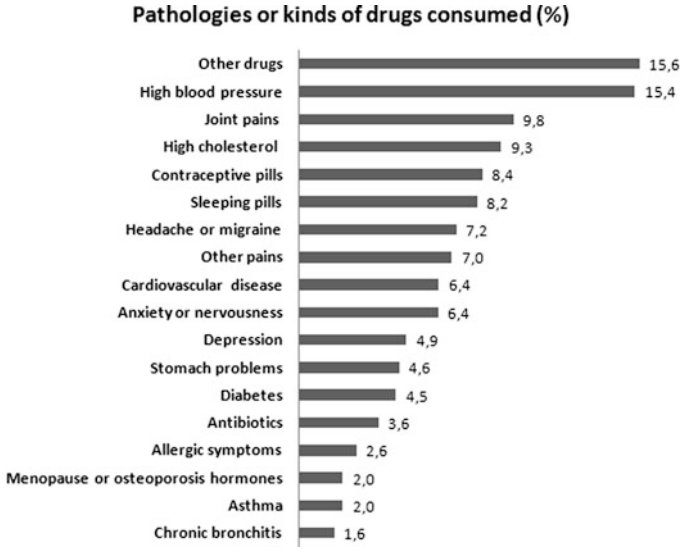

Fig. 5 Pathologies or kinds of drugs

4.7 Kinds of Drugs Consumed

"High blood pressure" was notoriously the most common pathology (17.7 %) that led to drug consumption (Fig. 5). Besides, "joint pains" (10.4 %), "high cholesterol" (10.0 %), "sleep disturbance" (8.3 %), and "cardiovascular disease" (7.4 %) frequently occurred. The use of drugs without specified reasons ("other drugs") accounted for 15.6 % of the total consumption.

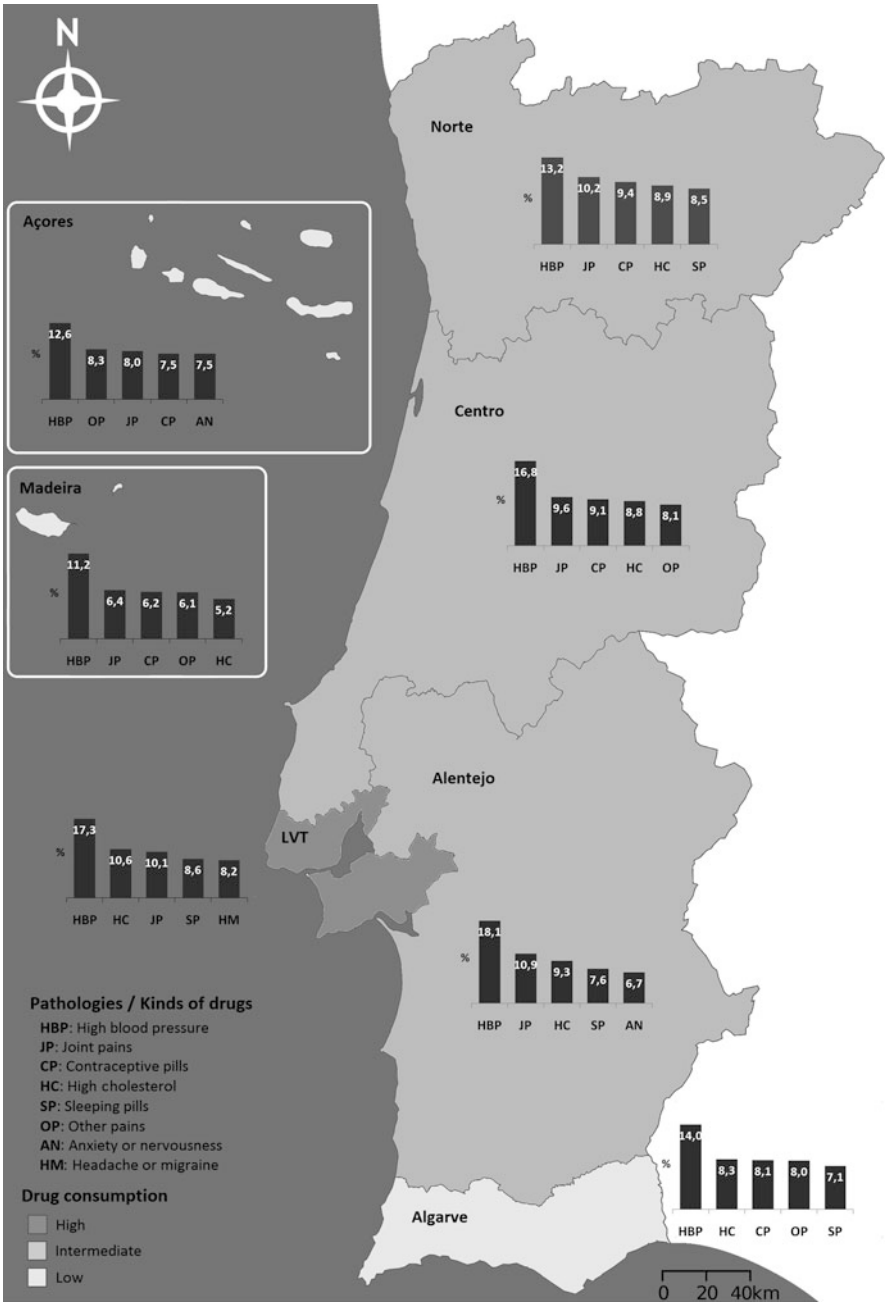

Fig. 6 Map of consumption level and the five most frequent pathologies or kinds of drugs in each NUTS II region

4.8 Kinds of Drugs Consumed per Region

For each NUTS II region, the five most frequent pathologies or kinds of drugs most consumed (excluding "other drugs") are presented in a map (Fig. 6). "Other drug" were not included because as they are not a specific pathology, they would add less interesting information. The colors on the map distinguish the consumption in different regions. Dark gray means that consumption is higher.

"High blood pressure" was the most frequent pathology in all regions. Alentejo reached the highest prevalence of "high blood pressure" (18.1 %). Madeira and Açores had the lowest percentage (11.2 % and 12.6 %, respectively). The other diseases or drugs more frequent were: "joint pains," "high cholesterol," "sleeping pills," "contraceptive pills," and "other pains."

5 Conclusions

From the study we concluded that:

- About half of the individuals consumed prescribed drugs.
- 9.4 % of the individuals used medication without prescription.
- Women and elderly were more likely to consume drugs either prescribed or not.
- Women and elderly were associated with polymedication.
- Men and young people were associated with lower drug consumption.
- LVT had the highest incidence of drug consumption, suggesting that individuals may have poorer health.
- Madeira had the lowest intakes of medication, suggesting that individuals may be healthier.
- High blood pressure was the most frequent cause of consumption in all NUTS II regions.

Our findings are consistent with what previous studies have shown: chronic use of medication and polymedication is more prevalent in women and elderly [3, 5, 10].

Our research points to the possibility that some differences may be associated with healthier lifestyles, demographic characteristics, medical care level and organization, in some regions such as Madeira, Açores, and Algarve. This shows a lower (ab)use of medication.

This work aims to warn about polymedication in Portugal and the need to study polymedication as a risk factor, especially in elderly people.

References

1. Abegunde, D.O., Mathers, C.D., et al.: The burden and costs of chronic diseases in low-income and middle-income countries. Lancet **370**, 1929–1938 (2007)
2. Alto Comissariado da Saúde: Consumo de medicamentos ansiolíticos, hipnóticos e sedativos e antidepressivos no mercado do SNS, em ambulatório http://www.acs.min-saude.pt/pns/acessibilidade-ao-medicamento/health-impact-assessment/ (2009). Accessed 05 April 2010

3. Antunes, P.: Medicação crónica numa lista de utentes. Rev. Port. Clin. Geral. **23**, 113–124 (2000)
4. Broeiro, P., Maio, I., et al.: Polifarmacoterapia: estratégias de racionalização. Rev. Port. Clin. Geral. **24**, 625–631 (2008)
5. Bushardt, R.L., Massey, E.B., et al.: Polypharmacy: misleading, but manageable. Clin. Interv. Aging **3**(2), 383–389 (2008)
6. Ferreira, R.: Consumo crónico de medicamentos na população de um Centro de Saúde. Rev. Port. Clin. Geral. **23**, 125–132 (2007)
7. Galvão, C.: O idoso polimedicado – estratégias para melhorar a prescrição. Rev. Port. Clin. Geral. **22**, 747–752 (2006)
8. INSA/INE: Inquérito Nacional de Saúde 2005/2006. Lisboa, INSA/INE (2010)
9. MAISCINCO: Polimedicação. http://www.maiscinco.com/a01.html (2008). Accessed 10 April 2010
10. Silva, P., Luís S., Biscaia A.: Polimedicação: um estudo de prevalência nos Centros de Saúde do Lumiar e de Queluz. Rev. Port. Clin. Geral. **20**, 323–336 (2004)
11. Wolter, K.M.: Introduction to Variance Estimation. Springer, New York (2007)

Part II

Estimation

Improved Shape Parameter Estimation in a Discrete Weibull Model

P. Araújo Santos and M.I. Fraga Alves

Abstract

A new shape parameter estimator for a discrete Weibull model is proposed. This estimator is based on an extension of the Khan et al. (IEEE Trans. Reliab. 38:348–350, 1989) method of proportions. Simulations are carried out to illustrate the improvement achieved in terms of bias and mean square error. The proposed estimator is applied on a financial dataset dealing with durations between violations in a quantitative risk management environment.

1 Introduction

In many reliability studies, data are measured as discrete random variables such as the number of copies made by a copying machine, number of cycles of a washing machine, and so on. Materials, equipment, devices, and structures are also frequently monitored only once per period rather than continuously, due to practical restrictions. In these types of reliability studies, the discrete Weibull model plays an important rule. For a survey on discrete lifetime distributions see Bracquemond and Gaudoin [2]. Moreover, the discrete Weibull can be applied to other problems, from political renewal analysis [6] to economic problems involving duration dependence. The "damaged goods" theory implies that the longer the period of unemployment,

P.A. Santos (✉)
Instituto Politécnico de Santarém, Departamento de Informática e Métodos Quantitativos, Escola Superior de Gestão e Tecnologia, Complexo Andaluz, Apartado 295, 2001-904 Santarém, Portugal
e-mail: paulo.santos@esg.ipsantarem.pt

M.I.F. Alves
Faculdade de Ciências, Departamento de Estatística e Investigação Operacional, Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
e-mail: isabel.alves@fc.ul.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_8, © Springer-Verlag Berlin Heidelberg 2013

the more likely the job seeker has some attribute that makes her unemployable, thus less likely to find a job. A discrete Weibull model with a shape parameter lower than one supports the "damaged goods" theory; see Lancaster [7] for econometric methods for the duration of unemployment. Other examples of duration dependence application are the study of speculative bubbles in stock markets [3] and backtesting Value-at-Risk [4].

In this work, we propose an improved estimator for the shape parameter of the discrete Weibull version of Nakagawa and Osaki [11], also known as type I discrete Weibull, with the following cumulative distribution function (cdf) and probability mass function (pmf):

$$F_D(d) = \begin{cases} 1 - q^{d^\theta}, & d = 1, 2, 3, \ldots \text{(jump points)} \\ 0, & d < 1 \end{cases} \tag{1}$$

$$f_D(d) = q^{(d-1)^\theta} - q^{d^\theta}, \qquad d = 1, 2, 3, \ldots \tag{2}$$

for $0 < q < 1$ and $\theta > 0$. The parameter q is the probability that the duration is greater than 1 and θ is the shape parameter: applying lifetime studies terminology, the distribution has increasing failure rate for $\theta > 1$, decreasing failure rate for $0 < \theta < 1$ and reduces to the geometric distribution when $\theta = 1$. If W is a continuous Weibull rv, then a type I discrete Weibull rv can be derived by time discretization $D = [W] + 1$, where $[W]$ denotes the integer part of W. Stein and Dattero [12] introduced a type II discrete Weibull and a type III was proposed by Padgett and Spurrier [8]. In a detailed study Bracquemond and Gaudoin [2] recommended the use of type I discrete Weibull and discuss the limitations of type II and type III.

The rest of this chapter is organized as follows. Section 2 provides a brief review of estimation methods. In Sect. 3 we present the new shape parameter estimator. In Sect. 4, through simulation experiments, we compare the performance of the new estimator with the method of moments and with the method of proportions. Finally, Sect. 5 presents an empirical application.

2 Estimation Methods

From the cdf (1), we have

$$\log[-\log(1 - F_D(d))] = \theta \log d + \log(-\log q).$$

Let $d^*_{1:v} < \ldots < d^*_{v:v}$, $v \leq n$, be the observed order statistics (o.s.'s) without ties of a sample $d_1, \ldots, d_n$ from the type I discrete Weibull distribution and $F_n(d) = n^{-1} \sum_{i=1}^{n} I_{\{d_i \leq d\}}$ the empirical cumulative distribution function (ecdf), associated with a random sample $D_1, \ldots, D_n$. If the points

$$\left\{ \log d^*_{i:v}, \log(-\log(1 - F_n(d^*_{i:v})) \right\}_{1 \leq i \leq v}$$

are approximately scattered around a straight line, it can be assumed that the underlying model is (1) and the parameters estimated by the Probability Plotting method, using these points.

Taking into account (2), we obtain the first two moments

$$\mu_1 = E[D] = \sum_{d=0}^{\infty}(d+1)q^{d^\theta} - \sum_{d=1}^{\infty} dq^{d^\theta} = 1 + \sum_{d=1}^{\infty} q^{d^\theta}$$

$$\mu_2 = 2\sum_{d=1}^{\infty} dq^{d^\theta} + E[D].$$

However, closed forms for these moments are not available. As in Khan et al. [9], for an observed sample $d_1, d_2, \ldots, d_n$, the Moments estimator, $\hat{\theta}_n^M$, is obtained by a numerical algorithm, which minimizes

$$M(q,\theta) = \left(\left(n^{-1}\sum_{i=1}^{n} d_i\right) - \mu_1\right)^2 + \left(\left(n^{-1}\sum_{i=1}^{n} d_i^2\right) - \mu_2\right)^2. \tag{3}$$

The method of proportions was proposed by Khan et al. [9]. Since $f_D(1) = 1-q$, the idea is to use the empirical frequency of observations greater than 1

$$\hat{q}^P = 1 - F_n(1). \tag{4}$$

In the same way, since $f_D(2) = q - q^{2^\theta}$ and using additionally, the empirical frequency of observations greater than 2

$$\hat{\theta}_n^P := \frac{1}{\log 2}\log\frac{\log(1-F_n(2))}{\log(1-F_n(1))}. \tag{5}$$

Denoting $p_d = f_D(d)$, $d = 1, 2, \ldots$, these authors noted that one may think of using other relation

$$\log q = k^{-\theta}\log(1 - p_1 - \cdots - p_k), \quad k = 1, 2, \ldots \tag{6}$$

In a simulation study it was concluded that the optimum choice of k is 2 and this lead us to (5).

The method of maximum likelihood considers the log-likelihood function $\log L(q,\theta; d_1, \ldots, d_n) = \sum_{i=1}^{n} \log\left\{q^{(d_i-1)^\theta} - q^{d_i^\theta}\right\}$; however, the ML equations $\partial L/\partial q = 0$ and $\partial L/\partial\theta = 0$ must be solved numerically; in this case, computational problems can occur, despite the good quality of estimates for high values of q [2]. Based on the method of proportions, approximate maximum likelihood estimators were proposed by Kulasekera [10], both for complete and type I censored data. Jazi et al. [5] present alternative approaches.

3 Improved Shape Parameter Estimation

The main drawback of Khan et al. [9] proportions estimator (5) is that it does not use all the observations but only a few of them, loosing a significant part of the available information. Here, we overcome this limitation.

Considering (6) for $d = 2, \ldots, k$ it is possible to write the following system of equations

$$\begin{cases} \theta &= \dfrac{1}{\log 2} \log \dfrac{\log(1 - F_D(2))}{\log q} \\ \cdots &= \qquad\qquad \cdots \\ \theta &= \dfrac{1}{\log k} \log \dfrac{\log(1 - F_D(k))}{\log q} \end{cases}$$

Now, adding all the equations and solving in order to θ

$$\theta = \frac{1}{(k-1)} \sum_{d=2}^{k} \frac{1}{\log d} \log \frac{\log(1 - F_D(d))}{\log q}. \tag{7}$$

Based on (7) we can derive estimators for θ. We suggest $c_d = 1$, for every d, and the estimation of $1 - \sum_{i=1}^{d} p_i$ and q by the ecdf, in the same line as Khan et al. [9], but considering all possible equations (6) to obtain one estimator.

Let $D_{1:n} \leq \cdots \leq D_{n:n}$ be the o.s.'s of a random sample $D_1, \ldots, D_n$ from the type I discrete Weibull distribution (1); whenever defined (i.e., $d_{n:n} > 2$), we propose the following class of ecdf based estimators

$$\hat{\theta}_n^{IP} := \frac{1}{(k-1)} \sum_{d=2}^{k} \frac{1}{\log d} \log \frac{\log(1 - F_n(d))}{\log(\hat{q}^P)}, \tag{8}$$

where k is chosen as large as possible for an observed data set (e.g., $k = d_{n:n} - 1$) and $\hat{q}$ is an estimator of q. With $\hat{q} = \hat{q}^P$ we denote (8) by $\hat{\theta}_n^{IP*}$. Since $F_n(d)$ are consistent estimators of $\sum_{i=1}^{d} p_i$, for $d = 2, \ldots, k$, $\hat{\theta}_n^{IP}$ is a consistent estimator of θ. Finally, notice that $F_n(k)$ is an estimator of $P[D < d_{n:n}]$ with complete data and with type I censored data, allowing us to use (8) in both cases.

4 Simulations

Here, we compare the moments estimator $\hat{\theta}_n^M$ (3), the proportions estimator $\hat{\theta}_n^P$ (5), and the proposed estimator $\hat{\theta}_n^{IP*}$. For the simulation study we have used the R language [13]. We present the simulated mean values and root of mean square errors (rmse) for $\theta = 0.5, 1, 1.5$ and $q = 0.5, 0.8$, using 5,000 simulations in each sample size ($n = 10, \ldots, 100$). It is possible to calculate estimates both with (5) and (8),

Table 1 Frequency of excluded samples

Sample	$\theta = 0.5$		$\theta = 1$		$\theta = 1.5$	
size n	$q = 0.5$	$q = 0.8$	$q = 0.5$	$q = 0.8$	$q = 0.5$	$q = 0.8$
10	0.0088	0.1020	0.0538	0.1094	0.2284	0.1072
20	0.0000	0.0116	0.0028	0.0114	0.0482	0.0116
30	0.0000	0.0022	0.0000	0.0018	0.0114	0.0002
40	0.0000	0.0000	0.0000	0.0004	0.0022	0.0000

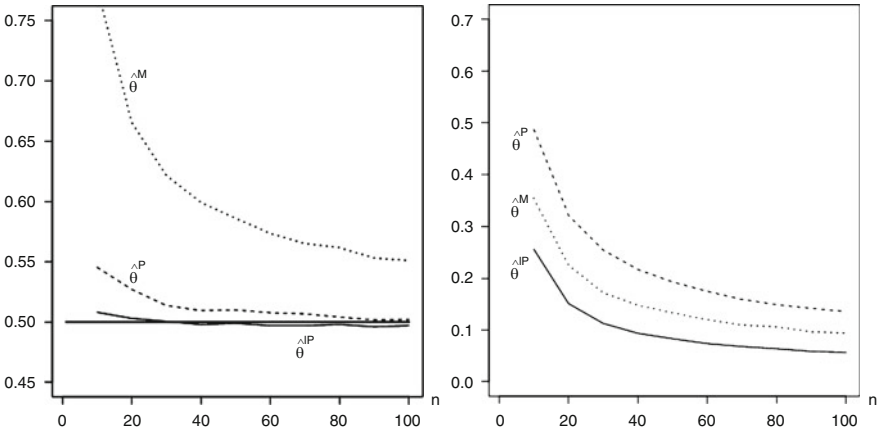

Fig. 1 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.5$ and $\theta = 0.5$ (decreasing failure rate)

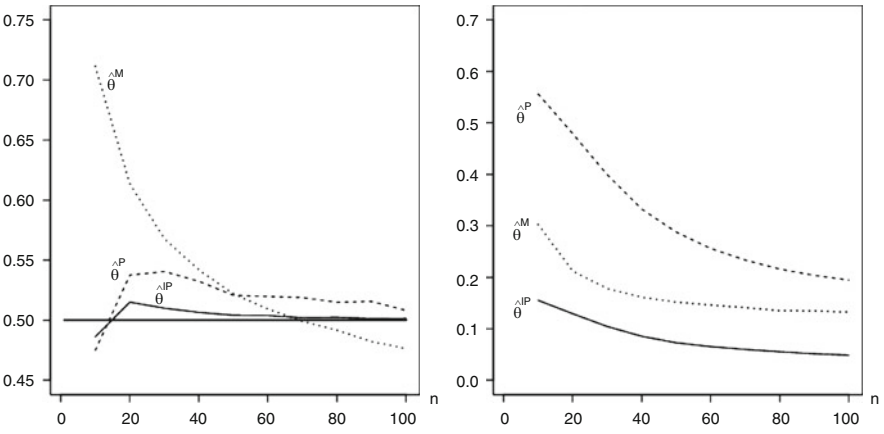

Fig. 2 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.8$ and $\theta = 0.5$ (decreasing failure rate)

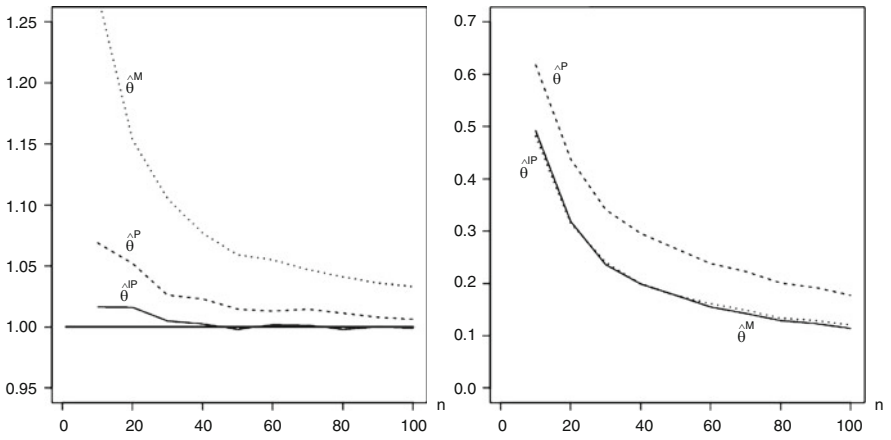

Fig. 3 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.5$ and $\theta = 1$ (geometric distribution)

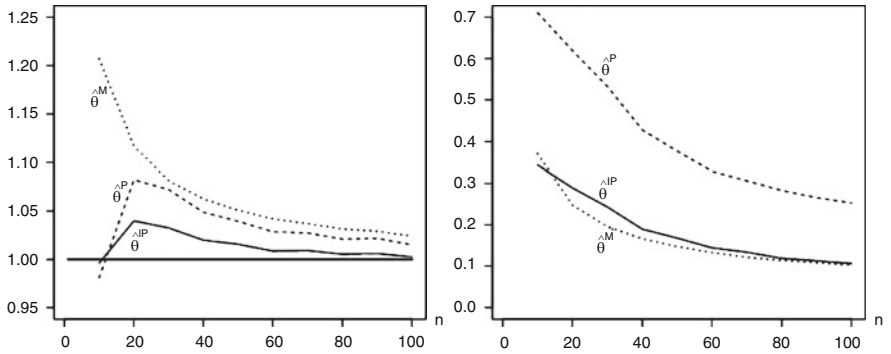

Fig. 4 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.8$ and $\theta = 1$ (geometric distribution)

only if $d_{1:n} = 1$ and $d_{n:n} > 2$. Based on these conditions, some samples were excluded (see Table 1). In terms of bias, for all cases, the estimator $\hat{\theta}_n^{IP*}$ performs much better than the other estimators under study. In terms of rmse, for $\theta < 1$ (decreasing failure rate), the estimator $\hat{\theta}_n^{IP*}$ performs better. For $\theta \geq 1$ (increasing failure rate or geometric distribution), the performance in terms of rmse of the moments estimator (3) and $\hat{\theta}_n^{IP*}$ is almost the same or in some cases ($q = 0.8$) (3) performs slightly better. The comparisons between the moment and proportion estimators are similar to the results obtained in simulations by Khan et al. [9] (Figs. 1–6).

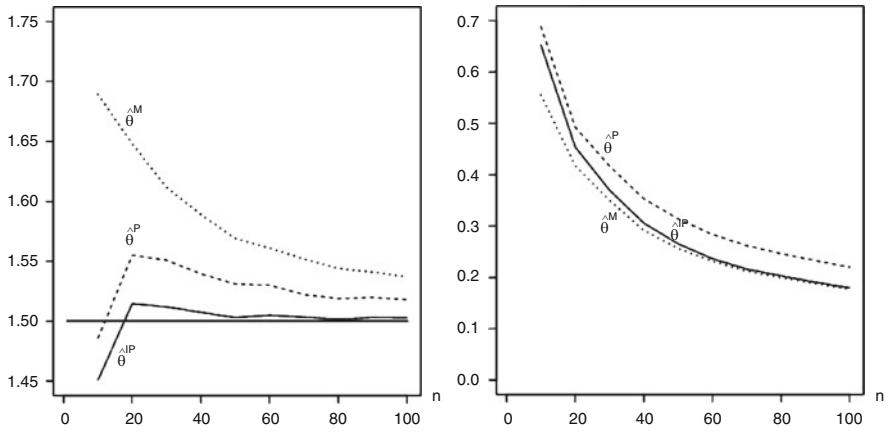

Fig. 5 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.5$ and $\theta = 1.5$ (increasing failure rate)

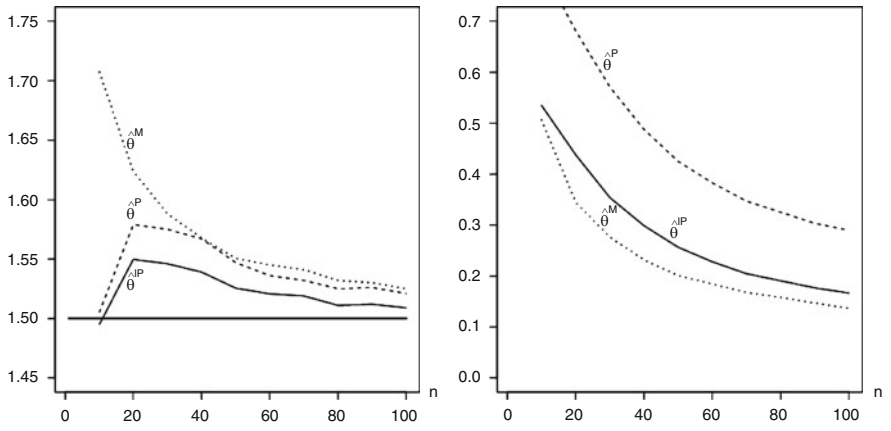

Fig. 6 Simulated mean values (*left*) and root mean squared errors (*right*) of $\hat{\theta}^M$, $\hat{\theta}^P$, and $\hat{\theta}^{IP*}$, from a discrete Weibull model with $q = 0.8$ and $\theta = 1.5$ (increasing failure rate)

5 Empirical Application

As an empirical example, we place ourselves in a context from quantitative risk management. We consider the Volkswagen share price from January 3, 2003 to January 29, 2010, and the daily log returns $R_{t+1} = \log(P_{t+1}/P_t)$, where P_t is the stock price at time t. The data come from web site http://chart.yahoo.com/ with ticker symbol *vow.de*. The corresponding one-day-ahead VaR forecasts made at time t for time $t + 1$, $\mathrm{VaR}_{t+1|t}(p)$, is defined by

$$P[R_{t+1} \leq \mathrm{VaR}_{t+1|t}(p)|\Omega_t] = p,$$

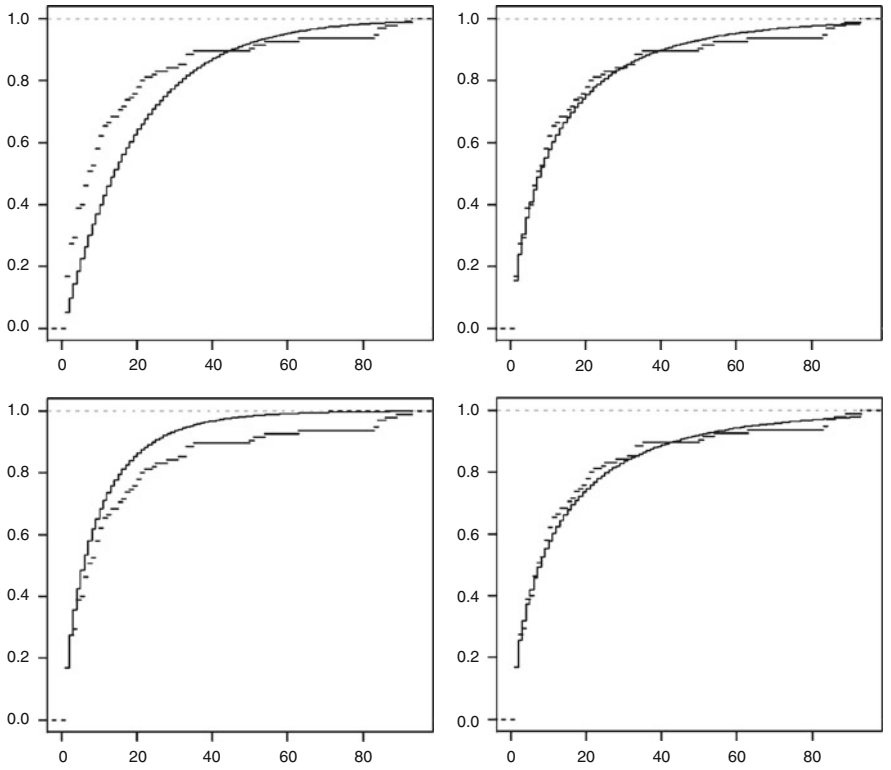

Fig. 7 Comparison of geometric (0.05), F_M, F_P, and F_{IP} cdfs and ecdf for Volkswagen durations data set. (**a**) Geometric(0.05) cdf (*gray*) and ecdf (*black*), (**b**) cdf F_M (*gray*) and ecdf (*black*), (**c**) cdf F_P (*gray*) and ecdf (*black*), (**d**) cdf F_{IP} (*gray*) and ecdf (*black*)

where Ω_t is the information set up to time-t and p is the *coverage rate*. Considering a *violation* the event that a return is lower than the reported VaR, we define the hit function

$$I_{t+1}(p) = \begin{cases} 1 & \text{if } R_{t+1} < \text{VaR}_{t+1|t}(p) \\ 0 & \text{if } R_{t+1} \geq \text{VaR}_{t+1|t}(p). \end{cases}$$

and the duration between two consecutive violations as $D_i := t_i - t_{i-1}$, where t_i denotes the day of violation number i. Christoffersen [1] showed that evaluating interval forecasts can be reduced to examining whether the hit sequence satisfies the unconditional coverage (UC) and independence (IND) properties. It is possible to write the IND property as $D_i \overset{iid}{\sim} D \sim$ discrete Weibull$(\theta = 1)$. A problematic non-verification of IND is the one that leads to clustering of violations, which corresponds to several large losses occurring in a short period. With clustering, we have an excessive number of very short durations and an excessive number of very long durations. The discrete Weibull with $\theta < 1$ will generate this pattern; for this

Table 2 Goodness-of-fit statistics for fits of the geometric (0.05), F_M, F_P, and F_{IP} distributions to the Volkswagen durations data set

Statistic	Geometric(0.05)	F_M	F_P	F_{IP}
Kolmogorov–Smirnov	0.221	0.052	0.095	0.0514
Chi-Square	46.7	6.16	35.8	5.84

reason, the estimate of the shape parameter can be used to identify a model that violates IND in this way. Using the popular Historical Simulation (HS) method for VaR(0.05), we calculate 95 durations with a moving window of size 250. The obtained estimates were $\hat{q}^M = 0.847$, $\hat{q}^P = 0.832$, $\hat{\theta}^M = 0.712$, $\hat{\theta}^P = 0.794$, and $\hat{\theta}^{IP*} = 0.67$. All estimators give evidence that the HS VaR method used leads to clustering of violations, with estimates of θ lower than 1. We consider three models, F_M, F_P, and F_{IP} fitted with the methods (3), (5), and (8). To assess how well these distributions fit the Volkswagen durations data set, Fig. 7 contains the plot of the ecdf along with F_M, F_P, and F_{IP} cdfs. We also plot the ecdf along with the cdf of the geometric (0.05) which corresponds to the UC and IND hypothesis. To measure the discrepancy between the ecdf and the cdfs, the Kolmogorov–Smirnov and chi-square statistics are given in Table 2. Clearly, the moments and the improved proportions methods provide much better fit than the proportions method. These two methods perform well with the real data set under study, but the improved proportions are based on a simple equation while the method of moments involves equations that cannot be solved easily by ordinary techniques.

Acknowledgements This research was partially supported by Fundação para a Ciência e Tecnologia (FCT/PROTEC, FCT/OE and PTDC/FEDER) and Center of Statistics and Applications of University of Lisbon (CEAUL).

References

1. Christoffersen, P.: Evaluating intervals forecasts. Int. Econ. Rev. **39**, 841–862 (1998)
2. Bracquemond, C., Gaudoin, O.: A survey on discrete lifetime distributions. Int. J. Reliab. Qual. Safe. Eng. **10**, 69–98 (2003)
3. Harman, Y.S., Zuehlke, T.W.: Duration dependence testing for speculative bubbles. J. Econ. Finance **28**, 17–36 (2004)
4. Haas, M.: Improved duration-based backtesting of Value-at-Risk. J. Risk **8**, 17–36 (2005)
5. Jazi, M.A., Lai, C., Alamatsaz, M.H.: A discrete inverse Weibull distribution and estimation of its parameters. Stat. Methodol. **7**, 121–132 (2010)
6. Lin, T., Guillén, M.: The rising hazards of party incumbency. A discrete renewal analysis. Polit. Anal. **7**, 31–57 (1998); An Annual Pulication of the Methodology Section of the American Political Science Association
7. Lancaster, T.: Econometric methods for the duration of unemployment. Econometrica **47**, 939–956 (1979)
8. Padgett, W.J., Spurrier, J.D.: Discrete failure models. IEEE Trans. Reliab. **34**, 253–256 (1985)
9. Khan, M.S.A., Khalique, A., Abouammoh, A.M.: On estimating parameters in a discrete weibull distribution. IEEE Trans. Reliab. **38**, 348–350 (1989)

10. Kulasekera, K.B.: Approximate MLE's of the parameters of a discrete weibull distribution with Type I censored data. Microelectron. Reliab. **34**, 1185–1188 (1994)
11. Nakagawa, T., Osaki, S.: The discrete Weibull distribution. IEEE Trans. Reliab. **24**, 300–301 (1975)
12. Stein, W.E., Dattero, R.: A new discrete Weibull distribution. IEEE Trans. Reliab. **R-33**, 196–197 (1984)
13. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org (2008)

DPOT Methodology: An Application to Value-at-Risk

M.I. Fraga Alves and P. Araújo Santos

Abstract

Threshold methods, based on fitting a stochastic model to the excesses over a threshold, were developed under the acronym POT (peaks over threshold). To eliminate the tendency to clustering of violations, a model-based approach within the POT framework, which uses the durations between excesses as covariates, is presented. Based on this approach we suggest models to forecast one-day-ahead Value-at-Risk and apply these models to the Standard & Poor's 500 Index. Out of sample results provide evidence that they can perform better than state-of-the art risk models.

1 Introduction

Value-at-Risk (VaR) aggregates several components of risk into a single number and has emerged as the standard measure in quantitative risk management (for a detailed discussion of VaR, see [12]. We will consider the symmetric of daily log returns, $R_{t+1} = -\log(V_{t+1}/V_t) \times 100$, where V_t is the value of the portfolio at time t. The one-day-ahead VaR forecast made at time t for time $t+1$, $\text{VaR}_{t+1|t}(p)$, is defined by

$$P[R_{t+1} > \text{VaR}_{t+1|t}(p)|\Omega_t] = p,$$

M.I.F. Alves (✉)
Faculdade de Ciências, Departamento de Estatística e Investigação Operacional,
Universidade de Lisboa, Campo Grande, 1749-016 Lisboa, Portugal
e-mail: isabel.alves@fc.ul.pt

P.A. Santos
Instituto Politécnico de Santarém, Departamento de Informática e Métodos Quantitativos,
Escola Superior de Gestão e Tecnologia, Complexo Andaluz,
Apartado 295, 2001-904 Santarém, Portugal
e-mail: paulo.santos@esg.ipsantarem.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_9, © Springer-Verlag Berlin Heidelberg 2013

where Ω_t is the information set up to time-t and p is the *coverage rate*. A *violation* occurs when the symmetric daily return exceeds the reported VaR.

We recall the Generalized Pareto Distribution (GPD)

$$G_{\gamma,\sigma}(y) = \begin{cases} 1 - (1 + \gamma y/\sigma)^{-1/\gamma}, & \gamma \neq 0 \\ 1 - \exp(-y/\sigma), & \gamma = 0, \end{cases} \tag{1}$$

where $\sigma > 0$, and the support is $y \geq 0$ when $\gamma \geq 0$ and $0 \leq y \leq -\sigma/\gamma$ when $\gamma < 0$. The probability that the random variable (r.v.) X assumes a value that exceeds a threshold u by at most y, given that it does exceed the threshold, is given by the *excess distribution*

$$F_u(y) = P[X - u \leq y | X > u] = \frac{F(y+u) - F(u)}{1 - F(u)}, \tag{2}$$

for $0 \leq y < x^F - u$, where x^F is the (finite or infinite) right endpoint of F, defined by $x^F := \sup\{x : F(x) < 1\}$. The Extreme Value Theory (EVT), with the following theorem, suggests the GPD (1) as an approximation for the excess distribution (2), for a sufficiently high threshold u.

Theorem 1.1. *(Balkema and de Haan [3] and Pickands [17]) It is possible to find a function $\beta(u)$ such that*

$$\lim_{u \to x_F} \sup_{0 \leq y < x_F - u} |F_u(y) - G_{\gamma,\beta(u)}(y)| = 0,$$

if and only if F is in the maximum domain of attraction of an extreme value distribution.

For a wide class of distributions, the excess distribution (2) over a high threshold u can be approximated by the GPD (1) and this result holds for all distributions for which the sequence of maxima linearly normalized converges to a non-degenerate limit law. Smith [19] proposed a tail estimator based on a GPD approximation to the excess distribution. We denote n the number of exceedances above u in a sample $X_1, \ldots, X_{n_x}$. Using n/n_x as estimator of $\bar{F}(u)$, the relation $\bar{F}_u(x-u) = \bar{F}(x)/\bar{F}(u)$ and $\bar{F}_u(x-u)$ estimated by a GPD approximation, we obtain the tail estimator and for $p < \bar{F}(u)$, inverting the tail estimator, we get the VaR POT estimator

$$\widehat{\text{VaR}}_{t+1}^{POT}(p) = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left(\left(\frac{n}{n_x p}\right)^{\hat{\gamma}} - 1\right). \tag{3}$$

However, this is an unconditional method and suffers from the problem of tendency to clustering of violations when applied to financial time series. In Sect. 2, in order to solve this problem, we propose risk models based on durations and within the POT framework. Comparisons between the proposed risk models and other models are made in Sect. 3. Finally, conclusions are given in Sect. 4.

2 A Duration-Based POT Method

Our main goal is to eliminate the tendency to clustering of violations that occurs with the POT method. To achieve this goal, within the POT framework we propose the presence of durations between excesses as covariates. Smith [20] develop ML and Least Squares estimation procedures under the POT framework with the shape and scale parameters dependent on covariates. For a general overview of EVT and its application to VaR, including the use of explanatory variables, see, for instance, Tsay [21]. For details about the mathematical theory of EVT and its applications to risk management, see Embrechts et al. [9]. Let $y_1, \ldots, y_n$ be excesses above a high threshold u, d_1 the duration until the first excess and $d_2, \ldots, d_n$, defined by

$$d_i = t_i - t_{i-1}, \tag{4}$$

where t_i denotes the day of excess i. We propose to use from the information set up to time t (Ω_t), the last v durations between excesses, $d_n, d_{n-1}, \ldots, d_{n-v+1}$ and the duration since the excess n which we define by d^t. With the durations $d_i, \ldots, d_{i-v+1}$, it is possible to consider at the time of excess number i, the duration since the preceding v excesses, defined by

$$d_{i,v} = d_i + \cdots + d_{i-v+1} = t_i - t_{i-v}. \tag{5}$$

At day t, after the excess n, we define $d_{t,1} = d^t$ and for $v = 2, 3, \ldots,$

$$d_{t,v} = d^t + d_{n,v-1}$$

which represents the duration until t since the preceding v excesses.

2.1 Empirical Motivation

The motivation for the presence of durations between excesses as covariates has mainly been based on the relation between the amount of the excess and durations which we observe in various financial time series. Figure 1 (left) presents for the S&P 500 Index, considering all the returns from January 4, 1950 through May 18, 2010, and a threshold such that $10\,\%$ of the values are larger than the threshold, the scatterplot of excesses (y_i) and durations since the preceding excess (d_i). Clearly, large excesses tend to be associated with short durations. In Fig. 1 (right) we plot the excesses (y_i) and the inverse of durations since the preceding excess ($1/d_i$). Table 1 gives Pearson correlations between excesses, durations, and the inverse of durations. The linear association between excesses and durations is weak, but increases when we take the inverse of durations, as expected. Adding durations we get the duration since the preceding v excesses defined in (5) and the correlation increases a little more when we compute the correlation between excesses and the inverse of these

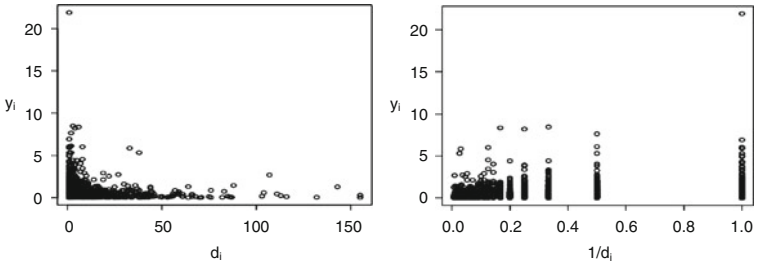

Fig. 1 S&P Index from January 4, 1950 through May 18, 2010. Scatter plot of excess above a high threshold ($u = 0.9897$) and duration since the preceding excess (*left*) and scatter plot of excess above a high threshold ($u = 0.9897$) and the inverse of duration since the preceding excess (*right*)

Table 1 S&P Index. Pearson correlation between y_i, d_{i-j}, $\frac{1}{d_{i-j}}$, and $\frac{1}{d_{i,v}}$

j	$\text{Corr}(y_i, d_{i-j})$	$\text{Corr}(y_i, \frac{1}{d_{i-j}})$	v	$\text{Corr}(y_i, \frac{1}{d_{i,v}})$
0	−0.123	0.193	2	0.284
1	−0.127	0.174	3	0.325
2	−0.096	0.149	4	0.335
3	−0.126	0.148	5	0.346

durations. We observe that the excesses have higher mean and higher variance with short durations, and lower mean and lower variance with long durations. Araújo Santos and Fraga Alves [1] obtained similar results for the Down Jones Average Index. Based on these empirical results, we propose to define the expected value and variance of the excesses dependent on the durations. This purpose can be achieved by modeling the parameter σ of the GPD model (1) as a function of parameters and the durations. We generically denote this function by g.

2.2 DPOT Model

With the durations (4) and the duration since the excess n, d^t, we assume the GPD for the excesses Y_i above u, such that

$$Y_t \sim GPD\Big(\gamma, \sigma_t = g(\alpha_1, \ldots, \alpha_k, \ldots, d^t, d_n, d_{n-1}, \ldots, d_{n-v+2})\Big),$$

where $\gamma, \alpha_1, \ldots, \alpha_k$, are parameters to be estimated. And we propose the following class of estimators

$$\widehat{\text{VaR}}_{t+1|t}^{DPOT}(p) = u + \frac{\hat{\sigma}_t}{\hat{\gamma}}\left(\left(\frac{n}{n_x p}\right)^{\hat{\gamma}} - 1\right), \tag{6}$$

with $\hat{\sigma}_t = g(\hat{\alpha_1}, \ldots, \hat{\alpha_k}, \ldots, d^t, d_n, d_{n-1}, \ldots, d_{n-v+2})$.

The proposed $DPOT$ method implies, for $\gamma < 1$, a conditional expected value for excesses, and for $\gamma < 1/2$, a conditional variance, both dependent on d^t and the last v durations between excesses,

$$E[Y_t|\Omega_t] = \frac{\sigma_t}{1-\gamma} \quad (\gamma < 1), \qquad V[Y_t|\Omega_t] = \frac{(\sigma_t)^2}{(1-2\gamma)} \quad (\gamma < 1/2).$$

The empirical results of Sect. 2.1 suggest an inverse relation between excesses and the durations since the preceding v excesses, with $1/(d_{i,v})^c$, $c > 0$, which leads to the specification $\sigma_t = \alpha \frac{1}{(d_{t,v})^c}$ and the VaR estimator

$$\widehat{\text{VaR}}_{t+1|t}^{DPOT(v,c)}(p) = u + \frac{\hat{\alpha}}{\hat{\gamma}(d_{t,v})^c}\left(\left(\frac{n}{n_x p}\right)^{\hat{\gamma}} - 1\right), \tag{7}$$

where $\hat{\gamma}$ and $\hat{\alpha}$ are estimators of the parameters γ and α. Applying the maximum likelihood theory to estimate the parameters, the log likelihood obtained is

$$\begin{aligned} \log \mathrm{L}(\gamma, \alpha) &= \log \prod_{i=v}^{n} f_{Y_i}(y_i) \\ &= -\sum_{i=v}^{n} \log\left(\frac{\alpha}{(d_{i,v})^c}\right) - \left(\frac{1}{\gamma} + 1\right) \sum_{i=v}^{n} \log\left(1 + \frac{\gamma}{\alpha} y_i (d_{i,v})^c\right). \end{aligned} \tag{8}$$

We choose $v = 3$ taking into account that the correlations increase only slightly for $v \geq 4$. We study the model with c estimated, but we achieve poor results. Empirical findings suggest that the method is robust for different values of c in the interval between 0.7 and 0.8. We present results for $v = 3$, $c \in \{0.8, 0.75, 0.7\}$ and apply an implementation of Nelder and Mead algorithm, using the stats package of R [18], to maximize (8).

3 Comparative Study

Using all the returns from S&P 500 Index, we compare the proposed DPOT method with a two-stage hybrid method which combines a time-varying volatility model with the EVT approach, known as Conditional EVT. We employ the R language in order to develop the programs. The data were obtained from the web site http://finance.yahoo.com/. Diebold et al. [8] proposed in a first step the standardization of the returns through the conditional means and variances estimated with a time-varying volatility model, and in a second step, estimation of a p quantile using EVT and the standardized returns. McNeil and Frey [15] combine a AR(1)-GARCH(1,1) process assuming normal innovations with the POT method from EVT. We will denote this model as CEVT-n. The filter with normal innovations, while capable of removing the majority of clustering, will frequently be a misspecified model

for returns. To accommodate this misspecification, Kuester et al. [13] suggested a filter with the skewed t distribution. We will denote this model as CEVT-sst. Several studies conclude that conditional EVT is the method with better out-of-sample performance, to forecast one-day-ahead VaR (e.g., [4, 6, 11, 13, 15, 16]), and this is the reason why we choose CEVT-n and CEVT-sst models for the comparative study. We compare the CEVT-sst, CEVT-n, and DPOT models with $v = 3$, $c \in \{0.8, 0.75, 0.7\}$, denoted respectively by DPOT(0.8), DPOT(0.75), and DPOT(0.7). We examine the one-day-ahead VaR(0.01) forecasts performance with the S&P, considering returns produced by all the historical data until May 18, 2010. Using a rolling window of size 1,000 we obtain 14,190 one-day-ahead VaR(0.01) forecasts for each model. As usual, the threshold u was chosen such that $10\,\%$ of the values are larger than the threshold (see [15]). The primary tool for assessing the accuracy of the interval forecasts is to monitor the binary sequence generated by observing if the return on day $t + 1$ is in the tail region specified by the VaR at time-t, or not. This is referred to as the hit sequence

$$I_{t+1}(p) = \begin{cases} 1 \text{if } R_{t+1} > VaR_{t+1|t}(p) \\ 0 \text{if } R_{t+1} \leq VaR_{t+1|t}(p). \end{cases}$$

Christoffersen [7] showed that evaluating interval forecasts can be reduced to examining whether the hit sequence satisfies the unconditional coverage (UC) and independence (IND) properties. To test the UC hypothesis, we apply the Kupiec test [14]. Engle and Manganelli [10] consider an autoregression for I_t and Berkowitz et al. [5] propose the logit model

$$\log\left(\frac{P[I_t = 1]}{1 - P[I_t = 1]}\right) = \alpha + \beta_1 I_{t-1} + \beta_2 \text{VaR}_{t|t-1}(p).$$

We can test the IND hypothesis with a likelihood ratio test considering for the null $\beta_1 = \beta_2 = 0$ and in this case the asymptotic distribution is chi-square with 2 degrees of freedom. We refer to this test as the CAViaR independence test of Engle and Manganelli (CAViaR). The other independence test applied was recently introduced in the literature [2] and is based on the ratio $(D_{N:N} - 1)/D_{[N/2]:N}$, where $D_{N:N}$ and $D_{[N/2]:N}$, are the maximum and the median of durations between consecutive violations and until the first violation. This new test is based on an exact distribution and outperforms, in terms of power, existing procedures in realistic settings. We refer to this test as MM ratio test. The empirical findings are presented in Table 2. In the case of the POT model both independence tests reject the IND hypothesis with very small p-values. With a violation frequency equal to 0.01367, the UC hypothesis is also clearly reject with this model. In terms of the UC hypothesis, both DPOT and CEVT models perform very well taking into account that in no case the hypothesis is rejected since all p-values are very high. It is interesting to note the impressive performance of CEVT models in terms of UC in Table 2, with 142 violations in 14,190 out-of-sample forecasts was impossible to obtain a better result

Table 2 Out-of-sample accuracy for VaR(0.01) applied to S&P 500 Index returns from January 4, 1950 until May 18, 2010, with a rolling window of size 1,000

Model	Violation frequencies	Kupiec p-value	CAViaR p-value	MM ratio p-value
POT	0.013672	0.0000	0.0000	0.0000
DPOT(0.8)	0.009725	0.7410	0.0189	0.7902
DPOT(0.75)	0.009443	0.5011	0.1018	0.1048
DPOT(0.7)	0.009443	0.5011	0.8659	0.0566
EVTC-n	0.010007	0.9933	0.0145	0.0166
EVTC-sst	0.010007	0.9933	0.0236	0.0314

Unconditional coverage and independence tests

(violation frequency equal to 0.010007). In terms of IND hypothesis, the DPOT models perform clearly better than the CEVT models. Considering the six cases with three DPOT models and two independence tests, with DPOT models the IND hypothesis is rejected only in one case. For the Conditional EVT models the IND hypothesis is rejected in all cases. This empirical evidence shows that the DPOT model can perform better than Conditional EVT in terms of removing the tendency to clustering of violations.

4 Final Remarks

In this work we presented a POT method that uses the durations between excesses as covariates. Based on this method, we suggest three DPOT models to forecast one-day-ahead VaR. Empirical findings, with the S&P 500 Index, show that they perform very well in terms of unconditional coverage and perform better than state-of-the art models in terms of removing the tendency to clustering of violations. Finally, we notice that in order to deal with the volatility clustering, the proposed models do not assume a parametric distribution for the entire distribution of the returns, as the CEVT models and other models widely used, but assume a parametric model only on the tail and based on solid asymptotic theory.

Acknowledgements This research was partially supported by National Funds through **FCT** - Fundação para a Ciência e a Tecnologia, FCT//PTDC/MAT/101736/2008, EXTREMA project.

References

1. Araújo Santos, P., Fraga Alves, M.I.: VaR prediction with a duration based POT method. In: Proceedings of the ISF2010, 30th International Symposium on Forecasting, San Diego, CA (2010)
2. Araújo Santos, P., Fraga Alves, M.I.: A new class of independence tests for interval forecasts evaluation. Comput. Stat. Data Anal. **56**, 3366–3380 (2012). doi:10.1016/j.csda2010.10.002
3. Balkema, A.A., de Haan, L.: Residual life time at great age. Ann. Probab. **2**, 792–804 (1974)

4. Bekiros, S.D., Georgoutsos, D.A.: Estimation of value-at-risk by extreme value and conventional methods: a comparative evaluation of their predictive performance. J. Int. Financ. Markets Institut. Money **15**(3), 2009–2228 (2005)
5. Berkowitz, J., Christoffersen P., Pelletier D.: Evaluating value-at-risk models with desk-level data. Management Science, Published online in Articles in Advance (2009)
6. Byström, H.: Managing extreme risks in tranquil and volatile markets using conditional extreme value theory. Int. Rev. Financ. Anal. **13**, 133–152 (2004)
7. Christoffersen P.: Evaluating intervals forecasts. Int. Econ. Rev. **39**, 841–862 (1998)
8. Diebold, F.X., Schuermann, T., Stroughair, J.D.: Pitfalls and opportunities in the use of extreme value theory in risk management. Working paper, Wharton School, University of Pennsylvania (1998), pp. 98–10
9. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modeling Extremal Events for Insurance and Finance. Springer, Berlin (1997)
10. Engel, R.F., Manganelli, S.: CAViaR: conditional autoregressive value-at-risk by regression quantiles. J. Business Econ. Stat. **22**, 367–381 (2004)
11. Ghorbel, A., Trabelsi, A.: Predictive performance of conditional extreme value theory in value-at-risk estimation. Int. J. Monet. Econ. Finance **1**, 121–147 (2008)
12. Jorian, P.: Value at Risk: The New Benchmark for Managing Financial Risk. McGraw-Hill, New York (2000)
13. Kuester, K., Mittik, S., Paolella, M.S.: Value-at-risk prediction: a comparison of alternative strategies. J. Financ. Econometrics **4**, 53–89 (2006)
14. Kupiec, P.: Techniques for verifying the accuracy of risk measurement models. J. Derivat. **3**, 73–84 (1995)
15. McNeil, A.J., Frey, R.: Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. J. Empirical Finance **7**, 271–300 (2000)
16. Ozun, A., Cifter, A., Yilmazer, S.: Filtered extreme value theory for value-at-risk estimation: evidence from Turkey. J. Risk Finance Incorporat. Balance Sheet **11**, 164–179 (2010)
17. Pickands III, J.: Statistical inference using extreme value order statistics. Ann. Stat. **3**, 119–131 (1975)
18. R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org (2010)
19. Smith, R.: Estimating tails of probability distributions. Ann. Stat. **15**, 1174–1207 (1987)
20. Smith, R.: Models for exceedances over high thresholds. J. R. Stat. Soc. B **52**, 393–442 (1990)
21. Tsay, R.: Analysis of Financial Time Series. Wiley Series in Probability and Statistics, John Wiley & Sons (2010)

Estimation of the Extremal Index Function in Case of Asymptotically Independent Markov Chains and Its Application to Stock Market Indices

Alexandra Ramos and Anthony Ledford

Abstract

To describe the behaviour of exceedances of a stationary process, attention has been given to the within-cluster behaviour of the extremes of Markov chains, which is determined by the short-range temporal dependence. These characterisations are usually based on the threshold dependence extremal index, $\theta(u)$, which measures the extent of clustering of exceedances of the process above the high threshold u. In this work, we intend to study the function $\theta(u)$, in case of asymptotically independent Markov chains. An estimator of $\theta(u)$ based on the methodology described in Ramos and Ledford (J. R. Stat. Soc. B 71(1):219–241, 2009) will be deduced and analysed. Applications of the new estimator to simulated data and to real data will be considered.

1 Introduction

Over the last decade or so, advances in multivariate extreme value theory have led to improved techniques for characterising the extremal behaviour of stationary time series. In particular, attention has been given to the within-cluster behaviour of the extremes of a series, which is determined by the short-range temporal dependence. Most of its characterisation has been done based on the assumption of Markovianity of the time series, as the class of dth-order Markov chains is sufficiently general and tractable. The simplest limiting quantifier of the extremal

A. Ramos (✉)
FEP and CMUP, Universidade do Porto, Rua Dr Roberto Frias, 4200-464-Porto, Portugal
e-mail: aramos@fep.up.pt

A. Ledford
AHL Research, Man Research Laboratory, Eagle House, Walton Well Road, Oxford, OX2 6ED, UK
e-mail: aledford@ahl.com

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_10, © Springer-Verlag Berlin Heidelberg 2013

within-cluster dependence for any stationary time series $\{X_n\}_{n\geq 1}$ satisfying the D condition in [4] is the extremal index θ with $0 \leq \theta \leq 1$. Values of θ close to 0 indicate a very strong short range extremal dependence, while values close to 1 suggest a rather weak dependence. In fact, θ is often defined by both the reciprocal of the limiting mean cluster size and the only parameter describing the effect of short-range extremal dependence on the limiting distribution of the componentwise maxima $M_N = \max\{X_1, \ldots, X_N\}$, in the sense that there exists a sequence of constants $a_N > 0$, b_N and $u_N = a_N x + b_N$ such that

$$\lim_{N\to\infty} \Pr(M_N \leq u_N) = \lim_{N\to\infty} \{F(u_N)\}^{N\theta},$$

where F is the common marginal distribution function and $\lim_{N\to\infty} \{F(u_N)\}^{N\theta} = G^{\theta}(x)$ for a non-degenerate distribution function G called the generalised extreme value distribution. However, some difficulties arise in the case $\theta = 1$, which implies that asymptotically extreme events occur singly. For example, for all asymptotically independent Markov processes $\{X_n\}_{n\geq 1}$, i.e. such that $\lim_{u\to x^*} \Pr(X_{i+1} > u | X_i > u) = 0$ where x^* is the upper limit of the support of the common marginal distribution, the extremal index is one, despite clustering of exceedances of high thresholds still occurring, see [2]. For such processes it is useful to characterise the extremal dependence by a penultimate version of the extremal index, $\theta(u) \equiv \theta(u, m)$ say, defined as a function of the high threshold u and $m = m(u)$ the length of the window of the clusters, and whose limit is θ as $u \to x^*$. Following [2], it is natural to define $\theta(u, m)$ as

$$\theta(u) \equiv \theta(u, m) = [E\{N(u, m) | N(u, m) \geq 1\}]^{-1}, \tag{1}$$

where $N(u, m) = \sum_{i=1}^{m} I(X_i > u)$ is the size of a cluster with window of length m initialised at $i = 1$, and where I is the indicator function.

Provided a suitable choice of m can be made, $\theta(u, m)$ can be estimated by an empirical version of (1), see [1] or [2]. Theoretical properties and estimation techniques of the extremal index and its penultimate version $\theta(u, m)$ have been studied fairly extensively. Their application in various areas include insurance, hydrology and telecommunications, see, e.g. [3]. Other estimators of $\theta(u, m)$ given in the literature include the runs estimator, the blocks estimator or a profile log-likelihood based estimator, see [1].

2 Fitting a Markov Model to the Tails of the Stationary Time Series

A common approach for statistically modelling the tail of a Markov chain consists of using a determined joint distribution of d-dimensional extreme values to model the dependence structure between consecutive variables $X_i, \ldots, X_{i+d}$ that exceed a fixed high threshold u, an approach which is based on the assumption that the

limiting behaviour of the chain holds exactly above the high threshold u. Assuming $\{X_n\}_{n\geq 1}$ to be a first-order stationary Markov chain, Ramos and Ledford in a work yet to be published suggest modelling the joint tail of the distribution of two consecutive pairs (X_i, X_{i+1}) with a bivariate joint tail distribution and extended their results to dth-order Markov chains, see [5] for details. Let $\{X_n\}_{n\geq 1}$ be a stationary first-order Markov chain with continuous state space and denote its joint distribution function by $F(x_i, x_{i+1})$, and its marginal distribution by $F(x) = \Pr(X_i \leq x)$ for all $i \geq 1$. Then, we consider the generalised Pareto distribution (GPD) to describe the univariate tail behaviour over a high threshold u_1, u_1 in the original scale of the variable X_i, which leads to

$$F(x) = \begin{cases} 1 - \lambda_1 \left\{1 + \xi(x - u_1)/\sigma\right\}_+^{-1/\xi}, & x \geq u_1 \\ 1 - \lambda_1, & x < u_1 \end{cases} \tag{2}$$

where $s_+ = \max(s, 0)$, ξ, and $\sigma > 0$ are shape and scale parameters respectively and λ_1 denotes the threshold exceedance probability. The joint distribution of (X_i, X_{i+1}) over a joint tail region $R_{11} = (u_1, \infty) \times (u_1, \infty)$ is treated in a similar way. We adopt the η-asymmetric logistic model ([6]) as a joint tail survivor model to describe the dependence between consecutive high observations of the chain, which is given by

$$\overline{F}(x_i, x_{i+1}) = \frac{\lambda}{N_\varrho}\left[\left(\frac{\varrho y_i}{u_f}\right)^{-\frac{1}{\eta}} + \left(\frac{y_{i+1}}{\varrho u_f}\right)^{-\frac{1}{\eta}} - \left\{\left(\frac{\varrho y_i}{u_f}\right)^{-\frac{1}{\alpha}} + \left(\frac{y_{i+1}}{\varrho u_f}\right)^{-\frac{1}{\alpha}}\right\}^{\frac{\alpha}{\eta}}\right] \tag{3}$$

for $x_i, x_{i+1} > u_1$ and where $\overline{F}(x, y) = \Pr(X > x, Y > y)$, $N_\varrho = \varrho^{-1/\eta} + \varrho^{1/\eta} - (\varrho^{-1/\alpha} + \varrho^{1/\alpha})^{\alpha/\eta}$, $\eta, \alpha \in (0, 1]$ and $\varrho > 0$, $y_j = -1/\log F(x_j)$ $(j = i, i+1)$, $F(x)$ is as defined in (2) and $u_f = -1/\log(1 - \lambda_1)$ is a high threshold in the unit Fréchet scale.

The fitting of the model is done through a threshold censored likelihood based on the approach described in [7], where data below a high threshold in each margin are censored.

3 Within-Clusters Behaviour at Extreme Levels

Having fitted the model, estimates of various extremal characteristics of the process can be obtained. In particular, we are here interested in studying the threshold dependent extremal index $\theta(u)$, which measures the extent of clustering of exceedances of the process above the high threshold u. Estimates of the threshold-dependent extremal index, as well as other quantities of interest, can be obtained by simulation of clusters of extreme events from the fitted model. Following [2] and [7], this can be done by first simulating the maximum value of the cluster and then simulating values backwards and forwards over the cluster window according

to the transition density of the fitted Markov chain. As the limiting distribution of the cluster maximum is the same as the limiting distribution of any other exceedance, it can be simulated from a GPD as in (2). The other cluster values are obtained from the following result, see [2]. Under regularity conditions and conditionally on the maximum $Y_1 > u$, the ratios $Y_2/Y_1, \ldots, Y_q/Y_{q-1}$ are independent as $u \to \infty$ and for fixed q, with common distribution function

$$H_F(y) = \lim_{u\to\infty} \Pr(Y_2 \le uy | Y_1 = u), \qquad y \ge u_f/u.$$

and so the ratios Y_i/Y_{i-1} can be generated from the distribution function above. Analogously, the observations preceding the maximum are generated by simulating the ratios Y_{i-1}/Y_i as independent points from

$$H_B(y) = \lim_{u\to\infty} \Pr(Y_1 \le uy | Y_2 = u), \qquad y \ge u_f/u.$$

Similar to [2] and using the η-asymmetric logistic model defined in (3), we consider the approximations

$$\begin{aligned}
\Pr(Y_2 \le uy | Y_1 = u) &\sim \exp\left(-\frac{\lambda\, u_f^{1/\eta} u^{1-1/\eta}}{\eta N_\varrho} \left[\varrho^{-1/\eta} - \varrho^{-1/\alpha} \left\{ \varrho^{-1/\alpha} + \left(\frac{y}{\varrho}\right)^{-1/\alpha} \right\}^{\alpha/\eta - 1} \right] \right) \\
&\equiv H_F^*(y;u) \\
\Pr(Y_1 \le uy | Y_2 = u) &\sim \exp\left(-\frac{\lambda\, u_f^{1/\eta} u^{1-1/\eta}}{\eta N_\varrho} \left[\varrho^{1/\eta} - \varrho^{1/\alpha} \left\{ (\varrho y)^{-1/\alpha} + \left(\frac{1}{\varrho}\right)^{-1/\alpha} \right\}^{\alpha/\eta - 1} \right] \right) \\
&\equiv H_B^*(y;u),
\end{aligned}$$

valid for $0 < \eta < 1$ and large u. These functions provide a threshold-dependent model for the transition distribution. Although they are valid only for $\eta < 1$, they have the advantage of producing approximations for H_F and H_B which are proper distribution functions, and since we are primarily interested in asymptotically independent Markov chains, we work with these functions for the remainder of this chapter. For asymptotically dependent Markov chains, i.e. when $\eta = 1$, we can either consider other approximations of H_F and H_B valid for this case under our framework or work with functions H_F and H_B defined under the framework in [7]. Then, the complete algorithm for the cluster simulation is as follows.

1. *Choice of threshold and length of cluster window.* Choose a high threshold u, in a unit Fréchet scale, such that $u > u_f$ and choose $m = 2r + 1$, where m and r are the length and the radius of the window, respectively.
2. *Simulation of the cluster maximum.* Simulate the cluster maximum, Y_0, from a GPD as defined in (2) but with u_1 replaced with u, $\sigma = u$, $\xi = 1$ and $\lambda_1 = 1 - \exp(-1/u)$.
3. *Simulation of the cluster.* Simulate the observations $Y_1, \ldots, Y_r$ iteratively from H_F^* with $u = Y_{i-1}$ $(i = 1, \ldots, r)$ and simulate the observations $Y_{-1}, \ldots, Y_{-r}$ iteratively from H_B^* with $u = Y_{-i+1}$ $(i = 1, \ldots, r)$. Each of these sequences will be rejected whenever a value $Y_i > Y_0$ is generated, and its simulation will be repeated.
4. *Estimation of functionals of interest.* Having repeated the previous steps to produce a large number k of clusters, several quantities of interest can be constructed empirically. For example, denoting by $N_j\left(u; \hat{\phi}\right)$ the number of exceedances of u (of $m(u)$) in the jth simulated cluster that has at least one exceedance of u, the threshold-dependent extremal index can be estimated by

$$\hat{\theta}\left(u, \hat{\phi}\right) = \left\{ \frac{1}{k} \sum_{j=1}^{k} N_j\left(u; \hat{\phi}\right) \right\}^{-1},$$

where $\hat{\phi}$ represents the model estimated parameters.

4 Application to Simulated Data from Gaussian Processes

In this section we examine the accuracy of the methodologies developed in Sects. 2 and 3 by studying the threshold-dependent extremal behaviour of simulated data consisting of a realisation of length $10,000$ generated from a stationary Gaussian autoregressive process of order 1, $\{X_n\}_n \geq 1$, with standard normal marginal distribution and lag τ correlation of ρ^τ for $\rho = 0.5$. Specifically, we consider $X_n = \rho X_{n-1} + (1 - \rho^2)^{1/2} \epsilon_n$ for $-1 < \rho < 1$, where ϵ_n are independent and identically distributed standard normal variables. As it was already mentioned in Sect. 1, this is a class of asymptotically independent Markov chains with extremal index $\theta = 1$, which exhibits dependence at sub-asymptotic levels. To restrict attention to the dependence features of the simulated data, the marginal distribution is treated as known and the points X_n are transformed to unit Fréchet variables Y_n $(n = 1, \ldots, 10,000)$, via $Y_i = -1/\log \Phi(X_i)$ where Φ denotes the standard normal cumulative distribution function.

The Markov tail model defined in Sect. 2 was fitted to the simulated data for the threshold $u_f = 11.99$ corresponding to the 0.92 quantile of the unit Fréchet marginal distribution, as stability of the model parameters estimates is observed over this level. The obtained coefficient of tail dependence estimate was $\hat{\eta} = 0.756\ (0.036)$, where standard error in parentheses are obtained by the delta method,

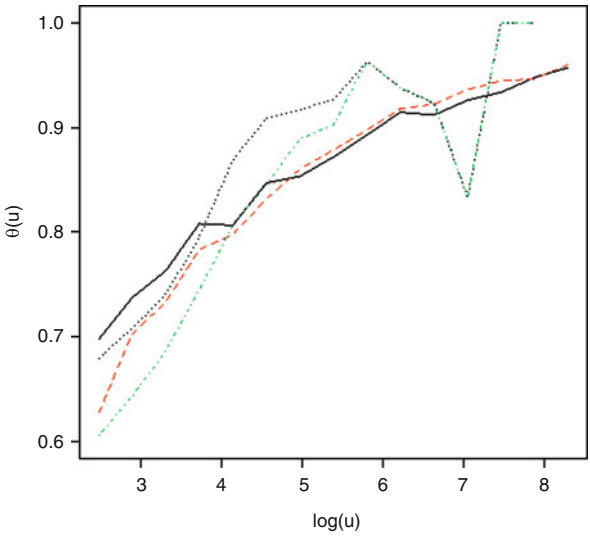

Fig. 1 Estimates of the extremal index function $\theta(u)$ against $\log u$ using the simulated Gaussian series and obtained by (—-) $\widehat{\theta}_1(u)$, $(\cdots)$ $\widehat{\theta}_2(u)$, (- · -) $\widehat{\theta}_3(u)$ and (- - -) its true behaviour

which is very close to the true value $\eta = 0.75$ and significantly different from 1, confirming the asymptotic independence of the series. The estimates obtained for the other parameters of the model were $\hat{\alpha} = 0.93\,(0.062)$ and $\hat{\varrho} = 1.22\,(0.135)$.

Figure 1 shows estimates of the extremal index $\theta(u)$ obtained by the simulation scheme of Sect. 3 with $m = 6$ and the fitted Markov tail model, which we denote by $\widehat{\theta}_1(u)$. For comparison, the true behaviour of $\theta(u)$ is also included in this figure. As suggested in [2], $\theta(u)$ may be evaluated empirically through simulation of clusters using the true transition densities. That is, the cluster maximum is simulated from a GPD with $u_1 = u$, $\xi = 0$, $\sigma = (1 - \Phi(u))/\Phi(u)$ and $\lambda_u = 1 - \Phi(u)$, and the distribution functions H_F^* and H_B^* are both equal to a normal distribution function with mean ρu and variance $1 - \rho^2$, evaluated at uy, and we use $m = 4$. Also shown in Fig. 1 are estimates of $\theta(u)$ obtained by the naive estimator $\widehat{\theta}_2(u) = \sum_{i=1}^{N} I(X_{i+1} \le u < X_i) / \sum_{i=1}^{N} I(X_i > u)$, and by the runs estimator $\widehat{\theta}_3(u) = \left[\sum_{i=1}^{N-1} I(X_i > u)\,(1 - I(X_{i+1} > u)) \cdots (1 - I(X_{i+k} > u))\right] / \sum_{i=1}^{N} I(X_i > u)$ for $k = 2$. From Fig. 1 we conclude that $\theta(u)$ is well approximated by $\widehat{\theta}_1(u)$, in particular when choosing a large u. In fact, for large values of u, the new estimator performs better than the naive or the runs estimators.

5 Application to Financial Data

In this section, we consider financial time series consisting of daily log-returns of closing stock index levels observed for the period of 1/1/00 to 30/9/09 for the Portuguese index PSI-20, the period of 2/4/84 to 30/9/09 for the English FTSE 100 and the period of 9/11/90 to 30/9/09 for the Swiss SSMI.

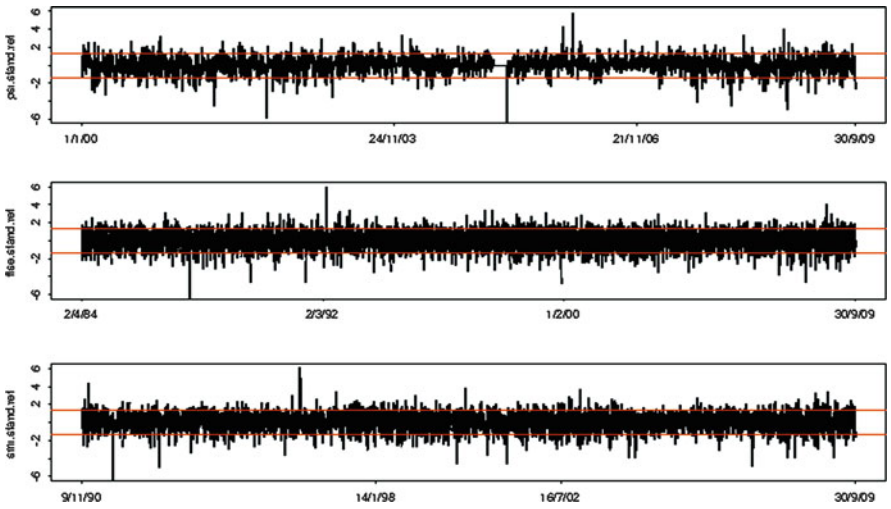

Fig. 2 Daily standardised log-returns of the PSI-20 (*top*), FTSE 100 and SSMI (*bottom*) series. The chosen thresholds u_1 are also identified

To remove the observed stock returns heteroskedasticity, we fit the volatility model GARCH(1,1) to each marginal series, and then apply the Markov chain modelling approach described in Sect. 2 to the positive/negative daily standardised log-returns of the series, which we denote by $\{X_n\}$. These are shown in Fig. 2. An evident temporal clustering of high and low extreme values can be seen, which indicates the presence of local dependence at extreme levels. This clustering behaviour has important implications in practice, as it corresponds to large consecutive changes associated with large financial gains or losses. Therefore, we examine both the upper and the lower joint tails of the pairs (X_i, X_{i+1}) for every data set, which requires the joint estimation of the extremal temporal-dependence structure as well as the tail of the marginal distribution.

The Markov tail model defined in Sect. 2 was fitted to these data sets for the thresholds u_f corresponding to the empirical 0.92 quantiles of the unit Fréchet transformed marginal data. The corresponding thresholds u_1 in the original scale, as in (2) and (3), are represented in Fig. 2. The obtained dependence parameter estimates are $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.31, 0.87, 0.65)$ for the upper tail and $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.56, 1.04, 2.9)$ for the lower tail of the standardised PSI-20 series, $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.61, 1.58, 1.10)$ for the upper tail and $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.59, 1.18, 0.35)$ for the lower tail of the standardised FTSE 100 series and $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.35, 1.08, 0.74)$ for the upper tail and $(\hat{\eta}, \hat{\alpha}, \hat{\varrho}) = (0.44, 0.72, 4.0)$ for the lower tail of the standardised SSMI series. Estimates for the coefficient of tail dependence are significantly lower than 1, indicating that all 3 series are asymptotically independent.

Figure 3 shows estimates of the extremal index function $\theta(u)$ for the fitted Markov tail model obtained by the simulation scheme of Sect. 3 with $m = 4$ and applied to the upper and lower tails of the 3 standardised series.

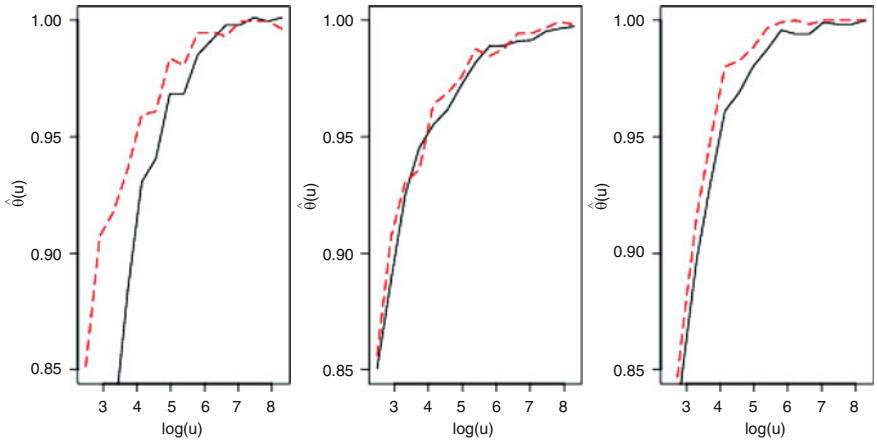

Fig. 3 Estimates of the extremal index function $\theta(u)$ against $\log u$ obtained by the new estimator $\widehat{\theta}_1(u)$ and applied to the (—-) upper and (- - -) lower tails of the standardised series of the stock Market indices PSI-20 (*left plot*), FTSE 100 (*middle plot*) and SSMI (*right plot*)

The estimates in Fig. 3 seem to suggest that, in general, for these standardised series, clusters tend to be longer in duration for the upper tail than for the lower tail, indicating that large falls tend to occur more in temporal isolation than large rises. The PSI 20 index seems to have bigger sized clusters than the other 2 indices, both for upper and lower tails.

Acknowledgements Alexandra Ramos was partially supported by FCT (Portugal) through the Centro de Matemática da Universidade do Porto.

References

1. Ancona-Navarrete, M., Tawn, J.A.: A comparison of methods for estimating the extremal index. Extremes **3**(1), 5–38 (2000)
2. Bortot, P., Tawn, J.A.: Models for the extremes of Markov chains. Biometrika **85**, 851–867 (1998)
3. Laurini, F.: Clusters of extreme observations and extremal index estimate in GARCH processes. Stud. Nonlin. Dyn. Econome. **8**(2), 1–21 (2004)
4. Ledbetter, M.R., Lindgren, G., Rootzén, H.: Extremes and Related Properties of Random Sequences and Series. Springer, New York (1983)
5. Ramos, A.: Modelos de Markov para a dependência extremal em séries temporais. In: Oliveira, I., Correia, E., Ferreira, F., Dias, S., Braumann, C. (eds.). Estatística: Arte de Explicar o Acaso. Edições SPE, pp. 549–558 (2009)
6. Ramos, A., Ledford, A.: A new class of models for bivariate joint tails. J. R. Stat. Soc. B **71**(1), 219–241 (2009)
7. Smith, R.L., Tawn, J.A., Coles, S.G.: Markov chain models for threshold exceedances. Biometrika **84**, 249–268 (1997)

Generated Covariates in Nonparametric Estimation: A Short Review

Enno Mammen, Christoph Rothe, and Melanie Schienle

Abstract

In many applications, covariates are not observed but have to be estimated from data. We outline some regression-type models where such a situation occurs and discuss estimation of the regression function in this context. We review theoretical results on how asymptotic properties of nonparametric estimators differ in the presence of generated covariates from the standard case where all covariates are observed. These results also extend to settings where the focus of interest is on average functionals of the regression function.

1 Introduction

Consider a nonparametric regression model of the form

$$Y = m_0(R) + \varepsilon,$$
$$E[\varepsilon|R] = 0,$$

where Y is a one-dimensional response variable and R is a q-dimensional covariate vector. The statistical goal is to nonparametrically estimate the regression function

E. Mammen (✉)
Department of Economics, University of Mannheim, D-68131 Mannheim, Germany
e-mail: emammen@rumms.uni-mannheim.de

C. Rothe
Toulouse School of Economics, 21 Allée de Brienne, F-31000 Toulouse, France
e-mail: rothe@cict.fr

M. Schienle
School of Business and Economics, Humboldt University Berlin, Spandauer Str. 1, D-10178 Berlin, Germany
e-mail: melanie.schienle@wiwi.hu-berlin.de

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_11, © Springer-Verlag Berlin Heidelberg 2013

$m_0 : \mathbf{R}^q \to \mathbf{R}$ or a functional of the regression function, e.g., a weighted average $T(m_0) = \int m_0(x)w(x)\mathrm{d}x$. We consider the case where the covariate R is unobserved but an estimator $\widehat{R}$ of R is available. In this note, we provide some examples where such a situation occurs. Furthermore, appropriate forms of nonparametric estimators of m_0 are discussed and results on their asymptotic distribution are reviewed. In particular, we analyze how the real feasible estimator of m_0 obtained via regression on $\widehat{R}$ differs from the infeasible one obtained by regressing on R. With stochastic expansions for the difference of these two estimators, the asymptotic distribution of the real estimator of m_0 can be accurately described.

The note is organized as follows. In the next section, some examples illustrate how and where generated covariates typically appear in practice. Section 3 provides an overview of the asymptotic theory when m_0 is estimated by local linear estimation. In particular, the theory can also be applied to cases where the main interest is in averages of the regression function m_0, which is also important for some of the stated examples.

2 Examples

2.1 Simultaneous Nonparametric Equation Models Without Additivity (Imbens and Newey [3])

In economic models, there are often unobserved covariates which affect both response and observed covariates. Generally, such covariates which are correlated with the disturbance are called endogenous. Imbens and Newey [3] propose a regression model with endogenous covariates where the error variable does not enter additively into the model. This allows for general forms of unobserved heterogeneity which has led to recent popularity of such nonseparable models among economists.

They consider a general regression relation of the form

$$Y = \mu(X_1, Z_1, e)$$

where X_1 and Z_1 are observed covariates and Y is a one-dimensional response. While Z_1 is independent of the error variable e, no assumptions are made on the dependence between X_1 and e at this stage. For identification, however, assume that the endogenous variable X_1 is generated as

$$X_1 = h(Z_1, Z_2, V),$$

where Z_2 is an observed so-called instrumental variable not contained in the original equation, and (Z_1, Z_2) is independent of the joint vector of errors (e, V).

If the function h is strictly monotone in V, one can set without loss of generality that the conditional distribution of V given (Z_1, Z_2) is the uniform law on [0,1]. This can be achieved by putting

$$V = F_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)$$

and choosing h as the inverse of $F_{X_1|Z_1,Z_2}$. Then by definition, the conditional distribution of V given (Z_1, Z_2) does not depend on (Z_1, Z_2). Thus, V is independent of (Z_1, Z_2). Note that the above independence assumption is slightly more restrictive, because it does not only require that (Z_1, Z_2) is independent of each e and V separately, but also of (e, V) jointly.

For fixed values of z_1, z_2 and v and for $x_1 = h(z_1, z_2, v)$ it is straightforward to show

$$\begin{aligned} &E[\mu(x_1, z_1, e)|V = v] \\ &\quad = E[\mu(X_1, Z_1, e)|Z_1 = z_1, Z_2 = z_2, V = v] \\ &\quad = E[\mu(X_1, Z_1, e)|X_1 = x_1, Z_1 = z_1, V = v] \\ &\quad = E[Y|Z_1 = z_1, Z_2 = z_2, V = v]. \end{aligned}$$

Thus we can write

$$Y = m_0(R) + \varepsilon,$$

where

$$\begin{aligned} S &= (X_1, Z_1, Z_2), \\ R &= r_0(S) = (X_1, Z_1, F_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)) = (X_1, Z_1, V), \\ m_0(x_1, z_1, v) &= E[\mu(x_1, z_1, e)|V = v], \\ \varepsilon &= Y - E[Y|S]. \end{aligned}$$

In this model, the covariate V is unobserved, but an estimate

$$\hat{V} = \hat{F}_{X_1|Z_1,Z_2}(X_1, Z_1, Z_2)$$

of V is available. Thus, instead of R also use the feasible $\hat{R} = (X_1, Z_1, \hat{V})$. Then the function m_0 can be estimated by regressing Y onto $\hat{R}$. Let us denote this estimator as real, feasible estimator $\hat{m}$. One may compare this estimator to the theoretical, infeasible estimator $\tilde{m}$ obtained from regressing Y onto R. If the asymptotics of the theoretical estimator $\tilde{m}$ are well understood, an asymptotic understanding of $\hat{m}$ can be based on a stochastic expansion of the difference of $\hat{m} - \tilde{m}$.

The function m_0 is not of direct interest because it contains the nuisance covariate V. In general, the focus is on the so-called average structural function

$E[\mu(x_1, z_1, e)]$, the expected response if one exogenously fixes X_1 at x_1 and Z_1 at z_1. This function can be estimated by

$$\int_0^1 \widehat{m}(x_1, z_1, v)\mathrm{d}v.$$

Other functionals of interest are averages of the derivative $\partial\mu(x_1, z_1, e)/\partial(x_1, z_1)$.

2.2 Simultaneous Nonparametric Equation Models with Additivity (Newey, Powell and Vella [9])

In Newey et al. [9] a submodel of the regression model of the last subsection is considered. The setup differs from the last subsection by assuming that the error enters additively into the regression function, i.e.

$$Y = \mu(X_1, Z_1) + e.$$

For the control equation also an additive specification is used:

$$X_1 = h(Z_1, Z_2) + V,$$

but one could also proceed with the control equation of the last section.

With (Z_1, Z_2) independent of (e, V) as before, it is

$$E[Y|X_1, Z_1, Z_2] = \mu(X_1, Z_1) + \lambda(V) = E[Y|X_1, Z_1, V]$$

with $\lambda(V) = E[e|V]$. Thus we get an additive model where the regressor in the second additive component is not observed. This additive model can also be obtained under slightly weaker conditions, namely that $E[e|Z_1, Z_2, V] = E[e|V]$ and $E[V|Z_1, Z_2] = 0$.

There are two major approaches to fit an additive nonparametric model: marginal integration and backfitting. In Marginal Integration [4,8,10], first a full dimensional regression function $E[Y|X_1 = x_1, Z_1 = z_1, V = v]$ is estimated. And then in a second step, v is integrated out to obtain an estimate of $\mu(x_1, z_1)$. The first step of this procedure can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where

$$\begin{aligned} S &= (X_1, Z_1, Z_2), \\ R &= r_0(S) = (X_1, Z_1, X_1 - h(Z_1, Z_2)) = (X_1, Z_1, V), \\ m_0(r) &= E[Y|R = r], \\ \varepsilon &= Y - E[Y|R]. \end{aligned}$$

A fit of the unobserved R is given by $\hat{R} = (X_1, Z_1, \hat{V})$ with $\hat{V} = X_1 - \hat{h}(Z_1, Z_2)$ where $\hat{h}$ is a (nonparametric) estimator of the control function h.

In the Smooth Backfitting approach [5] for an additive model, estimates are obtained by iteration. As ingredients for the iteration algorithm, one needs estimators of the marginal expectations $E[Y|X_1, Z_1]$, $E[Y|V]$, and of the joint density of (X_1, Z_1, V). Here estimation of $E[Y|V]$ can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where now

$$\begin{aligned} S &= (X_1, Z_1, Z_2), \\ R &= r_0(S) = X_1 - h(Z_1, Z_2) = V, \\ m_0(v) &= E[Y|V = v], \\ \varepsilon &= Y - E[Y|V]. \end{aligned}$$

2.3 Marginal Treatment Effects (Heckman and Vytlacil [1, 2])

In Heckman and Vytlacil [1, 2] the following model for treatment effects is discussed: we observe D, Y_D, X, Z in

$$\begin{aligned} Y_d &= \rho(X, U_d, \theta_d) \quad \text{for } d = 0, 1 \\ D &= 1, \qquad\qquad \text{if } V \leq \mu(Z), \text{ and } D = 0, \text{ otherwise.} \end{aligned}$$

Here θ_0 and θ_1 are unknown parameters that are finite or infinite-dimensional. Furthermore, ρ is a known function. An example for a specification would be $\rho(X, U_d, \theta_d) = m_d(X) + U_d$ with a "nonparametric parameter" $\theta_d = m_d$. The variable D is a dummy variable that indicates if a person is treated or not. The model contains counterfactual outcomes. If a person is treated ($D = 1$) the outcome Y_1 is observed, assuming that there also exists an unobserved outcome Y_0 that would have been observed if the person had not been treated. The participation of the person in the treatment is driven by an unobserved variable V. Without loss of generality, set V as uniform distribution on $[0, 1]$. For identification of the model the following condition is required:

(U_0, V) and (U_1, V) are conditionally independent of Z given X.

Note that the norming of V implies that $P(D = 1|Z) = \mu(Z)$.

Here, a function of interest is the Marginal Treatment Effect $MTE(x, v) = E[Y_1 - Y_0|X = x, V = v]$, the expected treatment effect for an individual with covariate $X = x$ that lies on the v-quantile of the unobserved propensity to participate in the treatment. It holds that

$$\begin{aligned} MTE(x, v) &= E[Y_1 - Y_0|X = x, V = v] \\ &= -\frac{\partial}{\partial v} E[Y_D|X = x, \mu(Z) = v]. \end{aligned}$$

This follows because for $\delta > 0$ small:

$$
\begin{aligned}
MTE\,(x,v) &= E[Y_1 - Y_0 | X = x, V = v] \\
&\approx \delta^{-1}\,(-E[Y_1 I[V \geq v + \delta]|X = x] - E[Y_0 I[V < v + \delta]|X = x] \\
&\quad + E[Y_1 I[V \geq v]|X = x] + E[Y_0 I[V < v]|X = x]) \\
&= \delta^{-1}\,(-E[Y_1 I[V \geq v + \delta]|X = x, \mu(Z) = v + \delta] \\
&\quad - E[Y_0 I[V < v + \delta]|X = x, \mu(Z) = v + \delta] \\
&\quad + E[Y_1 I[V \geq v]|X = x, \mu(Z) = v] + E[Y_0 I[V < v]|X = x, \mu(Z) = v]) \\
&= \delta^{-1}\,(-E[Y_D I[V \geq v + \delta]|X = x, \mu(Z) = v + \delta] \\
&\quad\quad - E[Y_D I[V < v + \delta]|X = x, \mu(Z) = v + \delta] \\
&\quad\quad + E[Y_D I[V \geq v]|X = x, \mu(Z) = v] + E[Y_D I[V < v]|X = x, \mu(Z) = v]) \\
&= \delta^{-1}\,(-E[Y_D | X = x, \mu(Z) = v + \delta] + E[Y_D | X = x, \mu(Z) = v]) \\
&\approx -\frac{\partial}{\partial v} E[Y_D | X = x, \mu(Z) = v].
\end{aligned}
$$

Here estimation of (the partial derivative of) $E[Y_D|X = x, \mu(Z) = v]$ can be rewritten as a regression problem $Y = m_0(R) + \varepsilon$ with unobserved regressor R where now

$$
\begin{aligned}
Y &= Y_D, \\
S &= (X, Z), \\
R &= r_0(S) = (X, \mu(Z)), \\
m_0(r) &= E[Y|(X, \mu(Z)) = r], \\
\varepsilon &= Y_D - E[Y_D|(X, \mu(Z))].
\end{aligned}
$$

Many treatment effects parameters and other parameters can be written as weighted averages of $MTE(x, v)$. Estimation of the MTE function is again based on a regression problem with an unobserved covariate $\mu(Z)$. Here interest is in a partial derivative of the regression function.

2.4 Further Examples

Further examples of regression problems with unobserved covariates are sample selection models, censored regression models, generalized Roy models, stochastic volatility models, and semiparametric GARCH-in-Mean models. For a discussion and/or references of these models we refer to Mammen et al. [6].

3 Nonparametric Regression with Nonparametrically Generated Covariates

In all examples of the last section, the fit $\hat{R}$ of the unobserved covariate is of the form $\hat{R} = \hat{r}(S)$, where $\hat{r}$ is an estimator of a function r_0 that fulfills $R = r_0(S)$ for an observed covariate S. Thus we have the following nonparametric regression model

$$Y = m_0(r_0(S)) + \varepsilon,$$
$$E[\varepsilon|S] = 0.$$

In this section, we give a brief description of the asymptotics of a nonparametric estimator $\hat{m}$ that is based on regressing Y onto the fitted covariate $\hat{R} = \hat{r}(S)$. For illustration, we restrict the discussion to the special case where $\hat{m} = \widehat{m}_{LL}$ is a local linear estimator for an i.i.d. sample (S_i, Y_i), i.e., $\widehat{m}_{LL}(x) = \widehat{\alpha}$, where $(\widehat{\alpha}, \widehat{\beta})$ minimizes

$$\sum_{i=1}^{n} \left[Y_i - \alpha - \beta^T (\widehat{R}_i - x) \right]^2 K_h(\widehat{R}_i - x).$$

Here $K_h(u)$ is a product kernel:

$$K_h(u) = (h_1 \times \cdots \times h_q)^{-1} K_1(u_1) \times \cdots \times K_q(u_q)$$

for kernel functions $K_1, \ldots, K_q$ and a bandwidth vector $h = (h_1, \ldots, h_q)$. We call this estimator also the real estimator, in contrast to the theoretical estimator $\widetilde{m}_{LL}$ which is defined as $\widetilde{m}_{LL}(x) = \widetilde{\alpha}$ where $(\widetilde{\alpha}, \widetilde{\beta})$ minimizes

$$\sum_{i=1}^{n} \left[Y_i - \alpha - \beta^T (R_i - x) \right]^2 K_h(R_i - x).$$

Since the R_is are unobserved, this theoretical estimator is infeasible. It is, however, introduced here because its asymptotic behaviour is well understood. Thus, for the asymptotic properties of the real estimator we only need a stochastic expansion of $\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x)$. Such an expansion was derived in Mammen et al. [6]. Tailored results for parameters obtained as functionals of m are derived in Mammen, Rothe and Schienle [7].

For the comparison of $\widehat{m}_{LL}$ and $\widetilde{m}_{LL}$, Mammen et al. [6] use three types of assumptions: besides standard smoothing assumptions, these are conditions on accuracy (A) and complexity (C) of the estimator $\widehat{r}$ of r_0. The assumption (A) requires that $\hat{r}$ converges to r_0 with a rate that is fast enough. The assumption (C) states that there exist sequences of sets $\mathscr{M}_n$ with two properties: (1) $\widehat{r} \in \mathscr{M}_n$ with probability tending to one. (2) The sets $\mathscr{M}_n$ are not too large, where size is measured by entropy. The main result in Mammen et al. [6] is the following expansion

$$\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x) \approx -m'(x) \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)}.$$

This result can be interpreted as follows: The real estimator $\widehat{m}_{LL}(x)$ and the oracle estimator $\widetilde{m}_{LL}(x)$ differ by a local weighted average of the estimator of r_0:

$$-m'(x) \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(\widehat{r}(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)}.$$

This local average is of the order of the bias of $\widehat{r}$ but it may have a faster rate as the variance part of $\widehat{r}$. Thus we can conclude that for achieving a certain rate of convergence for estimating m_0, it is not necessary that an estimator of r_0 has the same or a faster rate. A similar result can be obtained for derivatives of the regression function.

We now shortly outline the main ideas of how the expansion of $\widehat{m}_{LL}(x) - \widetilde{m}_{LL}(x)$ was obtained in Mammen et al. [6]. We want to compare:

real estimator $\widehat{m}_{LL}$ = SMOOTH of $\widehat{r}(S)$ versus $m_0(r_0(S)) + \varepsilon$,
oracle estimator $\widetilde{m}_{LL}$ = SMOOTH of $r_0(S)$ versus $m_0(r_0(S)) + \varepsilon$.

Now, because of additivity of the operator SMOOTH, it is

$$\begin{aligned}\widehat{m}_{LL} = \ & \text{SMOOTH of } \widehat{r}(S) \quad \text{versus} \quad m_0(\widehat{r}(S)) + \varepsilon \\ & + \text{SMOOTH of } \widehat{r}(S) \quad \text{versus} \quad m_0(r_0(S)) - m_0(\widehat{r}(S)).\end{aligned}$$

If $\widehat{r}$ was nonrandom we get, because $|\widehat{r}(S) - r_0(S)|$ is small by assumption (A),

$$\begin{aligned}\widehat{m}_{LL} \approx \ & \text{SMOOTH of } \quad r_0(S) \text{ versus } m_0(r_0(S)) + \varepsilon \\ & + \text{SMOOTH of } \widehat{r}(S) \quad \text{versus} \quad m_0(r_0(S)) - m_0(\widehat{r}(S)) \\ \approx \ & \widetilde{m}_{LL} \\ & + \text{SMOOTH of } r_0(S) \quad \text{versus} \quad m_0'(r_0(S))(r_0(S) - \widehat{r}(S)).\end{aligned}$$

This is (nearly) the formula of the desired expansion.

It remains to take care of the fact that $\widehat{r}$ is random and not purely deterministic. In order to do so, the argument must be uniform over the set of possible realizations of $\widehat{r}$. This can be achieved by an empirical process worst case analysis. We must show that

$$\left|\widehat{m}_{LL,r}(x) - \widetilde{m}_{LL}(x) + m'(x) \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)(r(S_i) - r_0(S_i))}{\frac{1}{n}\sum_{i=1}^{n} K_h(r_0(S_i) - x)}\right|$$

is small uniformly for r in $\mathscr{M}_n$. Here $\widehat{m}_{LL,r}$ is the local linear estimator based on regressing Y onto $r(S)$. At this stage of the proof one makes use of Assumption (C).

References

1. Heckman, J.J., Vytlacil, E.J.: Structural equations, treatment effects, and econometric policy evaluation. Econometrica **73**, 669–738 (2005)
2. Heckman, J.J., Vytlacil, E.J.: Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, J.J., Leamer, E.E. (eds.) Handbook of Econometrics, vol. 6, chapter 71. Elsevier, Amsterdam (2007)
3. Imbens, G.W., Newey, W.K.: Identification and estimation of triangular simultaneous equations models without additivity. Econometrica **77**, 1481–1512 (2009)
4. Linton, O., Nielsen, J.P.: Kernel estimation of partial means and a general variance estimator. Biometrika **82**, 93–100 (1995)
5. Mammen, E., Linton, O., Nielsen, J.P.: The existence and asymptotic properties of a backfitting algorithm under weak conditions. Ann. Stat. **27**, 1443–1490 (1999)
6. Mammen, E., Rothe, C., Schienle, M.: Nonparametric regression with nonparametrically generated covariates. Ann. Stat. **40**, 1132–1170 (2012)
7. Mammen, E., Rothe, C., Schienle, M.: Semiparametric estimation with generated covariates. Preprint (2012)
8. Newey, W.K.: A kernel method of estimating structured nonparametric regression based on marginal integration. Econometric Theory **10**, 233–253 (1994)
9. Newey, W.K., Powell, J.L., Vella, F.: Nonparametric estimation of triangular simultaneous equations models. Econometrica **67**, 565–603 (1999)
10. Tjøstheim, D., Auestad, B.H.: Nonparametric identification of nonlinear time series: selecting significant lags. J. Am. Stat. Assoc. **89** (1994)

A Generator of Heavy-Tailed Search Trees

Alda Carvalho and Carlos Santos

Abstract

In theoretical computational statistics, the detection of heavy tails in computing costs of some random search algorithms generated an increased interest in the modeling of this heavy-tailed behavior. We propose a general model for these algorithms that reproduces this characteristic. The model represents the search for a solution as the descent of a search tree with regular branching factors and equiprobable nodes. In this work, we present a way to generate heavy-tailed search trees using very simple rules, and we show how a particular case relates to the famous Wallis product.

1 Introduction

Constraint Satisfaction Problems (CSP) are the subject of intense research in computer science. They often exhibit high complexity, requiring a combination of heuristics and combinatorial search methods to be solved in a reasonable time. In the last decade we have witnessed tremendous progress in the design and development of complete, backtrack style search algorithms, for CSP.

Randomized search methods have greatly extended our ability to solve hard computational problems. In many cases, the incorporation of randomness into the computational process leads to a significant speed up over purely deterministic methods.

A. Carvalho (✉)
ISEL and CEMAPRE, Rua Conselheiro Emídio Navarro, 1, 1959-007, Lisboa, Portugal
e-mail: acarvalho@adm.isel.pt

C. Santos
ISEC, Alameda das Linhas de Torres, 179, 1750-142, Lisboa, Portugal
e-mail: carlos.santos@isec.universitas.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_12, © Springer-Verlag Berlin Heidelberg 2013

Backtrack random search can be viewed as a search tree. When a CSP is solved very quickly, a bad decision can lead to the exploration of a large sub-tree. The computing costs associated with these problems have been measured by the number of backtracks necessary to find a solution or to prove no solution exists. Such measure is a proxy for computing times and is independent of the particular machine used to perform the simulations.

Sometimes the number of bad choices has a strong impact in the computing costs. In these cases, the variance is very high and a power law decay can be observed. It is now a well-established fact that search algorithms can exhibit heavy-tailed behavior [1–5]. However, the reasons behind this fact are not well understood. So, a better understanding of this kind of distributions is well appreciated and it is already an important research subject.

Chen et al. [4] suggested random search trees as models for random search algorithms. These models are necessary, as actual algorithms follow very complex rules and follow ad hoc heuristics, making it difficult to understand the reasons behind the heavy-tailed behavior. However, these first tree models were just indicative and did not explain rules for generating the adequate trees.

Simple search trees that reproduce the desired distributions are difficult to set with easy and plausible rules. If the nodes are equiprobable and the branching factor constant, then one is led to set trees with an exponential behavior. If the nodes are not equiprobable, then it is trivial to set trees that generate any desired distribution. However, such tree models are not realistic and do not provide any insight into the algorithms behavior. Starting with a tree related to the Wallis product, a tree search class model with very simple rules is presented. In this model, decisions are equiprobable at each node; therefore it mimics a key characteristic of real random algorithms. Also, it follows very simple rules and is capable of displaying a heavy-tailed behavior.

2 The "Wallis Forest"

Consider the following "game" (not a game in the true sense of the term but an allegory of a random search tree). The Little Red Riding Hood enters in a dark forest where it is known that there are ravenous wolves. We know where they are, but the Little Red Riding Hood does not know, so she chooses her path at random. The dark forest, which can be considered infinite, is full of crosses. There is a first bifurcation in which one way runs into a hungry wolf and the other way leads to a new crossing. This new crossover has three new ways, one going to meet another wolf, while the other two lead to four new crosses... The process continues and is illustrated in the figures below (Figs. 1–3):

If X is the random variable counting the number of crosses visited by the Little Red Riding Hood, the rising question is *what is* $P(X > n)$*?* In other words, considering the starting level $n = 1$, *what is the probability that Little Red Riding Hood survive to the wolves of the first* n *levels?* With simple calculations we obtain the following responses:

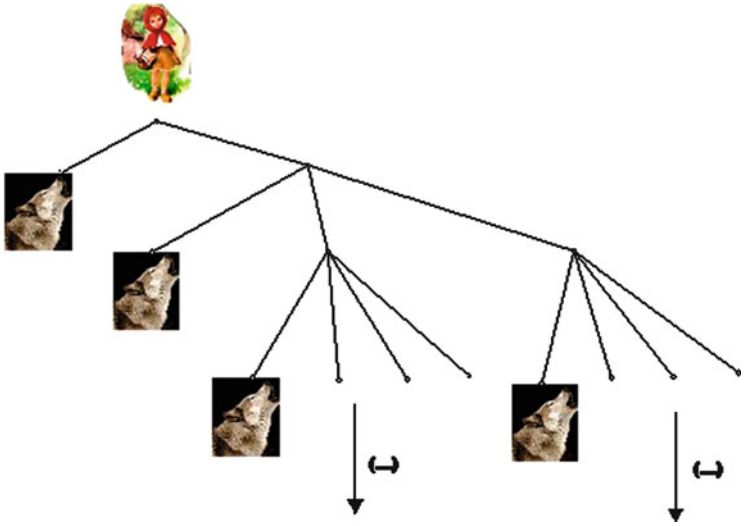

Fig. 1 Dark forest

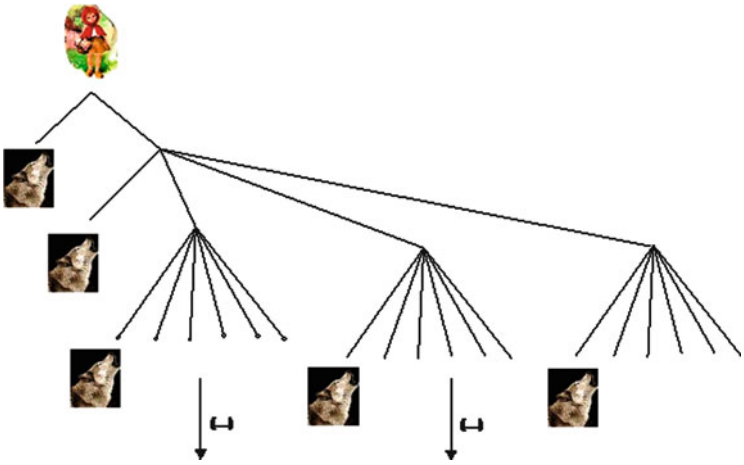

Fig. 2 Wallis forest

1. The probability that Little Red Riding Hood be eaten at the level $n > 1$ is equal to $P(X = n) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \cdots \times \frac{n-2}{n-1} \times \frac{1}{n} = \frac{1}{n-1} \times \frac{1}{n}$. We also have $P(X = 1) = 0$ (there are no wolves in the first level).
2. The probability that Little Red Riding Hood survive to the wolves of the first $n + 1$ levels is $P(X > n + 1) = \frac{1}{2} \times \frac{2}{3} \times \frac{3}{4} \times \cdots \times \frac{n}{n+1} = \frac{1}{n+1}$. We have $P(X > 1) = 1$ because there are no wolves in the first level.

Fig. 3 John Wallis (1616–1703)

$$\frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \frac{4}{5} \times \ldots = \prod_{n=1}^{\infty} \frac{4n^2}{4n^2-1} = \frac{\pi}{2}$$

Note that, in this case, the tail is not negligible. The more the Little Red Riding Hood survives, the greater her chance of surviving the next level. This simple model corresponds to a hyperbolic decay in the tail of the distribution of X, with $\alpha = 1$.

Let us consider a more interesting tree based on the idea exposed in the introduction. In this new tree, the number of paths at each crossing is no longer $(2, 3, 4, 5 \ldots)$ as in the previous case and becomes $(2, 4, 6, 8 \ldots)$.

In this case, we obtain,

$$P(X > n+1) = \frac{1}{2} \times \frac{3}{4} \times \frac{5}{6} \times \cdots \times \frac{2n-1}{2n}. \tag{1}$$

We have a nice way to estimate this value dated back to 1655, when John Wallis proposed his famous product ([6] is a very recent reference about this subject).

The key to understand the behavior of $P(X > n+1)$ is considering $P^2(X > n+1)$ instead. We obtain

$$P^2(X > n+1) = \frac{1}{2} \times \frac{1}{2} \times \frac{3}{4} \times \frac{3}{4} \times \frac{5}{6} \times \frac{5}{6} \times \cdots \times \frac{2n-1}{2n} \times \frac{2n-1}{2n}. \tag{2}$$

We know that $w(2n) \to \frac{\pi}{2}$ where $w(2n) = \frac{2}{1} \times \frac{2}{3} \times \cdots \times \frac{2n-2}{2n-1} \times \frac{2n}{2n-1}$ ($w(2) = \frac{2}{1}$, $w(4) = \frac{2}{1} \times \frac{2}{3} \times \frac{4}{3} \times \cdots$). So, the last expression leads to

$$P^2(X > n+1) = \frac{1}{w(2n)} \times \frac{1}{2n} \Leftrightarrow P(X > n+1) = \frac{1}{\sqrt{w(2n)}} \times \frac{1}{\sqrt{2n}}. \tag{3}$$

The immediate consequence is $P(X > n+1) \sim \sqrt{\frac{1}{\pi}} \times \frac{1}{\sqrt{n}}$. This model corresponds to a hyperbolic decay in the tail of the distribution of X, with $\alpha = \frac{1}{2}$.

3 The General Model

The previous examples suggest a possible generalization. In fact, it is possible to construct a random search tree with hyperbolic decay with rational exponent α. For an arbitrary $\alpha = \frac{p}{q}$, we present such a construction, showing that $P(X > n+1) \sim Cn^{-\alpha}$.

In the general case, a law stating the number of solution-nodes and no-solution nodes at each level is defined. We can generalize the example of the previous section fixing the number of solution nodes at each level and increasing the number of edges linearly with the level. The fixed number of node solutions is p. The linearly increasing branching factor is q.

The first level node has k edges and p solutions ($p < k$). The second level no-solution nodes have each $k+q$ edges and p solutions, the third level no-solution nodes have each $k+2q$ edges with p solutions, and so on. As an example, we can see on Fig. 4 the three first levels for $k = 3$, $p = 2$, $q = 3$ ($\alpha = \frac{2}{3}$).

Theorem 3.1. *Let X be the random variable that counts the number of visited nodes on the random search tree until a solution is found. For natural numbers k, p, q and the random search tree constructed above, we have*

$$P(X > n+1) \sim Cn^{-\frac{p}{q}},$$

where $C = \frac{\Gamma(\frac{k}{q})}{\Gamma(\frac{k-p}{q})}$, and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t$, $x > 0$ is the Gamma function.

Proof. We have, for $n > 1$,

$$P(X > n+1) = \frac{k-p}{k} \times \frac{k-p+q}{k+q} \times \frac{k-p+2q}{k+2q} \times \cdots \times \frac{k-p+(n-1)q}{k+(n-1)q}$$

and then

$$P(X > n+1) = \frac{q^n \times \frac{\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)}}{q^n \times \frac{\Gamma\left(n+\frac{k}{q}\right)}{\Gamma\left(\frac{k}{q}\right)}} = \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} \times \frac{\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)},$$

where we used the fact $\Gamma(n+z) = (n-1+z) \times \cdots \times z \times \Gamma(z)$.

To prove that $P(X > n+1) \sim n^{-\frac{p}{q}}$, we calculate $\lim_{n\to\infty} P(X > n+1) \times n^{\frac{p}{q}}$. We have

$$P(X > n+1) \times n^{\frac{p}{q}} = \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} \times \frac{n^{\frac{p}{q}} \times \Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)}.$$

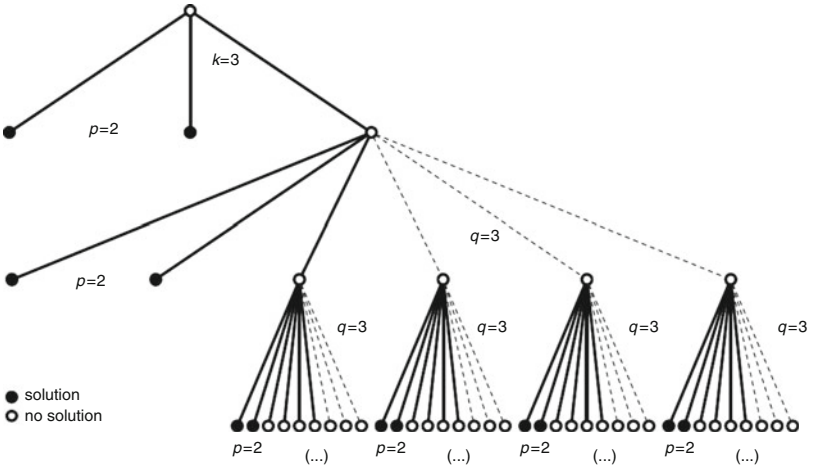

Fig. 4 Random search tree with power law decay with $\alpha = \frac{2}{3}$

Also, because $\lim_{n\to\infty} \frac{\Gamma(n+a)}{\Gamma(n+b)} n^{b-a} = 1$, we have $n^{\frac{p}{q}} \frac{\Gamma\left(n+\frac{k-p}{q}\right)}{\Gamma\left(n+\frac{k}{q}\right)} \to 1$ so,

$$P(X > n+1) \times n^{\frac{p}{q}} \to \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} \qquad \text{and} \qquad P(X > n+1) \sim \frac{\Gamma\left(\frac{k}{q}\right)}{\Gamma\left(\frac{k-p}{q}\right)} n^{-\frac{p}{q}}.$$

□

4 Final Remarks

We gave a first example for a tree that generates a distribution with index of stability $\alpha = 1$ and showed how is connected with the Wallis product. We have exhibited a class of random search trees with equiprobable nodes displaying heavy tails for the number of nodes visited until a solution is reached and showed that this class of trees is able to generate any rational index of stability.

References

1. Adler, R., Feldman, R., Taqqu, M.: A Practical Guide to Heavy Tails. Birkauser, Boston (1998)
2. Carvalho, A.: Probabilistic models for a class of computing cost distributions. PhD Thesis, ISEG, Thecnical University of Lisbon (2010)
3. Carvalho, A., Crato, N., Gomes, C.: A generative power-law search tree model. Comput. Oper. Res., Elsevier Science Ltd. **36**(8), 0305–0548 (2009)

4. Chen, H., Gomes, C., Selman, B.: Formal Models of Heavy-Tailed Behaviour in Combinatorial Search. Lecture Notes in Computer Science, vol. 2239, pp. 408–422. Springer, Berlin (2001)
5. Gomes, C., Selman, B., Crato, N., Kautz,H.: Heavy-tailed phenomena in satisfiability and constraint satisfaction problems. J. Automat. Reas. **24**, 67–100 (2000)
6. Wästlund, J.: An elementary proof of the Wallis product formula for pi. Am. Math. Mon. **114**, 914–917 (2007)

Part III

Extremes

Adaptive PORT-MVRB Estimation of the Extreme Value Index

M. Ivette Gomes and Lígia Henriques-Rodrigues

Abstract

In this chapter, we consider an application to environmental data of a bootstrap algorithm for the adaptive estimation of the extreme value index (EVI), the primary parameter in *Statistics of Extremes*. The EVI estimation is performed through the recent *Peaks Over Random Threshold Minimum-Variance Reduced-Bias* (PORT-MVRB) estimators, which apart from scale invariant, like the classical ones, are also location invariant. These estimators depend not only on an integer *tuning* parameter k, the number of top order statistics involved in the estimation, but also on an extra *control* real parameter q, $0 \leq q < 1$, which makes them highly flexible.

1 Introduction

Slightly more restrictively than working in the whole domain of attraction (for maxima), $\mathscr{D}_{\mathscr{M}}\left(EV_\gamma\right)$, of the *extreme value distribution* (EVD),

$$EV_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right),\ 1+\gamma x \geq 0 & \text{if } \gamma \neq 0 \\ \exp(-\exp(-x)),\ x \in \mathbb{R} & \text{if } \gamma = 0, \end{cases} \tag{1}$$

we shall consider a positive *extreme value index* (EVI), i.e. $\mathscr{D}_{\mathscr{M}}\left(EV_\gamma\right)_{\gamma>0}$.

M.I. Gomes (✉)
CEAUL and FCUL, Campo Grande, 1749-016, Lisboa, Portugal
e-mail: ivette.gomes@fc.ul.pt

L. Henriques-Rodrigues
CEAUL and Instituto Politécnico de Tomar, Portugal
e-mail: Ligia.Rodrigues@aim.estt.ipt.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_13, © Springer-Verlag Berlin Heidelberg 2013

Let us introduce the notations $F^{\leftarrow}(y) := \inf\{x : F(x) \geq y\}$ for the generalised inverse function of F, and RV_α for the class of regularly varying functions at infinity with an index of regular variation α, i.e. positive measurable functions g such that $\lim_{t\to\infty} g(tx)/g(t) = x^\alpha$, for some $\alpha \in \mathbb{R}$ and every $x > 0$. Then,

$$F \in \mathscr{D}_{\mathscr{M}}\left(EV_\gamma\right)_{\gamma>0} \iff U(t) := F^{\leftarrow}\big(1 - 1/t\big) \in RV_\gamma. \tag{2}$$

For these *heavy-tailed* models, given a sample $\underline{\mathbf{X}}_n = (X_1, \ldots, X_n)$, and the associated sample of ascending order statistics (os's), $(X_{1:n} \leq \cdots \leq X_{n:n})$, the classical EVI estimators are Hill estimators [14],

$$H_{k,n} \equiv H_{k,n}(\underline{\mathbf{X}}_n) := \frac{1}{k}\sum_{i=1}^{k}\left\{\ln\ \frac{X_{n-i+1:n}}{X_{n-k:n}}\right\}, \tag{3}$$

the average of the k log-excesses over a random level $X_{n-k:n}$, which needs to be an *intermediate* os, i.e. we need to have $k = k_n \to \infty$ and $k/n \to 0$, as $n \to \infty$, so that $H_{k,n}$ is consistent for the estimation of γ.

In this chapter, we apply the *Peaks Over Random Threshold* (PORT) methodology, recently introduced in [1] for classical estimators, to *Minimum-Variance Reduced-Bias* (MVRB) EVI estimators. The EVI, the shape parameter γ in (1), is the crucial parameter in *Statistics of Extremes*. The PORT EVI estimators, to be defined in Sect. 2, depend upon an extra *tuning parameter* q, $0 \leq q < 1$, which makes them highly flexible. Moreover, they are invariant for changes in location and scale, a property that does not hold for most of the classical estimators, which are invariant only for scale changes. In Sect. 3, we describe the most simple MVRB EVI estimator, the one studied in [3]. In Sect. 4, we shall provide a few details on an adaptive MVRB EVI estimation, taking into account the developments in [11] and [12], and provide an algorithm for the adaptive PORT-MVRB EVI estimation. Finally, in Sect. 5, we apply the methodology to a set of environmental data.

2 PORT EVI Estimation

Hill estimators are scale-invariant, but not location-invariant. And they can suffer drastic changes whenever we induce an arbitrary shift in the data. We can then easily obtain pictures similar to the so-called *Hill horror plots* [15], associated with a regularly varying function $L_U(t) = t^{-\gamma}U(t)$ going to infinity or zero, with $U(t)$ the (reciprocal) quantile function defined in (2). Such a kind of plots led the authors in [1] to introduce the so-called PORT estimators.

The PORT EVI estimators are functionals of a sample of excesses over a random level $X_{[nq]+1:n}$, i.e. functionals of

$$\underline{\mathbf{X}}_{k,n}^{(q)} := \left(X_{n:n} - X_{[nq]+1:n}, \ldots, X_{n-k:n} - X_{[nq]+1:n}\right), \tag{4}$$

with $1 \leq k < n - [nq] - 1$, and where

- $0 < q < 1$, for any $F \in \mathscr{D}_{\mathscr{M}}(EV_\gamma)_{\gamma>0}$ (*the random level is an empirical quantile*);
- $q = 0$, for distribution functions with a finite left endpoint x_F (*the random threshold is the minimum*).

If we think, for instance, on the Hill estimators, the class of associated PORT EVI estimators, the so-called PORT-Hill estimators, asymptotically studied in [1], and for finite samples in [9], are given by

$$H_{k,n}^{(q)} := H_k(\underline{\mathbf{X}}_{k,n}^{(q)}) = \frac{1}{k}\sum_{i=1}^{k}\Big\{\ln\frac{X_{n-i+1:n}-X_{[nq]+1:n}}{X_{n-k:n}-X_{[nq]+1:n}}\Big\}, \quad 0 \le q < 1.$$

Remark 2.1. The PORT EVI estimators are now both location and scale invariant.

Remark 2.2. Moreover, they depend on a *tuning parameter* q, which provides a highly flexible class of EVI estimators. Indeed, the PORT-Hill estimators can even overpass the MVRB EVI estimators, to be introduced next in Sect. 3, provided that we chose adequately the control parameter q (see [13]).

Remark 2.3. The choice $q = 0$ is appealing in practice, but needs to be carefully used, since it can induce a drastic sub-estimation and even inconsistency (see [9]).

3 MVRB EVI Estimation

The main reasons for the need of reduced-bias (RB) EVI estimation are sketched in the following. Hill estimators, $H_k \equiv H_{k,n}$, in (3), reveal usually a high asymptotic bias, i.e. $\sqrt{k}\,(H_k - \gamma)$ is asymptotically normal with variance γ^2 and a non-null mean value equal to $\lambda/(1-\rho)$, whenever $\sqrt{k}\,A(n/k) \to \lambda \neq 0$, finite. This non-null asymptotic bias and a rate of convergence of the order of $1/\sqrt{k}$ lead to sample paths with a high variance for small k, a high bias for large k, and a *very sharp* mean square error (MSE) pattern, as a k-function.

Recently, several authors have dealt with bias reduction in the field of *Extremes* and a simple class of MVRB EVI estimators is the one introduced in [3] and used for the MVRB semi-parametric estimation of $\ln VaR_p$ in [8]. Such a class, denoted $\overline{H} \equiv \overline{H}_k$, depends on the adequate estimation of second-order parameters (β, ρ), in a class slightly more restrict than $\mathscr{D}_{\mathscr{M}}(EV_\gamma)_{\gamma>0}$. More specifically, we shall assume that the regularly varying function $L_U(t) = t^{-\gamma}U(t)$ tends to a finite and non-null constant, i.e. with $\gamma > 0$, $\rho < 0$, and $\beta \neq 0$,

$$U(t) = Ct^{\gamma}\big(1 + A(t)/\rho + o(t^{\rho})\big), \qquad A(t) =: \gamma\beta t^{\rho}. \tag{5}$$

The functional form of the above-mentioned class of MVRB EVI estimators is

$$\overline{H}_k \equiv \overline{H}_{k,n} \equiv \overline{H}_{k,n}(\hat{\beta},\hat{\rho}) \equiv \overline{H}_{\hat{\beta},\hat{\rho}}(k) := H_k\big(1 - \hat{\beta}(n/k)^{\hat{\rho}}/(1-\hat{\rho})\big), \tag{6}$$

where $(\hat{\beta}, \hat{\rho})$ is a consistent adequate estimator of (β, ρ), with $\hat{\beta}$ and $\hat{\rho}$ based on a number of top order statistics, k_1, of a higher order than the number of upper order statistics, k, usually considered for the EVI estimation.

Remark 3.1. Algorithms for the adequate estimation of (β, ρ), as well as heuristic algorithms for the choice of k in the estimation of γ through $\overline{H}_k$, are provided in [7].

Remark 3.2. The bootstrap methodology in [11] and [12] can also lead us to the estimation of the optimal sample fraction (OSF), $k_0^{\overline{H}}(n)/n = \arg\min_k MSE\{\overline{H}_k\}/n$.

Remark 3.3. We shall again consider the class of ρ estimators introduced in [4], denoted $\widehat{\rho}_{k,n}(\tau)$. They are parameterised in a *tuning parameter* $\tau \in \mathbb{R}$. For an appropriate choice of $\tau =: \tau^*$, these ρ estimators exhibit stable sample paths, as functions of k, for high values of k. Consistency is attained provided that k is *intermediate* and such that $\sqrt{k}\, A(n/k) \to \infty$, as $n \to \infty$.

Remark 3.4. Regarding the estimation of β, we shall use the class of estimators $\widehat{\beta}_{k,n}(\hat{\rho})$ introduced in [5], further studied in [10] and in [2]. Consistency of $\widehat{\beta}_{k,n}(\hat{\rho})$ is achieved for *intermediate* k such that $\sqrt{k}\, A(n/k) \to \infty$, as $n \to \infty$, and ρ estimators, $\hat{\rho}$, such that $\hat{\rho} - \rho = o_p(1/\ln n)$.

Remark 3.5. In order to guarantee the validity of this condition on $\widehat{\rho}$, we compute $\widehat{\rho}_{k,n}(\tau)$ at $k = k_1 = \left[n^{1-\epsilon}\right]$, $\epsilon = 0.005$. Then, and provided that $\sqrt{k_1}\, A(n/k_1) \to \infty$, we get $\hat{\rho} - \rho := \hat{\rho}_{k_1,n}(\tau) - \rho = o_p(1/\ln n)$, a condition also needed so that $\overline{H}_{k,n}$ keeps the same asymptotic variance of the Hill estimator. Note that with such a choice of k_1, $\sqrt{k_1}\, A(n/k_1) \to \infty$ if and only if $\rho > 1/2 - 1/(2\epsilon) = -99.5$, an irrelevant restriction from a practical point of view, due to the fact that the Hill estimator has practically no bias if $|\rho| > 1$.

Based on these comments, and on all algorithms so far proposed, we shall now propose (and implement) the following:

Simplified Algorithm (second-order estimates):

1. Given the sample $(x_1, \ldots, x_n)$, consider $\{\hat{\rho}_{k,n}(\tau)\}_{k \in \mathscr{K}}$, $\mathscr{K} = ([n^{0.99}], [n^{0.999}])$, and compute the median value, denoted χ_τ, for $\tau = 0$ and $\tau = 1$.
2. Compute $I_\tau := \sum_{k \in \mathscr{K}} \left(\hat{\rho}_{k,n}(\tau) - \chi_\tau\right)^2$, $\tau = 0, 1$.
   Next choose the *tuning parameter* $\tau^* = 0$ if $I_0 \leq I_1$; otherwise, choose $\tau^* = 1$.
3. Work with $\hat{\rho}^* \equiv \hat{\rho}_{\tau^*} = \hat{\rho}_{k_1,n}(\tau^*)$ and $\hat{\beta}^* \equiv \hat{\beta}_{\tau^*} := \hat{\beta}_{k_1,n}(\hat{\rho}_{\tau^*})$, $k_1 = [n^{0.995}]$.

This algorithm leads in most of the cases to the *tuning parameter* $\tau = 0$ when $|\rho| \leq 1$ and $\tau = 1$, otherwise.

3.1 A Few Comments on the Asymptotic Behaviour of MVRB EVI Estimators

Let $k = k_n$ be an intermediate sequence of integers in $[1, n)$, such that $\sqrt{k}\, A(n/k) \to \lambda$, finite, as $n \to \infty$. Let us denote $\hat{\gamma}_{k,n}$ either $H_{k,n}$, in (3), or $\overline{H}_{k,n}$, in (6). Then, for the class of models under study, i.e. the class in (5),

$$\sqrt{k}\big(\hat{\gamma}_{k,n} - \gamma\big) \xrightarrow[n\to\infty]{d} \text{Normal}(\lambda\, b_{\hat{\gamma}},\ \sigma^2_{\hat{\gamma}}),$$

even when we work with RB EVI estimators, provided that $(\hat{\beta}, \hat{\rho})$ is consistent for the estimation of (β, ρ), and $(\hat{\rho} - \rho)\ln(n/k) = o_p(1)$, as $n \to \infty$. Whereas $b_H = 1/(1-\rho)$, we get $b_{\overline{H}} = 0$. Moreover, we have $\sigma^2_{H} = \sigma^2_{\overline{H}} = \gamma^2$. Consequently, for every k, the $\overline{H}$ estimators outperform the H estimators.

Asymptotic normality and the asymptotic bias behaviour of the MVRB estimators are easier to refer if we again slightly restrict the class of models under consideration, assuming that, $U(t) = Ct^{\gamma}\big(1 + A(t)/\rho + \beta' t^{2\rho} + o(t^{2\rho})\big)$, $A(t) = \gamma\beta t^{\rho}$, as $t \to \infty$, with $C, \gamma > 0$, $\beta, \beta' \neq 0$, $\rho < 0$. Then, under slight restrictions on k (see [2]), and with $\overline{Z}_k$ asymptotically standard normal, we can guarantee that

$$\overline{H}_{k,n}(\hat{\beta}, \hat{\rho}) \overset{d}{=} \gamma + \frac{\gamma\, \overline{Z}_k}{\sqrt{k}} + \Big(b_{\overline{H}}\, A^2(n/k) + O_p\Big(\frac{A(n/k)}{\sqrt{k}}\Big)\Big)(1 + o_p(1)),$$

where, with $\xi = \beta'/\beta$, $b_{\overline{H}} = \big(\xi/(1-2\rho) - 1/(1-\rho)^2\big)/\gamma$.

Consequently, even if $\sqrt{k}\, A(n/k) \to \infty$, with $\sqrt{k}\, A^2(n/k) \to \lambda_A$, finite, the type of k values where the MSE of $\overline{H}_{k,n}$ is minimum,

$$\sqrt{k}\,\big(\overline{H}_{k,n}(\hat{\beta}, \hat{\rho}) - \gamma\big) \xrightarrow[n\to\infty]{d} \text{Normal}\big(\lambda_A\, b_{\overline{H}},\ \sigma^2_{\overline{H}} = \gamma^2\big).$$

The bootstrap methodology enables us to estimate $k_0^{\overline{H}}(n) = \arg\min_k MSE\big\{\overline{H}_{k,n}\big\}$, in a way similar to the one used for the classical EVI estimators (see [6], among others), now through the use of an auxiliary statistic like $T_n^{\overline{H}}(k) := \overline{H}_{[k/2],n} - \overline{H}_{k,n}$, $k = 2, \ldots, n-1$. Such a method, detailed in [11], is now generalised in Sect. 4. Further results can be found in [12].

4 Adaptive PORT-MVRB EVI Estimation

The PORT EVI estimators in [1] were the PORT-Hill and the PORT-moment estimators. However, and despite of the fact that we have not yet developed all non-degenerate asymptotic theory related to PORT-ρ, PORT-β, or perhaps PORT-$A(\cdot)$, followed by PORT-MVRB EVI estimators, we think sensible to advance with the PORT-MVRB EVI estimators,

$$\overline{H}_{k,n}^{(q)} = \overline{H}_{k,n}\big(\hat{\beta}^{*(q)}, \hat{\rho}^{*(q)}\big),$$

based on PORT-ρ estimators, $\hat{\rho}^{*(q)}$, and PORT-β estimators, $\hat{\beta}^{*(q)}$, based on the sample of excesses in (4), as well as associated PORT-MVRB estimators of any parameter of extreme events. If we implement steps from 4. until 10. on the estimator $\overline{H}_k$, in (6), we shall obtain the adaptive MVRB estimator, denoted $\overline{H}^* \equiv \overline{H}^*_{n,n_1|T}$.

Algorithm (cont.) – adaptive bootstrap PORT-MVRB estimation of γ:

4. For $q = 0(0.05)0.95$, compute $\overline{H}_{k,n}^{(q)}$, $k = 1, 2, \ldots$.
5. Consider subsample sizes $n_1 = o(n)$ and $n_2 = [n_1^2/n] + 1$, *a fortiori* a $o(n)$.
6. For l from 1 until $B = 250$, generate independently, from the empirical distribution function, $F_n^*(x) = \sum_{i=1}^n I_{[X_i \le x]}/n$, associated with the observed sample $(x_1, x_2, \ldots, x_n)$, B bootstrap samples $(x_1^*, \ldots, x_{n_2}^*)$ and $(x_1^*, \ldots, x_{n_2}^*, x_{n_2+1}^*, \ldots, x_{n_1}^*)$, of sizes n_2 and n_1, respectively.
7. Denote $T_n^*(k)$ the bootstrap counterpart of the auxiliary statistic $T_n(k) := \overline{H}_{[k/2],n}^{(q)} - \overline{H}_{k,n}^{(q)}$, obtain $(t_{n_1,l}^*(k),\ t_{n_2,l}^*(k))$, $1 \le l \le B$, observed values of the statistic $T_{n_i}^*(k)$, $i = 1, 2$, and compute

$$MSE^*(n_i, k) = \frac{1}{B}\sum_{l=1}^{B}\big(t_{n_i,l}^*(k)\big)^2, \qquad k = 1, 2, \ldots, n_i - 1, \quad i = 1, 2.$$

8. Obtain $\hat{k}_{0T}^*(n_i) := \arg\min_{1 \le k \le n_i - 1} MSE^*(n_i, k)$, $i = 1, 2$.
9. Compute

$$\hat{k}_0^{(q)} \equiv \hat{k}_{0|\overline{H}^{(q)},T} := \min\left(n - 1, \left[(1 - 2^{\hat{\rho}^*})^{2/(1-2\hat{\rho}^*)}\,\frac{\big(\hat{k}_{0T}^*(n_1)\big)^2}{\hat{k}_{0T}^*(n_2)}\right] + 1\right).$$

10. Obtain $\overline{H}_{n,n_1|T}^{*(q)} := \overline{H}_{\hat{k}_0^{(q)},n}^{(q)}$.
11. Consider the estimate

$$\widehat{AMSE}(k; q) := \frac{\big(\overline{H}_{n,n_1|T}^{*(q)}\big)^2}{k} + \frac{\big(MSE^*(n_1, k)\big)^2}{\big(2^{2\hat{\rho}^*} - 1\big)^2 MSE^*(n_2, k)},$$

with $\hat{\rho}^*$, $MSE^*(n_i, k)$, $i = 1, 2$ and $\overline{H}_{n,n_1|T}^{*(q)}$ obtained in steps 3., 7. and 10. in the Algorithm, respectively.
12. Compute $\hat{q} := \arg\min_q \widehat{AMSE}(\hat{k}_0^{(q)}; q)$.
13. With the notation $\hat{k}_0^{(\hat{q})} \equiv \hat{k}_{0|\overline{H}^{(\hat{q})},T}$, obtain the final adaptive EVI estimate,

$$\overline{H}^{**} \equiv \overline{H}_{n,n_1|T}^{*(\hat{q})} := \overline{H}_{\hat{k}_0^{(\hat{q})},n}^{(\hat{q})}.$$

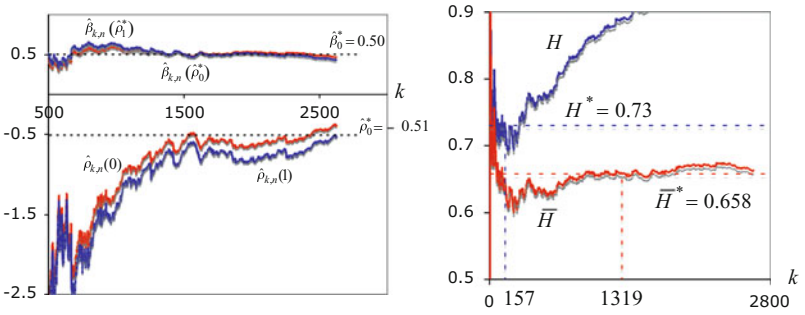

Fig. 1 Estimates of the second-order parameters β and ρ (*left*), and *bootstrap* EVI estimates based on the Hill estimator, H, and on the MVRB estimator, $\overline{H}$, (*right*) for the burned areas

5 An Application to Burned Areas

5.1 MVRB EVI Estimation of Burned Areas

Most of the wildfires in Portugal are controlled in a short period of time, with almost negligible effects. However, some of the wildfires become out-of-control, burn hectares of land and cause significant environmental and economic impacts, frankly negative. The data to be analysed are related with the number of burned hectares, above 100 ha, for all registered wildfires in Portugal in a period of 14 years (1990–2003). The sample, with a size $n = 2,627$, seems not to have any significant temporal structure and has been globally used. A preliminary data analysis provides immediately evidence that data have been censored at the left and that the right tail is quite heavy. In Fig. 1, we exhibit, at the left, the sample paths, as function of k, of the estimators of β and ρ for $\tau = 0$ and $\tau = 1$, and, at the right, the behaviour of the non-adaptive and adaptive MVRB EVI estimators under study in this chapter.

- The Algorithm in this chapter led us to the estimate $\hat{\rho}_0^* = -0.51$, obtained at the level $k_1 = [n^{0.995}] = 2524$. The associated β estimate is $\hat{\beta}_0^* = 0.50$.
- This methodology is quite resistant to different choices of k_1, which produce practically no changes in the samples paths of the MVRB EVI estimates.

Prior to [11] we could not adaptively estimate the OSF associated with MVRB estimators. The above-mentioned Algorithm, up to Step 10, and for $q = -1/n$, so that we are working with $\overline{H}$, helps us to obtain such an adaptive estimate. For a subsample of size $n_1 = [n^{0.975}] = 2157$, and B = 250 bootstrap generations, we have got $\hat{k}_0^{\overline{H}} = 1319$ and the MVRB EVI estimate, $\overline{H}^* = 0.658$, the value presented in Fig. 1, right, jointly with the bootstrap adaptive Hill estimate, $H^* = 0.73$, associated with $\hat{k}_0^H = 157$. Note also that in this case, the MVRB EVI estimators appear to be practically "unbiased".

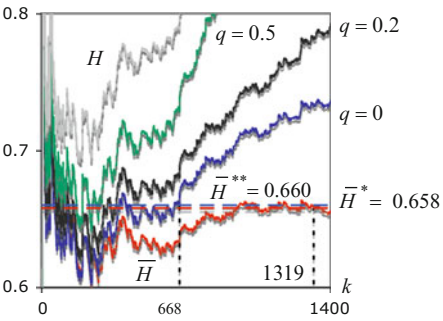

Fig. 2 PORT-MVRB EVI estimation for the burned areas

5.2 PORT-MVRB EVI Estimation for the Burned Areas

The Algorithm described in this chapter led us to the choice of either $\overline{H}$ or $\overline{H}^{(0)}$. We have been led to the same value for the minimum MSE, and, up to 2 decimal figures, to the same EVI estimate, the value 0.66, as illustrated in Fig. 2.

Acknowledgements This research was partially supported by National Funds through **FCT** — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011, PTDC/FEDER and grant SFRH/BPD/77319/2011. We also would like to thank João Carreiras for permission to use this data set.

References

1. Araújo Santos, P., Fraga Alves, M.I., Gomes, M.I.: Peaks over random threshold methodology for tail index and quantile estimation. Revstat **4**(3), 227–247 (2006)
2. Caeiro, F., Gomes, M.I., Henriques-Rodrigues, L.: Reduced-bias tail index estimators under a third order framework. Comm. Stat. Theor Methods **38**(7), 1019–1040 (2009)
3. Caeiro, F., Gomes, M.I., Pestana, D.: Direct reduction of bias of the classical Hill estimator. Revstat **3**(2), 111–136 (2005)
4. 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)
5. 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)
6. Gomes, M.I., Oliveira, O.: The bootstrap methodology in Statistical Extremes – choice of the optimal sample fraction. Extremes **4**(4), 331–358 (2001)
7. Gomes, M.I., Pestana, D.: A simple second order reduced-bias' tail index estimator. J. Stat. Comput. Simul. **77**(6), 487–504 (2007)
8. Gomes, M.I., Pestana, D.: A sturdy reduced-bias extreme quantile (VaR) estimator. J. Am. Stat. Assoc. **102**(477), 280–292 (2007)
9. Gomes, M.I., Fraga Alves, M.I., Araújo Santos, P.: PORT Hill and moment estimators for heavy-tailed models. Comm. Stat. Simul. Comput. **37**(6), 1281–1306 (2008)
10. Gomes, M.I., de Haan, L., Henriques-Rodrigues, L.: Tail Index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. J. Roy. Stat. Soc. B **70**(1), 31–52 (2008)
11. Gomes, M.I., Mendonça, S., Pestana, D.: The bootstrap methodology and adaptive reduced-bias tail index and Value-at-Risk estimation. In: Sakalauskas, L., Skiadas, C., Zavadskas, E.K. (eds.) ASMDA 2009, IMI and VGTU Editions, pp. 41–44 (2009)

12. Gomes, M.I., Figueiredo, F., Neves, M.M.: Adaptive estimation of heavy right tails: the bootstrap methodology in action. Extremes. DOI: 10.1007/s10687-011-0146-6 (2011)
13. Gomes, M.I., Henriques-Rodrigues, L., Miranda, C.: Reduced-bias location-invariant extreme value index estimation: a simulation study. Comm. Stat. Simul. Comput. **40**(3), 424–447 (2011)
14. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
15. Resnick, S.: Heavy tail modelling and teletraffic data. Ann. Stat. **25**(5), 1805–1869 (1997)

A Note on the Port Methodology in the Estimation of a Shape Second-Order Parameter

L. Henriques-Rodrigues and M. Ivette Gomes

Abstract

Under a semi-parametric framework and for *heavy right tails*, we introduce a class of *location invariant* estimators of an adequate shape *second-order parameter*, also ruling the rate of convergence of a normalized sequence of maximum values to a nondegenerate limit. This class is based on the PORT methodology, with PORT standing for *peaks over random thresholds*. Consistency of such estimators is achieved under a second-order condition on the right-tail of the underlying model F and for large intermediate ranks.

1 Introduction and Scope of This Chapter

Let $\underline{X}_n = (X_1, \ldots, X_n)$ denote a random sample of n independent, identically distributed (i.i.d.) random variables (r.v.'s) with distribution function (d.f.) F. We are interested in heavy-tailed models, i.e., d.f.'s F with a regularly varying *right-tail* function $\overline{F} := 1 - F$, with a negative index of regular variation $-1/\gamma$, i.e., a right-tail function $\overline{F}$ such that, for every $x > 0$, $\lim_{t\to\infty} \overline{F}(tx)/\overline{F}(t) = x^{-1/\gamma}$. Then, and denoting $G_\gamma(x) = \exp\left\{-(1+\gamma\, x)^{-1/\gamma}\right\}$, $1 + \gamma x > 0$, $\gamma \in \mathbb{R}$, the general *extreme-value* (EV) d.f., we are in the domain of attraction for maxima of G_γ, with $\gamma > 0$, and we write $F \in \mathscr{D}_{\mathscr{M}}(G_{\gamma>0})$. The parameter γ is the *extreme-value index* (EVI) or *tail index*, one of the primary parameters of extreme events. This type of heavy-tailed models appears often in practice, in fields like finance, insurance, and ecology, among others (see [11]). Moreover, let $F^{\leftarrow}$ denote the generalized inverse

L. Henriques-Rodrigues
CEAUL and Instituto Politécnico de Tomar, Portugal
e-mail: Ligia.Rodrigues@aim.estt.ipt.pt

M.I. Gomes (✉)
CEAUL and FCUL, Campo Grande, 1749-016, Lisboa, Portugal
e-mail: ivette.gomes@fc.ul.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_14, © Springer-Verlag Berlin Heidelberg 2013

function of F, defined by $F^{\leftarrow}(t) := \inf\{x : F(x) \geq t\}$, and let U be a (reciprocal) quantile function of the r.v. X, defined as $U(t) := F^{\leftarrow}(1 - 1/t),\ t \geq 1$.

First and Second-Order Conditions for Heavy-Tailed Models

In a context of heavy tails, with the usual notation RV_a for the class of regularly varying functions with an index of regular variation equal to a (see [2] for details on regular variation), we have the validity of the so-called first-order condition, $F \in \mathscr{D}_{\mathscr{M}}(G_{\gamma>0})$ if and only if $1 - F \in RV_{-1/\gamma}$ if and only if $U \in RV_{\gamma}$.

The second-order parameter rules the rate of convergence in the first-order condition, and it is the non-positive value $\rho\ (\leq 0)$ which appears in

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln x}{A(t)} = \begin{cases} \left(x^{\rho} - 1\right)/\rho & \text{if } \rho < 0 \\ \\ \ln x & \text{if } \rho = 0, \end{cases} \tag{1}$$

which is often assumed to hold for every $x > 0$, and where $|A|$ must then be in RV_{ρ}. For technical simplicity, we shall assume that $\rho < 0$. The adequate estimation of the second-order parameter ρ, in (1), is of primordial importance in the adaptive choice of the best number of top order statistics (o.s.'s) to be considered in the estimation of the EVI. Also, most of the research devised to improve the classical EVI estimators try to reduce the dominant component of their asymptotic bias, and deal with second-order reduced-bias (RB) EVI estimators. An overview of the subject can be found in Chap. 6 of [11]. See also [8].

The PORT Methodology

Let $X_{i:n}$, $1 \leq i \leq n$, denote the sample of ascending o.s.'s associated with the available random sample $\underline{X}_n = (X_1, \ldots, X_n)$. The class of estimators suggested here is a function of a sample of excesses over a random threshold $X_{n_q:n}$, $n_q = [nq] + 1$, with $[x]$ denoting, as usual, the integer part of x. Such a sample is denoted by

$$\underline{X}_n^{(q)} := \left(X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \ldots, X_{n_q+1:n} - X_{n_q:n}\right), \tag{2}$$

where we can have $0 < q < 1$, for any $F \in D_{\mathscr{M}}(G_{\gamma>0})$ (*the random threshold, $X_{n_q:n}$, is an empirical quantile*), and $q = 0$, for d.f.'s with a finite left endpoint $x_{_F} := \inf\{x : F(x) > 0\}$, (*the random threshold is the minimum, $X_{1:n}$*). In what follows, we use the notation χ_q for the q quantile of the d.f. F. Then (see [11], among others),

$$X_{n_q:n} \xrightarrow[n\to\infty]{p} \chi_q = F^{\leftarrow}(q), \quad \text{for} \quad 0 \leq q < 1 \quad \left(F^{\leftarrow}(0) = x_{_F}\right). \tag{3}$$

Any statistical inference methodology based on the sample of excesses $\underline{X}_n^{(q)}$, defined in (2), will be called a PORT-methodology, with PORT standing for *peaks over random thresholds*, named after [1], where this methodology easily enabled the study of classical location/scale invariant EVI estimators, studied for finite samples in [7].

Scope of This Chapter

In this chapter, making use of the above-mentioned PORT methodology, and denoting ρ_0 the second-order parameter for the unshifted model, we shall introduce a class of location-invariant semi-parametric estimators of

$$\rho_q := \begin{cases} -\gamma, & \text{if } \gamma + \rho_0 < 0 \ \wedge \ \chi_q \neq 0 \\ \rho_0, & \text{otherwise.} \end{cases} \tag{4}$$

The main motivation for such a class of PORT estimators of the shape second-order parameter ρ_q, in (4), is related to its possible use, concomitantly with an adequate estimator of the scale second-order parameter of the function A, in (1), in the construction of second-order PORT-MVRB EVI estimators, invariant for changes in location, another challenging and theoretical open subject, out of the scope of this chapter, but already dealt by simulation in [9].

For the construction of estimators of the shape second-order parameter ρ_q, in (4), we consider, for any real $\alpha > 0$, the same type of statistics used in [4] and [5], among others, i.e., the moment statistics

$$M_{n,k}^{(\alpha)} \equiv M_{n,k}^{(\alpha)}(\underline{X}_n) := \frac{1}{k} \sum_{i=1}^{k} \left(\ln X_{n-i+1:n} - \ln X_{n-k:n} \right)^{\alpha}, \tag{5}$$

now applied to the sample of excesses $\underline{X}_n^{(q)}$, $0 \leq q < 1$, in (2). For *intermediate* k sequences, i.e., sequences of integer values $k = k(n)$, between 1 and n, such that $k = k(n) \to \infty$, and $k = o(n)$, as $n \to \infty$, we shall thus consider the location and scale-invariant statistics,

$$M_{n,k}^{(\alpha,q)} \equiv M_{n,k}^{(\alpha)}(\underline{X}_n^{(q)}) := \frac{1}{k} \sum_{i=1}^{k} \left(\ln \frac{X_{n-i+1:n} - X_{n_q:n}}{X_{n-k:n} - X_{n_q:n}} \right)^{\alpha}, \tag{6}$$

defined for $k < n - n_q$, with $M_{n,k}^{(\alpha)}(\underline{X}_n) \equiv M_{n,k}^{(\alpha)}$ given in (5), $\alpha > 0$.

In Sect. 2 of this chapter, we present a few technical results related to the PORT methodology. Next, we introduce, in Sect. 3, the new class of PORT estimators of the parameter ρ_q, in (4), a shape second-order parameter associated with heavy-tailed models under a shift $\chi_q = F^{\leftarrow}(q)$, $0 \leq q < 1$, and analyze the conditions needed for consistency of this new class of estimators.

2 Technical Results Related to the PORT Methodology

If we induce any arbitrary shift, s, in the model X underlying our data, with quantile function $U_X(t)$, the transformed r.v., $Y = X - s$, has an associated quantile function given by $U_s(t) \equiv U_Y(t) = U_X(t) - s$. Consequently, in (1), $\rho = \rho_s$, with $\rho_s = -\gamma$ if $\gamma + \rho_0 < 0$, $s \neq 0$, and $\rho_s = \rho_0$, otherwise. When applying the PORT methodology,

we are working with the sample of excesses in (2), or equivalently, we are inducing a random shift, strictly related to χ_q, in (3). We shall thus use the subscript q instead of the subscript s, whenever we think of a shift χ_q.

Then, for all $x > 0$, the second-order condition in (1) can be rewritten as

$$\lim_{t\to\infty} \frac{\ln U_q(tx) - \ln U_q(t) - \gamma \ln x}{A_q(t)} = \frac{x^{\rho_q} - 1}{\rho_q}, \tag{7}$$

and $|A_q| \in RV_{\rho_q}$. Under the validity of (7), with $|A_0| \in RV_{\rho_0}$ and assuming $\chi_q \neq 0$,

$$A_q(t) := \begin{cases} \frac{\gamma\chi_q}{U_0(t)} & \text{if } \gamma + \rho_0 < 0 \\ A_0(t) + \frac{\gamma\chi_q}{U_0(t)} & \text{if } \gamma + \rho_0 = 0 \\ A_0(t), & \text{if } \gamma + \rho_0 > 0, \end{cases}$$

and

$$\rho_q := \begin{cases} -\gamma, & \text{if } \gamma + \rho_0 < 0 \\ \rho_0 & \text{if } \gamma + \rho_0 \geq 0. \end{cases}$$

Consequently, the introduction of a shift in the model underlying the data can possibly change the shape second-order parameter ρ, which indeed is equal to $-\gamma$ whenever we induce a non-null shift in any unshifted model with $\gamma + \rho_0 < 0$, like $X \frown$ Fréchet ($\gamma = 0.25$), for which $\rho_0 = -1$. Then, and for $Y = X - \chi_q$, $\chi_q \neq 0$, the second-order parameter ρ, in (1), becomes equal to $-\gamma$. Hence $\rho_q = -\gamma \neq \rho_0$ if and only if $\chi_q \neq 0$ and the underlying model is a non-shifted model with $\gamma + \rho_0 < 0$.

Asymptotic Behavior of Auxiliary Statistics

We next present the asymptotic behavior, as $n \to \infty$, of the statistics $M_{n,k}^{(\alpha,q)}$, in (6), based on the sample of excesses $\underline{X}_n^{(q)}$, $0 \leq q < 1$, in (2), and obviously independent on any shift s imposed to the data. We can thus assume that $s = 0$. Let $\mathbb{E}$ and $\mathbb{V}ar$ denote the mean value and variance operators, respectively, let E denote a unit exponential r.v. and let $\Gamma(t)$ denote the complete Gamma function. For any real $\alpha > 0$, with $\gamma \geq 0$ and $\rho < 0$, let us define

$$\mu_\alpha^{(1)}(\gamma) := \mathbb{E}\Big(E^\alpha \mathrm{e}^{-\gamma E}\Big) = \tfrac{\Gamma(\alpha+1)}{(1+\gamma)^{\alpha+1}} \quad \mu_\alpha^{(1)} := \mu_\alpha^{(1)}(0) = \Gamma(\alpha+1), \tag{8}$$

$$\sigma_\alpha^{(1)} := \sqrt{\mathbb{V}\mathrm{ar}(E^\alpha)} = \sqrt{\Gamma(2\alpha+1) - \Gamma^2(\alpha+1)}, \tag{9}$$

$$\mu_\alpha^{(2)}(\gamma,\rho) := \mathbb{E}\big(E^{\alpha-1}\, \mathrm{e}^{-\gamma E}\, \frac{(e^{\rho E} - 1)}{\rho}\big) = \tfrac{\Gamma(\alpha)}{\rho}\Big(\tfrac{(1+\gamma)^\alpha - (1+\gamma-\rho)^\alpha}{(1+\gamma-\rho)^\alpha (1+\gamma)^\alpha}\Big),$$

$$\mu_\alpha^{(2)}(\rho) := \mu_\alpha^{(2)}(0,\rho) = \tfrac{\Gamma(\alpha)}{\rho}\Big(\tfrac{1-(1-\rho)^\alpha}{(1-\rho)^\alpha}\Big),$$

$$\overline{\mu}_\alpha^{(2)}(\rho) := \frac{\mu_\alpha^{(2)}(\rho)}{\mu_\alpha^{(1)}}, \qquad \overline{\sigma}_\alpha^{(1)} := \frac{\sigma_\alpha^{(1)}}{\mu_\alpha^{(1)}}, \tag{10}$$

and for any $\theta_1,\ \theta_2 > 0$,

$$d_{\alpha,\theta_1,\theta_2}(\rho) := \overline{\mu}^{(2)}_{\alpha\theta_1}(\rho) - \overline{\mu}^{(2)}_{\alpha\theta_2}(\rho). \tag{11}$$

Moreover, with E_i i.i.d. unit exponential r.v.'s, let us define

$$Z_k^{(\alpha)} := \frac{\sqrt{k}}{\sigma^{(1)}_\alpha}\Big(\frac{1}{k}\sum_{i=1}^{k} E_i^\alpha - \Gamma(\alpha+1)\Big), \tag{12}$$

asymptotically standard normal r.v.'s, and

$$W_k^{(\alpha,\theta_1,\theta_2)} := \frac{\overline{\sigma}^{(1)}_{\alpha\theta_1}}{\theta_1} Z_k^{(\alpha\theta_1)} - \frac{\overline{\sigma}^{(1)}_{\alpha\theta_2}}{\theta_2} Z_k^{(\alpha\theta_2)}, \tag{13}$$

with $\sigma^{(1)}_\alpha$ and $\overline{\sigma}^{(1)}_\alpha$ given in (9) and (10), respectively.

We first state the following result, related to the behavior of $M^{(\alpha)}_{n,k}$, in (5), now needed only for $s = 0$ ($\rho = \rho_0$), proved in [6] under a third-order framework.

Proposition 2.1. *Under the validity of the second-order condition in (1), and unshifted models with $\rho_0 < 0$, for intermediate sequences $k = k_n$, and with $\mu^{(1)}_\alpha$, $\big(\overline{\mu}^{(2)}_\alpha(\rho), \overline{\sigma}^{(1)}_\alpha\big)$ and $Z_k^{(\alpha)}$ defined in (8), (10) and (12), respectively,*

$$M^{(\alpha)}_{n,k} \overset{d}{=} \gamma^\alpha \mu^{(1)}_\alpha \Big\{1 + \overline{\sigma}^{(1)}_\alpha \frac{Z_k^{(\alpha)}}{\sqrt{k}} + \frac{\alpha}{\gamma}\,\overline{\mu}^{(2)}_\alpha(\rho_0) A_0(n/k)(1+o_p(1))\Big\}. \tag{14}$$

We next provide for any α, a result already proved in [10] for $\alpha = 1, 2$.

Proposition 2.2. *For intermediate k, let us assume the validity of the second-order condition in (1). We then get for $M^{(\alpha,q)}_{n,k}$, in (6), $\alpha > 0$, $k < n - n_q$, with χ_q, $M^{(\alpha)}_{n,k}$ (for $s = 0$) and $\overline{\mu}^{(2)}_\alpha(\rho)$ given in (3), (5) and (10), respectively, the distributional representation,*

$$M^{(\alpha,q)}_{n,k} \overset{d}{=} M^{(\alpha)}_{n,k} + \frac{\Gamma(\alpha+1)\gamma^{\alpha-1}\chi_q}{U_0(n/k)}\Big(\alpha\,\gamma\,\overline{\mu}^{(2)}_\alpha(-\gamma) + \frac{h^{(\alpha)}(\gamma,\rho_0)\ A_0(n/k)}{\gamma\rho_0}\Big) \times (1+o_p(1)), \tag{15}$$

where

$$h^{(\alpha)}(\gamma,\rho) = (\alpha-1)\rho\,\overline{\mu}^{(2)}_{\alpha-1}(\rho) + \frac{1+2\gamma}{(1+\gamma)^\alpha} - \frac{1+2\gamma-\rho}{(1+\gamma-\rho)^\alpha}. \tag{16}$$

Proof. If $k = k_n$ is an intermediate sequence of integers, with $\{Y_i\}_{i=1,\ldots,k}$ i.i.d. unit Pareto r.v.'s, $Y_{n-k:n} \overset{p}{\sim} (n/k)$, and on the basis of the weak law of large numbers, we get, for a general α,

$$
\begin{aligned}
&M_{n,k}^{(\alpha,q)} \\
&\quad \overset{d}{=} \frac{1}{k}\sum_{i=1}^{k}\Big\{\ln\frac{X_{n-i+1:n}}{X_{n-k:n}} \\
&\qquad + \frac{\chi_q}{U_0(n/k)}\Big(1 - Y_i^{-\gamma} - Y_i^{-\gamma}\frac{Y_i^{\rho_0}-1}{\rho_0}A_0(n/k)\Big)(1+o_p(1))\Big\}^{\alpha} \\
&\quad \overset{d}{=} M_{n,k}^{(\alpha)} + \frac{\alpha\chi_q}{U_0(n/k)}\frac{1}{k}\sum_{i=1}^{k}\Big\{\Big(\gamma\ln Y_i + \Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)A_0(n/k)(1+o_p(1))\Big)^{\alpha-1} \\
&\qquad \times\Big(1 - Y_i^{-\gamma} - Y_i^{-\gamma}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)A_0(n/k)\Big)(1+o_p(1))\Big\} \\
&\quad \overset{d}{=} M_{n,k}^{(\alpha)} + \frac{\alpha\chi_q}{U_0(n/k)}\frac{1}{k}\sum_{i=1}^{k}\Big\{\gamma^{\alpha-1}(\ln Y_i)^{\alpha-1}\Big(1 + \frac{\alpha-1}{\gamma\ln Y_i}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big) \\
&\qquad \times A_0(n/k)(1+o_p(1))\Big) \\
&\qquad \times\Big(1 - Y_i^{-\gamma} - Y_i^{-\gamma}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)A_0(n/k)\Big)(1+o_p(1))\Big\} \\
&\quad \overset{d}{=} M_{n,k}^{(\alpha)} + \frac{\alpha\chi_q\gamma^{\alpha-1}}{U_0(n/k)}\frac{1}{k}\sum_{i=1}^{k}(\ln Y_i)^{\alpha-1}\Big\{\big(1 - Y_i^{-\gamma}\big) \\
&\qquad + A_0(n/k)\left(\frac{\alpha-1}{\gamma\ln Y_i}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)\right. \\
&\qquad \left. - \frac{\alpha-1}{\gamma\ln Y_i}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)Y_i^{-\gamma} - Y_i^{-\gamma}\Big(\frac{Y_i^{\rho_0}-1}{\rho_0}\Big)\right)(1+o_p(1))\Big\}.
\end{aligned}
$$

Consequently, and since $\ln Y \overset{d}{=} E$,

$$
\begin{aligned}
M_{n,k}^{(\alpha,q)} \overset{d}{=} M_{n,k}^{(\alpha)} + \frac{\alpha\chi_q\gamma^{\alpha-1}}{U_0(n/k)}\Big\{\gamma\,\mu_\alpha^{(2)}(-\gamma) + \frac{(\alpha-1)\mu_{\alpha-1}^{(2)}(\rho_0)}{\gamma}A_0(n/k)(1+o_p(1)) \\
- \Big(\frac{(\alpha-1)\mu_{\alpha-1}^{(2)}(\gamma,\rho_0)}{\gamma} + \mu_\alpha^{(2)}(\gamma,\rho_0)\Big)A_0(n/k)(1+o_p(1))\Big\}.
\end{aligned}
$$

Taking into account the fact that $\mu_\alpha^{(2)}(\rho) = \alpha\Gamma(\alpha)\,\overline{\mu}_\alpha^{(2)}(\rho)$ (see (8) and (10)), that

$$\frac{(\alpha-1)\mu_{\alpha-1}^{(2)}(\gamma,\rho)}{\gamma}+\mu_{\alpha}^{(2)}(\gamma,\rho)=\frac{\Gamma(\alpha)}{\gamma\,\rho}\Big(\frac{1+2\gamma-\rho}{(1+\gamma-\rho)^{\alpha}}-\frac{1+2\gamma}{(1+\gamma)^{\alpha}}\Big),$$

and the validity of (14), we are led to the result in (15) for any $\alpha \geq 1$, with $h^{(\alpha)}(\gamma,\rho)$ given in (16). □

3 The New Class of PORT-ρ Estimators

For $\alpha > 0$, let us consider the statistics $M_{n,k}^{(\alpha,q)} = M_{n,k}^{(\alpha)}(\underline{X}_n^{(q)})$, in (6). Under the second-order framework in (1), for intermediate k, on the basis of the results in Propositions 2.1 and 2.2, similar to the developments in [5], and for real *tuning parameters* $\tau_q \in \mathbb{R}$ and $\theta \neq 0$,

$$\Big(\frac{M_{n,k}^{(\alpha\theta,q)}}{\mu_{\alpha\theta}^{(1)}\gamma^{\alpha\theta}}\Big)^{\tau_q/\theta}\overset{d}{=}1+\frac{\tau_q}{\theta}\,\frac{\overline{\sigma}_{\alpha\theta}^{(1)}}{\sqrt{k}}\,Z_k^{(\alpha\theta)}+\frac{\alpha\tau_q\,\overline{\mu}_{\alpha\theta}^{(2)}(\rho_0)}{\gamma}A_0(n/k)$$

$$+\frac{\alpha\,\tau_q\chi_q\,\overline{\mu}_{\alpha\theta}^{(2)}(-\gamma)(1+o_p(1))}{U_0(n/k)}+\frac{\tau_q\,\chi_q\,h^{(\alpha\theta)}(\gamma,\rho_0)}{\theta\gamma^2\rho_0}\,\frac{A_0(n/k)}{U_0(n/k)}(1+o_p(1)),\tag{17}$$

i.e.

$$\Big(\frac{M_{n,k}^{(\alpha\theta,q)}}{\mu_{\alpha\theta}^{(1)}\gamma^{\alpha\theta}}\Big)^{\tau_q/\theta}\overset{d}{=}\Big(\frac{M_{n,k}^{(\alpha\theta)}}{\mu_{\alpha\theta}^{(1)}\gamma^{\alpha\theta}}\Big)^{\tau_q/\theta}+\frac{\tau_q\chi_q}{\theta\gamma U_0(n/k)}\Big(\alpha\theta\gamma\,\overline{\mu}_{\alpha\theta}^{(2)}(-\gamma)$$
$$+\frac{h^{(\alpha\theta)}(\gamma,\rho_0)}{\gamma\rho_0}A_0(n/k)\Big)(1+o_p(1)),$$

with $\big(\overline{\mu}_{\alpha}^{(2)}(\rho),\overline{\sigma}_{\alpha}^{(1)}\big)$, $Z_k^{(\alpha)}$ and $h^{(\alpha)}(\gamma,\rho)$, given in (10), (12) and (16), respectively.

We now refer the class of estimators in [5], based on the behavior of the statistics

$$T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)}:=\frac{D_{n,k}^{(\alpha,1,\theta_1,\tau)}(\gamma)}{D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)}(\gamma)}=\frac{\Big(M_{n,k}^{(\alpha)}/\mu_{\alpha}^{(1)}\Big)^{\tau}-\Big(M_{n,k}^{(\alpha\theta_1)}/\mu_{\alpha\theta_1}^{(1)}\Big)^{\tau/\theta_1}}{\Big(M_{n,k}^{(\alpha\theta_1)}/\mu_{\alpha\theta_1}^{(1)}\Big)^{\tau/\theta_1}-\Big(M_{n,k}^{(\alpha\theta_2)}/\mu_{\alpha\theta_2}^{(1)}\Big)^{\tau/\theta_2}},\tag{18}$$

obviously independent of the *nuisance* parameter γ, with $M_{n,k}^{(\alpha)}$ and $\mu_{\alpha}^{(1)}$ defined in (5) and (8), respectively. The admissible values of the *tuning* parameters are $\tau \in \mathbb{R}$ (as noticed in [3]), $\alpha, \theta_1, \theta_2 > 0$, $\theta_1 \neq 1$, $\theta_2 \neq \theta_1$. Under adequate conditions on k, and with $d_{\alpha,\theta_1,\theta_2}(\rho)$ defined in (11), the statistics $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)}$, in (18), converge in probability, as $n \to \infty$, to

$$t_{\alpha,\theta_1,\theta_2}(\rho) := \frac{d_{\alpha,1,\theta_1}(\rho)}{d_{\alpha,\theta_1,\theta_2}(\rho)} \tag{19}$$
$$= \theta_2 \frac{(\theta_1-1)(1-\rho)^{\alpha\theta_2} - \theta_1(1-\rho)^{\alpha(\theta_2-1)} + (1-\rho)^{\alpha(\theta_2-\theta_1)}}{(\theta_2-\theta_1)(1-\rho)^{\alpha\theta_2} - \theta_2(1-\rho)^{\alpha(\theta_2-\theta_1)} + \theta_1},$$

which enables, through inversion, the general class of estimators introduced and studied in [5], which can be written as $\widehat{\rho}_{n,k|T}^{(\alpha,\theta_1,\theta_2,\tau)} := -\left|t^{\leftarrow}_{\alpha,\theta_1,\theta_2}(T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)})\right|$, $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)}$ given in (18).

The class of consistent ρ_q-estimators, invariant for changes in location, and named PORT-ρ, is given by

$$\widehat{\rho}_{n,k|T}^{(\alpha,\theta_1,\theta_2,\tau_q,q)} := -\left|t^{\leftarrow}_{\alpha,\theta_1,\theta_2}(T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)})\right|, \tag{20}$$

where the statistics $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau)}$, in (18), are thus replaced by their PORT-versions,

$$T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)} := \frac{D_{n,k}^{(\alpha,1,\theta_1,\tau_q,q)}(\gamma)}{D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma)} = \frac{\left(M_{n,k}^{(\alpha,q)}/\mu_{\alpha}^{(1)}\right)^{\tau_q} - \left(M_{n,k}^{(\alpha\theta_1,q)}/\mu_{\alpha\theta_1}^{(1)}\right)^{\tau_q/\theta_1}}{\left(M_{n,k}^{(\alpha\theta_1,q)}/\mu_{\alpha\theta_1}^{(1)}\right)^{\tau_q/\theta_1} - \left(M_{n,k}^{(\alpha\theta_2,q)}/\mu_{\alpha\theta_2}^{(1)}\right)^{\tau_q/\theta_2}}. \tag{21}$$

On the basis of (17), using the notations $d_{\alpha,\theta_1,\theta_2}(\rho)$ and $W_k^{(\alpha,\theta_1,\theta_2)}$ introduced in (11) and (13), respectively, and with

$$H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho) := \frac{1}{\theta_1} h^{(\alpha\theta_1)}(\gamma,\rho) - \frac{1}{\theta_2} h^{(\alpha\theta_2)}(\gamma,\rho), \tag{22}$$

$h^{(\alpha)}(\gamma,\rho)$ defined in (16), we can write

$$D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma) \stackrel{d}{=} \frac{\tau_q}{\sqrt{k}} W_k^{(\alpha,\theta_1,\theta_2)} + \frac{\alpha\tau_q d_{\alpha,\theta_1,\theta_2}(\rho_0) A_0(n/k)}{\gamma}$$
$$+ \frac{\alpha\tau_q\chi_q d_{\alpha,\theta_1,\theta_2}(-\gamma)}{U_0(n/k)}(1+o_p(1))$$
$$+ \frac{\tau_q\,\chi_q\,H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho_0)}{\gamma^2\rho_0}\,\frac{A_0(n/k)}{U_0(n/k)}(1+o_p(1)). \tag{23}$$

The dominant component of the right hand-side of (23) depends on the behavior of the functions $A_0(t)$ and $1/U_0(t)$. We shall thus consider three different regions related with the vector (γ,ρ_0,χ_q):

(i) $\mathscr{R}_1 := \{\gamma+\rho_0 < 0 \wedge \chi_q \neq 0\}$
(ii) $\mathscr{R}_2 := \{\gamma+\rho_0 > 0 \vee (\gamma+\rho_0 \leq 0 \wedge \chi_q = 0)\}$
(iii) $\mathscr{R}_3 := \{\gamma+\rho_0 = 0 \wedge \chi_q \neq 0\}$.

We next state the following:

Theorem 3.1. *Under the validity of the second-order condition in* (1)*, with* $\rho < 0$*,* ρ_q *defined in* (4)*,* $\widehat{\rho}_{n,k|T}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}$ *defined in* (20)*,*

$$\widehat{\rho}_{n,k|T}^{(\alpha,\theta_1,\theta_2,\tau_q,q)} \xrightarrow[n\to\infty]{p} \rho_q,$$

for any real $\alpha > 0$*,* $\tau_q \in \mathbb{R}$*,* $\theta_1, \theta_2 \in \mathbb{R}^+ \backslash \{1\}, \theta_1 \neq \theta_2$ *and* $0 < q < 1$ *or* $q = 0$ *if* χ_0 *is finite, provided that* k *is an intermediate sequence, and moreover*

(i) *The underlying model lies in* $\mathscr{R}_1$*, and*

$$\sqrt{k}/U_0(n/k) \to \infty, \text{ as } n \to \infty. \tag{24}$$

(ii) *The underlying model lies in* $\mathscr{R}_2$*, and*

$$\sqrt{k}A_0(n/k) \to \infty, \quad \text{as } n \to \infty. \tag{25}$$

(iii) *The underlying model lies in* $\mathscr{R}_3$*, and*

$$\frac{1}{A_0(n/k)U_0(n/k)} \longrightarrow \widetilde{\lambda} \neq 0 \quad \text{and} \quad \sqrt{k}A_0(n/k) \to \infty, \quad \text{as} \quad n \to \infty. \tag{26}$$

Proof. (i) In the region $\mathscr{R}_1$, $A_0(t) = o(1/U_0(t))$, as $t \to \infty$, the second last term of the right-hand side of (23) is the dominant one, and the statistic $D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma)/(1/U_0(n/k))$ converges in probability to $\alpha\ \tau_q\chi_q\ d_{\alpha,\theta_1,\theta_2}(-\gamma)$ provided that (24) holds. Moreover, with $H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho)$ defined in (22), we get

$$\frac{D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma)}{1/U_0(n/k)} \overset{d}{=} \alpha\,\tau_q\chi_q\,d_{\alpha,\theta_1,\theta_2}(-\gamma) + \frac{\tau_q\chi_q}{\gamma^2\rho_0}H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho_0)A_0(n/k)$$
$$+ \frac{\tau_q}{\sqrt{k}}\,W_k^{(\alpha,\theta_1,\theta_2)}\,U_0(n/k) + \frac{\alpha\tau_q d_{\alpha,\theta_1,\theta_2}(\rho_0)}{\gamma}A_0(n/k)U_0(n/k).$$

Consequently, as $n \to \infty$, the statistic $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}$, in (21), converges to $t_{\alpha,\theta_1,\theta_2}(-\gamma)$, with $t_{\alpha,\theta_1,\theta_2}(\rho)$, defined in (19), and consistency follows.

(ii) In the region $\gamma + \rho_0 > 0$, where $1/U_0(t) = o(A_0(t))$, as $t \to \infty$, or more generally in the region $\mathscr{R}_2$, the second term of the right-hand side of (23) is the dominant one. If we assume that (25) holds,

$$\frac{D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma)}{A_0(n/k)} \overset{d}{=} \frac{\alpha\tau_q\,d_{\alpha,\theta_1,\theta_2}(\rho_0)}{\gamma} + \frac{\tau_q}{\sqrt{k}A(n/k)}\,W_k^{(\alpha,\theta_1,\theta_2)}$$
$$+\frac{\alpha\,\tau_q\,\chi_q d_{\alpha,\theta_1,\theta_2}(-\gamma)}{A_0(n/k)U_0(n/k)}(1+o_p(1))+\frac{\tau_q\chi_q}{\gamma^2\rho_0U_0(n/k)}H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho_0)(1+o_p(1)).$$

In this case, and as $n \to \infty$, the statistic $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}$, in (21), converges in probability to $t_{\alpha,\theta_1,\theta_2}(\rho_0)$, with $t_{\alpha,\theta_1,\theta_2}(\rho)$ defined in (19), and consistency follows.

(iii) In the region $\mathscr{R}_3$, $A_0(t)$ and $1/U_0(t)$ are of the same order, i.e., the dominant terms of the right-hand side of (23) are the second and the second last. If we assume that (26) holds,

$$\frac{D_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}(\gamma)}{A_0(n/k)} \stackrel{d}{=} \frac{\alpha\tau_q}{\gamma}\Big\{ d_{\alpha,\theta_1,\theta_2}(\rho_0) + \gamma\, \widetilde{\lambda}\chi_q d_{\alpha,\theta_1,\theta_2}(-\gamma)\Big\}$$

$$+\frac{\tau_q}{\sqrt{k}A_0(n/k)}\, W_k^{(\alpha,\theta_1,\theta_2)}\,(1+o_p(1))+\frac{\tau_q\chi_q H^{(\alpha,\theta_1,\theta_2)}(\gamma,\rho_0)}{\gamma^2\rho_0\, U_0(n/k)}(1+o_p(1)).$$

Then, as $n \to \infty$, the statistics $T_{n,k}^{(\alpha,\theta_1,\theta_2,\tau_q,q)}$, in (21), converge in probability to

$$\widetilde{t}_{\alpha,\theta_1,\theta_2}(\rho_0) = \frac{d_{\alpha,1,\theta_1}(\rho_0) + \gamma\, \widetilde{\lambda}\chi_q d_{\alpha,1,\theta_1}(\rho_0)}{d_{\alpha,\theta_1,\theta_2}(\rho_0) + \gamma\, \widetilde{\lambda}\chi_q d_{\alpha,\theta_1,\theta_2}(\rho_0)} = \frac{d_{\alpha,1,\theta_1}(\rho_0)}{d_{\alpha,\theta_1,\theta_2}(\rho_0)} = t_{\alpha,\theta_1,\theta_2}(\rho_0),$$

with $t_{\alpha,\theta_1,\theta_2}(\rho)$ defined in (19), and consistency follows again. □

Acknowledgements This research was partially supported by National Funds through **FCT** — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011, PTDC/FEDER and grant SFRH/BPD/77319/2011.

References

1. Araújo Santos, P., Fraga Alves, M.I., Gomes, M.I.: Peaks over random threshold methodology for tail index and quantile estimation. Revstat **4**(3), 227–247 (2006)
2. Bingham, N.H., Goldie, C.M., Teugels, J.L.: Regular Variation. Cambridge University Press, New York (1989)
3. Caeiro, F., Gomes, M.I.: A new class of estimators of a "scale" second order parameter. Extremes **9**, 193–211 (2006)
4. Dekkers, A.L.M., Einmahl, J.H.J., de Haan, L.: A Moment estimator for the index of an extreme value distribution. Ann. Stat. **17**, 1833–1855 (1989)
5. 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)
6. Gomes, M.I., de Haan, L., Peng, L.: Semi-parametric estimation of the second order parameter – asymptotic and finite sample behaviour. Extremes **5**(4), 387–414 (2002)
7. Gomes, M.I., Fraga Alves, M.I., Araújo Santos, P.: PORT Hill and moment estimators for heavy-tailed models. Comm. Stat. Simul. Comput. **37**(6), 1281–1306 (2008)
8. 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)
9. Gomes, M.I., Henriques-Rodrigues, L., Miranda, C.: Reduced-bias location-invariant extreme value index estimation: a simulation study. Comm. Stat. Simul. Comput. **40**, 424–447 (2011)

10. Henriques-Rodrigues, L., Gomes, M.I.: High quantile estimation and the PORT methodology. Revstat **7**(3), 245–264 (2009)
11. Reiss, R.-D., Thomas, M.: Statistical Analysis of Extreme Values, with Application to Insurance, Finance, Hydrology and Other Fields, 2nd edn; 3rd edn. Birkhäuser Verlag, Basel, Boston, Berlin (2001; 2007)

A Class of Semi-parametric Probability Weighted Moment Estimators

Frederico Caeiro and M. Ivette Gomes

Abstract

In this paper we deal with the semi-parametric estimation of the right tail $1 - F$. Through the use of probability weighted moments based on the largest observations, we study a class of estimators for the extreme value index γ, the scale parameter C, and the *Value-at-Risk* at a level p.

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be a set of n independent and identically distributed (i.i.d.) random variables (r.v.'s), from a population with distribution function (d.f.) F. We assume that F is a heavy tailed model with a Pareto-type tail, i.e.,

$$\overline{F}(x) := 1 - F(x) \;\sim\; (C/x)^{1/\gamma}, \qquad x \to \infty, \tag{1}$$

where C and γ are unknown scale and shape parameters, respectively. Then $\overline{F}$ is a regularly varying function with a negative index of regular variation equal to $-1/\gamma$, and we are in the max-domain of attraction of the Extreme Value distribution

$$EV_\gamma(x) = \exp\{-(1+\gamma x)^{-1/\gamma}\}, \quad 1 + \gamma x > 0.$$

Although we deal with the right tail $\overline{F}$, the results here presented are applicable to the left tail F, after the change of variable $Y = -X$.

F. Caeiro (✉)
CMA and Faculdade de Ciências e Tecnologia da UNL, 2829–516 Caparica, Portugal
e-mail: fac@fct.unl.pt

M.I. Gomes
CEAUL and FCUL, Campo Grande, 1749–016 Lisboa, Portugal
e-mail: ivette.gomes@fc.ul.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_15, © Springer-Verlag Berlin Heidelberg 2013

Suppose that we are interested in the estimation of a *high quantile* of probability $1-p$, or equivalently, in the estimation of the *Value-at-Risk* (VaR) at a level p, the size of the loss occurred with a small probability p,

$$\text{VaR}_p \equiv \chi_{1-p} := F^{\leftarrow}(1-p) = \inf\{x : F(x) \geq 1-p\}, \tag{2}$$

with the notation $F^{\leftarrow}$ standing thus for the generalized inverse function of F.

In Sect. 2 we present some already studied semi-parametric estimators and introduce a new class, to be studied in this chapter. In Sect. 3, after a few technical details, we study the asymptotic behavior of the estimators under consideration. Finally, Sect. 4 is dedicated to a small-scale simulation study.

2 Estimators Under Study

Under the largest observations framework, and whenever dealing with Pareto-type tailed models, the classical semi-parametric estimators of γ and C are the Hill estimator [7] and Weissman estimator [10], with functional expressions

$$\hat{\gamma}^{H}_{k,n} := \frac{1}{k}\sum_{i=1}^{k}\left(\ln\frac{X_{n-i+1:n}}{X_{n-k:n}}\right), \quad \hat{C}^{H}_{k,n} := X_{n-k:n}\left(\frac{k}{n}\right)^{\hat{\gamma}^{H}_{k,n}}, \quad k = 1,2,\ldots,n-1, \tag{3}$$

where $X_{i:n}$ denotes the *i-th* ascending order statistic. These estimators are pseudo-maximum likelihood estimators, obtained from an approximate likelihood, and have usually a high asymptotic bias which makes the choice of k very difficult. This problem led researchers to deal with bias reduction and study new estimators with smaller *mean squared error* (MSE).

Since heavy-tailed models only have mean value if $\gamma < 1$, methods based on sample moments are rarely considered when we work with such distributions. But in many practical fields like in finance or insurance, for example, we usually have a positive scale parameter γ smaller than 1, and even smaller than $1/2$. In this chapter, we again consider the *probability weighted moments* (PWM) method, a generalization of the *method of moments* [4]. This method is known for being more efficient than the maximum likelihood method for small to moderate sample sizes [8,9]. The PWM of a r.v. X are defined by $M_{p,r,s} := E(X^p(F(X))^r(1-F(X))^s)$, where p, r, and s are any real numbers. When $r = s = 0$, $M_{p,0,0}$ are the usual noncentral moments. It is usual to work with one of the two particular and simple cases:

$$a_r := M_{1,0,r} = E(X(1-F(X))^r) \quad \text{or} \quad b_r := M_{1,r,0} = E(X(F(X))^r). \tag{4}$$

Given a sample of size n, the unbiased estimators of a_r and b_r are, respectively,

$$\hat{a}_r = \frac{1}{n}\sum_{i=1}^{n-r}\frac{\binom{n-i}{r}}{\binom{n-1}{r}}X_{i:n} = \frac{1}{n}\sum_{i=1}^{n}\frac{(n-i)(n-i-1)\ldots(n-i-r+1)}{(n-1)(n-2)\ldots(n-r)}X_{i:n}, \tag{5}$$

and

$$\hat{b}_r = \frac{1}{n}\sum_{i=r+1}^{n}\frac{\binom{i-1}{r}}{\binom{n-1}{r}}X_{i:n} = \frac{1}{n}\sum_{i=1}^{n}\frac{(i-1)(i-2)\cdots(i-r)}{(n-1)(n-2)\cdots(n-r)}X_{i:n}. \tag{6}$$

The PWM method has already been used in extreme value theory (see Dielbolt et al. [3], Caeiro and Gomes [1], Cai et al. [2], among others). Caeiro and Gomes [1] studied the PWM estimators for the parameters of a Pareto tail, based on the top k largest observations, $X_{n:n} \geq X_{n-1:n} \geq \cdots \geq X_{n-k+1:n}$. The PWM estimators, valid for $\gamma < 1$, and based on the largest values, are

$$\hat{\gamma}_{k,n}^{PWM} = 1 - \frac{\hat{a}_1(k)}{\hat{a}_0(k) - \hat{a}_1(k)}, \qquad \hat{C}_{k,n}^{PWM} = \frac{\hat{a}_0(k)\,\hat{a}_1(k)}{\hat{a}_0(k) - \hat{a}_1(k)}\left(\frac{k}{n}\right)^{\hat{\gamma}_{k,n}^{PWM}}, \tag{7}$$

with $k = 2, \ldots, n$ and $\hat{a}_s(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i-1}{k-1}\right)^s X_{n-i+1:n}, \quad s = 0, 1.$

To overcome the restriction $\gamma < 1$ of the previous estimators, we will study the PWM estimators based on the moments

$$a_{r,s} := E(X^r(1 - F(X))^s). \tag{8}$$

The constant r will be a parameter that allows us to extend the domain of validity of the previous PWM estimators. The PWM estimators, valid for $\gamma < \frac{1}{r}$, and based on the largest values, are

$$\hat{\gamma}_{k,n}^{\mathrm{PWM}(r)} = \frac{1}{r}\left(1 - \frac{\hat{a}_{r,1}(k)}{\hat{a}_{r,0}(k) - \hat{a}_{r,1}(k)}\right), \quad k = 2, \cdots, n, \tag{9}$$

and

$$\hat{C}_{k,n}^{\mathrm{PWM}(r)} = \left(\frac{\hat{a}_{r,0}(k)\,\hat{a}_{r,1}(k)}{\hat{a}_{r,0}(k) - \hat{a}_{r,1}(k)}\right)^{\frac{1}{r}}\left(\frac{k}{n}\right)^{\hat{\gamma}_{k,n}^{\mathrm{PWM}(r)}}, \quad k = 2, \cdots, n, \tag{10}$$

with $\hat{a}_{r,s}(k) := \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i-1}{k-1}\right)^s (X_{n-i+1:n})^r, \; s = 0, 1$. The parameter r can also be used as a tuning parameter, controlled at our ease to reduce the bias or the MSE.

Since $\chi_{1-p} := F^{\leftarrow}(1-p) \sim Cp^{-\gamma}$, as $p \to 0$, the classical and PWM high quantile estimators, based on the largest values, are, respectively

$$\hat{Q}_{k,n}^{H}(p) := X_{n-k:n}\left(\frac{k}{np}\right)^{\hat{\gamma}_{k,n}^{H}}, \tag{11}$$

and

$$\hat{Q}_{k,n}^{\mathrm{PWM}(r)}(p) = \left(\frac{\hat{a}_{r,0}(k)\, \hat{a}_{r,1}(k)}{\hat{a}_{r,0}(k) - \hat{a}_{r,1}(k)} \right)^{\frac{1}{s}} \left(\frac{k}{np} \right)^{\hat{\gamma}_{k,n}^{\mathrm{PWM}(r)}}. \tag{12}$$

3 Asymptotic Properties

To guarantee the consistency of many semi-parametric estimators, we usually need to assume that k is intermediate, i.e.,

$$k = k_n \to \infty \quad \text{and} \quad k/n \to 0, \quad \text{as } n \to \infty. \tag{13}$$

To obtain information on the nondegenerate distributional behaviour of semi-parametric estimators, we assume a second-order condition,

$$\lim_{t\to\infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln x}{A(t)} = \frac{x^\rho - 1}{\rho} \quad \Leftrightarrow \quad \lim_{t\to\infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{A(t)} = x^\gamma\, \frac{x^\rho - 1}{\rho}, \tag{14}$$

valid for all $x > 0$, where $\rho \le 0$ is a second-order parameter controlling the speed of convergence of $U(tx)/U(t)$ to x^γ.

Hill's estimator is well studied in the literature. Under the above second-order condition in (14) and for intermediate k, we get [6]:

$$\hat{\gamma}_{k,n}^H \overset{d}{=} \gamma + \frac{\gamma}{\sqrt{k}}\, Z_k^H + \frac{A(n/k)}{1-\rho}(1 + o_p(1)), \tag{15}$$

with $Z_k^H = \sqrt{k}\left(\sum_{i=1}^k E_i/k - 1 \right)$, and $\{E_i\}$ i.i.d. standard exponential r.v.'s.

More generally than Theorem 3.1 in Caeiro and Gomes [1], but with a similar proof, we now state the following theorem.

Theorem 3.1. *Under the second-order framework, in (14), and for intermediate k, the asymptotic distributional representation*

$$\hat{\gamma}_{k,n}^{\mathrm{PWM}(r)} \overset{d}{=} \gamma + \frac{\sigma_{\mathrm{PWM}(r)}}{\sqrt{k}}\, Z_k^{\mathrm{PWM}(r)} + b_{\mathrm{PWM}(r)} A(n/k)(1 + o_p(1)) \tag{16}$$

holds, for $\gamma < 1/2r$, $r > 0$, where $Z_k^\bullet$ is a standard normal r.v., and

$$\sigma^2_{\mathrm{PWM}(r)} = \frac{\gamma^2 (1-r\gamma)(2-r\gamma)^2}{(1-2r\gamma)(3-2r\gamma)}, \qquad b_{\mathrm{PWM}(r)} = \frac{(1-r\gamma)(2-r\gamma)}{(1-r\gamma-\rho)(2-r\gamma-\rho)}. \tag{17}$$

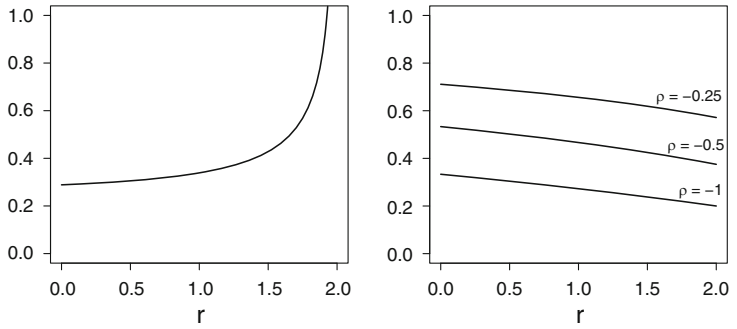

Fig. 1 *Left*: $\sigma_{\mathrm{PWM}(r)}$, as function of r for $\gamma = 0.25$, *Right*: $b_{\mathrm{PWM}(r)}$, as function of r for $\gamma = 0.25$ and three different values of ρ

Corollary 3.1. *If we further assume that* $\sqrt{k}\,A(n/k) \to \lambda$, *finite and not necessarily null,*

$$\sqrt{k}(\hat{\gamma}^{\bullet}_{k,n} - \gamma) \xrightarrow{d} N\left(\lambda\, b_{\bullet}, \sigma^2_{\bullet}\right), \qquad as \quad n \to \infty, \tag{18}$$

with $\bullet$ *denoting* H *or* $PWM(r)$, $\sigma^2_H = \gamma^2$ *and* $b_H = 1/(1-\rho)$.

Remark 3.1. Notice that $\sigma^2_H < \sigma^2_{PPWM(r)}$, for every $\gamma > 0$ and $r > 0$. On the other hand, $b_{PPWM(r)} < b_H$, unless $\rho = 0$.

In Fig. 1 we provide a picture of $\sigma_{\mathrm{PWM}(r)}$ (left) and $b_{\mathrm{PWM}(r)}$ (right) as function of r. These functions have opposite behaviour: The variance increases with r, but the bias decreases with r. The choice of the "optimal" r that minimizes the MSE is not obvious and is a subject outside the scope of this chapter.

With $\bullet$ denoting again H or $PWM(r)$, with $r < 1/\gamma$, we now state the following two theorems, related to the asymptotic behavior of the estimators of the scale parameter C and high quantiles. The proof of the first theorem can be found in de Haan and Ferreira [5], Theorem 4.3.8, and Caeiro and Gomes [1], Theorem 3.1. The second theorem follows straightforwardly from the relation $\hat{Q}^{\bullet}_{k,n}(p) = \hat{C}^{\bullet}_{k,n} p^{-\hat{\gamma}^{\bullet}_{k,n}}$.

Theorem 3.2. *Under the conditions of* Theorem 3.1, but with $\rho < 0$, if we further assume $\ln(k/n) = o(\sqrt{k})$ and $\sqrt{k}A(n/k) \longrightarrow \lambda$, as $n \to \infty$, then

$$\frac{\sqrt{k}}{\ln(k/n)}\left(\frac{\hat{C}^{\bullet}_{k,n}}{C} - 1\right) \overset{d}{=} \sqrt{k}\left(\hat{\gamma}^{\bullet}_{k,n} - \gamma\right)(1 + o_p(1)). \tag{19}$$

Theorem 3.3. *Under the conditions of* Theorem 3.1, *but with* $\rho < 0$, *if we assume that* $p = p_n$ *is a sequence of probabilities such that* $c_n = k/(np) \to \infty$, $\ln c_n = o(\sqrt{k})$ *and* $\sqrt{k}A(n/k) \longrightarrow \lambda$, *as* $n \to \infty$, *then,*

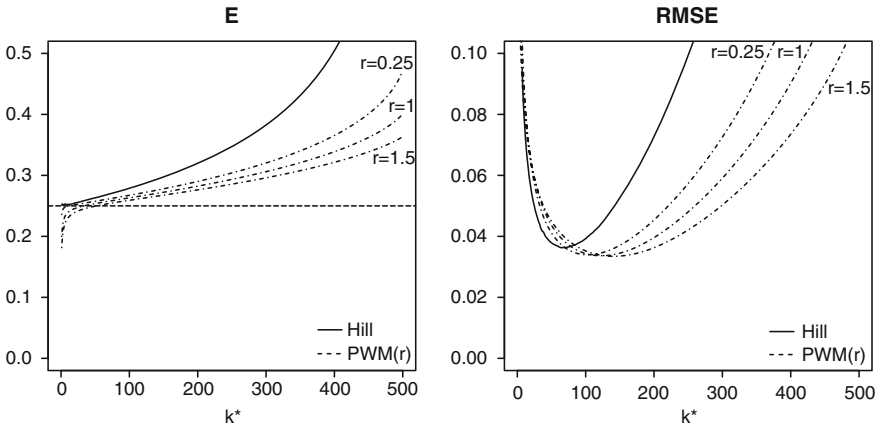

Fig. 2 Simulated mean values (*left*) and root mean squared errors (*right*), as functions of k^*, of the estimators $\hat{\gamma}^{H}_{k^*,n}$ and $\hat{\gamma}^{\mathrm{PWM}(r)}_{k^*,n}$, $n = 500$, from a Burr$(0.25, -1)$ parent

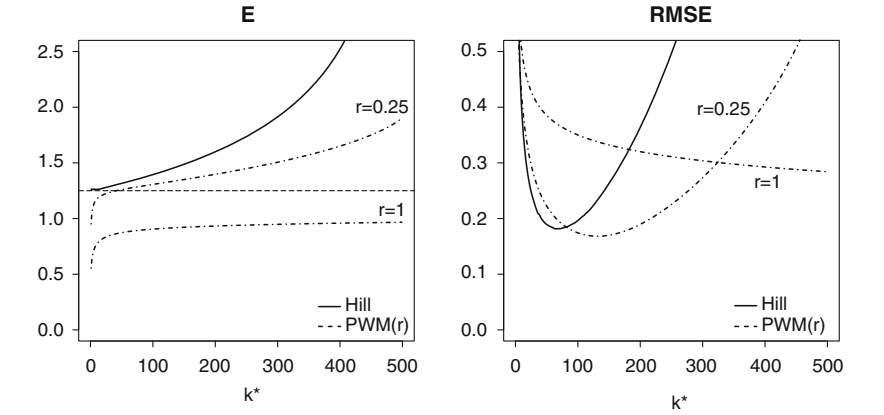

Fig. 3 Simulated mean values (*left*) and root mean squared errors (*right*), as functions of k^*, of the estimators $\hat{\gamma}^{H}_{k^*,n}$ and $\hat{\gamma}^{\mathrm{PWM}(r)}_{k^*,n}$, $n = 500$, from a Burr$(1.25, -1)$ parent

$$\frac{\sqrt{k}}{\ln c_n}\left(\frac{\hat{Q}^{\bullet}_{k,n}(p)}{\chi_{1-p}}-1\right)\stackrel{d}{=}\sqrt{k}\left(\hat{\gamma}^{\bullet}_{k,n}-\gamma\right)(1+o_p(1)). \tag{20}$$

Remark 3.2. The previous theorems allow us to conclude that the asymptotic dominant behavior of the scale parameter and high quantiles estimators is thus fully determined by the asymptotic behaviour of $\hat{\gamma}^{\bullet}_{k,n}$.

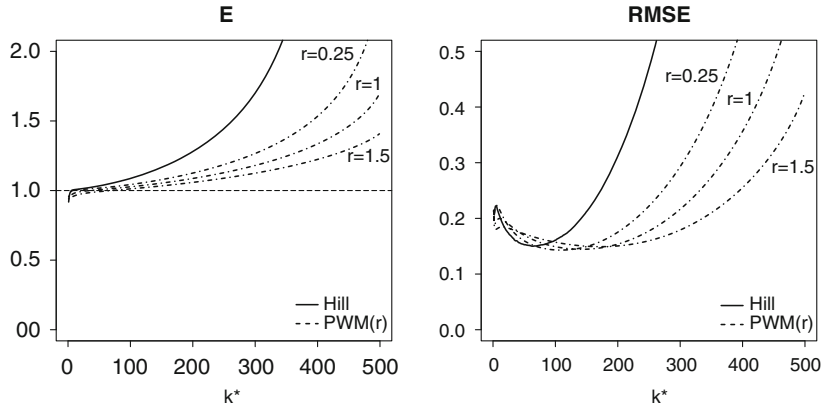

Fig. 4 Simulated mean values (*left*) and root mean squared errors (*right*), as functions of k^*, of the normalized estimators $\bar{Q}^{H}_{k^*,n}(p)$ and $\bar{Q}^{\mathrm{PWM}(r)}_{k^*,n}(p)$, $n = 500$, from a Burr$(0.25, -1)$ parent

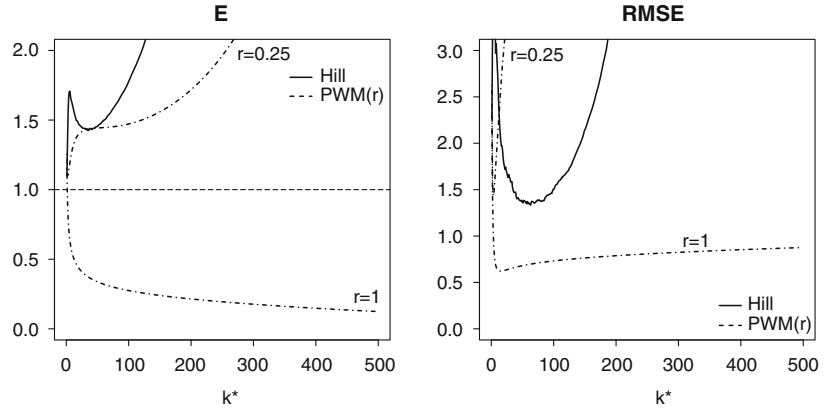

Fig. 5 Simulated mean values (*left*) and root mean squared errors (*right*), as functions of k^*, of the normalized estimators $\bar{Q}^{H}_{k^*,n}(p)$ and $\bar{Q}^{\mathrm{PWM}(r)}_{k^*,n}(p)$, $n = 500$, from a Burr$(1.25, -1)$ parent

4 Finite Sample Behaviour: Small-Scale Simulation Study

We have implemented a Monte Carlo simulation of 5,000 runs for the shape parameter estimators $\gamma^{H}_{k^*,n}$, $\hat{\gamma}^{\mathrm{PWM}(r)}_{k^*,n}$ and high quantile normalized estimators $\bar{Q}^{H}_{k^*,n}(p) := \hat{Q}^{H}_{k^*,n}(p)/\chi_{1-p}$ and $\bar{Q}^{\mathrm{PWM}(r)}_{k^*,n}(p) := \hat{Q}^{\mathrm{PWM}(r)}_{k^*,n}(p)/\chi_{1-p}$ with $p = 1/n$ and $r \in \{0.25, 1, 1.25\}$. To work with the exact same number of top o.s., we have considered $k^* = k$ for the classical estimators and $k^* = k + 1$ for the PWM estimators. We have only considered Burr underlying parents with d.f. $F(x) = 1 - (1 + x^{-\rho/\gamma})^{1/\rho}$, $x > 0$ with $(\gamma, \rho) \in \{(0.25, -1.0), (1.25, -1.0)\}$. For every

Table 1 Burr parent: Simulated optimal sample fraction and mean values/RMSE, at their simulated optimal levels for the high quantile normalized estimators $\bar{Q}^{H}_{k^*,n}(p)$ and $\bar{Q}^{\mathrm{PWM}(r)}_{k^*,n}(p)$ with $p = 1/n$

	$(\gamma,\rho) = (0.25,-1.0)$				$(\gamma,\rho) = (1.25,-1.0)$			
	n	k_0^*/n	E	$RMSE$	n	k_0^*/n	E	$RMSE$
	50	0.0400	0.9569	0.2086	50	0.0400	1.3431	3.1957
	100	0.2100	1.0471	0.1984	100	0.2000	1.7620	2.4345
	200	0.1750	1.0505	0.1800	200	0.1750	1.6919	1.9448
H	500	0.1280	1.0460	0.1499	500	0.1260	1.5083	1.3346
	1000	0.1100	1.0479	0.1313	1000	0.1100	1.4479	1.0342
	2000	0.0905	1.0443	0.1128	2000	0.0755	1.3237	0.8255
	5000	0.0710	1.0407	0.0926	5000	0.0698	1.2974	0.6254
	50	0.5600	1.0295	0.1889	50	0.4200	0.3722	0.7233
	100	0.4600	1.0413	0.1813	100	0.2200	0.4005	0.6844
	200	0.3750	1.0471	0.1674	200	0.0900	0.4421	0.6446
PWM(1)	500	0.2840	1.0487	0.1445	500	0.0300	0.4738	0.6222
	1000	0.2310	1.0483	0.1280	1000	0.0170	0.4648	0.6248
	2000	0.1845	1.0449	0.1111	2000	0.0100	0.4471	0.6336
	5000	0.1374	1.0399	0.0912	5000	0.0050	0.4274	0.6483
	50	0.0400	0.9674	0.2098	50	0.1000	1.1564	1.8126
	100	0.3500	1.0491	0.1895	100	0.0300	1.0993	1.6142
	200	0.2950	1.0527	0.1703	200	0.0150	1.1045	1.5402
PWM(0.25)	500	0.2160	1.0480	0.1430	500	0.0060	1.1096	1.4340
	1000	0.1790	1.0478	0.1264	1000	0.0040	1.1548	1.4836
	2000	0.1420	1.0428	0.1085	2000	0.2600	1.4928	1.0606
	5000	0.1120	1.0400	0.0890	5000	0.1588	1.3177	0.6786
	50	0.8600	1.0402	0.1621				
	100	0.6000	1.0346	0.1681				
	200	0.4650	1.0398	0.1629				
PWM(1.25)	500	0.3440	1.0439	0.1492				
	1000	0.2860	1.0473	0.1356				
	2000	0.2335	1.0465	0.1206				
	5000	0.1722	1.0415	0.0984				

estimator we have simulated, the *mean value*, the *root mean squared error* (RMSE) the *optimal level*, $k_0^* = \arg\min_k RMSE$, and the *optimal sample fraction*, k_0^*/n. parameter estimators by H and PWM(r) and the normalized high quantile estimators by W-H and PWM(r).

To illustrate the finite sample behavior of the estimators, we present, in Figs. 2–5, the simulated mean values (E) and root mean squared errors (RMSE) patterns as functions of k^* for a sample size $n = 500$. In Table 1 we present the simulated optimal sample fraction and the simulated mean values and RMSE of the above-mentioned high quantile normalized estimators, at their simulated optimal levels. For Burr models with $(\gamma,\rho) = (1.25,-1)$, we present the simulated results of the PWM(r) estimators for $r = 0.25$ and 1. But for this second model, we need to have $r < 0.8$ to assure the consistency of the PWM(r) estimators.

Acknowledgements This research was partially supported by FCT – Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2011 (CEAUL), PEst-OE/MAT/UI0297/2011 (CMA/UNL) and PTDC/FEDER.

References

1. Caeiro, F., Gomes, M.I.: Semi-parametric tail inference through probability-weighted moments. J. Stat. Plann. Infer. **141**, 937–950 (2011)
2. Cai, J., de Haan, L., Zhou, C.: Bias correction in extreme value statistics with index around zero. Technical report. (2011)
3. Diebolt, J., Guillou, A., Worms, R.: Asymptotic behaviour of the probability-weighted moments and penultimate approximation. ESAIM: Probab. Stat. **7**, 219–238 (2003)
4. 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)
5. de Haan, L., Ferreira, A.: Extreme Value Theory: an Introduction. Springer Science + Business Media, LLC, New York (2006)
6. de Haan, L., Peng, L.: Comparison of tail index estimators. Stat. Neerlandica **52**, 60–70 (1988)
7. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
8. Hosking, J., Wallis, J.: Parameter and quantile estimation for the Generalized Pareto distribution. Technometrics **29**(3), 339–349 (1987)
9. Landwher, J., Matalas, N., Wallis, J.: Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. Water Resour. Res. **15**, 1055–1064 (1979)
10. Weissman, I.: Estimation of parameters of large quantiles based on the k largest observations. J. Am. Stat. Assoc. **73**, 812–815 (1978)

Third Order Conditions and Max-semistability

Sandra Dias and Maria da Graça Temido

Abstract

A third order condition for distribution functions that belong to max-semistable domain of attraction is established. An extension of the third order condition introduced in Fraga Alves et al. (Math. Methods Stat. 12:155–176, 2003) is obtained for the max-semistable context.

1 Introduction

In Probability Theory the concept of max-semistability appeared initially in the literature of sums. The genesis of max-semistable laws is due to Pancheva [8] and Grinevich [7]. This class of laws, the MSS class, includes the class of max-stable laws (MS), and also includes discrete laws and multimodal laws that do not belong to the MS class. According to Pancheva [8], a distribution function (d.f.) G is max-semistable (MSS) if there are reals $r > 1$, $a > 0$ and b such that $G(x) = G^r(ax + b)$, $x \in \mathbb{R}$, or equivalently, if exists a d.f. F, an integer nondecreasing sequence $\{k_n\}$ satisfying

$$\lim_{n\to+\infty} \frac{k_{n+1}}{k_n} = r \geq 1\,, \tag{1}$$

S. Dias
Departamento de Matemática, Ed. de Ciências Florestais, Apartado 1013, 5001-801 Vila Real, Portugal
e-mail: sdias@utad.pt

M. da Graça Temido (✉)
CMUC and DMUC, University of Coimbra, 3001-454 Coimbra, Portugal
e-mail: mgtm@mat.uc.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_16, © Springer-Verlag Berlin Heidelberg 2013

and real sequences $\{a_n > 0\}$ and $\{b_n\}$ for which we have

$$\lim_{n\to+\infty} F^{k_n}\left(a_n\, x\, + b_n\right) = G(x)\,,$$

for all continuity points of G. In this case we say that F belongs to the domain of attraction of G and write $F \in \mathscr{D}(MSS)$. The numerical expression of the elements of this class is $G_{\gamma,\nu}(\frac{x-\mu}{\sigma})$, with $\mu \in \mathbb{R}$ and $\sigma > 0$, where

$$G_{\gamma,\nu}(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right)\nu\left(\log(1+\gamma x)^{-1/\gamma}\right), & 1+\gamma x > 0\, x \in \mathbb{R},\, \gamma \neq 0 \\ \mathbb{I}_{]-\infty,0[}(\gamma), & 1+\gamma x \leq 0\, x \in \mathbb{R}\, \gamma \neq 0 \\ \exp\left(-\mathrm{e}^{-x}\nu(x)\right) & \gamma = 0,\, x \in \mathbb{R}, \end{cases}$$

being ν a positive, bounded, and periodic function with period $p = \ln r$. When $\nu \equiv 1$ we obtain the MS class.

The presence of this periodic function and of the parameter r (or any of its positive integer powers) makes the statistical inference in this class more difficult. However, there are several results obtained, from which we mention Canto e Castro and Dias [1], and Canto e Castro et al. [2] and [4].

In Extremes Statistics, asymptotic normality of a wide range of functions of intermediate order statistics is a primordial result and well known when the underlying d.f. F satisfies, in addition to the first order condition, a second and a third order condition. We refer, respectively, to conditions 1.1.20 and 2.3.3 in de Haan and Ferreira [5] and condition 3.1 introduced in Fraga Alves, de Haan and Lin [6]. That first order condition is a necessary and sufficient condition so that a d.f. F belongs to some domain of attraction in the MS class.

The possibility of being able to improve the estimation of r and ν and to establish a test statistic (as well as its asymptotic law) for the hypothesis $H : F \in \mathscr{D}(MSS)$, motivated us to obtain a third order condition for max-semistable laws. Thus, we follow a study analogous to the one done for MS domains of attraction (although compromised by the presence of r and ν). Indeed, for d.f.'s belonging to some MSS domain of attraction, in Canto e Castro and Temido [3] are introduced first and second order conditions that extend this class to theirs counterparts established for the MS class.

2 First and Second Order Conditions

For a d.f. F define the inverse function $U = (1/(1-F))^{\leftarrow}$. Canto e Castro and Temido [3] proved that a d.f. F belongs to the domain of attraction of G if and only if there exist an integer sequence $\{k_n\}$, under the mentioned conditions, a real sequence $\{a_n\}$ and a real γ such that is verified the first order condition

$$\lim_{n\to+\infty} \frac{U(k_n\,x)-U(k_n)}{a_n} = L(x) := \begin{cases} \dfrac{h^\gamma(x)-h^\gamma(1)}{\gamma} & \gamma \neq 0 \\ g(x)-g(1) & \gamma = 0, \end{cases} \tag{2}$$

for each continuity point of L. The functions h and g are defined by $h^{\leftarrow}(x) = \dfrac{x}{\nu(-\ln x)}$ and $g^{\leftarrow}(x) = \dfrac{e^x}{\nu(x)}$ and satisfy $h(z\,y) = z\,h(y)$ and $g(z\,y) = g(y) + \ln z$ if and only if $z \in \{r^m,\, m \in \mathbb{Z}\}$. According to Canto e Castro and Temido [3], for a d.f. F satisfying (2) the respective second order condition is described in the following way:

$$\lim_{n\to+\infty} \frac{U(k_n\,x)-U(k_n)-a_n L(x)}{A_n} = H(x), \quad x > 0,$$

where $\{A_n\}$ is a real sequence satisfying $A_n/a_n \to 0, n \to +\infty$. Considering that H is not a multiple of L, those authors proved that there exist periodic functions ξ and ξ_1, with period $\ln r$, and a second order parameter $\zeta > 0$, such that

$$H(x) = \begin{cases} x^{\frac{\ln\zeta}{\ln r}}\,\xi(\ln x) + L(x) & \gamma \neq 0 \ \vee \ \zeta \neq 1 \\ \xi_1(\ln x) + \left(g^2(x) - g^2(1)\right)/\ln r^2 & \gamma = 0,\ \zeta = 1. \end{cases}$$

3 Third Order Condition

The appropriate third order condition for max-semistable distributions appears from the first and second order conditions and was, obviously, motivated by the third order condition introduced in Fraga Alves et al. [6]. For d.f.'s that satisfy the first and second order conditions, this third order condition assumes the existence of real sequences $\{a_n\}$, $\{A_n\}$ and $\{B_n\}$ for which there exists the limit function R in

$$\lim_{n\to+\infty} \frac{U(k_n\,x)-U(k_n)-a_n L(x)-A_n H(x)}{B_n} = R(x), \quad x > 0. \tag{3}$$

In order to establish the limit function R we present the following lemmas, where the first one was presented in Canto e Castro and Temido [3].

Lemma 3.1. *Let $\{k_n\}$ be a nondecreasing real sequence satisfying (1), with $r > 1$, and F a d.f. that satisfies (2). Then*

$$\lim_{n\to+\infty} \frac{U(k_n\,r^m\,x)-U(k_n\,r^m)}{a_n} = \lim_{n\to+\infty} \frac{U(k_{n+m}\,x)-U(k_{n+m})}{a_n}$$

and

$$\lim_{n\to+\infty} \frac{U(k_n\, r^m\, x) - U(k_n\, r^m)}{U(k_{n+m}\, x) - U(k_{n+m})} = 1.$$

Lemma 3.2. *Let $\{a_n\}$, $\{A_n\}$ and $\{B_n\}$ be positive real sequences such that*

$$\lim_{n\to+\infty} \frac{a_{n+1}}{a_n} = \beta, \qquad \lim_{n\to+\infty} \frac{A_{n+1}}{A_n} = \eta, \qquad \textit{and} \qquad \lim_{n\to+\infty} \frac{B_{n+1}}{B_n} = \theta.$$

Then there are constants C and C_1 such that

$$\lim_{n\to+\infty} \frac{a_{n+m} - \beta^m a_n + A_{n+m} - \beta^m A_n}{B_n} = \begin{cases} C(\beta^m - \theta^m) & \theta \neq \beta \\ Cm\theta^m & \theta = \beta, \end{cases}$$

$$\lim_{n\to+\infty} \frac{A_{n+m} - \eta^m A_n}{B_n} = \begin{cases} C(\eta^m - \theta^m) & \theta \neq \eta \\ Cm\theta^m & \theta = \eta \end{cases}$$

and

$$\lim_{n\to+\infty} \frac{a_{n+m} - a_n - mA_n}{B_n} = \begin{cases} C_1(1-\theta^m) + Cm & \theta \neq 1 \\ C_1 m + C\dfrac{m^2}{2} & \theta = 1. \end{cases}$$

Proof. Let us consider $s_m = \lim_{n\to+\infty} \frac{a_{n+m} - \beta^m a_n + A_{n+m} - \beta^m A_n}{B_n}$. Then

$$s_{m+p} = s_p\theta^m + \beta^p s_m = s_m\theta^p + \beta^m s_p.$$

If $\theta \neq \beta$, since

$$s_p\theta^m + \beta^p s_m = s_m\theta^p + \beta^m s_p \iff s_m(\beta^p - \theta^p) = s_p(\beta^m - \theta^m),$$

we conclude that there exists a constant C such that $s_m = C(\beta^m - \theta^m)$.

If $\theta = \beta$, this leads to $s_{m+p} = s_m\theta^p + \theta^m s_p$ which solution is $s_m = Cm\theta^m$ for some constant C.

In what concerns the second limit, write

$$\frac{A_{n+m} - \eta^m A_n}{B_n} = \eta^m \frac{\eta^{-n-m} A_{n+m} - \eta^{-n} A_n}{\eta^{-n} B_n} = \eta^m \frac{D_{m+n} - D_n}{E_n},$$

where $D_n = \eta^{-n} A_n$ and $E_n = \eta^{-n} B_n$, and consider

$$s_m = \lim_{n\to+\infty} \frac{D_{m+n} - D_n}{E_n},\ n \to +\infty.$$

Since

$$\frac{E_{m+n}}{E_n} = \frac{\eta^{-n-m} B_{m+n}}{\eta^{-n} B_n} \to \eta^{-m}\theta^m,$$

we deduce $s_{n+m} = s_p \eta^{-m}\theta^m + s_m$ and similarly $s_{m+p} = s_m \eta^{-p}\theta^p + s_p$, for every $m, p \in \mathbb{Z}$.

If $\theta \neq \eta$, there exists a real constant C such that $s_m = C(1 - \eta^{-m}\theta^m)$, which enables us to conclude that

$$\lim_{n\to+\infty} \frac{A_{n+m} - \eta^m A_n}{B_n} = \eta^m C(1 - \eta^{-m}\theta^m) = C(\eta^m - \theta^m).$$

If, otherwise $\theta = \eta$, we have $s_{n+m} = s_p + s_m$, for every $m, p \in \mathbb{Z}$. So, there exists a real constant C such that $s_m = Cm$, and the desired result is obtained.

For the last limit, taking

$$s_m = \lim_{n\to+\infty} \frac{a_{n+m} - a_n - mA_n}{B_n},$$

we obtain

$$\begin{aligned} s_{m+p} = & \lim_{n\to+\infty} \frac{a_{n+m+p} - a_{m+n} - pA_{n+m}}{B_{m+n}} \lim_{n\to+\infty} \frac{B_{n+m}}{B_n} \\ & + \lim_{n\to+\infty} \frac{a_{n+m} - a_n - mA_n}{B_n} + p \lim_{n\to+\infty} \frac{A_{n+m} - A_n}{B_n}. \end{aligned}$$

Take $\theta \neq 1$. Using the result obtained for the second limit, we get

$$s_{m+p} = s_p\theta^m + s_m + pC(1 - \theta^m).$$

Considering $s^*_m = s_m - Cm$, the previous equation is equivalent to $s^*_{m+p} = \theta^m s^*_p + s^*_m$ and in the same way we get $s^*_{m+p} = \theta^p s^*_m + s^*_p$. Therefore, there exists a constant C_1 such that $s^*_m = C_1(1 - \theta^m)$, which implies $s_m = Cm + C_1(1 - \theta^m)$.

If $\theta = 1$, from the second limit we obtain $s_{m+p} = s_p + s_m + Cpm$. Now we consider $s^*_m = s_m - Cm^2/2$, which allows to get $s^*_{m+p} = s^*_p + s^*_m$. Then $s^*_m = C_1 m$, for some constant C_1, and so $s_m = Cm^2/2 + C_1 m$. □

Theorem 3.1. *Let $\{k_n\}$ be a nondecreasing real sequence satisfying (1) with $r > 1$. If the third order condition (3) holds, where R is not a multiple of H, then there exist periodic functions ξ_i, $i = 2, 4, \ldots, 8$ with period $\ln r$, and a third order parameter χ such that the function R is given by*

$$R(x) =$$

$$= \begin{cases} x^{\frac{\ln\chi}{\ln r}}\xi_2(\ln x) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + L(x) & r^\gamma \neq \zeta \neq \chi, \chi \neq r^\gamma \\ x^{\frac{\ln\chi}{\ln r}}\xi_3(\ln x) + \dfrac{\ln x}{\chi \ln r} x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + L(x) & \zeta = \chi \neq r^\gamma \\ x^{\frac{\ln\chi}{\ln r}}\xi_4(\ln x) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + \dfrac{\ln x}{\chi \ln r}\dfrac{h^\gamma(x)}{\gamma} & \zeta \neq \chi = r^\gamma \neq 1 \\ x^{\frac{\ln\chi}{\ln r}}\xi_5(\ln x) + \dfrac{\ln x}{\chi \ln r}\left(x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + \dfrac{h^\gamma(x)}{\gamma}\right) & \chi = r^\gamma = \zeta \neq 1 \\ \xi_6(\ln x) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + \dfrac{g^2(x) - g^2(1)}{2\ln r} & \zeta \neq r^\gamma = \chi = 1 \\ x^{\frac{\ln\chi}{\ln r}}\xi_7(\ln x) + \xi_1(\ln x) + \dfrac{g^2(x) - g^2(1)}{2\ln r} + L(x) & \chi \neq r^\gamma = \zeta = 1 \\ \xi_8(\ln x) + \dfrac{\ln x}{\ln r}\xi_1(\ln x) + \dfrac{g^2(x) - g^2(1)}{2\ln r} + \dfrac{g^3(x) - g^3(1)}{6\ln^2 r} & r^\gamma = \zeta = \chi = 1, \end{cases}$$

with $x > 0$. *Furthermore,* $\xi_i(0) = 0$, $i = 2, 4, \ldots, 8$ *and, as* $n \to +\infty$, $B_{n+1}/B_n \to \chi$.

Proof. Due to the expected repetition we do not present all the steps. From (3) and considering that $L(r^m x) - L(r^m) = r^{m\gamma} L(x)$, we have

$$\begin{aligned} & R(r^m x) - R(r^m) \\ = & = \lim_{n\to+\infty} \frac{U(k_n r^m x) - U(k_n r^m) - a_n\left(L(r^m x) - L(r^m)\right) - A_n\left(H(r^m x) - H(r^m)\right)}{B_n} \\ & = \lim_{n\to+\infty} \frac{U(k_{n+m}x) - U(k_{n+m}) - a_{n+m}L(x) - A_{n+m}H(x)}{B_{n+m}}\frac{B_{n+m}}{B_n} \\ & \quad + \frac{a_{n+m}L(x) - a_n r^{m\gamma}L(x) - A_n\left(H(r^m x) - H(r^m)\right) + A_{n+m}H(x)}{B_n} \qquad (4) \\ & = R(x)(1 + o_n(1)) \lim_{n\to+\infty} \frac{B_{n+m}}{B_n} \\ & + \begin{cases} \dfrac{a_{n+m} - a_n r^{m\gamma} + A_{m+n} - r^{m\gamma}A_n}{B_n} L(x) + \dfrac{A_{n+m} - \zeta^m A_n}{B_n} x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) & r^\gamma \neq 1 \vee \zeta \neq 1 \\ \dfrac{a_{n+m} - a_n - mA_n}{B_n} L(x) + \dfrac{A_{n+m} - A_n}{B_n} H(x) & r^\gamma = \zeta = 1. \end{cases} \end{aligned}$$

Case A $(r^\gamma \neq \zeta \neq \chi, r^\gamma \neq \chi \neq 1)$: From (5) and Lemma 3.2 we obtain the functional equation

$$R(r^m x) - R(r^m) = R(x)\chi^m + L(x)(r^{m\gamma} - \chi^m) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x)(\zeta^m - \chi^m),$$

for all positive integer m. So, we solve the linear equation

$$R(rx) - R(r) = R(x)\chi + L(x)(r^\gamma - \chi) + x^{\frac{\ln\zeta}{\ln r}}(\zeta - \chi), \tag{5}$$

starting by its homogeneous part $R_h(rx) = R_h(x)\chi$. Considering $x = \mathrm{e}^z$ and multiplying by $\mathrm{e}^{-z\frac{\ln\chi}{\ln r}}$ we get

$$\mathrm{e}^{-(z+\ln r)\frac{\ln\chi}{\ln r}} R(\mathrm{e}^{z+\ln r}) = \mathrm{e}^{-z\frac{\ln\chi}{\ln r}} R(\mathrm{e}^z).$$

From the last equation we conclude that $\xi_2(x) := \mathrm{e}^{z+\frac{\ln\chi}{\ln r}} R(e^z)$ is a periodic function with period $\ln r$. Then a general solution of the homogeneous equation is $R_h(x) = x^{\frac{\ln\chi}{\ln r}}\xi_3(\ln x)$. Taking into account that a particular solution of the complete equation (5) is $R_p(x) = x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + L(x)$, we obtain

$$R(x) = R_h(x) + R_p(x) = x^{\frac{\ln\chi}{\ln r}}\xi_3(\ln x) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + L(x).$$

Case B $(r^\gamma \neq \zeta \neq \chi, r^\gamma \neq \chi = 1)$: The limit (5) and Lemma 3.2 give us the equation

$$R(r^m x) - R(r^m) = R(x) + L(x)(r^{m\gamma} - 1) + x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x)(\zeta^m - 1).$$

In this case $R_h(x) = \xi_2(\ln x)$ and $R_p(x) = x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x) + L(x)$.

Case C $(r^\gamma \neq \zeta = \chi = 1)$: From (5) we get the functional equation

$$R(r^m x) - R(r^m) = R(x) + L(x)(r^{m\gamma} - 1) + m\xi(\ln x),$$

for which $R_h(x)$ is equal to the previous one and $R_p(x) = \frac{\ln x}{\ln r}\xi(\ln x) + L(x)$.

Case D $(r^\gamma = \zeta = \chi \neq 1)$: Using the results from Lemma 3.2 in (5), we obtain the functional equation

$$R(r^m x) - R(r^m) = R(x)\chi + m\chi^{m-1}L(x) + m\chi^{m-1}x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x).$$

Since, in this case $R_h(x)$ is equal to the one in Case A and $R_p(x) = \frac{\ln x}{\chi \ln r}\frac{h^\gamma(x)}{\gamma} + \frac{\ln x}{\chi \ln r}x^{\frac{\ln\zeta}{\ln r}}\xi(\ln x)$ we conclude the desired result.

Case E $(\chi \neq r^\gamma = \zeta = 1)$: From (5) we obtain the equation

$$R(r^m x) - R(r^m) = R(x)\chi^m + L(x)\left((1-\chi^m) + m\right) + H(x)(1-\chi^m)$$

for which $R_p(x) = H(x) + L(x)$ is a particular solution of the complete equation.

Case F $(r^\gamma = \zeta = \chi = 1)$: In this case limit (5) gives

$$R(r^m x) - R(r^m) = R(x) + L(x)\left(m + \frac{m^2}{2}\right) + mH(x)$$

with

$$R_p(x) = \frac{\ln x}{\ln r}\xi_2(\ln x) + \frac{g^2(x) - g^2(1)}{2\ln r} + \frac{g^3(x) - g^3(1)}{6\ln^2 r}.$$

□

Acknowledgements Sandra Dias was funded by the Portuguese Government through the Fundação para a Ciência e a Tecnologia (FCT) under the project PEst-OE/MAT/UI4080/2011. The research of Maria da Graça Temido was supported by the Centre for Mathematics of the University of Coimbra and FCT, through European program COMPETE/FEDER.

References

1. Canto e Castro, L., Dias, S.: Asymptotic distribution of certain statistics relevant to the fitting of max-semistable models. Portugaliae Mathematica **66**, 401–412 (2009)
2. Canto e Castro, Dias, S., Temido, M.G.: Tail inference for a law in a max-semistable domain of attraction. Pliska Studia Mathematica Bulgarica **19**, 83–96 (2009)
3. Canto e Castro, L., Temido, M.G.: On a second order condition for max-semistable laws. Pliska Studia Mathematica Bulgarica **19**, 97–110 (2009)
4. Canto e Castro, L., Dias, S., Temido, M.G.: Looking for max-semistability: a new test for the extreme value condition. J. Stat. Plan. Infer. **141**, 3005–3020 (2011)
5. de Haan, L., Ferreira, A.: Extreme Value Theory: An Introduction. Springer Series in Operational Research and Financial Engineering, Springer, New York (2006)
6. Fraga Alves, I., de Haan, L., Lin, T.: Estimation of the parameter controlling the speed of convergence in extreme value theory. Math. Methods Stat. **12**, 155–176 (2003)
7. Grinevich, I.V.: Max-semistable limit laws under linear and power normalizations. Theory Probab. Appl. **38**, 640–650 (1992)
8. Pancheva, E.: Multivariate max-semistable distributions. Theory Probab. Appl. **18**, 679–705 (1992)

Part IV

Testing Statistical Hypothesis

Generalized p Values and Random p Values When the Alternative to Uniformity Is a Mixture of a Beta(1,2) and Uniform

M.F. Brilhante

Abstract

Combining p values methods and uniformity tests are closely related subjects in meta analysis. In this context it is also known that publication bias can seriously impair an overall decision. The recent concepts of generalized p values and of random p values emphasize that, when faced with a significant number of results that casts some doubt on the null hypothesis, the correct approach to the problem should be to combine evidence under the alternative hypothesis. Following previous research, we investigate generalized p values and random p values for testing uniformity when the alternative is a mixture of a Beta(1,2) and standard uniform random variables.

1 Introduction

The main goal of meta analysis is to provide methods for combining results from independent tests, in order to obtain a valid conclusion about a common null hypothesis. One way of combining information is through the reported p values, and therefore uniformity tests play a crucial role, since the observed p values are, under the null hypothesis, observations from the standard uniform distribution. There are, however, two major concerns in meta analysis: the small number of reported p values, which ultimately affects the power of any combined test; and the phenomenon known as publication bias, which is a direct consequence of the reported p values being usually smaller than 0.05, and thus creating a biased sample.

M.F. Brilhante (✉)
Universidade dos Açores (DM) and CEAUL, Campus de Ponta Delgada, Apartado 1422, 9501-801 Ponta Delgada, Portugal
e-mail: fbrilhante@uac.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_17, © Springer-Verlag Berlin Heidelberg 2013

In order to overcome the issues mentioned above, Gomes et al. [4] investigated the effects of using computationally augmented samples to test uniformity, and considered the family of random variables X_m with probability density function (pdf)

$$f_m(x) = \left(mx + 1 - \tfrac{m}{2}\right) \mathrm{I}_{(0,1)}(x)\,, \quad m \in [-2, 0]\,, \tag{1}$$

as an alternative ($m = 0$ corresponds to uniformity). They observed that the computational augmentation of samples had the nasty effect of decreasing the power of combined uniformity tests, contrarily to what was naïvely expected. Their results showed that data augmentation techniques, which has been used with relative success by other authors (e.g., [8]), are in fact quite arguable. Posterior investigations of Brilhante et al. [1, 2] confirmed the attraction role of the uniform distribution in this context and provided some explanations for the occurrence.

Following previous research on the topic, we decided to take a new approach to the problem. We apply with some success the concepts of generalized p values and of random p values to the specific uniformity test problem $H_0 : m = 0$, when compared to Fisher's method for combining evidence, and continue to consider family (1) as an alternative. Observe that densities defined by (1), which are a mixture of a $\mathrm{Beta}(1, 2)$ and $\mathrm{Uniform}(0, 1)$ with mixing weights $-\frac{m}{2}$ and $1 + \frac{m}{2}$, respectively, are quite attractive to work with, primarily because of their simplicity[1] and the fact that they tackle the problem of publication bias. Fisher's test is also used to explore the concept of expected p value with family (1). As we shall see, an interesting conclusion is derived here. Note that other models with support on $(0, 1)$ can be considered, e.g., Brilhante et al. [2] used the model $\mathrm{Beta}(1, q)$, $q \in [0.5, 3]$, when combining p values.

This chapter is organized as follows. In Sect. 2 we start by discussing the concepts of generalized p values and then applying the main ideas to obtain a generalized variable test for testing uniformity. In Sect. 3 we explore the concepts of random p values and expected p values using Fisher's method for combining evidence, always using family (1) as an alternative to uniformity.

2 Generalized p Values with Alternative in the Class X_m

The concept of generalized p values was introduced by Tsui and Weerahandi [10] to deal with the challenging, and quite often impossible, task of finding test statistics when nuisance parameters are involved. A brief overview of the subject follows.

[1] The family (1) results from tilting the standard uniform density using the point (0.5,1) as a rotation center.

Suppose that X is a random variable whose distribution depends on the parameter of interest θ and the nuisance parameter η (real or vector), and that we are interested in testing

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta > \theta_0 \,. \tag{2}$$

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the population. Since we are dealing with p values, we shall consider a significance testing setting. The usual approach to the problem is to find a test statistic $T = T(\boldsymbol{X}; \theta, \eta)$ whose distribution is free of η. If T is stochastically increasing with θ, i.e., $\Pr[T > t; \theta]$ is a nondecreasing function of θ for any given t and fixed η, then the p value for problem (2) is computed as

$$p = \Pr[T(\boldsymbol{X}; \theta, \eta) \geq T(\boldsymbol{x}; \theta, \eta) | \theta = \theta_0] \,,$$

where $\boldsymbol{x}$ is the observed value of $\boldsymbol{X}$.[2]

Tsui and Weerahandi [10] extended the conventional definition of p values in order to eliminate their dependence on nuisance parameters. First they began by defining the concept of generalized test variables.

Definition 2.1. A random variable of the form $T = T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta)$ is a generalized test variable for the parameter θ if it satisfies the following properties:
1. The observed value of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta)$, i.e., $T(\boldsymbol{x}; \boldsymbol{x}, \theta, \eta)$, is free of θ and η.
2. When θ is specified, the distribution of $T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta)$ is free of η.
3. For fixed $\boldsymbol{x}$ and η, $\Pr[T \leq t; \theta]$ is a monotonic function of θ for any given t.

If $T = T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta)$ is a generalized test variable stochastically increasing with θ, then the generalized extreme region for problem (2) is defined as

$$\mathscr{C}_{\boldsymbol{x}}(\theta, \eta) = \{\boldsymbol{X} : T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta) \geq T(\boldsymbol{x}; \boldsymbol{x}, \theta, \eta)\} \,,$$

which does not depend on the nuisance parameter η. Therefore, the generalized p value is computed as

$$p_G = \Pr[T(\boldsymbol{X}; \boldsymbol{x}, \theta, \eta) \geq T(\boldsymbol{x}; \boldsymbol{x}, \theta, \eta) | \theta = \theta_0].$$

The advantage of using generalized p values instead of ordinary p values in testing problems with nuisance parameters is that the former enables the problems to be solved exactly. An additional advantage is that they tend to behave quite well with respect to the usual frequentist approach.

[2] If T is stochastically decreasing with θ, then $p = \Pr[T(\boldsymbol{X}; \theta, \eta) \leq T(\boldsymbol{x}; \theta, \eta) | \theta = \theta_0]$.

Our main interest here is the testing problem

$$H_0 : m = 0 \quad vs. \quad H_1 : m < 0\,, \tag{3}$$

i.e., testing uniformity against a density more prone to generate values close to zero, as suggested by publication bias. Observe that (1) remains a pdf if $m \in [0, 2]$, and therefore a right unilateral test can also be considered (this case augments the probability near one).[3] Although no nuisance parameters are involved in problem (3), the main ideas supporting the concept of generalized p values can still be applied.

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from the population with pdf (1) and $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the observed value of $\boldsymbol{X}$. Ideally we should base our test statistics on sufficient statistics. However, the application of the factorization theorem does not provide a sufficient statistic for the parameter m outside the trivial solutions. Therefore, we decided to use the probability integral transform theorem and apply the recipe for constructing generalized pivot quantities proposed by Iyer and Patterson [5]—note that there exists no general method that will yield generalized test variables for a testing problem.

Basically the recipe can be stated as follows. Suppose that we have a problem involving k unknown parameters $\theta_1, \ldots, \theta_k$ and that the parameter of interest is $\theta = f(\theta_1, \ldots, \theta_k)$. If there exists a set of statistics $(U_1, \ldots, U_k)$ and a set of invertible pivots $(V_1, \ldots, V_k)$ relating $(U_1, \ldots, U_k)$ to $(\theta_1, \ldots, \theta_k)$, then by expressing the parameter of interest θ in terms of the U_i and V_i's, i.e., $\theta = g(U_1, \ldots, U_k, V_1, \ldots, V_k)$, an expression for a generalized pivot quantity, which can be used for inferences about θ, is $R = \theta - g(u_1, \ldots, u_k, V_1, \ldots, V_k)$, where u_i denotes the observed value of U_i, $i = 1, \ldots, k$.

Let F_m denote the distribution function of X_m. Then from the probability integral transform theorem it follows that $F_m(X_i) \sim \text{Uniform}(0, 1)$, $i = 1, \ldots, n$, and consequently $V = -2 \sum_{i=1}^{n} \ln[F_m(X_i)] \sim \chi^2_{2n}$ is a pivot quantity.

On the other hand,

$$\sum_{i=1}^{n} \ln[F_m(X_i)] = \sum_{i=1}^{n} \ln\left[X_i\left(\tfrac{m}{2}X_i + \tfrac{2-m}{2}\right)\right] = \sum_{i=1}^{n} \ln X_i + \sum_{i=1}^{n} \ln\left[\tfrac{m}{2}(X_i - 1) + 1\right].$$

Since $0 < \frac{m}{2}(x - 1) < 1$ for $m \in [-2, 0]$ and $x \in (0, 1)$—when $m \in [0, 2]$, we have $-1 < \frac{m}{2}(x - 1) < 0$—we can use as an approximation to $\ln\left[\frac{m}{2}(X_i - 1) + 1\right]$ the first term of the function's expansion in Maclaurin series, i.e., $\ln\left[\frac{m}{2}(X_i - 1) + 1\right] \approx \frac{m}{2}(X_i - 1)$, $i = 1, \ldots, n$. Hence,

[3] For $m \in [0, 2]$, note that (1) is a mixture of a Beta(2,1) and Uniform(0,1) with mixing weights $\frac{m}{2}$ and $1 - \frac{m}{2}$, respectively.

$$V \approx -2\sum_{i=1}^{n} \ln X_i - mn\left(\overline{X} - 1\right) . \tag{4}$$

Solving (4) with respect to m, we obtain $m \approx \frac{V+2\sum_{i=1}^{n} \ln X_i}{n(1-\overline{X})}$. We are now in condition to apply the method discussed above, and define

$$T(\boldsymbol{X}; \boldsymbol{x}, m) = m - \frac{V + 2\sum_{i=1}^{n} \ln x_i}{n(1-\overline{x})} \tag{5}$$

as a generalized test variable for m (note that this *modus operandi* guarantees that (5) is in fact a generalized test variable and that it is stochastically increasing with m). Since $T(\boldsymbol{x}; \boldsymbol{x}, m) \approx 0$, the generalized p value for problem (3) is computed as

$$p_G = \Pr\left[T(\boldsymbol{X}; \boldsymbol{x}, m) \leq 0 | m = 0\right] = 1 - \Pr\Big(V \leq -2\sum_{i=1}^{n} \ln x_i\Big) ,$$

with $V \sim \chi^2_{2n}$.

On the other hand, the generalized data-based power function of the generalized test variable (5) is defined as

$$\pi(\boldsymbol{x}; m) = \Pr\left[T(\boldsymbol{X}; \boldsymbol{x}, m) \leq 0 | m\right] = 1 - \Pr\Big[V \leq mn(1-\overline{x}) - 2\sum_{i=1}^{n} \ln x_i\Big] .$$

If we were interested in the right unilateral test $H_0 : m = 0$ vs. $H_1 : m > 0$, then $p_G = \Pr\left[T(\boldsymbol{X}; \boldsymbol{x}, m) \geq 0 | m = 0\right] = \Pr\Big(V \leq -2\sum_{i=1}^{n} \ln x_i\Big)$.

A commonly applied method for combining p values from independent tests is Fisher's method, which is based on the test statistic $T_n = -2\sum_{i=1}^{n} \ln X_i$. Under the null hypothesis of uniformity we have $T_n \sim \chi^2_{2n}$, and therefore H_0 should be rejected if $T_n \geq \chi^2_{2n;1-\alpha}$, where $\chi^2_{2n;1-\alpha}$ denotes the quantile of order $1-\alpha$ of the distribution. (For more details on combining p values see, e.g., Pestana [7].) Although there is no absolute better way of combining evidence, simulation studies have shown that Fisher's test is one with the best performance in a majority of situations. Thus its normal use as a comparison with other possible tests.

Observe that (5) is, in a certain way, a generalization of Fisher's test statistic for problem (3). This is quite understandable given the way how both tests are derived. On the other hand, it is also obvious that the two tests are equally powerful for problem (3). However, the generalized test variable (5) has here a slighter advantage over Fisher's test, in the sense that it allows to test nonuniformity null hypotheses $H_0 : m = m_0 \neq 0$ if we desire so. Figure 1 shows the simulated generalized power function (based on 5,000 runs) of the generalized test variable (5) for sample sizes $n = 5, 10, 20, 30, 50, 100$ and $\alpha = 0.05$.

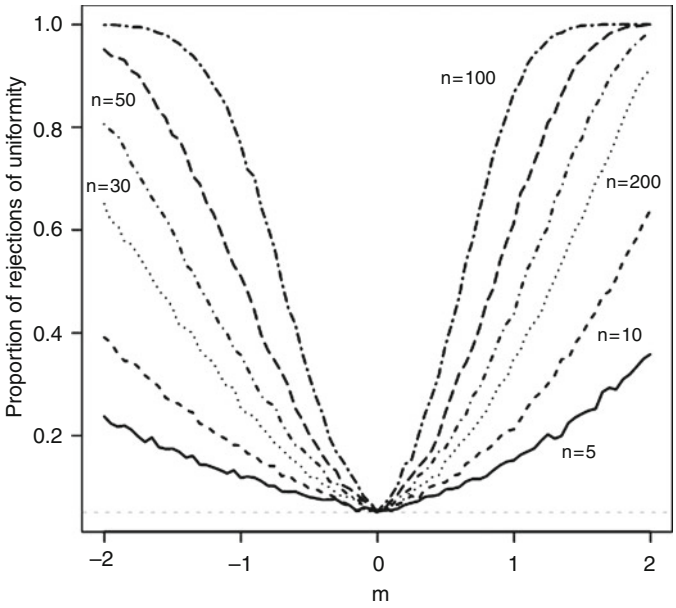

Fig. 1 Proportion of rejections of uniformity using the generalized test variable $T(\boldsymbol{X}; \boldsymbol{x}, m)$

3 Random p Values and Fisher's Method with Alternative in the Class X_m

An important aspect of p values, which is often forgotten, is its stochastic behavior. Dempster and Schatzoff [3] were the first to explore this behavior, and more recently other authors have rekindled the subject (e.g., [6, 9]). A small introduction to the topic follows.

Let T be a continuous test statistic to test hypotheses about a parameter of interest θ. Let F_0 denote the distribution function (df) of T under the null hypothesis H_0 and F_θ the df under some alternative H_1. If large values of T give evidence in favor of H_1, then the observed p value is given by $p = 1 - F_0(t)$, where t is the observed value of T. The random p value associated with the test is simply defined as the random variable $P = 1 - F_0(T)$. From the previous definition it follows immediately that P is a standard uniform random variable under H_0 (a direct consequence from the probability integral transform theorem and the uniform distribution properties). On the other hand, the df of P under the alternative is given by

$$\Pr_\theta[P \leq p] = 1 - F_\theta(F_0^{-1}(1-p)), \quad 0 < p < 1. \tag{6}$$

Some authors advocate the use of the expected p value, EPV, under the alternative as a measure of performance of a test (e.g., [3, 9]). The smaller the value, the better the test. Other reasons in favor of use of the EPV in these authors opinion are: (1) it depends on the alternative, not on the significance level; (2) it allows to determine the sample size; and (3) it allows to determine which alternative the observed p value represents. However, other authors have a different understanding on the matter, they argue that the random p value's distribution under alternatives is usually highly skewed, and therefore is not a good measure to represent its distribution ([6]).

From a computational perspective it can be easier to determine the EPV by computing $\Pr[T^* \geq T]$, where T^* and T are independent random variables with T^* having df F_θ and T df F_0 (cf. [9]). Note that

$$\mathrm{EPV}(\theta) = \Pr[T^* \geq T] = \int_{-\infty}^{\infty} \Pr[T^* \geq T \big| T = t] f_T(t) dt = \mathrm{E}[1 - F_\theta(T)],$$

which under the null hypothesis takes the value 0.5, as expected.

For problem (3) it is not difficult to derive the distribution of Fisher's statistic T_n under the alternative, specially for small sample sizes, and use this knowledge to determine the distribution of $P = 1 - F_0(T_n)$ under a particular alternative m.

For example, T_2 has pdf

$$f_{T_2}(t) = \frac{1}{2}\left[\left(m^2 \mathrm{e}^{-t/2} + \frac{(m-2)^2}{4}\right)\frac{t}{2} - m(m-2)(1-\mathrm{e}^{-t/2})\right]\mathrm{e}^{-t/2}, \quad t > 0,$$

and T_3

$$\begin{aligned} f_{T_3}(t) = \frac{1}{128}\mathrm{e}^{-t}\Big\{ & -12m[-16 + \mathrm{e}^{t/2}(t-4)^2] + 8t^2\mathrm{e}^{t/2} + \\ & + 6m^2[\mathrm{e}^{t/2}(t-8)^2 - 16t - 64] + \\ & + m^3[-\mathrm{e}^{t/2}(t-12)^2 + 144 + 48t + 8t^2]\Big\}, \quad t > 0. \end{aligned}$$

The general expression for P's pdf under the alternative m can easily be derived from (6), and is given by

$$f_P(p) = \frac{f_m[F_0^{-1}(1-p)]}{f_0[F_0^{-1}(1-p)]} = \frac{f_m(\chi^2_{2n;1-p})}{f_0(\chi^2_{2n;1-p})}, \quad 0 < p < 1,$$

where f_m denotes now the pdf of T_n under the alternative m.

In Fig. 2 we plot the pdf of P for some values of m and for $n = 3$ (the pattern for $n = 2$ is similar). In Fig. 3 we plot the simulated EPV under the alternative hypothesis. It is interesting to observe here that for sample sizes smaller than 20, we are not likely to observe p values under 0.05.

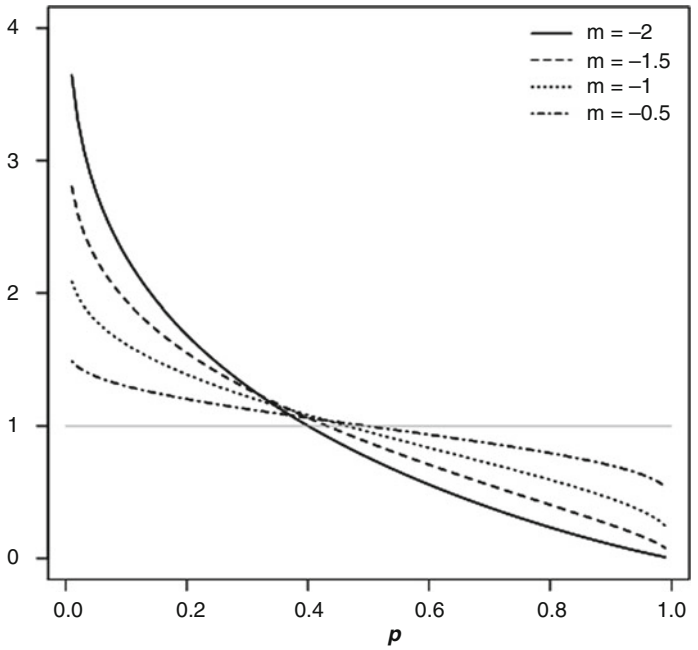

Fig. 2 Probability density function for the random p value P under some alternatives for $n = 3$

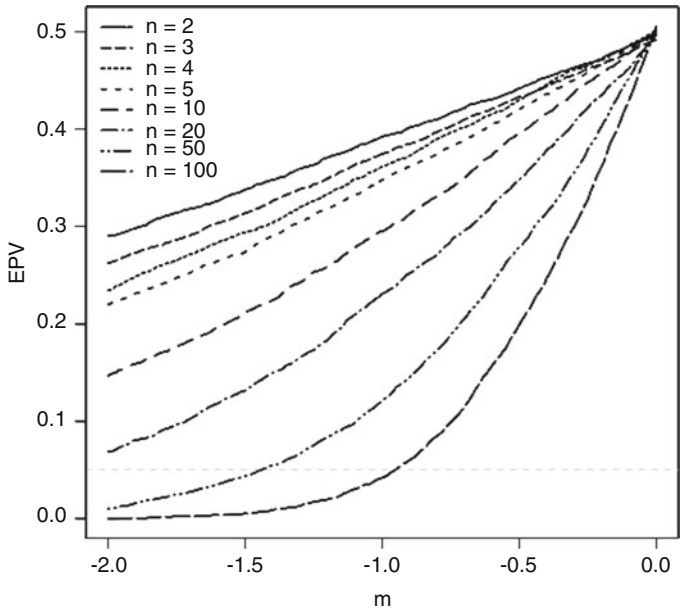

Fig. 3 EPV under the alternative hypothesis for Fisher's test statistic

4 Conclusions

The concepts of generalized p values and of random p values can offer attractive tools when dealing with meta-analytical syntheses of evidence, specially in studies showing grounds to reject the overall null hypothesis. Aside from the philosophical appeal of these concepts, their construction is straightforward, and theoretical and computational results can be sometimes quite convincing.

Following previous work, we explored these relatively new concepts using the family of pdf's (1) and proposed in Sect. 2 a generalized test variable for problem (3), comparable to Fisher's test in this particular case. On the other hand, applying the concept of random p values to Fisher's statistic in Sect. 3, and computing the associated EPV for some sample sizes, revealed that some caution is needed on our part in meta-analysis syntheses when dealing with small sample sizes (a recommendation closely related to power considerations). However, it is advisable to use EPVs with some prudence, specially in the presence of highly skewed distributions for P under the alternative.

Acknowledgements This research has been supported by National Funds through FCT—Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011.
The author would also like to thank Professor Dinis Pestana for the stimulating conversations held about the topics discussed here, and the reviewers for their comments which led to a substantial improvement of the chapter.

References

1. Brilhante, M.F., Mendonça, S., Pestana, D., Sequeira, F.: Using products and powers of products to test uniformity. In: Luzar-Stiffler, V., et al. (eds.) Proceedings of the 32nd International Conference on Information Technology Interfaces, SRCE, Zagreb pp. 509–514 (2010)
2. Brilhante, M.F., Pestana, D., Sequeira, F.: Combining p-values and random p-values. In: Luzar-Stiffler, V., et al. (eds.) Proceedings of the 32nd International Conference on Information Technology Interfaces, pp. 515–520 (2010)
3. Dempster, A.P., Schatzoff, M.: Expected significance level as a sensibility index for test statistics. J. Am. Stat. Assoc. **60**, 420–436 (1965)
4. Gomes, M.I., Pestana, D.D., Sequeira, F., Mendonça, S., Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. In: Luzar-Stiffler, V., et al. (eds.) Proceedings of the 31st International Conference on Information Technology Interfaces, pp. 243–248 (2009)
5. Iyer, H.K., Patterson, P.D.: In a recipe for constructing generalized pivotal quantities and generalized confidence intervals, Colorado State University Department of Statistics Technical Report 2002/10, http://www.state.colostate.edu/research/2002_2010.pdf (2002/2010)
6. Kulinskaya, E., Morgenthaler, S., Staudte, R.G.: Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence. Wiley, Chichester (2008)
7. Pestana, D.: Combining p-values. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science, pp. 1145–1147. Springer, Berlin (2011)
8. Royle, J.A., Dorazio, R.M., Link, W.A.: Analysis of multinomial models with unknown index using data augmentation. J. Comput. Grap. Stat. **16**, 67–85 (2007)
9. Sackrowitz, H., Samuel-Cahn, E.: p-Values as Random Variables – Expected P Values. J. Am. Stat. Assoc. **53**, 326–331 (1999)
10. Tsui, K., Weerahandi, S.: Generalized p-values in significance testing of hypotheses. J. Am. Stat. Assoc. **84**, 602–607 (1989)

The Block-Matrix Sphericity Test: Exact and Near-Exact Distributions for the Test Statistic

Filipe J. Marques, Carlos A. Coelho, and Paula Marques

Abstract

In this work near-exact distributions for the likelihood ratio test (l.r.t.) statistic to test the one sample block-matrix sphericity hypothesis are developed under the assumption of multivariate normality. Using a decomposition of the null hypothesis in two null hypotheses, one for testing the independence of the k groups of variables and the other one for testing the equality of the k block diagonal matrices of the covariance matrix, we are able to derive the expressions of the l.r.t. statistic, its h-th null moment, and the characteristic function (c.f.) of its negative logarithm. The decomposition of the null hypothesis induces a factorization on the c.f. of the negative logarithm of the l.r.t. statistic that enables us to obtain near-exact distributions for the l.r.t. statistic. Numerical studies using a measure based on the exact and approximating c.f.'s are developed. This measure is an upper bound on the distance between the exact and approximating distribution functions, and it is used to assess the performance of the near-exact distributions and to compare these with the Box type asymptotic approximation developed by Chao and Gupta (Commun. Stat. Theory Methods 20:1957–1969, 1991).

F.J. Marques (✉) · C.A. Coelho
Faculdade de Ciências e Tecnologia, Departamento de Matemática and Centro de Matemática e Aplicações, Universidade Nova de Lisboa, Quinta da Torre 2829-516, Caparica, Portugal
e-mail: fjm@fct.unl.pt; cmac@fct.unl.pt

P. Marques
Instituto Superior Dom Dinis, Av. 1.º Maio, 2430, Marinha Grande, Portugal
e-mail: paulafartaria@gmail.com

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_18, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

The one sample block-matrix sphericity test is of great interest when we wish to test, under multivariate normality, if in a sequence of p random variables (r.v.'s) $X_1, \ldots, X_p$ we have k independent groups of p^* variables and if all of the k covariance matrices are equal. We show that we can split the null hypothesis of the block-matrix test in two null hypotheses, one for testing the independence among the k groups of variables and the other for testing the equality of the k covariance matrices. The exact distribution of the likelihood ratio test (l.r.t.) statistic has a very complicated expression which makes its use very difficult in practice. Therefore the development of easy to use and yet highly accurate approximations becomes a good target. Our aim is to show that, based on a the decomposition of the null hypothesis of the one sample block-matrix sphericity, we are able to derive the expressions of the l.r.t statistic and also of its h-th null moment, and the characteristic function (c.f.) of its negative logarithm. The factorization induced on the c.f. of the logarithm of the l.r.t. statistic, by the decomposition of the null hypothesis, together with the results in [7] and [6] will allow us to develop very accurate near-exact distributions for the l.r.t statistic (see the foundations of this methodology developed Coelho and Marques in [8]). In [4] the exact null distribution of the l.r.t. statistic when $k = 2$ is obtained using the inverse Mellin transform and the Meijer G-function what renders the quantile computations too hard even for small values of p^*, reinforcing the need for good manageable approximations. In [3] the authors present an asymptotic approximation based on Box's method (see [2]) which we will use to compare with the new approximations proposed.

2 The Decomposition of the Test Null Hypothesis

Let us consider a sample of size N taken from a p-variate normal population $N_p(\underline{\mu}, \Sigma)$. We intend to test the following null hypothesis:

$$H_0 : \Sigma = \begin{pmatrix} \Delta & 0 & \ldots & 0 \\ 0 & \Delta & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Delta \end{pmatrix} \left(= I_k \otimes \Delta \right), \quad (\Delta \text{ not specified}) \tag{1}$$

where Δ is of order p^*, with $p = kp^*$.

The null hypothesis in (1) may be decomposed in two null hypotheses, more precisely

$$H_0 = H_{0b|0a} \circ H_{0a} \tag{2}$$

where, for

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1k} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{k1} & \Sigma_{k2} & \dots & \Sigma_{kk} \end{pmatrix}, \tag{3}$$

we have

$$H_{0a} : \Sigma = bdiag(\Sigma_{11}, \Sigma_{22}, \dots, \Sigma_{kk}), \tag{4}$$

the null hypothesis to test the independence among the k groups of p^* variables and

$$H_{0b|0a} : \Sigma_{11} = \Sigma_{22} = \cdots = \Sigma_{kk} (= \Delta) \ , \ (\Delta \text{ not specified}) \quad \text{assuming } H_{0a} \text{ true} \tag{5}$$

the null hypothesis to test the equality of the k covariance matrices of order p^* .

3 The l.r.t. Statistic, λ^*, and the h-th Null Moment of λ^*

The expressions of the l.r.t statistics, λ_a^* and $\lambda_{b|a}^*$, to test the null hypotheses in (4) and (5) respectively are given by (see Chaps. 9 and 10 in [1])

$$\lambda_a^* = \frac{|A|^{n/2}}{\prod_{j=1}^{k} |A_{jj}|^{n/2}} \quad \text{and} \quad \lambda_{b|a}^* = \frac{(kn)^{knp^*/2}}{\prod_{j=1}^{k} n^{p^* n/2}} \frac{\prod_{j=1}^{k} |A_{jj}|^{n/2}}{|A^*|^{nk/2}} \tag{6}$$

where $n = N - 1$, $A = \sum_{i=1}^{N} \left(X_i - \overline{X}\right) \left(X_i - \overline{X}\right)'$, A_{jj} is the j-th diagonal matrix of order p^* of A and $A^* = A_{11} + \cdots + A_{kk}$. Using the decomposition in (2) we may obtain the expression for the l.r.t. statistic, λ^*, to test H_0 in (1) as the product of the expressions of the l.r.t. statistics in (6) (see Lemma 10.3.1 in [1])

$$\lambda^* = \lambda_a^* \lambda_{b|a}^* = \frac{|A|^{n/2}}{\left| \frac{1}{k} \sum_{j=1}^{k} A_{jj} \right|^{nk/2}} . \tag{7}$$

Given the independence of the l.r.t. statistics, λ_a^* and $\lambda_{b|a}^*$, in (6), under H_0 in (1), the expression of the h-th null moment of λ^* may be obtained as the product of the expressions of the h-th null moments of λ_a^* and $\lambda_{b|a}^*$ (see Chaps. 9 and 10 in [1]), therefore

$$E\left[(\lambda^*)^h\right] = E\left[\left(\lambda^*_{b|a}\lambda^*_a\right)^h\right] = E\left[\left(\lambda^*_{b|a}\right)^h\right] \times E\left[(\lambda^*_a)^h\right]$$

$$= \frac{(nk)^{nkp^*h/2}}{\prod_{j=1}^{k} n^{p^*nh/2}} \frac{\Gamma_{p^*}\left(\frac{nk}{2}\right)}{\Gamma_{p^*}\left(\frac{nk}{2}(1+h)\right)} \prod_{j=1}^{k} \frac{\Gamma_{p^*}\left(\frac{n}{2}(1+h)\right)}{\Gamma_{p^*}\left(\frac{n}{2}\right)} \tag{8}$$

$$\times \frac{\Gamma_{p^*k}(\frac{1}{2}n + \frac{1}{2}hn)}{\Gamma_{p^*k}(\frac{1}{2}n)} \prod_{i=1}^{k} \frac{\Gamma_{p^*}\left(\frac{n}{2}\right)}{\Gamma_{p^*}\left(\frac{n}{2}(1+h)\right)} \tag{9}$$

$$= k^{\frac{1}{2}p^*knh} \prod_{j=1}^{p^*k} \frac{\Gamma\left(\frac{1}{2}(n+nh-j+1)\right)}{\Gamma\left(\frac{1}{2}(n-j+1)\right)} \prod_{i=1}^{p^*} \frac{\Gamma\left(\frac{1}{2}(nk-i+1)\right)}{\Gamma\left(\frac{1}{2}(nk+nkh-i+1)\right)},$$

where $\Gamma_{p^*}(.)$ represents the multivariate p^* gamma function (see [1]).

4 The c.f. of $W = -\log \lambda^*$

Noticing that $E\left(\mathrm{e}^{\mathrm{i}tW}\right) = E\left((\lambda^*)^{-\mathrm{i}t}\right)$, the expression of the c.f. of the r.v. $W = -\log\lambda^*$ may be derived from the expression of the h-th null moment of λ^* as

$$\Phi_W(t) = k^{-\frac{1}{2}p^*kn\mathrm{i}t} \prod_{j=1}^{p^*k} \frac{\Gamma\left(\frac{1}{2}(n-n\mathrm{i}t-j+1)\right)}{\Gamma\left(\frac{1}{2}(n-j+1)\right)} \prod_{j=1}^{p^*} \frac{\Gamma\left(\frac{1}{2}(nk-j+1)\right)}{\Gamma\left(\frac{1}{2}(nk-nk\mathrm{i}t-j+1)\right)}.$$

However, if we consider the decomposition in (8) and (9) we may rewrite the c.f. of W as the product of the c.f's of $W_1 = -\log\lambda^*_a$ and $W_2 = -\log\lambda^*_{b|a}$

$$\Phi_W(t) = \underbrace{\frac{\Gamma_{p^*k}(\frac{1}{2}n - \frac{1}{2}\mathrm{i}tn)}{\Gamma_{p^*k}(\frac{1}{2}n)} \prod_{i=1}^{k} \frac{\Gamma_{p^*}\left(\frac{n}{2}\right)}{\Gamma_{p^*}\left(\frac{n}{2}(1-\mathrm{i}t)\right)}}_{\Phi_{W_1}(t)}$$

$$\times \underbrace{\frac{(nk)^{-nkp^*\mathrm{i}t/2}}{\prod_{j=1}^{k} n^{-p^*n\mathrm{i}t/2}} \frac{\Gamma_{p^*}\left(\frac{nk}{2}\right)}{\Gamma_{p^*}\left(\frac{nk}{2}(1-\mathrm{i}t)\right)} \prod_{j=1}^{k} \frac{\Gamma_{p^*}\left(\frac{n}{2}(1-\mathrm{i}t)\right)}{\Gamma_{p^*}\left(\frac{n}{2}\right)}}_{\Phi_{W_2}(t)},$$

which is a much more useful form for the development of near-exact distributions.

4.1 Factorizations of the c.f.'s of $W_1 = -\log \lambda_a^*$ and $W_2 = -\log \lambda_{b|a}^*$

With the final goal of developing near-exact distributions for the l.r.t. statistic, λ^*, in (7) (see Sect. 5 for details) we will use factorizations of the c.f.'s of $W_1 = -\log \lambda_a^*$ and $W_2 = -\log \lambda_{b|a}^*$. These factorizations, already obtained in [6] and [7], show that the exact distribution of both W_1 and W_2 may be represented in the form of the sum of two independent r.v.'s, one with a Generalized Integer Gamma (GIG) distribution (see [5]) and the other with the distribution of the sum of independent Logbeta r.v.'s, eventually multiplied by a constant. These similarities in the structure of the c.f.'s of W_1 and W_2 will be of great use to achieve our goal.

4.1.1 The c.f. of $W_1 = -\log \lambda_a^*$

Coelho in [6] shows a possible factorization for the c.f. of $W_1 = -\log \lambda_a^*$, in the form

$$\Phi_{W_1}(t) = \underbrace{\prod_{j=2}^{p-1} \left(\frac{n-j}{n}\right)^{r_j^*} \left(\frac{n-j}{n} - \mathrm{i}t\right)^{-r_j^*}}_{\Phi_{1,1}(t)} \underbrace{\left\{\frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n}{2} - \frac{1}{2} - \frac{n}{2}\mathrm{i}t\right)}{\Gamma\left(\frac{n}{2} - \frac{n}{2}\mathrm{i}t\right)\Gamma\left(\frac{n}{2} - \frac{1}{2}\right)}\right\}^{m^*}}_{\Phi_{1,2}(t)} \quad (10)$$

with $m^* = k$ if p^* is odd and $m^* = 0$ if p^* is even; the parameters r_j^* are the parameters r_j given by expressions (A.3) and (A.4) in [9]. The c.f. $\Phi_{1,1}(t)$ is the c.f. of the sum of $p-2$ r.v.'s with Gamma distribution, with integer shape parameters, r_j^*, and with rate parameters $\frac{n-j}{n}$ $(j = 2, \ldots, p-1)$, that is, the c.f. of a GIG distribution with depth $p-2$. The c.f. $\Phi_{1,2}(t)$ is the c.f. of the sum of m^* independent Logbeta distributed r.v.'s, multiplied by $\frac{n}{2}$, with parameters $\frac{n}{2} - \frac{1}{2}$ and $\frac{n}{2}$.

4.1.2 The c.f. of $W_2 = -\log \lambda_{b|a}^*$

Coelho and Marques in [7] derive the following factorization for the c.f. of $\Phi_{W_2}(t)$

$$\begin{aligned}\Phi_{W_2}(t) = &\underbrace{\prod_{j=1}^{p-1} \left(\frac{n-j}{n}\right)^{r_j} \left(\frac{n-j}{n} - \mathrm{i}t\right)^{-r_j}}_{\Phi_{2,1}(t)} \\ &\times \prod_{j=1}^{\lfloor p/2 \rfloor} \prod_{k=1}^{q} \frac{\Gamma(a_j + b_{jk})}{\Gamma(a_j + b_{jk}^*)} \frac{\Gamma(a_j + b_{jk}^* - n\mathrm{i}t)}{\Gamma(a_j + b_{jk} - n\mathrm{i}t)} \\ &\underbrace{\times \left(\prod_{k=1}^{q} \frac{\Gamma(a_p + b_{pk})}{\Gamma(a_p + b_{pk}^*)} \frac{\Gamma(a_p + b_{pk}^* - \frac{n}{2}\mathrm{i}t)}{\Gamma\left(a_p + b_{pk} - \frac{n}{2}\mathrm{i}t\right)}\right)^{p \perp\!\!\!\perp 2}}_{\Phi_{2,2}(t)}\end{aligned} \quad (11)$$

where a_j, b_{jk}, b^*_{jk}, a_p, b_{pk}, and b_{pk^*} are given in (3) and (4) in [7], r_j are given in expressions (5)–(7) also in [7] and where

$$p \perp\!\!\!\perp 2 = \left\lfloor \frac{p+1}{2} \right\rfloor - \left\lfloor \frac{p}{2} \right\rfloor = \begin{cases} 0\,, & \text{for } p \text{ even} \\ 1\,, & \text{for } p \text{ odd.} \end{cases}$$

The c.f. $\Phi_{2,1}(t)$ is the c.f. of the sum of $p-1$ r.v.'s with Gamma distribution, with integer shape parameters, r_j, and with rate parameters $\frac{n-j}{n}$ $(j = 1, \ldots, p-1)$, that is, the c.f. of a GIG distribution with depth $p-1$. The c.f. $\Phi_{2,2}(t)$ is the c.f. of the sum of $\lfloor p/2 \rfloor \times q + q \times p \perp\!\!\!\perp 2$ independent Logbeta distributed r.v.'s, the first $\lfloor p/2 \rfloor \times q$ multiplied by n and the remaining by $\frac{n}{2}$.

5 Near-Exact Distributions for W and λ^*

Using the similarities observed on the factorizations of the c.f.'s of W_1 and W_2 given in Sect. 4.1, we may rewrite the c.f. of $W = -\log \lambda^*$.

Theorem 5.1. *The c.f. of $W = -\log \lambda^*$ may be represented in the form*

$$\Phi_W(t) = \underbrace{\prod_{j=1}^{p-1} \left(\frac{n-j}{n}\right)^{\upsilon_j} \left(\frac{n-j}{n} - \mathrm{i}t\right)^{-\upsilon_j}}_{\Phi_{W_1^*}(t)} \underbrace{\Phi_{1,2}(t) \times \Phi_{2,2}(t)}_{\Phi_{W_2^*}(t)} \tag{12}$$

with $\Phi_{1,2}(t)$ and $\Phi_{2,2}(t)$ respectively in (10) and (11),

$$\upsilon_j = \begin{cases} r_j & j = 1 \\ r_j + r_j^* & j = 2, \ldots, p-1\,, \end{cases} \tag{13}$$

and with r_j^* equal to the parameters r_j in (A.3) and (A.4) of [9] and r_j given by expressions (5)–(7) in [7].

The near-exact c.f.'s will thus have the form

$$\underbrace{\Phi_{W_1^*}(t)}_{\text{GIG distribution}} \times \Phi_{ne}(t) \tag{14}$$

where $\Phi_{W_1^*}(t)$ is given by (12) and $\Phi_{ne}(t)$ is the c.f. that we will use to approximate the c.f. $\Phi_{W_2^*}(t)$ in (12). Since a Logbeta distribution is indeed an infinite mixture of Gamma distributions, we propose $\Phi_{ne}(t)$ to be the c.f. of a single Gamma r.v. or a mixture of two or three Gamma r.v.'s. The parameters in $\Phi_{ne}(t)$ are evaluated ensuring that $\Phi_{W_2^*}(t)$ and $\Phi_{ne}(t)$ have the same first two, four, or six derivatives at $t = 0$; that is the same as to ensure that the exact and approximating distributions have the same first two, four, or six moments. Thus, we will have

$$\Phi_{ne}(t) = \sum_{\ell=1}^{h/2} \theta_\ell \, \mu^{\delta_\ell} (\mu - \mathrm{i}t)^{-\delta_\ell} \tag{15}$$

with $\theta_\ell > 0$ ($\ell = 1, \ldots, h/2$ for $h = 2,\ 4$ or 6), $\sum_{\ell=1}^{h/2} \theta_\ell = 1$, and

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} \Phi_{W_2^*}(t)\bigg|_{t=0} = \frac{\mathrm{d}^j}{\mathrm{d}t^j} \Phi_{ne}(t)\bigg|_{t=0}, \quad j = 1, \ldots, h \tag{16}$$

for $h = 2, 4$, or 6, depending on $\Phi_{ne}(t)$ being the c.f. of a single r.v. or a mixture of two or three Gamma r.v.'s with the same rate parameters. Using this approach we obtain, for $h = 2$, as near-exact distribution for W a single Generalized Near-Integer Gamma ($GNIG$) distribution (see [6]) or, for $h = 4$ or 6, a mixture of two or three $GNIG$ distributions. By simple transformation it is easy to obtain near-exact distributions for λ^*.

Theorem 5.2. *The near-exact distributions for λ^* are either an exponential GNIG distribution or a mixture of 2 or 3 exponential GNIG distributions of depth p respectively for $h = 2, 4$ or 6, with p.d.f. given by (using the notation of Appendix B in [9])*

$$\sum_{\ell=1}^{h/2} \theta_\ell \, f^{\,GNIG}\Big(-\log w | \upsilon_1, \ldots, \upsilon_{p-1}, \delta_\ell \,;\, \frac{n-1}{n}, \ldots, \frac{n-p+1}{n}, \mu \,;\, p\Big) \frac{1}{w}$$

and c.d.f given by

$$1 - \sum_{\ell=1}^{h/2} \theta_\ell \, F^{\,GNIG}\Big(-\log w | \upsilon_1, \ldots, \upsilon_{p-1}, \delta_\ell \,;\, \frac{n-1}{n}, \ldots, \frac{n-p+1}{n}, \mu \,;\, p\Big)$$

with $\upsilon_1, \ldots, \upsilon_{p-1}$ as in (13), and where, for $h = 2, 4$ or 6, the parameters θ_ℓ, δ_ℓ and μ are obtained as the numerical solution of the system of equations in (16), with $\theta_{h/2} = 1 - \sum_{\ell=1}^{h/2-1} \theta_\ell$.

6 Numerical Studies

To assess the accuracy of the near-exact distributions we will use the measure

$$\frac{1}{2\pi} \int_{-\infty}^{+\infty} \left| \frac{\Phi_W(t) - \Phi_{\mathrm{app}}(t)}{t} \right| \mathrm{d}t \,, \tag{17}$$

where $\Phi_W(t)$ and $\Phi_{\mathrm{app}}(t)$ represent respectively the exact and the approximate c.f. of the r.v. $W = -\log \lambda^*$. For further details on this measure see [7]. We will denote

Table 1 Values for the measure in (17) of the approximating distributions of $W = -\log \lambda^*$ for $p^* = 3$ and increasing values of p, k and n

p	p^*	k	n	GNIG	M2GNIG	M3GNIG	Box
9	3	3	11	4.5×10^{-6}	3.6×10^{-9}	4.0×10^{-12}	5.8×10^{-2}
12	3	4	14	1.2×10^{-6}	4.3×10^{-10}	2.9×10^{-13}	1.2×10^{-1}
15	3	5	17	4.3×10^{-7}	8.2×10^{-11}	3.8×10^{-14}	2.1×10^{-1}
21	3	7	23	8.3×10^{-8}	5.1×10^{-12}	9.0×10^{-16}	4.1×10^{-1}
27	3	9	29	2.5×10^{-8}	6.3×10^{-13}	5.2×10^{-17}	6.4×10^{-1}

Table 2 Values for the measure in (17) of the approximating distributions of $W = -\log \lambda^*$ for $p = 8$, $p^* = 4$, $k = 2$ and increasing values of n

p	p^*	k	n	GNIG	M2GNIG	M3GNIG	Box
8	4	2	10	3.1×10^{-6}	2.9×10^{-10}	9.1×10^{-13}	3.3×10^{-2}
8	4	2	50	2.1×10^{-7}	6.9×10^{-13}	1.9×10^{-14}	3.2×10^{-5}
8	4	2	100	5.5×10^{-8}	4.3×10^{-14}	3.2×10^{-18}	3.3×10^{-6}

Table 3 Values for the measure in (17) of the approximating distributions of $W = -\log \lambda^*$ for $p = 9$, $p^* = 3$, $k = 3$ and increasing values of n

p	p^*	k	n	GNIG	M2GNIG	M3GNIG	Box
9	3	3	11	4.5×10^{-6}	3.6×10^{-9}	4.0×10^{-12}	5.8×10^{-2}
9	3	3	50	4.6×10^{-7}	2.7×10^{-11}	1.2×10^{-14}	9.0×10^{-5}
9	3	3	100	1.2×10^{-7}	2.0×10^{-12}	1.5×10^{-15}	9.2×10^{-6}

by $GNIG$, $M2GNIG$ and $M3GNIG$ the near-exact distributions corresponding, respectively, to a GNIG or to a mixture of 2 or 3 GNIG distributions and by Box the asymptotic approximation in [3].

From Tables 1–3 we may observe the very good asymptotic properties of the near-exact distributions for increasing values of n and p. These distributions display very good results for the measure in (17) specially when compared with the ones obtained for the asymptotic approximation denoted by Box.

We may point out the excellent performance of the $M3GNIG$ near-exact distribution and the remarkable behavior of all near-exact distributions for small sample sizes.

All values of the measure in (17) were computed using the numerical integration functions of Mathematica, version 7, and were calculated with at least eight digits of precision, out of which only two are shown.

7 Conclusions

We have shown that, based on a decomposition of the null hypothesis of the one sample block-matrix sphericity, we may derive the expressions of the l.r.t statistic, its h-th null moment, and also of the c.f. of its logarithm. This decomposition induces a factorization on the c.f. of the l.r.t. statistic which together with the results obtained

in [6] and [7] allow us to develop very accurate near-exact distributions for the l.r.t statistic and for the logarithm of the l.r.t. statistic. Numerical studies show that the near-exact distributions are very accurate approximations exhibiting at the same time very good asymptotic properties.

Acknowledgements This research was partially financially supported by the Portuguese Foundation for Science and Technology, through Centro de Matemática e Aplicações of Universidade Nova de Lisboa (CMA/FCT/UNL), under the project PEst-OE/MAT/UI0297/2011.

References

1. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, New York (2003)
2. Box, G.E.P.: A general distribution theory for a class of likelihood criteria. Biometrika **36**, 317–346 (1949)
3. Chao, C.C., Gupta, A.K.: Testing of homogeneity of diagonal blocks with blockwise independence. Commun. Stat. Theor Methods **20**, 1957–1969 (1991)
4. Cardeño, L., Nagar, D.K.: Testing block sphericity of a covariance matrix. Divulgaciones Matemáticas **9**, 25–34 (2001)
5. Coelho, C.A.: The generalized integer gamma distribution – a basis for distributions in multivariate statistics. J. Multivar. Anal. **64**, 86–102 (1998)
6. Coelho, C.A.: The generalized near-integer gamma distribution: a basis for 'near-exact' approximations to the distributions of statistics which are the product of an odd number of independent Beta random variables. J. Multivar. Anal. **89**, 191–218 (2004)
7. Coelho, C.A., Marques F.J.: Near-exact distributions for the likelihood ratio test statistic to test equality of several variance-covariance matrices in elliptically contoured distributions. Comput. Stat., article in press, (2012) doi: 10.1007/s00180-011-0281-1
8. Coelho, C.A., Marques, F.J.: The advantage of decomposing elaborate hypotheses on covariance matrices into conditionally independent hypotheses in building near-exact distributions for the test statistics. Lin. Algebra Appl. **430**, 2592–2606 (2009)
9. Marques, F.J., Coelho, C.A.: A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Statistics. Test **20**, 180–203 (2011)

Extensions of Dorfman's Theory

Rui Santos, Dinis Pestana, and João Paulo Martins

Abstract

Economic impact of composite sampling is investigated in the realistic framework of tests with positive probability of false positive and of false negative results. Sensitivity and specificity when pooling samples are also discussed, using rarefaction as a framework.

1 Introduction

During the Second World War, Dorfman [2] used composite sampling in order to get a more efficient method of identifying syphilis infected soldiers: testing a group of n members is feasible and saves resources. A negative result of a compound test means that no one is infected and a positive result means that at least one of the members is infected. In this last case, individual tests to all the group soldiers would be mandatory, in order to identify all infected members. As the main goal is to minimize the number of expected tests to identify all the infected members, the main issue is to find the optimal group dimension n. The compound analyses allow us to save resources in a variety of problems aside from blood testing, cf. [1]. Besides, the optimal n depends on the researcher's main goal: to identify all the infected individuals (classification problem) or to estimate the prevalence rate (estimation problem).

R. Santos (✉) · J.P. Martins
School of Technology and Management, Polytechnic Institute of Leiria, CEAUL — Center of Statistics and Applications of University of Lisbon, Portugal
e-mail: rui.santos@ipleiria.pt; jpmartins@estg.ipleiria.pt

D. Pestana
Faculty of Sciences of Lisbon, University of Lisbon, CEAUL — Center of Statistics and Applications of University of Lisbon, Portugal
e-mail: dinis.pestana@fc.ul.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_19, © Springer-Verlag Berlin Heidelberg 2013

Dorfman's original idea was applied to qualitative analyses (presence or absence of the infection), without measuring any quantitative variable (antigenes or antibodies or bacteria counts, or proportion of specific cells, or weight or volume of some chemical compound). In his methodology, it was implicitly assumed that the blood tests had no errors. In the present work, we assume that each test may return a false negative or a false positive, i.e., the conditional probability of returning the true diagnostic in the presence or absence of the disease, known respectively as sensitivity and specificity, are essential operating characteristics of the test. Some previous works, concerned mainly with prevalence rate estimation, discuss misclassification issues. Nevertheless, most of these studies assume that pooling does not affect misclassification (cf. [8, 9, 12, 13]). Others studies, such as [7], provide simple models for modeling sensitivity without taking into account the number of infected members in the group. However, an infected member in the pooled sample can be excessively diluted and become undetectable in the compound test. Furthermore, the sensitivity and specificity of a compound test must depend of the number of infected individual on the group—the dilution effect, cf. [6]. Hierarchical models to capture the dilution effect in prevalence estimation were used in [14, 16], but they do not measure the probability of misclassification. Thus, the main goal of this work is to analyze the influence of dilution and rarefaction on sensitivity and specificity in the use of compound tests and Dorfman's classification methodology. Moreover, for low prevalence rates, it is shown that the dilution effect for just one (or two) infected element in the group is sufficient to evaluate misclassification in the use of pooled samples.

2 Dorfman's Theory

Let p denote the prevalence rate of the infection and the independent Bernoulli random variables X_i, with $i = 1, \ldots, N$, represent the presence $(X_i = 1)$ or absence $(X_i = 0)$ of the infection in the i-th population unit. A compound test on a sample of size n has negative result if none of those units is infected, i.e., $\sum_{i=1}^{n} X_i = 0$, and it has positive result if at least one element is infected, i.e., $\sum_{i=1}^{n} X_i \geq 1$. In this case individual tests have to be carried out to identify all infected members. In this chapter we follow Dorfman's procedure, although our methods can be applied to others methodologies, such as the ones proposed in [3,4,8,10] or [11]. Hence, the expected number of tests needed to identify infected units in the population using groups of n members is $\mathbb{E}[T] = N\left(\frac{1}{n} + 1 - (1-p)^n\right)$ (for simplicity, assume that $\frac{N}{n} \in \mathbb{N}$). Thus a simple quantification of the efficiency of compound versus individual tests is the relative cost $\mathbf{RC} = \frac{\mathbb{E}[T]}{N} = \frac{n+1}{n} - (1-p)^n$, $n \geq 2$, which can be used to find the most efficient value for n knowing the infection prevalence p. Dorfman [2] analyzed different prevalence rates and concluded that the compound test is more efficient if p is lower than approximately 0.3066 and the efficient value for n (represented by n^*) will decrease with p, without attaining the value $n = 2$, as described in Table 1. Observe that for $0.123 \leq p \leq 0.307$ we have $n^* = 3$

Table 1 Relative cost and n optimum for some prevalence rates p

p	0.31	0.30	0.13	0.12	0.07	0.05	0.025	0.01	0.005	10^{-3}	10^{-4}	10^{-5}	10^{-6}
n^*	1	3	3	4	4	5	7	11	15	32	100	317	1000
RC	1	0.990	0.675	0.650	0.502	0.426	0.305	0.196	0.139	0.063	0.020	0.006	0.002

Table 2 Joint probabilities of an individual test

	Individual test result		
Truth	X_i^+	X_i^-	
$X_i = 1$	$\varphi_s p$	$(1-\varphi_s)p$	p
$X_i = 0$	$(1-\varphi_e)q$	$\varphi_e q$	q
	$\varphi_s p + (1-\varphi_e)q$	$(1-\varphi_s)p + \varphi_e q$	1

and for $p > 0.307$ the compound test is not efficient ($n^* = 1$). For $p < 0.123$ the approximation $(1-p)^n \approx 1 - np$ (cf. [15]) provides the approximate optimal solution $n^* \approx p^{-0.5}$, whose maximum error is 1, with no significant influence on the relative cost. Thus $\mathbf{RC} \approx \frac{1}{n} + np$ and, for $n = n^*$, $\mathbb{E}[T] \approx 2N\sqrt{p}$. Finucan [3] proposes the smallest integer not less than $p^{-0.5} + 0.5$ for n^*. In all examples we will use the most efficient value for n, i.e. $n = n^*$.

3 The Inclusion of Errors in the Tests Results

Now consider that the tests results are subject to some sources of error. We shall denote $\varphi_s \in (0, 1]$ the sensitivity of one single test—probability of one positive test (X_i^+) in one infected sample ($X_i = 1$), i.e., $\mathbb{P}\left(X_i^+ | X_i = 1\right)$—and $\varphi_e \in (0, 1]$ the specificity of the test—probability of getting a negative result (X_i^-) from a not infected sample ($X_i = 0$), i.e., $\mathbb{P}\left(X_i^- | X_i = 0\right)$. Then, $1-\varphi_s$ will be the probability of one false negative test and $1-\varphi_e$ the probability of one false positive. Assuming that the results of the test and the condition of the tested units are independent, denoting $q = 1 - p$, we obtain the probabilities shown in Table 2.

3.1 Sensitivity and Specificity in Compound Tests

Let $I^{[n]} = \sum_{i=1}^{n} X_i$ represent the number of infected elements in a sample of size n and $I^{[i,n]} = \mathbb{P}(I^{[n]} = i) = \binom{n}{i} p^i q^{n-i}, i = 0, \ldots, n$. Let $X^{[+,n]}$ [resp. $X^{[-,n]}$] represent a positive [resp. negative] result on the compound test. The purpose of this section is to compare the sensitivity φ_s and the specificity φ_e of one simple test ($n = 1$) with the pooled sensitivity $\varphi_s^{[n]} = \mathbb{P}(X^{[+,n]} | I^{[n]} \geq 1)$ and pooled specificity $\varphi_e^{[n]} = \mathbb{P}(X^{[-,n]} | I^{[n]} = 0)$ of one compound test in a sample of size n.

Observe that the value of n does not affect the specificity; in fact, if we dissolve one cm^3 of blood from n not infected elements and then we choose, at random, one cm^3 from this mixture to test, then this sample of blood will not be infected, and

this framework is obviously equivalent to using just one cm^3 of blood from one non infected person. So $\varphi_e^{[n]} = \varphi_e$.

On the other hand, the sensitivity of the compound test $\varphi_s^{[n]}$ is a function of the number m of infected members in the group. Denote $\varphi_s^{[m,n]} = \mathbb{P}(X^{[+,n]}|I^{[n]} = m)$, and assume that $\varphi_s^{[n,n]} = \varphi_s$, i.e., that the sensitivity of the compound test when all members are infected is equal to the sensitivity of one simple test. Under these assumptions we shall get $0 < \varphi_s^{[1,n]} \leq \varphi_s^{[2,n]} \leq \cdots \leq \varphi_s^{[n,n]} = \varphi_s$, as a consequence of the dilution of the fluid and its rarefaction, and

$$\varphi_s^{[n]} = \frac{\sum_{j=1}^{n} \mathbb{P}\left(X^{[+,n]}|I^{[n]} = j\right) \mathbb{P}\left(I^{[n]} = j\right)}{\mathbb{P}\left(I^{[n]} \geq 1\right)} = \sum_{j=1}^{n} \varphi_s^{[j,n]} \frac{I^{[j,n]}}{1-q^n} = \sum_{j=1}^{n} \varphi_s^{[j,n]} \lambda_j, \tag{1}$$

where $\sum_{j=1}^{n} \lambda_j = 1$ and therefore $\varphi_s^{[n]}$ is a weighted mean of the sensitivities $\varphi_s^{[j,n]}$; so $\varphi_s^{[n]} \leq \varphi_s$ as a consequence of $\varphi_s^{[m,n]} \leq \varphi_s$, for every m. Hung and Swallow [6] applied a similar approach in order to analyze the robustness of prevalence estimation under misclassification concerned with the dilution effect. Nevertheless, they considered $\varphi_s = \varphi_e = 1$ (misclassification is only due the dilution effect) and proposed two different specific forms to $\varphi_s^{[j,n]}$ without providing general formulas for sensitivity and specificity in compound tests.

In fact, the expected number of infected members in the n elements, supposing $p \approx 0$ (thus $n^* \approx \frac{1}{\sqrt{p}}$), is $\frac{1}{n}$ and therefore $\lambda_i > \lambda_j$ to $i < j$. Moreover, the first weight λ_1 decreases with p but the remaining λ_i increase with p. For $p = 0.1$, $\lambda_1 \approx 0.85$ and $\lambda_2 \approx 0.14$; for $p = 0.01$ $\lambda_1 \approx 0.951$ and $\lambda_2 \approx 0.048$. In general, the $\varphi_s^{[j,n]}$ for $j > 2$ can be disregarded, and in most cases $\varphi_s^{[2,n]}$ can also be discarded (for $p = 0.001$, for instance, $\lambda_1 \approx 0.985$). So, the value $\varphi_s^{[1,n]}$ of the sensitivity of one compound test when just one of the n members is infected has a large impact on the total sensitivity of the compound test (without requiring any assumptions concerning the dilution effect). If the sensitivity $\varphi_s^{[1,n]}$ is low (compared with φ_s), we do not recommend using the compound test.

In order to compare $\varphi_s^{[m,n]}$ with φ_s we can use the weights $k_m^{[n]} = 1 - \varphi_s^{[m,n]} \varphi_s^{-1}$ where $1 > k_1^{[n]} \geq k_2^{[n]} \geq \cdots \geq k_n^{[n]} = 0$. Thus $\varphi_s^{[n]} = \varphi_s - \sum_{j=1}^{n} \lambda_j k_j^{[n]} \varphi_s$, and the difference between the sensitivity φ_s and the pooled sensitivity $\varphi_s^{[n]}$ is given by $\varphi_s - \varphi_s^{[n]} = \sum_{j=1}^{n} \lambda_j k_j^{[n]} \varphi_s$. So, when $p \approx 0$, we have $\varphi_s - \varphi_s^{[n]} \approx (\lambda_1 k_1^{[n]} + \lambda_2 k_2^{[n]}) \varphi_s$. We illustrate the use of those weights in Sect. 4.1 using the Poisson distribution.

In a compound test with n elements, the probability of each outcome is given in Table 3.

3.2 Sensitivity and Specificity in Dorfman's Methodology

Using Dorfman's methodology, for one infected person being identified by the test, the compound test must be positive (this happens with probability $\varphi_s^{[n]}$ which depends on $I^{[n]}$) and then, in the simple test, we also need to get a positive result

Table 3 Joint probabilities of a compound test

	Compound test result		
Truth	$X^{[+,n]}$	$X^{[-,n]}$	
$I^{[n]} \geq 1$	$\varphi_s^{[n]}\,(1-q^n)$	$\left(1-\varphi_s^{[n]}\right)(1-q^n)$	$1-q^n$
$I^{[n]} = 0$	$(1-\varphi_e)q^n$	$\varphi_e\, q^n$	q^n
	$\varphi_s^{[n]} - q^n\left(\varphi_s^{[n]} + \varphi_e - 1\right)$	$1-\varphi_s^{[n]} + q^n\left(\varphi_s^{[n]} + \varphi_e - 1\right)$	1

(with probability φ_s). Therefore, supposing that the tests results are independent, the sensitivity φ_{s_n} of the process will be given by

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

and $\varphi_{s_n} \leq \varphi_s^2 \leq \varphi_s$, i.e., the sensitivity in Dorfman's methodology is smaller than the sensitivity of one simple test. There are two different possibilities for one not infected person being identified by the test: either the compound test is negative or the compound test is positive and the simple test is negative. So the specificity φ_{e_n} of the process will be given by

$$\begin{aligned}
\varphi_{e_n} &= \mathbb{P}\left(X_1^- | X_1 = 0\right) = \sum_{i=0}^{n-1} \mathbb{P}\left(X_1^- | X_1 = 0, I^{[n-1]} = i\right) \mathbb{P}\left(I^{[n-1]} = i\right) \\
&= \sum_{i=0}^{n-1} \left[\mathbb{P}\left(X_1^- | X_1 = 0\right) \mathbb{P}\left(X^{[+,n]} | I^{[n]} = i\right) + \mathbb{P}\left(X^{[-,n]} | I^{[n]} = i\right)\right] I^{[i,n-1]} \\
&= \left[\varphi_e + \varphi_e(1-\varphi_e)\right] q^{n-1} + \sum_{i=1}^{n-1} \left[\varphi_e \varphi_s^{[i,n]} + \left(1-\varphi_s^{[i,n]}\right)\right] I^{[i,n-1]} \\
&= 1 - (1-\varphi_e)\left[(1-\varphi_e)q^{n-1} + \sum_{i=1}^{n-1} \varphi_s^{[i,n]} I^{[i,n-1]}\right] \\
&= 1 - (1-\varphi_e)\xi \qquad (3)
\end{aligned}$$

where $\xi = \xi(\varphi_s^{[i,n]}, \varphi_e, p, n)$ is a weighted mean of $(1-\varphi_e)$ and $\varphi_s^{[i,n]}$, $i = 1, \ldots, n-1$, therefore $\xi \leq 1$ and consequently $\varphi_{e_n} \geq \varphi_e$ (the specificity in Dorfman's methodology is greater than in a simple test). If we do not include

the rarefaction factor (i.e., supposing $\varphi_s^{[m,n]} = \varphi_s, 1 \leq m \leq n$, cf. [8]), then $\varphi_{s_n} = \varphi_s^2 \leq \varphi_s$, which does not depend on n and corresponds to the maximum value for φ_{s_n} (best situation for φ_{s_n}). In this case the specificity φ_{e_n} is given by

$$\varphi_{e_n} = 1 - (1 - \varphi_e)\varphi_s - (1 - \varphi_e)(1 - \varphi_e - \varphi_s)q^{n-1}, \tag{4}$$

which depends on n, increases with q^{n-1} (for suitable values for φ_s and φ_e, i.e., $\varphi_s + \varphi_e > 1$) and corresponds to the minimum value for φ_{e_n} (worst case for φ_{e_n}). If we use the approximation $n \approx \frac{1}{\sqrt{p}}$, then $q^{n-1} \approx (1-p)^{\frac{1}{\sqrt{p}}-1}$ decreases with $p \in (0, 0.31)$ and, therefore, φ_{e_n} also decreases with p. Nevertheless, if φ_e is close to one, the value of $(1 - \varphi_e)(1 - \varphi_e - \varphi_s)$ will be very close to zero and, therefore, the value of p will not imply great changes in the specificity φ_{e_n}.

Hence, in this simplified situation, the gain of specificity using Dorfman's methodology is generally negligible, but the loss of sensitivity can be large. For instance, for $\varphi_s = 0.9$ we get $\varphi_{s_n} = 0.81$ and the probability of a false negative (a most serious error, as an infected person is not detected) increases 9 % taking into account the rarefaction factor, the loss would be even greater.

4 The Number of Bacteria

Let us consider the problem of detecting the presence of some type of bacteria to test the contamination of some substance, for instance of yogurt at an industrial unit.

First, let us assume that the contamination rate is equal to p. Thus, if we choose n yogurts at random to test, the number of contaminated ones is a random variable $I \frown \mathbf{B}(n, p)$. Let us assume further that the number of bacteria Y_i^* in one ml taken from the i-th yogurt, chosen at random, is zero if it is not contaminated. Otherwise, it is modeled by some random variable Y_i with discrete distribution $\mathbf{D}$ with parameters vector θ and support in $\mathbb{N}_0$, that is $Y_i \frown \mathbf{D}(\theta)$.

So, when pooling together one ml from n yogurts, the number of bacteria present in the pooled sample will be $B_n = \sum_{i=1}^n Y_i^* = \sum_{i=1}^I Y_i$. If the compound test uses one ml of the mixture of yogurts, as in the simple test, i.e., if we choose, at random, one ml from the n ml of the amalgamated sample, assuming that the mixture is homogeneous, the number of bacteria in one ml of this compound is $B_1 \frown \mathbf{B}\left(B_n, \frac{1}{n}\right)$ (we can get the same result by considering $B_1 = \sum_{i=1}^n W_i^* = \sum_{i=1}^I W_i$ where the random variable W_i is described by the hierarchic model $W_i \sim \mathbf{B}\left(T, \frac{1}{n}\right)$ where $T \sim \mathbf{D}(\theta)$). Hierarchical pooling models using continuous distributions were used in [14, 16] to develop procedures for HIV prevalence estimation with the dilution effect.

Table 4 exhibits the results for some commonly used count distributions, namely the main "basic count distributions" with extended Panjer's recursion, cf. [5].

Table 4 Example with some common count distributions

	$\mathbf{D}(\theta)$	$B_1\vert_{I=i}$
Poisson	$Y_i \sim \mathbf{P}(\lambda)$, $\lambda > 0$	$B_1\vert_{I=i} \sim \mathbf{P}\left(i\frac{\lambda}{n}\right)$
Binomial	$Y_i \sim \mathbf{B}(m, p_1)$	$B_1\vert_{I=i} \sim \mathbf{B}\left(im, \frac{p_1}{n}\right)$
Negative binomial	$Y_i \sim \mathbf{NB}(r, p_1)$, $y \in \mathbb{N}_0$ $\mathbb{P}(Y_i = y) = \binom{y+r-1}{r-1} p_1^r (1-p_1)^y$	$B_1\vert_{I=i} \sim \mathbf{NB}\left(ir, \frac{np_1}{np_1+1-p_1}\right)$
Logarithmic	$Y_i \sim \mathbf{Log}(\theta)$, $\theta \in (0,1)$, $\mathbb{P}(Y_i = y) = -\frac{1}{\ln(1-\theta)} \frac{\theta^y}{y}$ $y \in \mathbb{N}$	$f_{W_i}(x) = \begin{cases} \frac{\ln\left(1-\theta+\frac{\theta}{n}\right)}{\ln(1-\theta)} & x = 0 \\ \frac{-1}{x\ln(1-\theta)} \left(\frac{\frac{\theta}{n}}{1-\theta\left(1-\frac{1}{n}\right)}\right)^x & x \in \mathbb{N} \end{cases}$

4.1 Sensitivity and Specificity

For illustration purposes, our analyses will be restricted to the case of the number of bacteria in one contaminated yogurt being modeled by the Poisson distribution, i.e., $Y_i \sim \mathbf{P}(\lambda)$ (the others cases are analogous).

Let us suppose that the test never returns wrong results (it identifies always the presence of bacteria if any exists in the ml analyzed). Hence, the sensitivity is the probability of existing at least one bacterium in one ml from on infected yogurt, and therefore $\varphi_s = \mathbb{P}(Y_i > 0) = 1 - \mathrm{e}^{-\lambda}$; and the specificity will be equal to certitude (one not contaminated yogurt does not have any bacterium). Under these conditions, the number of bacteria Y^* in one ml is modeled by a zero-inflated Poisson distribution with probability mass function

$$f_{Y^*}(x) = \begin{cases} (1-p) + p\mathrm{e}^{-\lambda} & x = 0 \\ p\,\dfrac{\mathrm{e}^{-\lambda}\,\lambda^x}{x!} & x \in \mathbb{N}. \end{cases} \tag{5}$$

In the compound test, the number of bacteria in 1 ml of yogurt will be characterized by a Poisson distribution with parameter $m\frac{\lambda}{n}$, where $m \leq n$ represents the number of infected yogurts in the group. As the number m of contaminated yogurt is $m \sim \mathbf{B}(n, p)$, the number of bacteria Y_n^* in 1 ml of the pooled sample has probability function given by

$$f_{Y_n^*}(x) = \begin{cases} \displaystyle\sum_{i=0}^{n} I^{[i,n]} \mathrm{e}^{-i\frac{\lambda}{n}} & x = 0 \\ \displaystyle\sum_{i=1}^{n} I^{[i,n]}\,\frac{\mathrm{e}^{-i\frac{\lambda}{n}}\,\left(i\frac{\lambda}{n}\right)^x}{x!} & x \in \mathbb{N}, \end{cases} \tag{6}$$

that is equal to $f_{Y^*}(x)$ when $n = 1$. So, even in the compound test, we still have $\varphi_{e_n} = 1$ (if none of the yogurts is contaminated then the test is always negative, as we obtain in (3) using $\varphi_e = 1$) and, denoting by $\mathbf{C}$ the contaminated yogurts,

$$\varphi_{s_n} = \mathbb{P}\left(Y_n^* \geq 1 | Y_1 \in \mathbf{C}\right) \times \mathbb{P}\left(Y_1 \geq 1\right) = \sum_{i=0}^{n-1} I^{[i,n-1]} \left(1 - \mathrm{e}^{-(i+1)\frac{\lambda}{n}}\right) \left(1 - \mathrm{e}^{-\lambda}\right),$$

which corresponds to formula (2) with $\varphi_s^{[i+1,n]} = 1 - \mathrm{e}^{-(i+1)\frac{\lambda}{n}}$ and $\varphi_s = 1 - e^{-\lambda}$.

Applying the weights $k_m^{[n]}$ computed in Sect. 3.1, we get $k_m^{[n]} = 1 - \frac{1-e^{-\frac{m}{n}\lambda}}{1-e^{-\lambda}}$, and the difference $\varphi_s - \varphi_s^{[n]}$ is approximately $\left(\lambda_1 k_1^{[n]} + \lambda_2 k_2^{[n]}\right) \varphi_s$. If we consider, as an example, a prevalence rate of $p = 0.01$ ($n^* = 11$) and a parameter $\lambda = 10$, then $\varphi_s - \varphi_s^{[n]} \approx 0.39\varphi_s$, so the difference between the single test sensitivity and the pooled sensitivity is approximately 40 %!

We can also include in the model the probability of having an extra source of error associated with the test itself, specifically $1 - \varphi_{e_T}$ of false positive and probability $1 - \varphi_{s_T}$ of false negative, the resulting specificity is

$$\begin{aligned}
\varphi_{e_n} &= \varphi_{e_T} \mathbb{P}\left(Y_n^* = 0 | Y_1 \notin \mathbf{C}\right) + (1 - \varphi_{s_T}) \mathbb{P}\left(Y_n^* > 0 | Y_1 \notin \mathbf{C}\right) + \\
&+ \varphi_{e_T} \left[(1 - \varphi_{e_T}) \mathbb{P}\left(Y_n^* = 0 | Y_1 \notin \mathbf{C}\right) + \varphi_{s_T} \mathbb{P}\left(Y_n^* > 0 | Y_1 \notin \mathbf{C}\right)\right] \\
&= \varphi_{e_T} \varphi_0 + (1 - \varphi_{s_T})(1 - \varphi_0) + \varphi_{e_T} \left[(1 - \varphi_{e_T})\varphi_0 + \varphi_{s_T}(1 - \varphi_0)\right] \\
&= 1 - (1 - \varphi_{e_T})\varphi_{s_T} - (1 - \varphi_{e_T})(1 - \varphi_{e_T} - \varphi_{s_T})\varphi_0, \qquad (7)
\end{aligned}$$

where $\varphi_0 = \sum_{i=0}^{n-1} I^{[i,n-1]} \left(\mathrm{e}^{-i\frac{\lambda}{n}}\right)$. This formula is analogous to the one previously deduced in Sect. 3.2 (but with φ_0 instead of q^{n-1}). The sensitivity (where both the compound and single tests have to be positive) is

$$\begin{aligned}
\varphi_{s_n} &= \left[(1 - \varphi_{e_T}) \mathbb{P}\left(Y_n^* = 0 | Y_1 \in \mathbf{C}\right) + \varphi_{s_T} \mathbb{P}\left(Y_n^* > 0 | Y_1 \in \mathbf{C}\right)\right] \times \\
&\quad \times \left[\varphi_{s_T} \mathbb{P}\left(Y_1 \geq 1\right) + (1 - \varphi_{e_T}) \mathbb{P}\left(Y_1 = 0\right)\right] \\
&= \left[(1 - \varphi_{e_T})\varphi_1 + \varphi_{s_T}(1 - \varphi_1)\right] \left[\varphi_{s_T}\left(1 - \mathrm{e}^{-\lambda}\right) + (1 - \varphi_{e_T})\mathrm{e}^{-\lambda}\right] \\
&= \varphi_{s_T}^2 + \varphi_1 \mathrm{e}^{-\lambda}(1 - \varphi_{e_T} - \varphi_{s_T})^2 + \varphi_{s_T}(\varphi_1 + \mathrm{e}^{-\lambda})(1 - \varphi_{e_T} - \varphi_{s_T}) \qquad (8)
\end{aligned}$$

where $\varphi_1 = \sum_{i=0}^{n-1} I^{[i,n-1]} \left(\mathrm{e}^{-(i+1)\frac{\lambda}{n}}\right)$. This incorporates the probabilities of erroneous outcomes due to the operational characteristics of the test himself (where we obtain $\varphi_{s_T}^2$, as in the cases previously analyzed) and of sampling errors (in one contaminated group, we get no bacterium in the compound sample).

Table 5 Relative cost, sensitivity, and specificity for two subpopulations

p_1	0.0075	0.005	0.001	0.0001	0.005	0.005	0.005	0.005	0.001	0.001	0.001	0.001
ω_1	0.5	0.5	0.5	0.5	0.25	0.10	0.75	0.9	0.25	0.10	0.75	0.9
p_2	0.0125	0.015	0.019	0.0199	0.011(6)	0.010(5)	0.025	0.055	0.013	0.011	0.037	0.091
$\mathbf{RC}$	0.1939	0.1887	0.1650	0.1467	0.1928	0.1945	0.1806	0.1698	0.1823	0.1905	0.1393	0.1132
$\varphi_{e_n}^{[S]}$	0.9950	0.9950	0.9953	0.9953	0.9949	0.9950	0.9951	0.9953	0.9951	0.9950	0.9956	0.9961
$\varphi_{s_n}^{[S]}$	0.5936	0.5606	0.4810	0.4048	0.5757	0.5935	0.5386	0.5111	0.5506	0.5746	0.4026	0.3388

5 Dealing with Subpopulations

Let us consider that our population is divided into k groups (subpopulations) with weights $\omega_1, \omega_2, \ldots, \omega_k$, where $\sum_{i=1}^{k} \omega_i = 1$, and prevalence rates $p_1, p_2, \ldots, p_k$. If the presence of subpopulations in the compound tests is disregarded, the results for efficiency, sensitivity, and specificity are the same as the ones obtained with one population with prevalence rate given by $p = \sum_{i=1}^{k} \omega_i p_i$.

On the other hand, if this information is not disregarded in the Dorfman's methodology, the expected number of tests will be $\mathbb{E}[T_S] = \sum_{i=1}^{k} \omega_i N \min_{n_i} \left\{ \frac{n_i+1}{n_i} - (1-p_i)^{n_i}, 1 \right\}$, $n_i \geq 2$, where n_i is chosen in function of the p_i (cf. Table 1). The relative cost, $\mathbf{RC} = \frac{\mathbb{E}[T_S]}{N}$, is the weighted mean of the relatives cost of each subpopulation, cf. [3].

For the simplified case (without rarefaction), the sensitivity $\varphi_{s_n}^{[S]} = \varphi_s^2 = \varphi_{s_n}$. But the specificity is $\varphi_{e_n}^{[S]} = 1-(1-\varphi_e)\varphi_s - (1-\varphi_e)(1-\varphi_e-\varphi_s)\sum_{i=1}^{k} \omega_i q_i^{n_i-1}$. For $p \approx 0$, we get $\varphi_{e_n}^{[S]} \approx 1-(1-\varphi_e)\varphi_s - (1-\varphi_e)(1-\varphi_e-\varphi_s)\sum_{i=1}^{k} \omega_i (1-p_i)^{\frac{1}{\sqrt{p_i}}-1}$. As $f(p) = (1-p)^{\frac{1}{\sqrt{p}}-1}$ is a convex function for $p \in (0,1)$, using Jensen's inequality $\sum_{i=1}^{k} \omega_i f(p_i) \geq f\left(\sum_{i=1}^{k} \omega_i p_i\right)$, and therefore ($p = \sum_{i=1}^{k} \omega_i p_i$),

$$\sum_{i=1}^{k} \omega_i (1-p_i)^{\frac{1}{\sqrt{p_i}}-1} \geq \left(1 - \sum_{i=1}^{k} \omega_i p_i\right)^{\frac{1}{\sqrt{\sum_{i=1}^{k} \omega_i p_i}}-1} = (1-p)^{\frac{1}{\sqrt{p}}-1}. \quad (9)$$

Hence, if we analyze the subpopulation strata separately, the specificity increases.

As an example, consider the Poisson example with $\lambda = 10$, $p = 0.01$, $\varphi_{s_T} = \varphi_{e_T} = 0.95$. Working with the population as a whole, using (7) and (8), we have $\varphi_{e_n} = 0.9949$, $\varphi_{s_n} = 0.5780$ and $\mathbb{E}[T] = 0.1956N$ (i.e., $\mathbf{RC} = 0.1956$). Table 5 shows the results testing the two subpopulations separately (with $k = 2$). Testing the subpopulations separately is slightly more efficient than otherwise, but the loss of sensitivity should be analyzed previously.

6 Conclusion

The use of compound tests optimizes the number of expected tests and consequently saves resources. On the other hand, the possible loss of sensitivity and specificity can restrict the usefulness of this methodology. Although the pooled sensitivity and specificity are widely studied, the rarefaction factor is often disregarded. In this work we stress the importance of determining the $\varphi_s^{[1,n]}$ sensitivity to control the total pooled sensitivity. Therefore, the rarefaction factor is crucial in pooled sample analysis, but for low prevalence rates the knowledge of the sensitivity $\varphi_s^{[1,n]}$ can be a simple alternative to measure the overall misclassification, compared to the use of more general dilution models (cf. [6, 7]).

Acknowledgements The authors thank the referees for their very useful comments. 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.

References

1. 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)
2. Dorfman, R.: The detection of defective members in large populations. Ann. Math. Stat. **14**, 436–440 (1943)
3. Finucan, H.M.: The blood testing problem. Appl. Stat. J. Roy. Stat. C **13**, 43–50 (1964)
4. Gill, A., Gottlieb, D.: The identification of a set by successive intersections. In: Information and Control, pp. 20–25. Ellis Horwood, Chichester (1974)
5. Hess, K.T., Liewald, A., Schmidt, K.D.: An extension of Panjer's recursion. Astin Bull. **32**(2), 283–297 (2002)
6. Hung, M., Swallow, W.: 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, C.: Comparison of group testing algorithms for case identification in the presence of testing errors. Biometrics **63**, 1152–1163 (2007)
9. 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)
10. Sterret, A.: On the detection of defective members of large populations. Ann. Math. Stat. **28**, 1033–1036 (1957)
11. Sobel, M., Elashoff, R.: Group testing with a new goal, estimation. Biometrika **62**, 181–193 (1975)
12. 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)
13. 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)
14. Wein, L., Zenios, S.: Pooled testing for HIV screening: capturing the dilution effect. Oper. Res. **44**, 543–569 (1996)

15. Williams, D.: Weighing the Odds: A Course in Probability and Statistics. Cambridge University Press, Cambridge (2001)
16. Zenios, S., Wein, L.: Pooled testing for HIV prevalence estimation exploiting the dilution effect. Stat. Med. **17**, 1447–1467 (1998)

Part V

Models with Stochastic Differential Equations

Prediction for Individual Growth in a Random Environment

Patrícia A. Filipe, Carlos A. Braumann, Nuno M. Brites, and Carlos J. Roquete

Abstract

In the literature it is usual to find classic regression models to describe the dynamics of a certain growth phenomenon. However, in phenomena of a dynamic nature, it is more appropriate to use models that are able to incorporate the dynamics of the growth process and the effect produced by the environmental random fluctuations on such dynamics. This can be done using stochastic differential equations (SDE) models. In this chapter, we start by comparing the quality of fitting and prediction using nonlinear regression models and SDE models. For the SDE models, we discuss the computation of asymptotic confidence intervals for prediction using simulation and the delta method. We show an application using cattle weight data from several females of the Mertolengo cattle breed.

1 Introduction

We study models for the growth of individual organisms in randomly fluctuating environments and show an application using cattle weight data from several females of Mertolengo cattle breed of the *rosilho* phenotype. Most commonly, the weight of an animal at age t, X_t, can be described by a differential equation of the form $\mathrm{d}Y_t = \beta(\alpha - Y_t)\mathrm{d}t$, where $Y_t = g(X_t)$ with g a strictly increasing C^1 function.

P.A. Filipe (✉) · C.A. Braumann · N.M. Brites
Universidade de Évora, Centro de Investigação em Matemática e Aplicações, Colégio Luís Verney, Rua Romão Ramalho 59, 7000-671 Évora, Portugal
e-mail: pasf@uevora.pt; braumann@uevora.pt; d7488@alunos.uevora.pt

C.J. Roquete
Universidade de Évora, Instituto de Ciências Agrárias e Ambientais Mediterrânicas, Núcleo da Mitra, Apartado 94, 7002-774 Évora, Portugal
e-mail: croquete@uevora.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_20, © Springer-Verlag Berlin Heidelberg 2013

For example, $g(x) = x$, $g(x) = -x^{-1}$, $g(x) = x^c$, $c > 0$, and $g(x) = \ln x$ are typical particular choices for g, corresponding to the *monomolecular model*, the *logistic model*, the *Bertalanffy–Richards model*, and the *Gompertz model*, respectively. The solution of the deterministic model is $Y_t = \alpha + \mathrm{e}^{-\beta(t-t_0)}(y_0 - \alpha)$, where $y_0 = g(x_0)$, x_0 being the weight at age t_0 (initial age). Notice that α is the asymptotic value of Y_t and therefore $A = g^{-1}(\alpha)$ is the asymptotic weight or weight at maturity. The parameter β is a growth coefficient, a rate of approach to maturity. Using this type of models, the fitting to data is usually made by nonlinear regression methods in which it is assumed independence between observed deviations from the regression curve at different times. This would be suitable if the deviations were due to measurement errors, but it is inappropriate if deviations are due to random changes on growth rates induced by environmental random fluctuations.

Stochastic differential equation models incorporate the dynamics of the growth process and the effect of random fluctuations on growth dynamics. To describe individual growth in a random environment, we use stochastic differential equations of the form

$$\mathrm{d}Y_t = \beta(\alpha - Y_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \quad Y_{t_0} = y_0, \tag{1}$$

where σ measures the intensity of the effect of environmental random fluctuations on growth and W_t is the standard Wiener process. We have considered the stochastic versions of the Gompertz model (SGM) and the Bertalanffy–Richards model with $c = 1/3$ (SBRM). This is a monophasic model, in which there is only one functional form describing the average dynamics of the complete growth curve (see, for instance, [1]). In [3] we presented the generalization of the above stochastic model to the multiphasic case, in which we consider that the growth coefficient assumes different values for different phases of the animals life. We have also extended the model to the case where the asymptotic size varies randomly from animal to animal (mixed-effects SDE model) (forthcoming paper). In this work we present a study of prediction issues using a monophasic SDE model. We compare classical regression models with stochastic differential equation models (Sect. 2), and focus on prediction using this last type of models (Sect. 3).

2 Classic Regression *vs* Stochastic Differential Equations

Assume that we pretend to model Y as a function of time t and observe (t_1, y_1), (t_2, y_2),..., (t_n, y_n), where y_n is the observed value of Y_{t_n}. A classical regression model can take the form $y_i = f(t_i, \theta) + \varepsilon_i$, where, in our case, we have $f(t_i, \theta) = \alpha + \mathrm{e}^{-\beta(t_i - t_0)}(y_0 - \alpha)$, with $\theta = (\alpha, \beta, y_0)$ and ε_i are $i.i.d.$ $\mathcal{N}(0, \sigma_\varepsilon^2)$. For both curve fitting and prediction we use $\hat{y} = \hat{\alpha} + \mathrm{e}^{-\hat{\beta}(t-t_0)}(\hat{y}_0 - \hat{\alpha})$ where $(\hat{\alpha}, \hat{\beta}, \hat{y}_0)$ are estimates obtained through the least square estimation method. One can find, for instance, in [4] the main theoretical aspects of classical regression as well as several application examples.

For the SDE model, one cannot truly speak of curve fitting for ages $t = t_k$, since Y_k is the true size (in the Y scale) of the animal, and not, as in the case of regression, a value measured with some error for which we want to estimate the true value. What we call here curve fitting is indeed our estimate of the deterministic curve that would result if, starting from the initial size Y_{t_0}, there were no environmental random fluctuations ($\sigma = 0$), which is in our case equivalent to the curve of the expected values of Y_t: $\hat{Y}_t = \hat{E}[Y_t|y_0] = \hat{\alpha} + (Y_{t_0} - \hat{\alpha}) \exp(-\hat{\beta}(t - t_0))$. Here $\hat{E}$ is the estimate of the expected value and $\hat{\alpha}$ e $\hat{\beta}$ are maximum likelihood estimates based on all the observations.

Concerning prediction in SDE models, given the values of the process until instant t_k, Y_{t_1}, Y_{t_2},..., Y_{t_k}, we wish to predict Y_t for $t > t_k$. The solution of the SDE model (1) given the value of Y_{t_k} is

$$Y_t = \alpha + (Y_{t_k} - \alpha)\, \mathrm{e}^{-\beta(t-t_k)} + \sigma \mathrm{e}^{-\beta t} \int_{t_k}^{t} \mathrm{e}^{\beta s} \mathrm{d}W(s). \tag{2}$$

Considering that Y_t is a Markov process, we obtain $E[Y_t|Y_{t_1}, \ldots, Y_{t_k}] = E[Y_t|Y_{t_k}]$, and using the fact that the process (2) follows a gaussian distribution with mean $\alpha + (Y_{t_k} - \alpha) \exp(-\beta(t - t_k))$ and variance $\frac{\sigma^2}{2\beta}(1 - \exp(-2\beta(t - t_k)))$, we can use as predictor

$$\hat{Y}_t = \hat{E}[Y_t|Y_{t_k}] = \hat{\alpha} + (Y_{t_k} - \hat{\alpha})\, \mathrm{e}^{-\hat{\beta}(t-t_k)}, \tag{3}$$

where $\hat{p} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$ are the maximum likelihood estimates based on observations up to time t_k. We can consider that $\hat{Y}_t - Y_t$ is approximately gaussian and obtain the asymptotic 95 % confidence interval for Y_t using the expression

$$\hat{Y}_t - E\left[\hat{Y}_t - Y_t\right] \pm 1.96 \sqrt{\mathrm{Var}\left[\hat{Y}_t - Y_t\right]}. \tag{4}$$

For illustration of the results, we have worked with data from one animal randomly selected between a set of 97 animals. For this animal we had available 51 observations of its weight from birth till approximately 12 years of age. We have started by comparing the non linear regression (NLR) model with the SDE model in terms of fitting. Table 1 shows estimation results based on the data of the full trajectory. Figure 1 shows the plots resulting from the fitting based on Gompertz and Bertalanffy–Richards models. Comparing the values of the root of the mean square error (RMSE), presented in Table 2, we can see that the NLR model is better than the SDE model. This result was expected, since the least square method was used to obtain the estimates of the NLR model parameters, which minimizes the sum of the square errors, consequently the RMSE. However, there is only a slight difference between the two models and the NLR model has an additional parameter, y_0, while in the SDE model we use the actual known value $y_0 = g(x_0)$, with $x_0 = 26$ kg for our trajectory.

Table 1 NLR vs SDE: parameters estimates based on full data of the chosen animal

	n		y_0	A	β
Gompertz	51	NLR	26.90	409.44	1.40
		SDE		406.13	1.49
B-R		NLR	30.75	416.41	1.04
		SDE		407.36	1.22
Gompertz	72	NLR	27.26	493.29	1.37
		SDE		493.39	1.40
B-R		NLR	34.98	500.90	0.96
		SDE		497.00	1.09
Gompertz	42	NLR	21.70	436.59	1.23
		SDE		415.47	1.38
B-R		NLR	21.66	441.83	0.84
		SDE		423.12	0.99

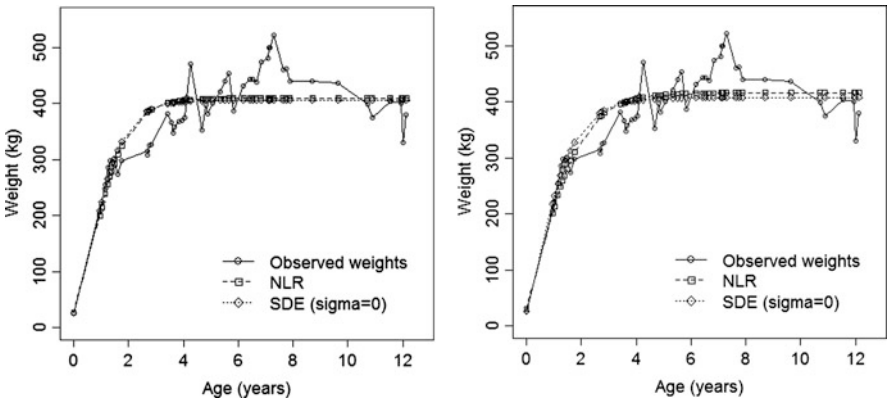

Fig. 1 NLR vs SDE: Fitting results (Gompert model on the left; Bertalanffy–Richards model on the right)

Table 2 NLR vs SDE: values of the root of the mean square error

RMSE	n		Fitting	LT prediction	SS prediction
Gompertz	51	NLR	45.4	68.9	59.9
		SDE	46.2	42.8	27.5
B-R		NLR	42.7	64.3	56.3
		SDE	44.9	38.9	27.3
Gompertz	72	NLR	46.3	45.1	39.0
		SDE	46.6	31.0	23.6
B-R		NLR	41.6	37.6	32.5
		SDE	43.6	27.4	23.8
Gompertz	42	NLR	39.0	55.1	42.6
		SDE	45.8	52.4	30.6
B-R		NLR	35.6	48.3	37.9
		SDE	40.6	42.7	29.4

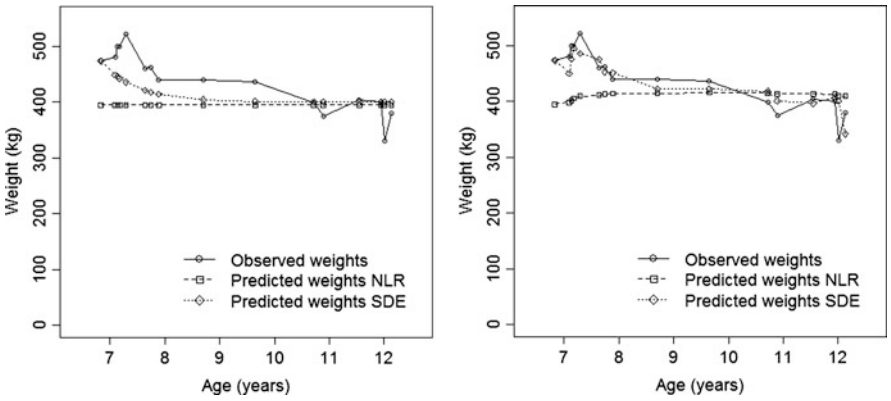

Fig. 2 NLR vs SDE: Prediction (LT on the left; SS on the right) based on Gompertz model

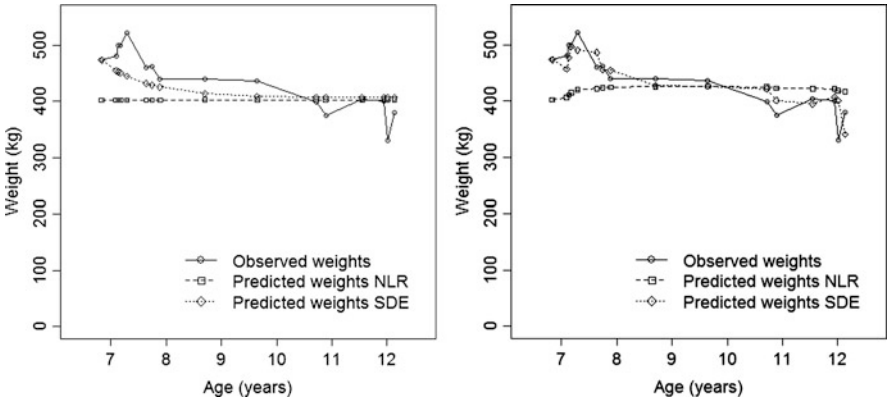

Fig. 3 NLR vs SDE: Prediction (LT on the left; SS on the right) based on Bertalanffy–Richards model

In terms of prediction, one can see that the SDE model is better than the NLR model. For further analysis of the quality of prediction, we used for estimation a subset of the data, leaving out the last 15 observations of the trajectory, (t_{36}, y_{36}), (t_{37}, y_{37}), …, (t_{50}, y_{50}). We present the results for long-term (LT) prediction and step-by-step (SS) prediction at the ages left out of the estimation procedure. In the first case, we predict future weights of the animal using estimates obtained based on the weights observed until age t_{35}. In each step of the second case, we just predict the animal size at next age using the current size as a starting point and using maximum likelihood updated parameter estimates based on all observations up to and including the present time. The RMSE values obtained are shown in Table 2. Figures 2 and 3 clearly reveal that the SDE models are more appropriate to predict future weights of an animal. Similar results were obtained with data from

other randomly selected animals; for illustration, besides the results concerning the initially randomly chosen animal (with $n = 51$ observations), Tables 1 and 2 show the results for two other randomly chosen animals (one with $n = 72$ and the other with $n = 42$ observations).

3 Confidence Intervals for Prediction

For the SDE model, we shall now discuss the computation of the asymptotic confidence limits for prediction given by (4). Since it was not possible to obtain explicit expressions for $E[\hat{Y}_t - Y_t]$ and $\mathrm{Var}[\hat{Y}_t - Y_t]$, we obtained these values using simulated data and a delta method procedure.

3.1 Simulation

In [2], we presented the first results related to the prediction of future weights of an animal based on SGM and SBRM. At that time we have used simulation methods to obtain the asymptotic confidence limits for prediction.

Since $Y_k|Y_{t_{k-1}}=y_{k-1}\sim\mathcal{N}\left(\alpha+(y_{k-1}-\alpha)\,\mathrm{e}^{-\beta(t_k-t_{k-1})},\frac{\sigma^2}{2\beta}\left(1-\mathrm{e}^{-2\beta(t_k-t_{k-1})}\right)\right)$, and using the maximum likelihood estimates, $\hat{\mathbf{p}}$, to approximate $\mathbf{p}$, we have simulated S full trajectories of Y_t at the same ages as the trajectory we wish to predict, $\mathbf{y}^{*\mathbf{i}} = \left(y_0^{*i}, y_1^{*i}, \ldots, y_n^{*i}\right)$ $(i = 1, \ldots, S)$. Based on data until age t_k, for each of the S simulated trajectories we have obtained the maximum likelihood estimates $\hat{\mathbf{p}}^{*\mathbf{i}}$ $(i = 1, \ldots, S)$, the predicted values $\hat{y}_t^{*i} = \hat{\alpha}^{*i} + (y_k^{*i} - \hat{\alpha}^{*i})\mathrm{e}^{-\hat{\beta}^{*i}(t-t_k)}$ $(i = 1, \ldots, S)$, the prediction errors $\hat{y}_t^{*i} - y_t^{*i}$ $(i = 1, \ldots, 1000)$, and their correspondent mean and variance for each $t > t_k$.

3.2 Delta Method

The delta method allows us to find confidence intervals for functions of maximum likelihood estimates. With this method we can approximate a function for which the computation of its mean and variance is complex, with a linear approximation and obtain the mean and variance of this function.

In our case, we define

$$f(\hat{\alpha}, \hat{\beta}, Z_\beta) = \hat{Y}_t - Y_t = \hat{\alpha} - \alpha + (Y_k - \hat{\alpha})\mathrm{e}^{-\hat{\beta}(t-t_k)} - (Y_k - \alpha)\mathrm{e}^{-\beta(t-t_k)} - \sigma\mathrm{e}^{-\beta t}Z_\beta,$$

where $Z_\beta = \int_{t_k}^{t} \mathrm{e}^{\beta s}\mathrm{d}W(s)$ follows a gaussian distribution with mean 0 and variance $(\mathrm{e}^{2\beta t} - \mathrm{e}^{2\beta t_k})/2\beta$. Applying the delta method, we obtain $E\left[\hat{Y}_t - Y_t\right] \simeq f(\alpha, \beta, 0) = 0$. The variance is given by

$$
\begin{aligned}
&\text{Var}\left[\hat{Y}_t - Y_t\right] \\
&\simeq E\left[\left((\hat{\alpha}-\alpha)\frac{\partial f(\alpha,\beta,0)}{\partial\hat{\alpha}} + (\hat{\beta}-\beta)\frac{\partial f(\alpha,\beta,0)}{\partial\hat{\beta}} + Z_\beta\frac{\partial f(\alpha,\beta,0)}{\partial Z_\beta}\right)^2\right] \\
&= \left(\frac{\partial f}{\partial\hat{\alpha}}(\alpha,\beta,0)\right)^2 \text{Var}[\hat{\alpha}]\left(\frac{\partial f}{\partial\hat{\beta}}(\alpha,\beta,0)\right)^2 \text{Var}[\hat{\beta}] \\
&+ \left(\frac{\partial f}{\partial Z_\beta}(\alpha,\beta,0)\right)^2 \text{Var}[Z_\beta] + 2\frac{\partial f}{\partial\hat{\alpha}}(\alpha,\beta,0)\frac{\partial f}{\partial\hat{\beta}}(\alpha,\beta,0)\text{Cov}[\hat{\alpha},\hat{\beta}].
\end{aligned}
$$

Note that the terms involving $\text{Cov}[\hat{\alpha}, Z_\beta]$ and $\text{Cov}[\hat{\beta}, Z_\beta]$ are null because $\hat{\alpha}$ and $\hat{\beta}$ are estimators based on observations previous to t_k and the stochastic integral, Z_β, is over a time interval after that instant, so there is no correlation between these estimators and Z_β. Then, the variance can be written as follows: $\text{Var}\left[\hat{Y}_t - Y_t\right] \simeq V_{EE} + V_{SI}$ where, with $E_k = \text{e}^{-\beta(t-t_k)}$,

$$
\begin{aligned}
V_{EE} = \quad & (1-E_k)^2\text{Var}[\hat{\alpha}] + (Y_k-\alpha)^2(t-t_k)^2E_k^2\text{Var}[\hat{\beta}] \\
& -2(Y_k-\alpha)(t-t_k)E_k(1-E_k)\text{Cov}[\hat{\alpha},\hat{\beta}]
\end{aligned}
$$

corresponds to the variability related to estimation errors (EE) of the parameters and

$$
V_{SI} = \frac{\sigma^2}{2\beta}\left(1-E_k^2\right)
$$

describes the variability caused by the random fluctuations of the environment (variance of the term with stochastic integral (SI)).

We have applied the simulation method and the delta method for the prediction of the last 15 observations of the referred trajectory, and we have observed that the results using simulated values ($S = 1,000$ runs) were very close to the ones resulting from the use of the delta method. This allows us to use the delta method for the computation of the asymptotic limits of the prediction.

Figure 4 shows the plots for predicted values and their asymptotic 95 % confidence limits, computed through the delta method, based on SGM and SBRM. The results were obtained for long-term prediction and step-by-step prediction. The values were computed on the scale of the variable Y and then inverted back to the scale of the weight X using the inverse function of g ($g^{-1}(y) = \exp(y)$ for SGM and $g^{-1}(y) = y^3$ for SBRM).

Sometimes, we wish to predict the future size of an animal for which we only know the current size. We can do that assuming the parameters are the same for all the animals of the same type for which we have available data. So, we have also obtained the long-term prediction results of the same weights of the animal, selected between the group of 97 animals, using data from all of the other 96

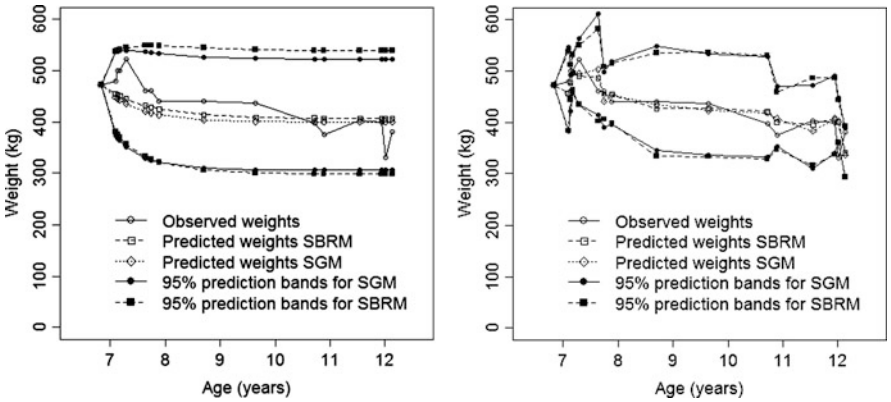

Fig. 4 Observed weights, predicted weights, and asymptotic 95 % confidence bands for prediction (LT on the left; SS on the right). Results based on SGM and SBRM

animals of the group. The values of the RMSE obtained using the estimates based on data from the 96 trajectories were 44.27 and 43.30 for the SGM and the SBRM, respectively. We would expect to obtain better results using the data from all the animals. However, we can justify this results with the fact that the animal selected did not match the mean growth pattern of the 96 animals. In fact, when this happens, we get better results using the data from the single trajectory we wish to predict, should it be available.

4 Conclusion

We have presented results for the prediction of future sizes of an individual in a random environment. We have compared fitting and prediction results when using NLR models against SDE models. We have seen that in terms of prediction, SDE models produce quite better results. For SDE models, we have seen that the delta method gives good approximations to the confidence limits for prediction. Even in the absence of past data from an individual, one can make predictions on its future size using data from a group of similar individuals. We have also studied prediction for a biphasic SDE models, not shown here due to space limitations. For all calculations, we used the R software and the code is available upon request.

Acknowledgements This work, within the research project FCOMP-01-0124-FEDER-007388, was financed by Fundação para a Ciência e Tecnologia (FCT) and is a QREN initiative, cofinanced by the European Regional Development Fund through the Operational Programme for Competitiveness Factors by the European Union. The first three authors are members of the Centro de Investigação em Matemática e Aplicações and the fourth author is member of the Instituto de Ciências Agrárias e Ambientais Mediterrânicas, both research centers of the Universidade de Évora financed by FCT.

References

1. Braumann, C.A., Filipe, P.A., Carlos C., Roquete, C.J.: Growth of individuals in randomly fluctuating environments. In: Vigo-Aguiar, J., Alonso, P., Oharu, S., Venturino, E., Wade, B. (eds.) Proceedings of the International Conference in Computational and Mathematical Methods in Science e Engineering, pp. 201–212, Gijon (2009)
2. Filipe, P.A., Braumann C.A., Roquete, C.J.: Modelos de crescimento de animais em ambiente aleatório. In: Ferrão, M.E., Nunes, C., Braumann, C.A. (eds.) Estatística Ciência Interdisciplinar, Actas do XIV Congresso Anual da Sociedade Portuguesa de Estatística, Edições SPE, pp. 401–410 (2007)
3. Filipe, P.A., Braumann C.A., Roquete, C.J.: Multiphasic individual growth models in random environments. Methodol. Comput. Appl. Probab. **14**(1), 49–56 (2012)
4. Seber, G.A.F., Wild, C.J.: Non-Linear Regression. Wiley, New York (1989)

A Note on (Dis)Investment Options and Perpetuities Under CIR Interest Rates

Manuela Larguinho, José Carlos Dias, and Carlos A. Braumann

Abstract

In this chapter, we discuss alternative ways of computing the options to invest in and divest from an investment project in a CIR economy (Cox et al., Econometrica 53(2):385–408, 1985). Moreover, different methods of determining CIR perpetuities will also be analyzed.

1 Introduction

The most realistic case of the capital theory of investment is characterized by investments with costly reversibility in which a firm can purchase capital at a given price (by paying an investment cost $\overline{I}$) and sell capital at a lower price (by receiving the divestment proceeds $\underline{I}$), i.e., there is a fraction α of the invested capital, $\alpha := \underline{I}/\overline{I}$ $(0 < \alpha < 1)$, that a firm can recoup when divesting.

Decisions made under an uncertain environment where it is costly to reverse economic actions lead to an intermediate range, called the hysteretic band, where the continuation is the optimal policy until some threshold is reached.

M. Larguinho (✉)
Department of Mathematics, ISCAC, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal
e-mail: mlarguinho@iscac.pt

J.C. Dias
BRU-UNIDE and ISCTE-IUL Business School, Av. Prof. Aníbal Bettencourt, 1600-189 Lisboa, Portugal
e-mail: jose.carlos.dias@iscte.pt

C.A. Braumann
Department of Mathematics, Centro de Investigação em Matemática e Aplicações, Universidade de Évora, Rua Romão Ramalho 59, 7000-671 Évora, Portugal
e-mail: braumann@uevora.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_21, © Springer-Verlag Berlin Heidelberg 2013

In this chapter, such as in [6], we consider the situation of a firm that can invest $\overline{I}$ at any time and receive a perpetuity (a project) with constant cash flow rate. Even though the project's cash flows are fixed, its perpetuity value (negatively related to rates) is stochastic because the interest rate r used to discount the perpetual flows is assumed to follow a CIR [3] diffusion process and thus r can change. Investment in this perpetual project will be triggered when interest rates are low (high perpetuity value) and in particular at a critical level $\underline{r}$.

Let us denote by $F_1(r)+P(r)$ and $F_0(r)$ the value functions for the active project (with perpetuity) which has the option to shut, and for the inactive project (without flow) but with the option to open, respectively. Investment will be triggered as the interest rate r falls to the lower threshold $\underline{r}$ while divestment will be triggered as rates rise to the upper trigger $\overline{r}$ ($\overline{r} > \underline{r}$). Thus, the decision to switch from the idle state to the operating state, and vice versa, can be described by the following value matching conditions

$$\begin{aligned} &\text{idlestate} \rightarrow \text{operatingstate, thatis} \quad F_0(\underline{r}) + \overline{I} \rightarrow F_1(\underline{r}) + P(\underline{r}) \\ &\text{idlestate} \leftarrow \text{operatingstate, thatis} \quad F_0(\overline{r}) + \underline{I} \leftarrow F_1(\overline{r}) + P(\overline{r}). \end{aligned}$$

Thus, an idle firm invests when rates fall to $\underline{r}$ and an operating firm will disinvest once the interest rate rise to $\overline{r}$. Denote by V the value premium of active to idle firms, this depend on current rates r:

$$V(r) = F_1(r) + P(r) - F_0(r), \tag{1}$$

and the value matching conditions can be coupled with two smooth pasting (first order) conditions

$$V(\underline{r}) = \overline{I},\ V'(\underline{r}) = 0,\ V(\overline{r}) = \underline{I},\ V'(\overline{r}) = 0. \tag{2}$$

The range $\underline{r}$, $\overline{r}$ is the hysteretic band of the problem since idle firms do not invest and operating firms do not suspend within this intermediate level of interest rate.

To find the entry and exit interest rate thresholds, and thus the optimal policy of the firm, we need to numerically evaluate a system of four highly nonlinear equations (i.e., with two value matching and two smooth pasting conditions).

In this chapter, we discuss alternative ways of computing the options to invest in and divest from a project in a CIR economy. Furthermore, different methods of determining CIR perpetuities will also be analyzed.

The remainder of this chapter is as follows. Section 2 discusses in detail the necessary components for analyzing the optimal entry and exit decision of a firm with CIR interest rates. Section 3 presents computational results comparing the alternative methods of computing the options components and perpetuities under CIR diffusions. Section 4 presents the concluding remarks.

2 Optimal Entry and Exit Decisions Under CIR Interest Rates

Under the risk-neutral measure Q, Cox et al. [3] modeled the evolution of the interest rate, r_t, by the stochastic differential equation (SDE):

$$\mathrm{d}r_t = \big[\kappa\theta - (\lambda+\kappa)r_t\big]dt + \sigma\sqrt{r_t}\mathrm{d}W_t^Q, \tag{3}$$

where W_t^Q is a standard Brownian motion under Q, κ, θ, and σ are positive constants representing reversion rate, asymptotic rate, and volatility parameters, respectively, and λ is the market risk. The condition $2\kappa\theta > \sigma^2$ has to be imposed to ensure that the interest rate remains positive.[1]

2.1 CIR General Claims

Following Cox et al. [3], the price of a general interest rate claim $F(r,t)$ with cash flow rate $C(r,t)$ satisfies the following partial differential equation

$$\frac{1}{2}\sigma^2 r\frac{\partial^2 F(r,t)}{\partial r^2} + \kappa(\theta - r)\frac{\partial F(r,t)}{\partial r} + \frac{\partial F(r,t)}{\partial t} - \lambda r\frac{\partial F(r,t)}{\partial r} - rF(r,t) + C(r,t) = 0. \tag{4}$$

The price of a zero coupon bond with maturity at T, $Z(r,t,T)$, satisfies (4) with $C(r,t) = 0$ subject to the boundary condition $Z(r,T,T) = 1$ and is given by

$$Z(r,t,T) = A(t,T)\mathrm{e}^{-B(t,T)r} \tag{5}$$

where $A(t,T) = \left(\frac{2\gamma\, e^{\left((\kappa+\lambda+\gamma)(T-t)\right)/2}}{(\kappa+\lambda+\gamma)\left(\mathrm{e}^{\gamma(T-t)}-1\right)+2\gamma}\right)^{\frac{2\kappa\theta}{\sigma^2}}$, $B(t,T) = \frac{2\left(\mathrm{e}^{\gamma(T-t)}-1\right)}{(\kappa+\lambda+\gamma)\left(\mathrm{e}^{\gamma(T-t)}-1\right)+2\gamma}$, and $\gamma = \left((\kappa+\lambda)^2 + 2\sigma^2\right)^{1/2}$.

2.2 CIR Perpetuity

In a CIR diffusion, the value of a perpetuity, denoted by $P(r)$, that pays coupons at a constant unit rate $C(r,t) = 1$, should satisfy the following ordinary differential equation (ODE):[2]

[1] See [7] for a complete description of the boundary conditions.

[2] We must note that since $\lim_{t\to\infty}\frac{\partial P(r,t)}{\partial t} = 0$, the value of a perpetuity is not time dependent.

$$\frac{1}{2}\sigma^2 r\frac{\mathrm{d}^2P(r)}{\mathrm{d}r^2}+\kappa(\theta-r)\frac{\mathrm{d}P(r)}{\mathrm{d}r}-\lambda r\frac{\mathrm{d}P(r)}{\mathrm{d}r}-rP(r)+1=0. \tag{6}$$

The solution of ODE (6) is the value of a perpetuity given by

$$P(r)=E_{t_0}^{Q}\left[\int_{t_0}^{\infty}\mathrm{e}^{-\int_{t_0}^{t}r_s ds}\,\mathrm{d}t\right]=\int_{t_0}^{\infty}Z(r,t_0,t)\,\mathrm{d}t, \tag{7}$$

where $Z(r,t,T)$ is the price of a zero coupon bond. There are several ways to evaluate the value of this perpetuity. In this chapter, we analyze two methods proposed by Delbaen [4] and one method proposed by Geman and Yor [8]. In the first method of Delbaen [4] (method D_1), the value of a perpetuity is computed in the following manner:

$$P(r)=\int_0^1\frac{2}{\eta}\mathrm{e}^{-z(2r/\eta)}(1+\beta z)^{(\eta/2\gamma)(2\kappa\theta/\sigma^2)-1}(1-z)^{[(\gamma-\kappa-\lambda)/2\gamma](2\kappa\theta/\sigma^2)-1}\,\mathrm{d}z, \tag{8}$$

where $\eta=\kappa+\lambda+\gamma$ and $\beta=(\gamma-\kappa-\lambda)/\eta$. In the second method of Delbaen [4] (method D_2), the value of a perpetuity is given by

$$P(r)=\frac{\gamma}{\kappa\theta}\Phi_1(a,b,c,x,y), \tag{9}$$

where Φ_1 is the degenerate hypergeometric function defined as

$$\Phi_1(a,b,c,x,y)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{1}{n!}\frac{1}{m!}\frac{(a)_{m+n}}{(c)_{m+n}}(b)_m x^m y^n, \tag{10}$$

where $(\alpha)_j$ is the Pochhammer symbol, and where $a=1$, $b=1-\frac{\kappa+\lambda+\gamma}{2\gamma}\frac{2\kappa\theta}{\sigma^2}$, $c=1+\frac{\gamma-\kappa-\lambda}{2\gamma}\frac{2\kappa\theta}{\sigma^2}$, $x=-\frac{\gamma-\kappa-\lambda}{\kappa+\lambda+\gamma}$, and $y=-\frac{2r}{\kappa+\lambda+\gamma}$.

The alternative formulation proposed by Geman and Yor [8] (method GY) is given by

$$P(r)=\frac{\psi}{\omega}\mathrm{e}^{\phi\psi r/2}\int_0^1\frac{(1+z)^p(1-z)^q\mathrm{e}^{\frac{r\omega(z+\varphi)}{2(1+\varphi z)}}}{(1+\varphi z)^{\delta\psi/2}}\,\mathrm{d}z, \tag{11}$$

where $\delta=\kappa\theta$, $\phi=\frac{\kappa+\lambda}{2}$, $\psi=\frac{4}{\sigma^2}$, $\omega=(2\psi+\phi^2\psi^2)^{1/2}$, $\varphi=\frac{\phi\psi}{\omega}$, $p=\frac{\phi\delta\psi^2}{4\omega}+\frac{\delta\psi}{4}-1$, and $q=\frac{\delta\psi}{4}-\frac{\phi\delta\psi^2}{4\omega}-1$.

2.3 Complementary Functions

Particularizing the time homogenous situation gives a simple ODE that determines the perpetual option to invest in or divest from a project. Furthermore for the options themselves no cash flows are present, so $C(r)=0$

$$\frac{1}{2}\sigma^2 r\frac{\mathrm{d}^2F(r)}{\mathrm{d}r^2}+\kappa(\theta-r)\frac{\mathrm{d}F(r)}{\mathrm{d}r}-\lambda r\frac{\mathrm{d}F(r)}{\mathrm{d}r}-rF(r)=0. \tag{12}$$

Dias and Shackleton [6] have shown that the ode (12) can be transformed into Kummer's equation given by:

$$z_0g''(z_0)+(b-z_0)g'(z_0)-a_0g(z_0)=0, \tag{13}$$

$$z_1g''(z_1)+(b-z_1)g'(z_1)-a_1g(z_1)=0, \tag{14}$$

with $a_{0,1}=\frac{\kappa\theta}{\sigma^2}\left(1\mp\frac{(\kappa+\lambda}{\sigma^2}\right)$, $b=\frac{2\kappa\theta}{\sigma^2}$, and $z_{0,1}=\pm 2\gamma r/\sigma^2$. The solutions to (13) and (14) correspond to the complementary functions to be used for an idle firm and an operating firm, respectively. Since there are many alternative solutions to Kummer's equation (see [1, Equations 13.1.12-19]) care must be taken to choose the ones that give simpler solutions and are easier to apply boundary conditions. In [6], the complete solutions of (13) and (14) are given by (see, [1, Equation 13.1.11])

$$y=A_0M(a_0,b,z_0)+B_0U(a_0,b,z_0), \tag{15}$$

$$y=A_1M(a_1,b,z_1)+B_1U(a_1,b,z_1), \tag{16}$$

where $A_{0,1}$ and $B_{0,1}$ are arbitrary constants, $M(a,b,z)$ is the Kummer's confluent hypergeometric function (e.g., [1, Equation 13.1.2]), and finally $U(a,b,z)$ is the Tricomi confluent hypergeometric function (e.g., [1, Equation 13.1.3]).

If we multiply the complete solutions (15) and (16) by $\mathrm{e}^{v_0 r}$ and $\mathrm{e}^{v_1 r}$, respectively, with $v_{0,1}=\frac{\kappa+\lambda\mp\gamma}{\sigma^2}$, and reverse the change of variables, we obtain

$$F_0(r)=C_1\mathrm{e}^{v_0 r}M(a_0,b,z_0)+C_2\mathrm{e}^{v_0 r}U(a_0,b,z_0), \tag{17}$$

$$F_1(r)=C_3\mathrm{e}^{v_1 r}M(a_1,b,z_1)+C_4\mathrm{e}^{v_1 r}U(a_1,b,z_1), \tag{18}$$

which are also solutions of the ode (12), and where C_{1-4} are constants to be determined from boundary conditions.

The solution to (12) can also be obtained via $Mathematica$ or $Maple$. For instance, Dias [5] provides the following solutions for an idle firm and an operating firm obtained through $Mathematica$:

$$F_0(r)=C_{1M}\mathrm{e}^{v_0 r+\mu\log(r)}U(a,1+\mu,z_0)+C_{2M}\mathrm{e}^{v_0 r+\mu\log(r)}L_n^{\beta}(z_0), \tag{19}$$

$$F_1(r)=C_{3M}\mathrm{e}^{v_0 r+\mu\log(r)}U(a,1+\mu,z_0)+C_{4M}\mathrm{e}^{v_0 r+\mu\log(r)}L_n^{\beta}(z_0), \tag{20}$$

with $\mu=1-\frac{2\kappa\theta}{\sigma^2}$, $a=-\frac{\kappa\theta(\kappa+\lambda+\gamma)-\sigma^2\gamma}{\sigma^2\gamma}$, $\beta=\mu$, $n=-a$, and where $L_n^{\beta}(z)$ is the Laguerre polynomial as defined in [1, Equation 22.5.54].

Carmona and León [2] provide also an alternative solution of (12) for the valuation of an investment project with the option to wait, $F_0(r)$, given by[3]

$$F_0(r) = C\mathrm{e}^{v_0 r}\left(M(a_0,b,z_0) - \frac{\Gamma(b)}{\Gamma(a_0)}\frac{\Gamma(1+a_0+b)}{\Gamma(2-b)} z_0^{1-b} M(a_0-b+1, 2-b, z_0)\right). \tag{21}$$

Even though (17–21) are all solutions to (12) we should choose the ones that are easier to use. For instance, to solve the entry and exit problems we need to numerically evaluate a system of four highly nonlinear equations. Thus, the appropriate choice of the options components solutions is relevant for simplifying numerical computations. The analytical study of the Kummer's confluent hypergeometric functions $M(a,b,z)$ and $U(a,b,z)$ allow us to conclude that the solution of Dias and Shackleton [6] is the easiest to apply and turns the economic hysteresis problem much more simple to analyze and understand. Based on these observations, we describe below the necessary boundary conditions that should be applied to (17) and (18).

2.4 Boundary and First Order Conditions

Given the near-zero and asymptotic behavior of the functions $M(a,b,z)$ and $U(a,b,z)$ and the necessary boundary conditions, Dias and Shackleton [6] found that the expected net present value in the idle state with the option to open and the option to switch out of the perpetuity are respectively given by:

$$F_0(r) = C_2\mathrm{e}^{v_0 r} U(a_0,b,z_0), \tag{22}$$

$$F_1(r) = C_3\mathrm{e}^{v_1 r} M(a_1,b,z_1), \tag{23}$$

which, using (1) and (2), lead to two value matching conditions

$$F_1(\underline{r}) + P(\underline{r}) - F_0(\underline{r}) = \overline{I} \qquad \text{and} \qquad F_1(\overline{r}) + P(\overline{r}) - F_0(\overline{r}) = \underline{I}. \tag{24}$$

The first order (smooth pasting) conditions are given by

$$\frac{\mathrm{d}F_0(\underline{r})}{\mathrm{d}\underline{r}} = \frac{\mathrm{d}F_1(\underline{r})}{\mathrm{d}\underline{r}} + \frac{\mathrm{d}P(\underline{r})}{\mathrm{d}\underline{r}} \qquad \text{and} \qquad \frac{\mathrm{d}F_0(\overline{r})}{\mathrm{d}\overline{r}} = \frac{\mathrm{d}F_1(\overline{r})}{\mathrm{d}\overline{r}} + \frac{\mathrm{d}P(\overline{r})}{\mathrm{d}\overline{r}}. \tag{25}$$

Thus, the solution of the two-sided control problem rests on the determination of the two embedded constants C_2, C_3, and two thresholds $\underline{r}$, $\overline{r}$.

[3] Even though Carmona and León [2] do not present solutions for the option to divest, $F_1(r)$ could also be easily obtained.

3 Numerical Analysis

This section aims to present computational comparisons of the alternative methods of computing the value of a perpetuity, and then using these results to analyze the investment hysteresis problem.

3.1 Perpetuities

To compare, in terms of speed and accuracy, the computation of a perpetuity under the CIR diffusion we need to choose a benchmark. An obvious candidate for a benchmark is to use numerical integration using, for instance, Gauss–Kronrod's method. In order to simplify the numerical computations, we start to analyze if the use of a fixed number T_{max} in the upper limit of the integral of the price of a CIR zero coupon bond $\int_{t_0}^{\infty} Z(r, t_0, t)dt$, instead of using infinity, will not generate any problem.[4] The results are shown in Table 1. Considering $T_{max} = 5,000$ seems to be quite reasonable for the analysis and it will simplify the numerical computations if we use this approach. Now we want to evaluate the differences in methods proposed by Delbaen [4] and Geman and Yor [8] compared against the selected benchmark. Table 2 reports such comparison results using the following set of parameters: $\sigma \in \{0.05, 0.07, \ldots, 0.17, 0.19\}$, $r \in \{0, 0.01, \ldots, 0.14, 0.15\}$, $\lambda \in \{-0.10, -0.05, 0\}$, $\kappa \in \{0.15, 0.20, \ldots, 0.85, 0.90\}$, and $\theta \in \{0.030, 0.045, \ldots, 0.885, 0.900\}$. This combination of parameters produce $55,296$ perpetuity values. The MaxAE, MaxRE, RMSE, MeanAE, and n denote, respectively, the maximum absolute error, the maximum relative error, the root mean squared error, the mean absolute error, and the number of times the absolute difference between the two methods exceeds one penny. The last column of the Table 2 reports the CPU time, in seconds, for computing $55,296$ values of perpetuities, using the function $Timing[.]$ available in *Mathematica 7.0*, running on a Pentium IV $(2.53GhZ)$ personal computer. The results from Table 2 indicate that the Method D_2 proposed by Delbaen [4] performs best in terms of accuracy, and the method D_1 proposed by Delbaen [4] is the most efficient in terms of computation time.

3.2 Entry and Exit Problem

Table 3 presents the results for investment options on a CIR economy using the three alternative formulations described above (and after applying the appropriate boundary conditions) and under the same set of parameters as in [2], i.e., $\kappa = 0.45$, $\theta = 0.03$, $\sigma = 0.15$, and $\lambda = 0$. In addition, we use an investment cost of $\overline{I} = 5$.

[4] We have tried other combinations and we have reached the same conclusions.

Table 1 Benchmark selection

		κ					θ		
T_{max}	θ	0.15	0.50	0.90	T_{max}	κ	0.03	0.09	0.15
100	0.09	15.983797	12.295701	11.795340	100	0.50	31.816054	12.295701	8.052518
500	0.09	15.986306	12.296527	11.706291	500	0.50	33.154585	12.296527	8.052518
5,000	0.09	15.986306	12.296527	11.706291	5,000	0.50	33.154593	12.296527	8.052518
∞	0.09	15.986306	12.296527	11.706291	∞	0.50	33.154593	12.296527	8.052518

Note: Parameters used in calculations: $\sigma = 0.125$, $\lambda = -0.05$, and $r = 0$.

Table 2 Differences in perpetuity values for each method compared against the benchmark

Methods	MaxAE	MaxRE	RMSE	MeanAE	n	CPU time
D_1	$2.97E-01$	$8.45E-03$	$3.43E-04$	$2.97E-03$	2,263	$275s$
D_2	$4.93E-03$	$1.52E-04$	$4.44E-10$	$9.41E-08$	0	$9,147s$
GY	$2.90E-01$	$8.27E-03$	$3.27E-04$	$2.89E-03$	2,238	$377s$

Table 3 Option to invest in a CIR economy

Equation	Constant	$\underline{r}$	$F_0(\underline{r})$
(17)	34.4582	0.1759	21.1574
(19)	16.1497	0.1759	21.1574
(21)	25.8925	0.1759	21.1574

Note: CIR parameters used in calculations: $\kappa = 0.45$, $\theta = 0.03$, $\sigma = 0.15$, and $\lambda = 0$. The investment cost is set at $\overline{I} = 5$

Table 4 Upper and lower interest rate thresholds for the entry and exit problem in a CIR economy

α	$\underline{r}$	$\overline{r}$
0.00	0.1759	$+\infty$
0.25	0.1759	0.9439
0.50	0.1759	0.8897
0.75	0.1759	0.5496
1.00	0.2000	0.2000

Note: CIR parameters used in calculations: $\kappa = 0.45$, $\theta = 0.03$, $\sigma = 0.15$, and $\lambda = 0$. The investment cost and the disinvestment proceeds are set at $\overline{I} = 5$ and $\underline{I} := \alpha\,\overline{I}$, respectively

The perpetuity value is computed using the method D_2 proposed by Delbaen [4]. As expected, the option value is the same under the three alternative solutions. However, given that the solution proposed by Dias and Shackleton [6] is much more simple, the preference is to use this one. Table 4 illustrates the upper and lower interest rate thresholds for the entry and exit problem in a CIR economy for the same parameters set defined in Table 3. When $\alpha = 0$, an operating firm never shuts its project. Therefore, the corresponding threshold represents the interest rate level that will induce an idle firm to enter in a project and continue its operations forever since the option to shut down is worthless. Table 4 reveals also that the increasing levels of flexibility reduce, as expected, the hysteric band. These results provide useful

insights for practitioners since they clearly highlight that managerial consideration of abandonment options at the time of project initiation can add value.

4 Conclusion

In this chapter, we analyzed the alternative methods to calculate the value of a perpetuity under a CIR diffusion, and we conclude that the method D_2 proposed by Delbaen [4] performs best in terms of accuracy. We also discussed the alternative ways to calculate the options to invest in or disinvest from a project under a CIR economy, and we conclude that the solution given in [6] is more simple for applying the smooth pasting conditions, particularly for entry and exit problems.

Acknowledgements Dias is member of the BRU-UNIDE, and Larguinho and Braumann are members of the Research Center Centro de Investigação em Matemática e Aplicações (CIMA), both centers financed by the Fundação para a Ciência e Tecnologia (FCT). Dias gratefully acknowledges the financial support from the FCTs grant number PTDC/EGE-ECO/099255/2008.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1972)
2. Carmona, J., León, A.: Investment option under CIR interest rates. Finance Res. Lett. **4**, 242–253 (2007)
3. Cox, J.C., Ingersoll, J.E., Ross, S.A.: A theory of the term structure of interest rate. Econometrica **53**(2), 385–408 (1985)
4. Delbaen, F.: Consols in the CIR model. Math. Finance **3**(2), 125–134 (1993)
5. Dias, J.C.: Essays in real options models under interest rate uncertainty. PhD Thesis, ISCTE Business School (2006)
6. Dias, J.C., Shackleton, M.B.: Hysteresis effects under CIR interest rates. Eur. J. Oper. Res. **211**, 594–600 (2011)
7. Feller, W.: Two singular diffusion problems. Ann. Math. **54**, 173–182 (1951)
8. Geman, J.H., Yor, M.: Bessel processes, Asian options, and perpetuities. Math. Finance **3**(4), 349–375 (1993)

Stochastic Runge–Kutta Schemes for Discretization of Hysteretic Models

Pedro Vieira, Paula Milheiro de Oliveira, and Álvaro Cunha

Abstract

The need to produce numerical solutions of stochastic differential equations (SDE) is present in problems arising in many areas. This is the case in Seismic Engineering where hysteretic models are used (see Wan et al., Soil Dyn Earthquake Eng 21:75–81, 2001 for an example of a problem involving a bridge column). The simulation of the solutions of these nonlinear equations is based on a discretization scheme. In the study of hysteretic models subjected to Gaussian white noise, we aim to compare the response obtained by using two schemes in the discretization of the SDE, in terms of the second statistical moments of the displacement, with that obtained from solving numerically the ODE system satisfied by the moments that arises after the use of adapted Monte Carlo simulation. We analyze the single degree of freedom Noori–Baber–Wen model for different values of (a) the parameters of the nonlinearity coefficient, (b) the parameters that characterize the type of hysteresis, (c) the parameters that take

P. Vieira (✉)
University of Trás–os–Montes e Alto Douro, Quinta de Prados, Vila Real, Portugal and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal
e-mail: pmvieira@fe.up.pt

P. Milheiro–Oliveira
CMUP and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal
e-mail: poliv@fe.up.pt

A. Cunha
CEC and Faculty of Engineering of the University of Porto, Rua Dr. Roberto Frias, Porto, Portugal
e-mail: acunha@fe.up.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_22, © Springer-Verlag Berlin Heidelberg 2013

into account with the degradation effect of resistance, stiffness, and the pinching effect. We conclude that when the discretization step is small, the estimates of the second moment are similar in both schemes meaning that the choice between the weakly convergency schemes is irrelevant. However the solutions obtained by using the Runge–Kutta schemes are different from those obtained by approximately solving the equations of the moments. This difference is more relevant in situations where the allowed contribution of the dissipated energy is larger.

1 Introduction

We are interested in a class of models that describe the nonlinear behavior of Civil Engineering structures subjected to seismic excitations named the nonlinear models with hysteretic behavior.

Let us consider a general nonlinear hysteretic model with a single degree of freedom as in Noori et al. [4]:

$$\begin{cases} m\ddot{x} + c\dot{x} + \alpha k x + (1-\alpha)\, k z = -m w\,(t) \\ \dot{z} = \left[1 - \xi_1 \mathrm{e}^{\frac{-\epsilon^2}{2\xi_2^2}}\right] \left[A\dot{x} - \nu\left(\beta\,|\dot{x}|\,|z|^{n-1}\, z + \gamma \dot{x}\,|z|^n\right)\right] \frac{1}{\eta} \\ \dot{\epsilon} = (1-\alpha)\, k z \dot{x} \end{cases} \tag{1}$$

which corresponds to a simple structure subjected to a Gaussian white noise, $w\,(t)$, of variance $2\pi S_w$. The response variables in the model are: x, the displacement; $\dot{x}$, the velocity; z, that defines the displacement component explained by the nonlinear part of the model and ϵ that characterizes the dissipated energy by hysteresis. We separate the parameters into two groups:

- The parameters that characterize the linear part of the model: the mass (m); the damping (c); the stiffness (k), and the nonlinearity coefficient (α).
- The parameters that characterize the nonlinear part of the model: A, β and γ (that characterize the hysteretic behavior); ν, η (that characterize, respectively, the effects of degradation of stiffness and resistance) and ξ_1, ξ_2 (that characterize the pinching effect).

In more sophisticated versions of this model, as the ones in Baber et al. [1] and Noori et al. [4], A, ν, η, ξ_1, and ξ_2 are assumed to be also of stochastic nature and ϵ-dependent (see expressions (14) and (15) in Sect. 4).

We first rewrite model (1) in the Ito form:

$$\mathrm{d}X_t = \Theta\,(X_t)\,\mathrm{d}t + \Psi \mathrm{d}W_t \tag{2}$$

where $X\,(t) = [x, \dot{x}, z, \epsilon]^T$ is the response vector and the drift and diffusion matrices are, respectively:

$$\Theta\left(X\left(t\right)\right)=\begin{bmatrix} x_2 \\ -m^{-1}\left(\alpha k x_1+c x_2+\left(1-\alpha\right)k x_3\right) \\ \left[1-\xi_1 \mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\right]\left[A x_2-\nu\left(\beta\left|x_2\right|\left|x_3\right|^{n-1}x_3+\gamma x_2\left|x_3\right|^n\right)\right]\frac{1}{\eta} \\ \left(1-\alpha\right)k x_3 x_2 \end{bmatrix}, \tag{3}$$

$$\Psi=\begin{bmatrix}0 & -1 & 0 & 0\end{bmatrix}^T. \tag{4}$$

Since the exact solution of (2) is not known, we compare the moments of the approximation given by using the stochastic schemes presented in the next section with moments computed approximately from applying the stochastic calculus on (2). As we will see the moments of the exact solution of (2) solve a set of ordinary differential equations whose solution can be approximated if one replaces high order expectation terms by their Monte Carlo counterparts. This approach is explained in Sect. 3.

2 Discretization Schemes

Using the fact that the diffusion matrix is independent of the process X, the schemes exposed by Tocino [5] result in two simple schemes, respectively

$$X_{k+1}=X_k+\frac{1}{2}\left(\Theta\left(X_k\right)+\Theta\left(X_k+\Theta\left(X_k\right)\Delta t+\Psi\Delta W_k\right)\right)\Delta t+\Psi\Delta W_k \tag{5}$$

and

$$X_{k+1}=X_k+\Theta\left(X_k+\frac{1}{2}\Theta\left(X_k\right)\Delta t+\Psi\left(\frac{2-\sqrt{6}}{4}\Delta W_k+\frac{\sqrt{6}}{12}\frac{\left(\Delta W_k\right)^3}{\Delta t}\right)\right)\Delta t+\Psi\Delta W_k, \tag{6}$$

where ΔW_k represents the increment of the Wiener process in $[t_k, t_{k+1}]$.

3 Moments of the Response of Hysteretic Models

Following the idea of Noori et al. [4], we apply the Ito formula to the function

$$h\left(x_1,x_2,x_3,x_4\right)=x_1^{p_1}x_2^{p_2}x_3^{p_3}x_4^{p_4} \tag{7}$$

where $X=\left[x_1,x_2,x_3,x_4\right]^T$ is the solution of (2) and p_1,p_2,p_3 and p_4 are nonnegative integers. This gives

$$\mathrm{d}h=h_X^T\mathrm{d}X+\frac{1}{2}\mathrm{Trace}\left(\Psi Q\Psi^T h_{XX}\right)\mathrm{d}t, \tag{8}$$

where $Q = 2\pi S_w$ and h_X, h_{XX} represent, respectively, the first and second derivative of h with respect to X. Taking the expected value in (8) we obtain the general expression for the ordinary differential equations for the time evolution of the moments of X_t in (2):

$$\begin{aligned}
\dot{m}_{p_1,p_2,p_3,p_4} = & \frac{1}{2} p_2 \left(p_2 - 1\right) 2\pi S_w m_{p_1,p_2-2,p_3,p_4} + p_1 m_{p_1-1,p_2+1,p_3,p_4} \\
& - m^{-1} \alpha k p_2 m_{p_1+1,p_2-1,p_3,p_4} - m^{-1} c p_2 m_{p_1,p_2,p_3,p_4} \\
& - m^{-1} \left(1-\alpha\right) k p_2 m_{p_1,p_2-1,p_3+1,p_4} \\
& + \frac{p_3 A}{\eta} m_{p_1,p_2+1,p_3-1,p_4} - \frac{p_3 \nu}{\eta} \beta E \left[\left|x_2\right| \left|x_3\right|^{n-1} x_1^{p_1} x_2^{p_2} x_3^{p_3} x_4^{p_4} \right] \\
& - \frac{p_3 \nu}{\eta} \gamma E \left[\left|x_3\right|^{n} x_1^{p_1} x_2^{p_2+1} x_3^{p_3-1} x_4^{p_4} \right] \\
& - \frac{p_3 \xi_1}{\eta} A E \left[\mathrm{e}^{-\frac{x_3^2}{2\xi_2^2}} x_1^{p_1} x_2^{p_2+1} x_3^{p_3-1} x_4^{p_4} \right] \\
& + \frac{p_3 \xi_1}{\eta} \nu \beta E \left[\mathrm{e}^{-\frac{x_3^2}{2\xi_2^2}} \left|x_2\right| \left|x_3\right|^{n-1} x_1^{p_1} x_2^{p_2} x_3^{p_3} x_4^{p_4} \right] \\
& + \frac{p_3 \xi_1}{\eta} \nu \gamma E \left[\mathrm{e}^{-\frac{x_3^2}{2\xi_2^2}} \left|x_3\right|^{n} x_1^{p_1} x_2^{p_2+1} x_3^{p_3-1} x_4^{p_4} \right] \\
& + p_4 \left(1-\alpha\right) k m_{p_1,p_2+1,p_3+1,p_4-1}
\end{aligned} \tag{9}$$

where m_{p_1,p_2,p_3,p_4} for $p_i \in 0, 1, 2 \ldots$ represents the $p = p_1 + p_2 + p_3 + p_4$ statistical moment of the response vector X. By assigning p_1, p_2, p_3, and p_4 the appropriate values, the following equations are derived:

$$\begin{cases}
\dot{m}_{1000} = m_{0100} \\
\dot{m}_{0100} = -m^{-1} \left(\alpha k m_{1000} + c m_{0100} + \left(1-\alpha\right) k m_{0010}\right) \\
\dot{m}_{0010} = \frac{1}{\eta} \left[A m_{0100} - \nu \left(\beta E_1 + \gamma E_2\right) - A \xi_1 E_3 + \nu \xi_1 \left(\beta E_4 + \gamma E_5\right)\right] \\
\dot{m}_{0001} = \left(1-\alpha\right) k m_{0110}
\end{cases} \tag{10}$$

and

$$\begin{cases}
\dot{m}_{2000} = 2 m_{1100} \\
\dot{m}_{0200} = 2\pi S_w - m^{-1} \left(2 c m_{0200} + 2\alpha k m_{1100} + 2\left(1-\alpha\right) k m_{0110}\right) \\
\dot{m}_{0020} = \frac{2}{\eta} \left[A m_{0110} - \nu \left(\beta E_6 + \gamma E_7\right) - A \xi_1 E_8 + \nu \xi_1 \left(\beta E_9 + \gamma E_{10}\right)\right] \\
\dot{m}_{1100} = m_{0200} - m^{-1} \left(\alpha k m_{2000} + c m_{1100} + \left(1-\alpha\right) k m_{1010}\right) \\
\dot{m}_{1010} = m_{0110} + \frac{1}{\eta} \left[A m_{1100} - \nu \left(\beta E_{11} + \gamma E_{12}\right) - A \xi_1 E_{13} \right. \\
\qquad \left. + \nu \xi_1 \left(\beta E_{14} + \gamma E_{15}\right)\right] \\
\dot{m}_{0110} = -m^{-1} \left(\alpha k m_{1010} + c m_{0110} + \left(1-\alpha\right) k m_{0020}\right) + \\
\qquad + \frac{1}{\eta} \left[A m_{0200} - \nu \left(\beta E_{16} + \gamma E_{17}\right) - A \xi_1 E_{18} + \nu \xi_1 \left(\beta E_{19} + \gamma E_{20}\right)\right].
\end{cases} \tag{11}$$

This approach differs from that of Noori et al. [4] in that the assumption of Gaussianity of the response variables, which is made in [4], is dropped. As a consequence the equations defined by (10) and (11) are different from those obtained in [4]. These equations can be solved for instance by applying an ordinary Runge–Kutta scheme of order four. We used the implementation of Dormand–Prince that is available in Matlab and, assuming that the type of distribution of the response vector is unknown, we considered the expected values $E_1, \ldots, E_{20}$ defined, respectively, as

$$E\left[|x_2|\,|x_3|^{n-1}\,x_3\right];\quad E\left[|x_3|^n\,x_2\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2}}\,x_2\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2}}\,|x_2|\,|x_3|^{n-1}\,x_3\right];$$
$$E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2}}\,|x_3|^{n-1}\,x_2\right];\quad E\left[|x_2|\,|x_3|^{n-1}\,x_3\right];\quad E\left[|x_3|^n\,x_2x_3\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,x_2x_3\right];$$
$$E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,|x_2|\,|x_3|^{n-1}\,x_3\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,|x_3|^n\,x_2x_3\right];\quad E\left[|x_2|\,|x_3|^{n-1}\,x_1x_3\right];$$
$$E\left[|x_3|^n\,x_1x_2\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,x_1x_2\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,x_1\,|x_2|\,|x_3|^{n-1}\,x_3\right];$$
$$E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,|x_3|^n\,x_1x_2\right];\quad E\left[|x_2|\,|x_3|^{n-1}\,x_2x_3\right];\quad E\left[|x_3|^n\,x_2^2\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,x_2^2\right];$$
$$E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,|x_2|\,|x_3|^{n-1}\,x_2x_3\right];\quad E\left[\mathrm{e}^{\frac{-x_3^2}{2\xi_2^2}}\,|x_3|^n\,x_2^2\right],$$

approximated by performing Monte Carlo simulations at each time step. In this work, $1,000$ simulations of the response vector X_t based on the schemes (5) and (6) were used to produce estimates of $E_1, \ldots, E_{20}$.

4 Numerical Results

We consider model (1) with $S_w = 0.1$, $m = 0.933$, $c = 0.5728$, $k = 35.2$, and $A = n = 1$. We simulate $1,000$ trajectories in the time interval $[0, 20]$ and we use discretization schemes (5) and (6) to approximate solutions of model (1). In the following figures these schemes will be named Tocino #1 and Tocino #2, respectively. We analyze three different cases.

Case 1: $\xi_1 = 0$, $\eta = \nu = 1$ and $\dot{\epsilon}$ negligible (the so called Bouc–Wen model, see [8])
Model (1) reduces to:

$$\begin{cases} m\ddot{x} + c\dot{x} + (1-\alpha)\,kz = -mw\,(t) \\ \dot{z} = \left[A\dot{x} - \left(\beta\,|\dot{x}|\,|z|^{n-1}\,z + \gamma\dot{x}\,|z|^n\right)\right] \end{cases}. \tag{12}$$

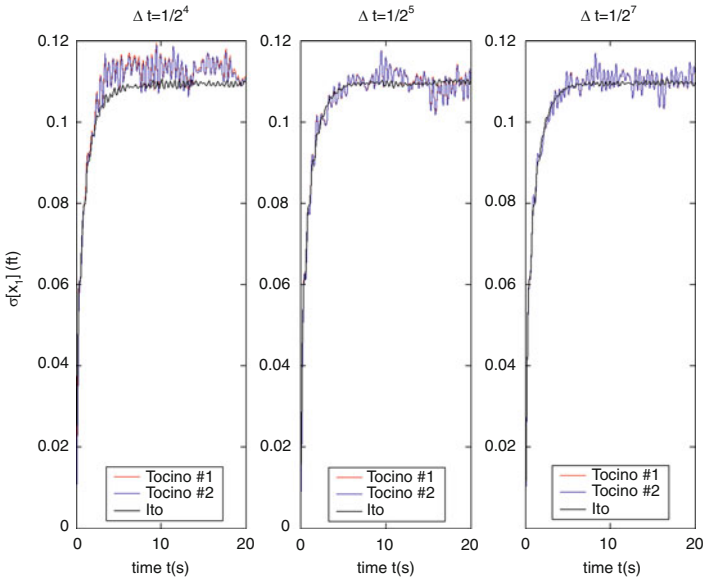

Fig. 1 Standard deviation of the displacement for $\alpha = 0.9$ and $\frac{\beta}{\gamma} = -1$

Figure 1 shows the estimates of the standard deviation of the displacement varying with time for a model with hardening tendency ($\beta/\gamma = -1$) for different choices of the discretization step, $\Delta t = 1/2^4, 1/2^5, 1/2^7$ and for $\alpha = 0.9$. This scenario is known to be close to the linear behavior. The figure illustrates the fact that for a fixed discretization step, the standard deviations of the displacement obtained by using different discretization schemes are similar. For large discretization steps these standard deviations are different from those obtained by solving the moment equations (10) and (11). The same happens for $\alpha = 0.05$ in Fig. 2.

Case 2: $\xi_1 = 0$ (the so called Baber–Wen model, see [1])
Model (1) reduces to:

$$\begin{cases} m\ddot{x} + c\dot{x} + (1-\alpha)\, kz = -mw\,(t) \\ \dot{z} = \left[A\dot{x} - \nu\left(\beta\,|\dot{x}|\,|z|^{n-1}\, z + \gamma\dot{x}\,|z|^{n}\right)\right]\frac{1}{\eta} \\ \dot{\epsilon} = (1-\alpha)\, kz\dot{x} \end{cases} \tag{13}$$

with

$$\begin{cases} A = 1 - \delta_A \epsilon\,(t) \\ \nu = 1 + \delta_\nu \epsilon\,(t) \\ \eta = 1 + \delta_\eta \epsilon\,(t) \end{cases} . \tag{14}$$

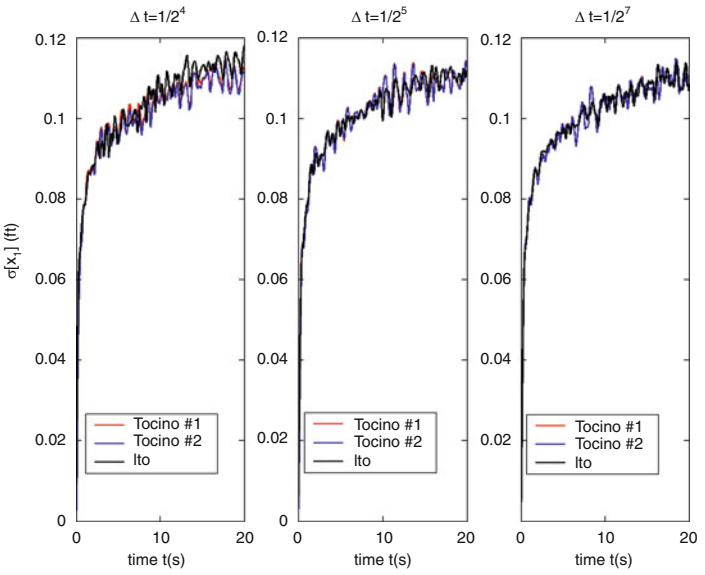

Fig. 2 Standard deviation of the displacement for $\alpha = 0.05$ and $\frac{\beta}{\gamma} = -1$

The parameters δ_A, δ_ν, and δ_η control the amount of the dissipated energy in the estimation of parameters A, ν, and η, at each time instant t. We use $\Delta t = 1/2^7$, $\beta/\gamma = -1$, and $\alpha = 0.5$ in order to avoid being close to the linear behavior ($\alpha = 1$) as well as to the pure nonlinear behavior ($\alpha = 0$). Figure 3 shows the estimates of the standard deviation of the displacement for $\delta_\eta = 0.5$ and different values of δ_ν. We notice that the lines corresponding to the Runge–Kutta schemes (5) and (6) are indistinguishable. Both schemes behave similarly but become more distant from the moments of (10)–(11) as δ_ν gets larger. On the other hand, considering $\delta_\nu = 0.5$ and different values of δ_η, we obtain that there is almost no difference in the results when δ_η changes (Fig. 4).

Case 3: The so called Noori–Baber–Wen model (see [4])

Model (1) is considered with A, ν, and η given by (14) and ξ_1 and ξ_2 given by:

$$\begin{cases} \xi_1 = \xi_{10}\left(1 - e^{-p\epsilon(t)}\right) \\ \xi_2 = \left(\xi_0 + \delta_\xi \epsilon\left(t\right)\right)\left(\lambda + \xi_1\right) \end{cases}. \tag{15}$$

These quantities ξ_1 and ξ_2 characterize, respectively, the severity and sharpness of the pinching behavior. The parameter ξ_{10} represents the maximum value allowed for ξ_1 and p controls the development rate of ξ_1.

We take $\alpha = 0.5$, $A = n = 1$ and $\delta_\nu = \delta_\eta = 0.5$, $\beta/\gamma = -1$. The results are shown in Fig. 5 for different pinching levels: a low pinching level ($\xi_{10} = 0.8$, $\lambda = 0.05$) and a high pinching level ($\xi_{10} = 0.9$, $\lambda = 0.15$); and for different

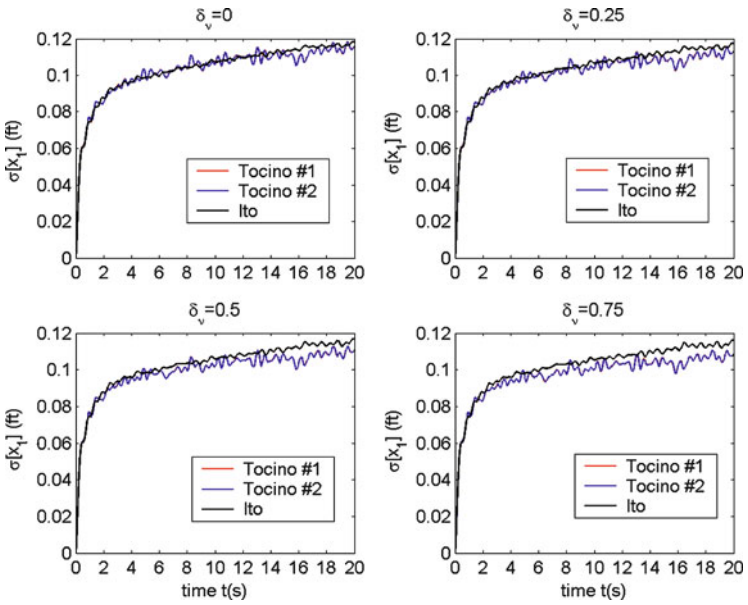

Fig. 3 Standard deviation of the displacement for $\alpha = 0.5$, $\frac{\beta}{\gamma} = -1$, $\delta_\eta = 0.5$ and for different values of δ_ν

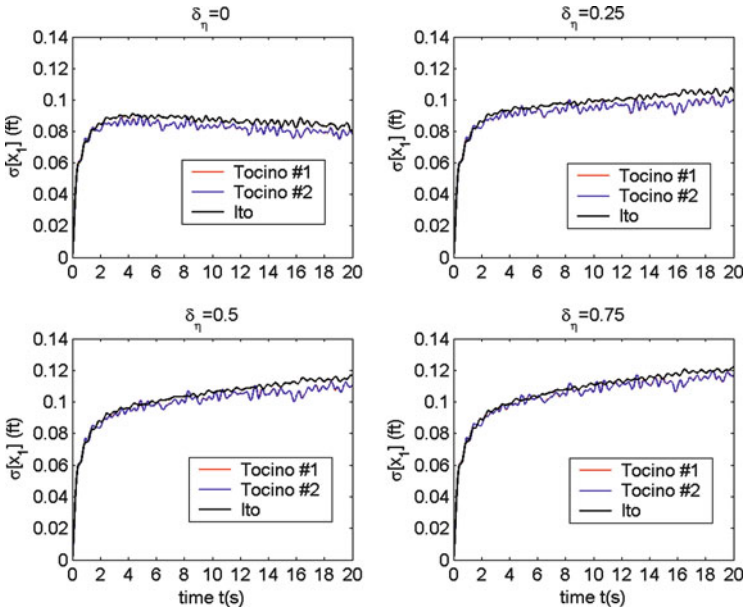

Fig. 4 Standard deviation of the displacement for $\alpha = 0.5$, $\frac{\beta}{\gamma} = -1$, $\delta_\nu = 0.5$ and for different values of δ_η

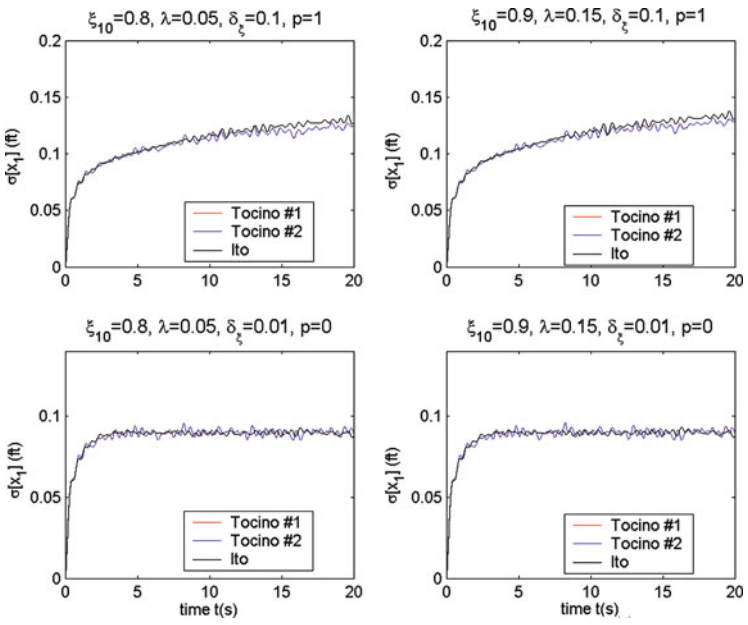

Fig. 5 Standard deviation of the displacement for $\alpha = 0.5$, $\frac{\beta}{\gamma} = -1$, two levels of pinching and different values of δ_ξ and p

contributions of the dissipated energy ϵ, respectively, for a high contribution ($\delta_\xi = 0.1, p = 1$) and a low contribution ($\delta_\xi = 0.01, p = 0$). We notice that the standard deviations obtained by using schemes (5) and (6) are indistinguishable by sight, but differ more from those obtained by solving the moment equations when the parameters that control the contribution of the dissipated energy are larger. Similar situations are observed for systems with no hardening tendency ($\beta/\gamma = 1$).

5 Conclusions

From the simulation study, we may expect that the Runge–Kutta discretization schemes (5) and (6) produce similar values of the second statistical moment of the response as far as Δt is kept small. The deviation of these results from those given by solving the moments equation gets larger as the non-stationarity gets larger. The procedure can be easily extended to the n-dimension case since model (2) presents the same matricial form in blocks of size 4, the hysteretic component and the dissipating energy component being characterized, for each dimension i, by the corresponding component of each parameter vector A_i, ν_i, η_i, β_i, γ_i, ξ_{1i}, and ξ_{2i} associated with the correspondent component of the response vector x_{2i} and x_{3i}. Simulations can be performed in parallel and benefit from the architecture of parallel machines.

The investigation of efficient discretization schemes of the type presented here is important as it provides tools for the simulation and analysis of nonlinear systems appearing in seismic engineering problems, as is the case of hysteretic systems with pinching effect (see [3] for the case of masonry structures). The inherent complexity of the drift function of these models highlights the importance of developing efficient discretization schemes that do not involve the derivatives of this function. The present work shows that the discretization schemes adapted from Tocino [6] that are shown in (5) and (6) are fitted for the purpose of simulating and analyzing hysteretic systems with pinching effect, in the usual region of values of the model parameters that appear in a set of engineering problems (see [2–4, 7] for examples) and as long as the time step is kept small enough, replacing the popular stochastic Euler scheme with benefit to the user in terms of accuracy.

Acknowledgements Supported by FCT under the Annual Financial Support Program for Research Units granted to CEC and CMUP project PEst-C/MAT/UI0144/2011. The authors are very grateful to the anonymous referees for their helpful comments and suggestions.

References

1. Baber, T.T., Wen, Y.K.: Random vibrations of hysteretic degrading systems. J. Eng. Mech. **132**(6), 610–618 (1981)
2. Ibarra, L.F., Medina, R.A., Krawinkler, H.: Hysteretic models that incorporate strength and stiffness deterioration. Earthquake Eng. Struct. Dyn. **34**(12), 1489–1511 (2005)
3. Madan, A., Reinhorn, A.M., Mander, J.: Modeling of masonry infill panels for structural analysis. J. Struct. Eng. **123**(10), 1295–1302 (1997)
4. Noori, M.N., Padula, M.D., Davoodi, H.: Application of an itô-based approximation method to random vibration of a pinching hysteretic system. Nonlin. Dyn. **3**, 305–327 (1992)
5. Tocino, A., Vigo-Aguiar, J.: Weak second order conditions for stochastic Runge–Kutta methods. SIAM J. Scientif. Comput. **24**(2), 507–523 (2002)
6. Tocino, A.: Mean-square stability of second-order Runge–Kutta methods for stochastic differential equations. J. Comput. Appl. Math. **175**, 355–367 (2005)
7. Wan, S., Loh, C.H., Peng, S.Y.: Experimental and theoretical study on softening and pinching effects of bridge column. Soil Dyn. Earthquake Eng. **21**, 75–81 (2001)
8. Wen, Y.K.: Method for random vibration of hysteretic systems. J. Eng. Mech. **102**(2), 249–263 (1976)

Part VI

Stochastic Processes

Misleading Signals in Simultaneous Schemes for the Mean Vector and Covariance Matrix of a Bivariate Process

Patrícia Ferreira Ramos, Manuel Cabral Morais, António Pacheco, and Wolfgang Schmid

Abstract

In a bivariate setting, misleading signals (MS) correspond to valid alarms which lead to the misinterpretation of a shift in the mean vector (resp. covariance matrix) as a shift in the covariance matrix (resp. mean vector). While dealing with bivariate output and two univariate control statistics (one for each parameter), MS occur when:

- The individual chart for the mean vector triggers a signal before the one for the covariance matrix, although the mean vector is on-target and the covariance matrix is off-target.
- The individual chart for the variance triggers a signal before the one for the mean, despite the fact that the covariance matrix is in-control and the mean vector is out-of-control.

Since MS can be rather frequent in the univariate setting, as reported by many authors, this chapter thoroughly investigates the phenomenon of MS in the bivariate case.

P.F. Ramos (✉)
CEMAT, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
e-mail: patriciaferreira@ist.utl.pt

M.C. Morais · A. Pacheco
CEMAT and Mathematics Department, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
e-mail: mjmorais@math.ist.utl.pt; apacheco@math.ist.utl.pt

W. Schmid
Department of Statistics, European University Viadrina, Grosse Scharrnstrasse 59,15230 Frankfurt (Oder), Germany
e-mail: schmid@euv-frankfurt-o.de

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_23, © Springer-Verlag Berlin Heidelberg 2013

Table 1 Definition of misleading signals of types III and IV in simultaneous schemes for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Type of MS	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	First chart to signal
III	On-target	Off-target	Chart for $\boldsymbol{\mu}$
IV	Off-target	On-target	Chart for $\boldsymbol{\Sigma}$

1 Introduction

When we intend to monitor both the mean vector and the covariance matrix of a bivariate process, it is common to run two individual charts at the same time, one for the mean vector ($\boldsymbol{\mu}$) and another one for the covariance matrix ($\boldsymbol{\Sigma}$). The schemes that make use of these two individual charts are the popular simultaneous (or joint) schemes. When we are using a simultaneous scheme, the process is suspected to be out of control whenever at least one of the individual charts triggers a signal. Similarly, a signal triggered by one of the individual charts might correspond to a shift in $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ or both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Moreover, it is possible that a shift in one parameter results in a signal triggered by the individual chart designed to monitor the other parameter leading to what [13, 20] called a misleading signal (MS). When we are using a simultaneous scheme to monitor the mean vector and the covariance matrix of a bivariate process, the MS likely to happen are listed in Table 1.

The phenomenon of misleading signals in simultaneous schemes for univariate i.i.d. processes has been addressed by several authors ([13, 14, 18–20]). More recently, [2,8,16,17] discussed this issue for univariate autocorrelated processes. As far as we have investigated, this phenomenon has not been studied for multivariate processes.

The remainder of this chapter is structured as follows. In Sect. 2 we make a brief review of existing simultaneous schemes for bivariate processes, present the Hotelling-$|\mathbf{S}|$ and the EWMA type schemes under study, and end with some numerical results concerning the PMS for these two types of schemes. In Sect. 3 we present some concluding remarks on the obtained results.

2 Simultaneous Schemes for Bivariate Processes

The first chart to monitor the mean vector of a multivariate process was proposed by [5]. Regarding the control of the covariance matrix, Alt [1] presented a control chart based on the sample generalized variance ($|\mathbf{S}|$). The simultaneous scheme resulting of using these two charts is the most commonly used and it is usually considered the Shewhart analogue for multivariate processes [9]. However, this simultaneous scheme has the disadvantage of being based only on the most recent observation, which led to the emergence of CUSUM and EWMA type schemes for both univariate and multivariate processes.

In the particular case of bivariate processes, some new schemes have been proposed recently by [4,7,10], just to name a few authors. References [4,7] propose

new EWMA type charts that monitor the mean vector and the covariance matrix simultaneously. In addition, [10] recommend the use of a scheme based on an univariate control statistic with noncentral chi-square distribution. In this chapter we shall consider two different simultaneous schemes: the Hotelling-|S| scheme and an EWMA-type scheme described below.

2.1 Output Process and Simultaneous Schemes

Let us denote by $\{\mathbf{Y}_N\}$ the target process which we assume to be i.i.d. with a bivariate normal distribution with mean vector $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$, both known. As for the observed process, denoted by $\{\mathbf{X}_N\}$, we shall assume that it is related to the target process as follows [12]:

$$\mathbf{X}_N = \boldsymbol{\mu}_0 + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\delta} + \boldsymbol{\Theta}(\mathbf{Y}_N - \boldsymbol{\mu}_0), \tag{1}$$

where n is the fixed sample size and $\boldsymbol{\delta} = (\delta_1, \delta_2)'$ and $\boldsymbol{\Theta} = \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{12} & \theta_{22} \end{bmatrix}$ represent the shifts in the mean vector and in the covariance matrix, respectively. Therefore, when a shift in the mean vector (resp. covariance matrix) occurs, the observed process has a bivariate normal distribution with mean vector $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_0 + \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\delta}$ (resp. covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Theta}\boldsymbol{\Sigma}_0\boldsymbol{\Theta}'$). In this chapter we shall focus on the detection of increases in the mean vector or in the covariance matrix, therefore we shall assume that $\boldsymbol{\delta}$ and $\boldsymbol{\Theta}$ are such that both the mean and the variances of the process increase. Without loss of generality we shall also assume that the components of $\boldsymbol{\delta}$ are equal, that is, $\delta_1 = \delta_2 = \delta$.

As previously mentioned, we shall consider two types of simultaneous schemes: the Hotelling-|S| scheme and an EWMA type scheme. Both schemes make use of the control statistics proposed by [1, 5] to monitor the mean vector and the covariance matrix. These control statistics have the following in-control properties when dealing with the N^{th} sample:

$$T_N^2 = n(\overline{\mathbf{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}_0^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}_0) \overset{i.i.d.}{\sim} \chi_2^2; \tag{2}$$

$$U_N = \frac{2(n-1)|\mathbf{S}|^{1/2}}{|\boldsymbol{\Sigma}_0|^{1/2}} \overset{i.i.d.}{\sim} \chi^2_{2(n-2)}, \tag{3}$$

where $\mathbf{S}$ is the sample covariance matrix.

The control statistics and control limits of these two simultaneous schemes are condensed in Table 2, where we note that under control $E(T_N^2) = 2$ and $E(U_N) = 2(n-2)$.

Capitalizing on the distributional properties of T_N^2 and U_N, we can conclude that the run length (RL) of the individual Hotelling chart for $\boldsymbol{\mu}$ ($RL_{H-\mu}$) and of the individual $|\mathbf{S}|$ chart for $\boldsymbol{\Sigma}$ ($RL_{H-\Sigma}$) have geometric distribution with parameters

Table 2 Control statistics and control limits for the Hotelling-|S| and EWMA simultaneous schemes

$\max\{2, T_N^2\}$	$LCL_{H-\mu} = 2$ $UCL_{H-\mu} = 2 + 2\gamma_{H-\mu}$
$\max\{2(n-2), U_N\}$	$LCL_{H-\Sigma} = 2(n-2)$ $UCL_{H-\Sigma} = 2(n-2) + 2\sqrt{n-2}\gamma_{H-\Sigma}$
$Z_N = \begin{cases} 2, N = 0 \\ \max\{2, \lambda T_N^2 + (1-\lambda) Z_{N-1}\}, N \geq 1 \end{cases}$	$LCL_{E-\mu} = 2$ $UCL_{E-\mu} = 2 + \gamma_{E-\mu}\sqrt{\frac{4\lambda}{2-\lambda}}$
$W_N = \begin{cases} 2(n-2), & N = 0 \\ \max\{2(n-2), \lambda U_N + (1-\lambda) W_{N-1}\}, & N \geq 1 \end{cases}$	$LCL_{E-\Sigma} = 2(n-2)$ $UCL_{E-\Sigma} = 2(n-2) + \gamma_{E-\Sigma}\sqrt{\frac{4\lambda(n-2)}{2-\lambda}}$

$$\begin{cases} \xi_{H-\mu}(\boldsymbol{\delta}, \boldsymbol{\Theta}) = 1 - F_{\chi^2_{2,\nu}}(2 + 2\gamma_{H-\mu}), & \boldsymbol{\Theta} = \mathbf{I} \\ \xi_{H-\mu}(\boldsymbol{\delta}, \boldsymbol{\Theta}) = 1 - F^*(2 + 2\gamma_{H-\mu}), & \boldsymbol{\Theta} \neq \mathbf{I} \end{cases} \tag{4}$$

and

$$\xi_{H-\Sigma}(\boldsymbol{\Theta}) = 1 - F_{\chi^2_{2(n-2)}}\left[\frac{2(n-2) + 2\sqrt{n-2}\gamma_{H-\Sigma}}{|\boldsymbol{\Theta}|}\right], \tag{5}$$

where $F_{\chi^2_{2,\nu}}$ represents the distribution function of a noncentral chi-square distribution with noncentrality parameter $\nu = \boldsymbol{\delta}'\boldsymbol{\delta}$ and F^* represents the distribution function of T_N^2, suggested by [11, p. 95] and given by

$$F^*(x) = \sum_{k=0}^{+\infty} (-1)^k c_k \frac{x^{k+1}}{(k+1)!}, \tag{6}$$

where

$$c_k = \begin{cases} (4\lambda_1\lambda_2)^{-1/2}, & k = 0 \\ \frac{1}{k}\sum_{r=0}^{k-1} d_{k-r} c_r, & k \geq 1 \end{cases} \tag{7}$$

$$d_k = \frac{1}{2}\left[\frac{1}{(2\lambda_1)^k} + \frac{1}{(2\lambda_2)^k}\right] \tag{8}$$

and λ_1, λ_2 are the eigenvalues of the matrix $n\mathbf{B}'\boldsymbol{\Sigma}_0^{-1}\mathbf{B}$ with $\mathbf{B}\mathbf{B}' = \boldsymbol{\Sigma}_1/n$.

Since $\overline{\mathbf{X}}$ and $\mathbf{S}$ are independent ([23, p. 48]), we can conclude, by the disjoint blocks theorem (see, e.g., [6, p. 76]), that T_N^2 and U_N are also independent.

Therefore, the RL of the simultaneous scheme ($RL_{H-\mu,\Sigma}$) also has geometric distribution with parameter

$$\xi_{H-\mu,\Sigma}(\boldsymbol{\delta},\boldsymbol{\Theta}) = \xi_{H-\mu}(\boldsymbol{\delta},\boldsymbol{\Theta}) + \xi_{H-\Sigma}(\boldsymbol{\delta},\boldsymbol{\Theta}) - \xi_{H-\mu}(\boldsymbol{\delta},\boldsymbol{\Theta}) \times \xi_{H-\Sigma}(\boldsymbol{\delta},\boldsymbol{\Theta}). \quad (9)$$

It is interesting to note that the RL of the individual chart for $\boldsymbol{\Sigma}$ does not depend on the shift in the mean vector.

As for the EWMA individual and simultaneous schemes, the Markov chain approach [3] provides the following approximations to the survival functions of the run lengths $RL_{E-\mu}$, $RL_{E-\Sigma}$ and $RL_{E-\mu,\Sigma}$

$$\overline{F}_{RL_{E-\mu}(\boldsymbol{\delta},\boldsymbol{\Theta})}(m) \simeq \mathbf{e}'_{\mu}[\mathbf{Q}_{\mu}(\boldsymbol{\delta},\boldsymbol{\Theta};x_{\mu})]^{m}\,\mathbf{1}_{\mu} \quad (10)$$

$$\overline{F}_{RL_{E-\Sigma}(\boldsymbol{\Theta})}(m) \simeq \mathbf{e}'_{\Sigma}[\mathbf{Q}_{\Sigma}(\boldsymbol{\Theta};x_{\Sigma})]^{m}\,\mathbf{1}_{\Sigma} \quad (11)$$

$$\overline{F}_{RL_{E-\mu,\Sigma}(\boldsymbol{\delta},\boldsymbol{\Theta})}(m) = \overline{F}_{RL_{E-\mu}(\boldsymbol{\delta},\boldsymbol{\Theta})}(m) \times \overline{F}_{RL_{E-\Sigma}(\boldsymbol{\Theta})}(m), \quad (12)$$

for $m = 0, 1, 2, \ldots$, where

- $(x_{\mu}+1)$ and $(x_{\Sigma}+1)$ are the number of transient states associated with the absorbing Markov chain.
- $\mathbf{e}_{\mu}$ (resp. $\mathbf{e}_{\Sigma}$) denotes de 1^{st} vector of the orthonormal basis for $\mathbb{R}^{x_{\mu}+1}$ (resp. $\mathbb{R}^{x_{\Sigma}+1}$), associated with the state related to the initial value of the control statistic.
- $\mathbf{1}_{\mu}$ (resp. $\mathbf{1}_{\Sigma}$) is a column vector of $(x_{\mu}+1)$ (resp. $(x_{\Sigma}+1)$) ones.
- $\mathbf{Q}_{\mu}(\boldsymbol{\delta},\boldsymbol{\Theta})$ and $\mathbf{Q}_{\Sigma}(\boldsymbol{\Theta})$ are the substochastic matrices that govern the transitions between the states of the Markov chains.

Moreover, the left partial sums of the entries of $\mathbf{Q}_{\mu}(\boldsymbol{\delta},\boldsymbol{\Theta})$ and $\mathbf{Q}_{\Sigma}(\boldsymbol{\Theta})$ are, respectively:

$$\begin{aligned} &a_{\mu,ij}(\boldsymbol{\delta},\boldsymbol{\Theta};x_{\mu}) \\ &= F_{T_N^2}\left(\frac{1}{\lambda}\left\{2+(j+1)\frac{\gamma_{E-\mu}\sqrt{4\lambda/(2-\lambda)}}{x_{\mu}+1}\right.\right. \\ &\quad \left.\left.-(1-\lambda)\left[2+\frac{\gamma_{E-\mu}(i+1/2)\sqrt{4\lambda/(2-\lambda)}}{x_{\mu}+1}\right]\right\}\right),\ i,j=0,1,\ldots,x_{\mu}; \end{aligned}$$

$$\begin{aligned} &a_{\Sigma,ij}(\boldsymbol{\Theta};x_{\Sigma}) \\ &= F_{\chi^2_{2(n-2)}}\left(\frac{1}{\lambda|\Theta|}\left\{2(n-2)+(j+1)\frac{\gamma_{E-\Sigma}\sqrt{4\lambda(n-2)/(2-\lambda)}}{x_{\Sigma}+1}\right.\right. \\ &\quad \left.\left.-(1-\lambda)\left[2(n-2)+\frac{\gamma_{E-\Sigma}(i+1/2)\sqrt{4\lambda(n-2)/(2-\lambda)}}{x_{\Sigma}+1}\right]\right\}\right), \\ &\qquad\qquad i,j=0,1,\ldots,x_{\Sigma}. \end{aligned}$$

2.2 Probability of a Misleading Signal

The frequency with which an MS occurs in a simultaneous scheme is a cause of concern. This obviously suggests the use of an additional performance measure—the probability of misleading signal (PMS)—whose definition depends on those of the RL of the two individual charts. According to the definition of MS of types III and IV, the corresponding PMS can be written as

$$\begin{aligned} PMS_{III}(\boldsymbol{\Theta}) &= P[RL_{\mu}(\mathbf{0},\boldsymbol{\Theta}) < RL_{\Sigma}(\boldsymbol{\Theta})] \\ &= \sum_{i=1}^{+\infty} \left[\overline{F}_{RL_{\mu}(\mathbf{0},\boldsymbol{\Theta})}(i-1) - \overline{F}_{RL_{\mu}(\mathbf{0},\boldsymbol{\Theta})}(i)\right] \times \overline{F}_{RL_{\Sigma}(\boldsymbol{\Theta})}(i); \quad (13) \end{aligned}$$

$$\begin{aligned} PMS_{IV}(\boldsymbol{\delta}) &= P[RL_{\Sigma}(\mathbf{I}) < RL_{\mu}(\boldsymbol{\delta},\mathbf{I})] \\ &= \sum_{i=1}^{+\infty} \left[\overline{F}_{RL_{\Sigma}(\mathbf{I})}(i-1) - \overline{F}_{RL_{\Sigma}(\mathbf{I})}(i)\right] \times \overline{F}_{RL_{\mu}(\boldsymbol{\delta},\mathbf{I})}(i). \quad (14) \end{aligned}$$

It is worth noting that we can derive exact expression for the PMS of the Hotelling-|S| simultaneous scheme. In fact, since the associated run lengths have geometric distribution with the parameters defined by (4) and (5), we get

$$PMS_{III-H}(\boldsymbol{\Theta}) = \frac{\xi_{H-\mu}(\mathbf{0},\boldsymbol{\Theta}) \times [1 - \xi_{H-\Sigma}(\boldsymbol{\Theta})]}{\xi_{H-\mu,\Sigma}(\mathbf{0},\Theta)}; \quad (15)$$

$$PMS_{IV-H}(\boldsymbol{\delta}) = \frac{[1 - \xi_{H-\mu}(\boldsymbol{\delta},\mathbf{I})] \times \xi_{H-\Sigma}(\mathbf{I})}{\xi_{H-\mu,\Sigma}(\boldsymbol{\delta},\mathbf{I})}. \quad (16)$$

As for the EWMA schemes, we use the approximations (10) and (11) of the survival functions of the RL and truncate the series (13) and (14).

2.3 Numerical Results

We now present some numerical results concerning the PMS in the Hotelling-|S| and EWMA schemes. These results were obtained by considering:

- Sample size $n = 5$.
- In control mean vector $\boldsymbol{\mu}_0 = (0, 0)$.
- In control covariance matrix either $\boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$ or $\boldsymbol{\Sigma}_0 = \mathbf{I}$.
- $\lambda_{\mu} = \lambda_{\Sigma} = \lambda \in \{1, 0.5, 0.05\}$, allowing the comparative assessment of the Hotelling-|S| and two EWMA simultaneous schemes.
- $x_{\mu} + 1 = x_{\Sigma} + 1 = 101$ transient states used in the Markov approach.

Table 3 Critical values of the individual schemes and associated in-control ARL

λ	$\gamma_{E-\mu}$	$\gamma_{E-\Sigma}$	$ARL_{E-\mu}(\mathbf{0},\mathbf{I})$	$ARL_{E-\Sigma}(\mathbf{I})$	$ARL_{E-\mu,\Sigma}(\mathbf{0},\mathbf{I})$
1.00	5.90776	4.75094	1000.00	999.999	500.000
0.50	5.46644	4.50619	999.412	999.421	500.002
0.05	3.33721	3.12739	987.181	987.217	500.016

The critical values in the control limits of the individual charts were determined such that the in-control average run length (ARL) of the simultaneous scheme would be approximately 500 samples and the ARL of the individual charts would be approximately the same. These values were calculated using the *regula falsi* method and they are condensed in Table 3.
To study the behavior of the PMS of Type III, we have contemplated four different scenarios involving the matrices

$$\boldsymbol{\Theta}_1 = \begin{bmatrix} \sqrt{\theta} & 0 \\ 0 & \sqrt{\theta} \end{bmatrix}, \boldsymbol{\Theta}_2 = \begin{bmatrix} 1 & \sqrt{\theta} \\ \sqrt{\theta} & 1 \end{bmatrix}, \boldsymbol{\Theta}_3 = \begin{bmatrix} \sqrt{\theta} & -\sqrt{\theta} \\ \sqrt{\theta} & \sqrt{\theta} \end{bmatrix} \text{ and } \boldsymbol{\Theta}_4 = \begin{bmatrix} \sigma^* \sqrt{\frac{1-\rho^*}{2}} & \sigma^* \sqrt{\frac{1+\rho^*}{2}} \\ -\sigma^* \sqrt{\frac{1-\rho^*}{2}} & \sigma^* \sqrt{\frac{1+\rho^*}{2}} \end{bmatrix}:$$

1. We assume that both variances increase equally and that the correlation coefficient remains the same. The matrix $\boldsymbol{\Theta}$ is of the form $\boldsymbol{\Theta}_1$ with $\theta > 1$.
2. We consider increases both in the variance and in the correlation coefficient. In this case we have $\boldsymbol{\Theta} = \boldsymbol{\Theta}_2$ where the values of θ were chosen such that the variances are the same as in Scenario 1.
3. We study the behavior of the PMS of Type III when the variances increase and the correlation shifts to zero. Here we have $\boldsymbol{\Theta} = \boldsymbol{\Theta}_3$ and, unlike the previous scenario, the values of θ can only be chosen in such way that the variances of the first component of the bivariate normal quality characteristic are the same as in Scenario 1.
4. Here we start with an in-control covariance matrix equal to the identity matrix and study the behavior of PMS of Type III when there are increases in the variances and a shift in the correlation coefficient. $\boldsymbol{\Theta}$ is of the form $\boldsymbol{\Theta}_4$ where σ^{*2} and ρ^* are the new values of the variances and of the correlation, respectively.

The results we have obtained for these four scenarios are summarized in Table 4. They surely deserve some comments. First of all, it is important to stress out that the MS of Type III are very likely to occur in these simultaneous schemes specially for very small shifts in the covariance matrix. The results concerning the first scenario indicate that the PMS of Type III appears to decrease as the shift in the variance becomes larger. It also seems to increase with the value of λ, suggesting that this type of MS is more likely to occur in the Hotelling-$|\mathbf{S}|$ simultaneous scheme than in the EWMA type scheme. As for scenario 2, on one hand we observe the opposite behavior in terms of the monotonic behavior with the magnitude of the shift, that is, as both the variances and the correlation increase, the PMS of Type III also increases. On the other hand, the MS of Type III become more likely as λ decreases, unlike in scenario 1. The PMS of Type III obtained for the third scenario show

Table 4 PMS of Type III for simultaneous Hotelling-|S| ($\lambda = 1$) and EWMA simultaneous schemes ($\lambda = 0.5, 0.05$)

	Σ_1					Σ_0	**Scenario**
	$\begin{bmatrix} 1.01 & 0.303 \\ 0.303 & 1.01 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.315 \\ 0.315 & 1.05 \end{bmatrix}$	$\begin{bmatrix} 1.2 & 0.36 \\ 0.36 & 1.2 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 0.45 \\ 0.45 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 2 & 0.6 \\ 0.6 & 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$	1
ρ_1	0.3	0.3	0.3	0.3	0.3		
1.00	0.493338	0.470243	0.402394	0.323562	0.262602		
λ0.50	0.491053	0.459151	0.367658	0.275227	0.223378		
0.05	0.478638	0.406067	0.275333	0.214696	0.188278		
	$\begin{bmatrix} 1.01 & 0.33 \\ 0.33 & 1.01 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.45 \\ 0.45 & 1.05 \end{bmatrix}$	$\begin{bmatrix} 1.2 & 0.79 \\ 0.79 & 1.2 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 1.3 \\ 1.3 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 2 & 1.95 \\ 1.95 & 2 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$	2
ρ_1	0.33	0.43	0.66	0.87	0.98		
1.00	0.502779	0.563216	0.861575	0.994459	–		
λ0.50	0.502946	0.565347	0.874792	0.996639	–		
0.05	0.503551	0.571285	0.931102	–	–		
	$\begin{bmatrix} 1.01 & 0 \\ 0 & 1.88 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0 \\ 0 & 1.95 \end{bmatrix}$	$\begin{bmatrix} 1.2 & 0 \\ 0 & 2.23 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 0 \\ 0 & 2.79 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 \\ 0 & 3.71 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$	3
ρ_1	0	0	0	0	0		
1.00	0.522976	0.498640	0.425560	0.335134	0.255899		
λ0.50	0.459357	0.432878	0.360581	0.285525	0.227996		
0.05	0.349150	0.332861	0.292123	0.252388	0.222915		
	$\begin{bmatrix} 1.01 & 0.505 \\ 0.505 & 1.01 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.525 \\ 0.525 & 1.05 \end{bmatrix}$	$\begin{bmatrix} 1.2 & 0.6 \\ 0.6 & 1.2 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 0.75 \\ 0.75 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$		
ρ_1	0.5	0.5	0.5	0.5	0.5		
1.00	0.925244	0.908116	0.832099	0.668480	0.474916		
λ0.50	0.939737	0.921881	0.832554	0.621058	0.406073		
0.05	0.986951	0.973323	0.820556	0.487846	0.335958	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	4
	$\begin{bmatrix} 1.01 & 0.707 \\ 0.707 & 1.01 \end{bmatrix}$	$\begin{bmatrix} 1.05 & 0.735 \\ 0.735 & 1.05 \end{bmatrix}$	$\begin{bmatrix} 1.2 & 0.84 \\ 0.84 & 1.2 \end{bmatrix}$	$\begin{bmatrix} 1.5 & 1.05 \\ 1.05 & 1.5 \end{bmatrix}$	$\begin{bmatrix} 2 & 1.4 \\ 1.4 & 2 \end{bmatrix}$		
ρ_1	0.7	0.7	0.7	0.7	0.7		
1.00	0.995131	0.993212	0.980580	0.922897	0.760349		
λ0.50	0.997253	0.995988	0.985959	0.923080	0.712014		
0.05	0.999951	0.999916	0.998541	0.919548	0.622448		

a similar behavior to the ones in scenario 1. The MS of Type III become more unlikely as the shift in the variances increases and λ decreases. Finally, in scenario 4 the numerical results suggest that the PMS of Type III decreases as λ and the shift in the variances increase, and it increases as the correlation increases. As a final comment to this results, we should add that the PMS of Type III can be larger

Table 5 PMS of Type IV for simultaneous Hotelling-|S| ($\lambda = 1$) and EWMA simultaneous schemes ($\lambda = 0.5, 0.05$)

	λ		
δ	1.00	0.50	0.05
0.05	0.495208	0.494691	0.490680
0.25	0.405094	0.394540	0.319622
0.50	0.232321	0.204998	0.087947
1.00	0.050466	0.032235	0.005031
2.00	0.003300	0.001662	0.000036

than 0.9 in some cases. These values are much larger than the one obtained in the univariate case. One possible explanation to this fact is that the PMS of Type III is a function of the determinant of the matrix $\boldsymbol{\Theta}$ and not directly a function of the shifts in the variances and in the correlation coefficient.

As for the PMS of Type IV, the results are condensed in Table 5 for several shifts in the mean vector. Like MS of Type III, we can see that MS of Type IV are also very likely to occur for very small shifts in this case in the mean vector. The values in Table 5 suggest that the MS of Type IV are more likely to occur in the Hotelling-|S| simultaneous scheme than in the EWMA simultaneous scheme. In fact the PMS of Type IV seems to increase with the smoothing parameter λ; in addition it tends to decrease with the value of δ. This apparent monotonic behavior agrees with the results previously obtained by [15] for univariate processes.

3 Concluding Remarks

The results presented in this chapter illustrate that MS are very likely to happen in bivariate schemes, specially for very small shifts in the mean vector or in the covariance matrix.

As for the monotonic behavior of the PMS of types III and IV, the numerical results suggest a similar behavior to the one observed in the univariate case, namely, the MS of Type III (resp. IV) are more likely to occur for small shifts in the covariance matrix (resp. mean vector) and occur less frequently in EWMA-type simultaneous schemes.

We strongly believe that these results show that the PMS is an important performance measure and should be taken into account, in addition to other relevant RL-related performance measures.

As a final note, we ought to point out that the Referee drew our attention to [21,22], who studied a few control charts to monitor the covariance matrix. These authors show that these control charts outperform some of the most commonly used ones, for example, the one that makes use of the generalized variance. However, the control statistic proposed by [21] has a crippling disadvantage: even though it has a well-known distribution under control, as far as we have investigated, it is very difficult to derive the correspondent out-of-control distribution, as a consequence we are not able to calculate the PMS, the goal of our study, unless we resort to simulation as [22] did. As for the other charts used in the simulation study by [22],

we ought to stress that for all of them it is very difficult to derive the out-of-control distribution; moreover, the use of the modified likelihood ratio test is questionable because, in practice, the monitoring is usually based on samples that are not large enough to justify the use of such an asymptotic test. In spite of these disadvantages, we plan to perform a simulation-based study to compare the PMS of a few control charts for the covariance matrix, such as the one proposed by [21].

Acknowledgements The authors thank the financial support of *Centro de Matemática e Aplicações* (CEMAT) and *Fundação para a Ciência e a Tecnologia* (FCT). The first author was also supported by grant SFRH/BD/35739/2007 of FCT and would like to thank all the members of the Department of Statistics of the European University Viadrina (Frankfurt Oder, Germany) for their hospitality.

References

1. Alt, F.: Multivariate quality control. In: Kotz, S., Johnson, N.L, Read, C.R. (eds.) The Encyclopedia of Statistical Sciences, pp. 110–122. Wiley, New York (1984)
2. Antunes, C.: Avaliação do impacto da correlação em sinais erróneos em esquemas conjuntos para o valor esperado e variância (Assessment of the impact of the correlation on misleading signals in joint schemes for the mean and variance). Instituto Superior Técnico, Universidade Técnica de Lisboa (2009)
3. Brook, D., Evans, D.A.: An approach to the probability distribution of CUSUM run length. Biometrika **59**, 539–549 (1972)
4. Chen, G., Cheng, S.W., Xie, H.: A new multivariate control chart for monitoring both location and dispersion. Comm. Stat. Simul. Comput. **34**, 203–217 (2005)
5. Hotelling, H.: Multivariate quality control illustrated by the air testing of sample bombsights. In: Eisenhart, C., Hastay, M.W., Wallis, W.A. (eds.) Techniques of Statistical Analysis, pp. 111–184. McGraw Hill, New York (1947)
6. Karr, A.F.: Probability. Springer, New York (1993)
7. Khoo, B.C.: A new bivariate control chart to monitor the multivariate process mean and variance simultaneously. Qual. Eng. **17**, 109–118 (2005)
8. Knoth, S., Morais, M.C., Pacheco, A., Schmid, W.: Misleading signals in simultaneous residual schemes for the mean and variance of a stationary process. Comm. Stat. Theory Methods **38**, 2923–2943 (2009)
9. Lowry, C.A., Montgomery, D.C.: A review of multivariate control charts. IIE Trans. **27**, 800–810 (1995)
10. Machado, M.A.G., Costa, A.F.B.: Monitoring the mean vector and the covariance matrix of bivariate processes. Brazilian J. Oper. Produc. Manag. **5**, 47–62 (2008)
11. Mathai, A.M., Provost, S.B.: Quadratic Forms in Random Variables. Marcel Dekker, New York (1992)
12. Morais, M.C., Okhrin, Y., Pacheco, A., Schmid, W.: EWMA charts for multivariate output: some stochastic ordering results. Comm. Stat. Theory Methods **37**, 2653–2663 (2008)
13. Morais, M.C., Pacheco, A.: On the performance of combined EWMA schemes for μ and σ: a Markovian approach. Commun. Stat. Simul. Comput. **29**, 153–174 (2000)
14. Morais, M.C., Pacheco, A.: Misleading signals in joint schemes for μ and σ. In: Lenz, H.J., Wilrich, P.T. (eds.) Frontiers in Statistical Quality Control, vol. 16, pp. 100–122. Physica-Verlag, Heidelberg (2006)
15. Morais, M.J.C.: Stochastic ordering in the performance analysis of quality control schemes. Ph.D. thesis, Instituto Superior Técnico, Technical University of Lisbon (2002)

16. Ramos, P.F., Morais, M.C., Pacheco, A.: Misleading signals in simultaneous residual schemes for the process mean and variance of AR(1) processes: a stochastic ordering approach (accepted for publication in the international book series "Studies in Theoretical and Applied Statistics") (2012)
17. Ramos, P.F., Morais, M.C., Pacheco, A., Schmid, W.: Assessing the impact of autocorrelation in misleading signals in simultaneous residual schemes for the process mean and variance: a stochastic ordering approach (accepted for publication in "Frontiers on Statistical Quality Control 10") (2011)
18. Reynolds, M.R. Jr, Stoumbos, Z.G.: Monitoring the process mean and variance using individual observations and variable sampling intervals. J. Qual. Technol. **33**, 181–205 (2001)
19. Reynolds, M.R. Jr, Stoumbos, Z.G.: Control charts and the efficient allocation of sampling resources. Technometrics **46**, 200–214 (2004)
20. St. John, R.C., Bragg, D.J.: Joint X-bar R charts under shift in mu or sigma. ASQC Quality Congress Transactions — Milwaukee, pp. 547–550 (1991)
21. Tang, P.F., Barnett, N.S.: Dispersion control for multivariate processes. Aust. J. Stat. **38**(3), 235–251 (1996)
22. Tang, P.F., Barnett, N.S.: Dispersion control for multivariate processes – some comparisons. Aust. J. Stat. **38**(3), 253–273 (1996)
23. Tong, Y.L.: The Multivariate Normal Distribution. Springer, New York (1990)

On the Finite Dimensional Laws of Threshold GARCH Processes

Esmeralda Gonçalves, Joana Leite, and Nazaré Mendes-Lopes

Abstract

In this chapter we establish bounds for the finite dimensional laws of a threshold GARCH process, X, with generating process Z. In this class of models the conditional standard deviation has different reactions according to the sign of past values of the process. So, we firstly find lower and upper bounds for the law of $\left(X_1^+, -X_1^+, \ldots, X_n^+, -X_n^+\right)$, in certain regions of $\mathbb{R}^{2n}$, and use them to find bounds of the law of $(X_1, \ldots, X_n)$. Some of these bounds only depend on the parameters of the model and on the distribution function of the independent generating process, Z. An application of these bounds to control charts for time series is presented.

1 Introduction

As the true theoretical law of conditional heteroskedastic models is difficult to find, the most part of the analysis undertaken for these models is dedicated to the study of properties or probabilistic summaries of those laws. The use of these models, for instance within the quality control theory, needs the assessment of the probability of certain regions depending on the process. So, contrary to that trend, Pawlak and Schmid [5], Gonçalves, Leite and Mendes-Lopes [1], and Gonçalves and Mendes-Lopes [3] develop studies to find bounds for the finite dimensional laws of certain transformations of ARCH and TGARCH processes, respectively, here denoted X.

E. Gonçalves · N. Mendes-Lopes
CMUC, Department of Mathematics, University of Coimbra, 3001-454 Coimbra, Portugal
e-mail: esmerald@mat.uc.pt; nazare@mat.uc.pt

J. Leite (✉)
Institute of Accounting and Administration of Coimbra, Quinta Agrícola, Bencanta, 3040-316 Coimbra, Portugal
e-mail: jleite@iscac.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_24, © Springer-Verlag Berlin Heidelberg 2013

These bounds are expressed in terms of the distribution function of the independent generating process, Z, and it becomes clear that the marginal law of X is, in certain regions, strongly controlled by that of the process Z. This fact is rather relevant as we know that these laws have in general quite different characteristics; for example, the marginal law of X is leptocurtic even if it does not happen with that of the independent generating process. The application in time series of this kind of bounds on control charts with symmetric control limits has been explored in the literature [2, 6], due to the relevance of the target process distribution in the alarm signal definition as it makes possible to analyze the probability of having the process out of control in a given moment.

For a real stochastic process $X = (X_t,\ t \in \mathbb{Z})$ let us define $X_t^+ = X_t \mathbb{I}_{\{X_t \geq 0\}}$, $X_t^- = X_t \mathbb{I}_{\{X_t < 0\}}$ and $\underline{X_t}$ the sigma field generated by $X_t, X_{t-1}, \ldots$.

The process X follows a generalized threshold auto-regressive conditionally heteroskedastic model with orders p and q, TGARCH(p, q), if for real constants $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_i \geq 0$, $\gamma_j \geq 0$, $(i = 1, \ldots, q,\ j = 1, \ldots, p)$ and a sequence of independent and identically distributed real random variables, $(Z_t,\ t \in \mathbb{Z})$, with zero mean, unit variance and Z_t independent of $\underline{X_{t-1}}$ we have, for every $t \in \mathbb{Z}$,

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i X_{t-i}^+ - \sum_{i=1}^{q} \beta_i X_{t-i}^- + \sum_{j=1}^{p} \gamma_j \sigma_{t-j}. \end{cases}$$

The process $Z = (Z_t,\ t \in \mathbb{Z})$ is called the generating process of X. If $\gamma_j = 0$, $j = 1, \ldots, p$, we say that X follows a TARCH(q) model.

The main characteristic of threshold conditionally heteroskedastic models is the fact that they allow to take into account different reactions in the volatility according to the sign of the process values even for values with the same absolute size. So, these models capture the so-called leverage effect very common in financial time series of daily returns [4]. A not so very common, but yet still present characteristic in some daily returns series is skewness, positive in some cases and negative in others [7]. For these financial series, only control charts designed with asymmetric control limits are appropriate. So, following the ideas of Gonçalves and Mendes-Lopes [3] for the finite dimensional laws of the absolute value process of a TGARCH model, we first establish in Sect. 2 lower and upper bounds for the law of $\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)$, in certain regions of $\mathbb{R}^{2n}$. A relationship between the finite dimensional laws of the process X and the corresponding laws of $|X|$ and (X^+, X^-) processes is obtained in Sect. 3 from which we bound the distribution function of $(X_1, \ldots, X_n)$. Some of these bounds only depend on the parameters of the model and on the distribution function of the error process Z. These new results on the distribution function of $\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)$, are essential to generalize the study developedin Gonçalves et al. [2] to control charts with

asymmetric limits. A preliminary contribution for the evaluation of the run length of this kind of control charts is presented in Sect. 4. In the last section some concluding remarks are included.

2 Bounds for the Distribution Function of $(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-)$

Let $X = (X_t,\ t \in \mathbb{Z})$ be a TGARCH(p,q) process. In the following, the distribution function of Z_t is denoted by F_Z and, if the law of Z_t is absolutely continuous, f_Z denotes its density of probability function. We also consider the conventions $\sum_{i=1}^{0}(.) = 0$ and, for $n < q+1$, $\prod_{t=q+1}^{n}(.) = 1$.

If the law of Z_t is diffuse, we have the following upper bound for the distribution function of $\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)$:

$$F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}(x_1, x_1^*, \ldots, x_n, x_n^*) \leq \prod_{t=1}^{n}\left[F_Z\left(\frac{x_t}{\theta}\right) - F_Z\left(-\frac{x_t^*}{\theta}\right)\right]$$

for every $(x_1, x_1^*, \ldots, x_n, x_n^*) \in [0, +\infty[^{2n}$ and where $\theta = \alpha_0\left(1 + \sum_{j=1}^{p}\gamma_j\right)$.

This result follows easily taking into account the equality

$$F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}(x_1, x_1^*, \ldots, x_n, x_n^*) = P\left(-\frac{x_1^*}{\sigma_1} \leq Z_1 \leq \frac{x_1}{\sigma_1}, \ldots, -\frac{x_n^*}{\sigma_n} \leq Z_n \leq \frac{x_n}{\sigma_n}\right),$$

the inequality $\sigma_t \geq \alpha_0 + \sum_{j=1}^{p}\gamma_j\sigma_{t-j}$ as well as the independence and identical law of $Z_1, \ldots, Z_n$.

The determination of a lower bound of $F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}$ is more complex and it is subject to more restrictive hypotheses. The following theorem states a result for TARCH(q) and TGARCH (1,1) models. The general case of TGARCH(p, q) models is studied in an analogous way, with increased calculations complexity. Let us define the real function $h(x) = xf_Z'(x) + 2f_Z(x)$, $x \in \mathbb{R}$.

Theorem 2.1. *Let $X = (X_t,\ t \in \mathbb{Z})$ be a stationary* TGARCH(p, q) *process, with variance σ_X^2, such that the law of Z_t is absolutely continuous with a differentiable density of probability. For $(x_1, x_1^*, \ldots, x_n, x_n^*) \in [0, +\infty[^{2n}$ such that*

$$\forall y \geq 0, \quad g(y) = x_t h\left(\frac{x_t}{\alpha_0 + y}\right) + x_t^* h\left(-\frac{x_t^*}{\alpha_0 + y}\right) \geq 0,$$

where

$$t = \begin{cases} 1, \ldots, \min\{q, n\}, & \text{if } p = 0 \\ 1, \ldots, n, & \text{if } p = 1, \end{cases}$$

we have:

a) *if* $p = 0$ *and* $1 \leq q < n$,

$$F_{\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)}\left(x_1, x_1^*, \ldots, x_n, x_n^*\right) \geq \prod_{t=1}^{q}\left[F_Z\left(\frac{x_t}{u_t}\right) - F_Z\left(-\frac{x_t^*}{u_t}\right)\right] \prod_{t=q+1}^{n}\left[F_Z\left(\frac{x_t}{v_t}\right) - F_Z\left(-\frac{x_t^*}{v_t}\right)\right]$$

where $u_t = \alpha_0 + \sum_{i=1}^{t-1}\left(\alpha_i x_{t-i} + \beta_i x_{t-i}^*\right) + E\left(X_t^+\right)\sum_{i=t}^{q}\left(\alpha_i + \beta_i\right)$, $t = 1, \ldots, q$, *and* $v_t = \alpha_0 + \sum_{i=1}^{q}\left(\alpha_i x_{t-i} + \beta_i x_{t-i}^*\right)$, $t = q+1, \ldots, n$;

b) *if* $p = 0$ *and* $q \geq n \geq 1$,

$$F_{\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)}\left(x_1, x_1^*, \ldots, x_n, x_n^*\right) \geq \prod_{t=1}^{n}\left[F_Z\left(\frac{x_t}{u_t}\right) - F_Z\left(-\frac{x_t^*}{u_t}\right)\right]$$

where $u_t = \alpha_0 + \sum_{i=1}^{t-1}\left(\alpha_i x_{t-i} + \beta_i x_{t-i}^*\right) + E\left(X_t^+\right)\sum_{i=t}^{q}\left(\alpha_i + \beta_i\right)$, $t = 1, \ldots, n$;

c) *if* $p = 1$ *and* $q = 1$,

$$F_{\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)}\left(x_1, x_1^*, \ldots, x_n, x_n^*\right) \geq \prod_{t=1}^{n}\left[F_Z\left(\frac{x_t}{w_t}\right) - F_Z\left(-\frac{x_t^*}{w_t}\right)\right]$$

where $w_t = \alpha_0 \sum_{j=1}^{t-1}\gamma_1^{j-1} + \sum_{i=1}^{t-1}\gamma_1^{j-1}\left(\alpha_1 x_{t-j} + \beta_1 x_{t-j}^*\right) + \gamma_1^{t-1} E\left(\sigma_1\right)$, $t = 1, \ldots, n$.

Proof. We present the proof for the situation considered in **a)**, $p = 0$ and $1 \leq q \leq n$. Let $\left(x_1, x_1^*, \ldots, x_n, x_n^*\right) \in [0, +\infty[^{2n}$. We note that if $X_1^+ \leq x_1, -X_1^- \leq x_1^*, \ldots, X_n^+ \leq x_n, -X_n^- \leq x_n^*$ then

(i) for $t \in \{2, \ldots, q\}$, $\sigma_t \leq \alpha_0 + \sum_{i=1}^{t-1}\left(\alpha_i x_{t-i} + \beta_i x_{t-i}^*\right) + \sum_{i=t}^{q}\left(\alpha_i X_{t-i}^+ - \beta_i X_{t-i}^-\right) = S_t$;

(ii) for $t = 1$, $\sigma_1 = \alpha_0 + \sum_{i=1}^{q}\left(\alpha_i X_{1-i}^+ - \beta_i X_{1-i}^-\right) = S_1$;

(iii) and for $t \in \{q+1, \ldots, n\}$, $\sigma_t = \alpha_0 + \sum_{i=1}^{q}\left[\alpha_i X_{t-i}^+ + \beta_i\left(-X_{t-i}^-\right)\right] \leq \alpha_0 + \sum_{i=1}^{q}\left(\alpha_i x_{t-i} + \beta_i x_{t-i}^*\right) = v_t$.

So,

$$F_{\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)}\left(x_1, x_1^*, \ldots, x_n, x_n^*\right)$$

$$= P\left(-\frac{x_t^*}{\sigma_t} \leq Z_t \leq \frac{x_t}{\sigma_t},\, t=1,\ldots,q, -\frac{x_t^*}{\sigma_t} \leq Z_t \leq \frac{x_t}{\sigma_t},\, t=q+1,\ldots,n\right)$$

$$\geq P\left(-\frac{x_t^*}{S_t} \leq Z_t \leq \frac{x_t}{S_t},\, t=1,\ldots,q, -\frac{x_t^*}{v_t} \leq Z_t \leq \frac{x_t}{v_t},\, t=q+1,\ldots,n\right).$$

As S_t is $\underline{X}_0$-measurable ($t=1,\ldots,q$), v_t is nonrandom ($t=q+1,\ldots,n$), Z_t is independent of $\underline{X}_{t-1}$ and $\underline{X}_0 \subset \underline{X}_{t-1}$, $Z_1,\ldots,Z_n$ are independent and identically distributed and absolutely continuous, we get, using the properties of the mean and conditional mean,

$$P\left(-\frac{x_t^*}{S_t} \leq Z_t \leq \frac{x_t}{S_t}, t=1,\ldots,q, -\frac{x_t^*}{v_t} \leq Z_t \leq \frac{x_t}{v_t}, t=q+1,\ldots,n\right)$$

$$= E\left[P\left(-\frac{x_t^*}{S_t} \leq Z_t \leq \frac{x_t}{S_t}, t=1,\ldots,q, -\frac{x_t^*}{v_t} \leq Z_t \leq \frac{x_t}{v_t}, t=q+1,\ldots,n \mid \underline{X}_0\right)\right]$$

$$= \prod_{t=1}^{q} E\left[F_Z\left(\frac{x_t}{S_t}\right) - F_Z\left(-\frac{x_t^*}{S_t}\right)\right] \prod_{t=q+1}^{n} E\left[F_Z\left(\frac{x_t}{v_t}\right) - F_Z\left(-\frac{x_t^*}{v_t}\right)\right].$$

For t arbitrarily fixed in $\{1,\ldots,q\}$, let us consider the function $R_t\colon [0,+\infty[\longrightarrow [-1,1]$ defined by $R_t(y) = F_Z\left(\frac{x_t}{\alpha_0+y}\right) - F_Z\left(-\frac{x_t^*}{\alpha_0+y}\right)$. We have

$$\frac{\mathrm{d}^2 R_t}{\mathrm{d}y^2}(y) = \frac{1}{(\alpha_0+y)^3} g(y).$$

So, for every $t \in \{1,\ldots,q\}$, if $g(y) \geq 0$, for every $y \geq 0$, then R_t is a convex function. In these conditions we may apply Jensen inequality and we obtain

$$\prod_{t=1}^{q} E\left[F_Z\left(\frac{x_t}{S_t}\right) - F_Z\left(-\frac{x_t^*}{S_t}\right)\right] \prod_{t=q+1}^{n} E\left[F_Z\left(\frac{x_t}{v_t}\right) - F_Z\left(-\frac{x_t^*}{v_t}\right)\right]$$

$$\geq \prod_{t=1}^{q}\left[F_Z\left(\frac{x_t}{E(S_t)}\right) - F_Z\left(-\frac{x_t^*}{E(S_t)}\right)\right] \prod_{t=q+1}^{n}\left[F_Z\left(\frac{x_t}{v_t}\right) - F_Z\left(-\frac{x_t^*}{v_t}\right)\right]$$

and so

$$F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}(x_1,x_1^*,\ldots,x_n,x_n^*)$$

$$\geq \prod_{t=1}^{q}\left[F_Z\left(\frac{x_t}{u_t}\right) - F_Z\left(-\frac{x_t^*}{u_t}\right)\right] \prod_{t=q+1}^{n}\left[F_Z\left(\frac{x_t}{v_t}\right) - F_Z\left(-\frac{x_t^*}{v_t}\right)\right],$$

taking into account that $E(S_t) = u_t$. □

Remark 2.1. We point out that $E\left(X_t^+\right)$ and $E\left(\sigma_1\right)$ only depend on the coefficients of the model and on Z law. In fact, as X is second order stationary, it is also strictly stationary; so, $E\left(\sigma_t\right)$ is independent of t and we obtain

$$E\left(X_t^+\right) = E\left(\sigma_t\right) E\left(Z_t^+\right) = \frac{\alpha_0}{1 - E\left(Z_t^+\right) \sum_{i=t}^{q} \left(\alpha_i + \beta_i\right) - \sum_{i=t}^{p} \gamma_i} E\left(Z_t^+\right).$$

□

The result presented is valid for a large class of probability laws of the process Z. In order to get some insight on the sets where the lower bounds obtained for the distribution function of $\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)$ are valid, related to the positivity of the g function, we analyze the positivity of the h function considering two distributions particularly useful in the applications.

Example 2.1. For Z_t distributed according to the standard Gaussian law, we have $h(x) = \frac{1}{\sqrt{2\pi}} f_Z\left(x\right)\left[-x^2 + 2\right], x \in \mathbb{R}$, and so $h(x) \geq 0$ if $x \in \left[-\sqrt{2}, \sqrt{2}\right]$.

Example 2.2. We consider now that Z_t follows a centered and reduced Laplace law, that is, Z_t is absolutely continuous with density $f_Z\left(y\right) = \frac{\sqrt{2}}{2} \exp\left(-\frac{\sqrt{2}}{2}\left|y\right|\right),\ y \in \mathbb{R}$. We obtain, for $y > 0$, $h(y) = f_Z\left(y\right)\left(-\frac{\sqrt{2}}{2} y + 2\right)$ and $h(y) \geq 0$ if $y \leq 2\sqrt{2}$.

3 Bounds for the Distribution Function of $(X_1, \ldots, X_n)$

As a probabilistic application of the results developed in Sect. 2, we analyze now some bounds of the distribution function of $(X_1, \ldots, X_n)$ in the regions $]0, +\infty[^n$ and $]-\infty, 0[^n$. Let us consider $\theta = \alpha_0\left(1 + \sum_{j=1}^{p} \gamma_j\right)$.

If $(x_1, \ldots, x_n) \in]0, +\infty[^n$, we easily conclude that

$$F_{(X_1,\ldots,X_n)}\left(x_1, \ldots, x_n\right) \leq \prod_{t=1}^{n} F_Z\left(\frac{x_t}{\theta}\right).$$

A lower bound for $F_{(X_1,\ldots,X_n)}\left(x_1, \ldots, x_n\right)$ is found as a consequence of the following theorem.

Theorem 3.1. *Let $X = (X_t,\ t \in \mathbb{Z})$ be a* TGARCH(p, q) *process such that the law of Z_t is absolutely continuous. Then, for every $(x_1, \ldots, x_n) \in]0, +\infty[^n$,*

$$F_{(X_1,\ldots,X_n)}\left(x_1, \ldots, x_n\right) \geq$$

$$\geq \left[F_Z\left(0\right)\right]^n + \sum_{t=1}^{n} F_{\left(X_1^+, -X_1^-, \ldots, X_t^+, -X_t^-\right)}\left(x_1, 0, x_2, 0, \ldots, x_t, 0\right)\left[F_Z\left(0\right)\right]^{n-t}.$$

Proof. We have

$$\begin{aligned}
&F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) = P(X_1 \le x_1, X_t \le x_t,\, t=2,\ldots,n)\\
&= P(0 \le X_1 \le x_1, X_t \le x_t,\, t=2,\ldots,n) + P(X_1 < 0, X_t \le x_t,\, t=2,\ldots,n)\\
&= P(0 \le X_1 \le x_1, 0 \le X_2 \le x_2, X_t \le x_t,\, t=3,\ldots,n) +\\
&\quad + P(0 \le X_1 \le x_1, X_2 < 0, X_t \le x_t,\, t=3,\ldots,n) +\\
&\quad + P(X_1 < 0, 0 \le X_2 \le x_2, X_t \le x_t,\, t=3,\ldots,n) +\\
&\quad + P(X_1 < 0, X_2 < 0, X_t \le x_t,\, t=3,\ldots,n).
\end{aligned}$$

Repeating this reasoning we obtain

$$\begin{aligned}
&F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)\\
&\ge P(0 \le X_t \le x_t,\, t=1,\ldots,n) + P(X_t < 0,\, t=1,\ldots,n)\\
&\qquad + \sum_{t=1}^{n-1} P(0 \le X_i \le x_i,\, i=1,\ldots,t,\, X_i < 0,\, i=t+1,\ldots,n).
\end{aligned}$$

But

$$\begin{aligned}
P(0 \le X_t \le x_t,\, t=1,\ldots,n) &= P\left(X_t^+ \le x_t, -X_t^- \le 0,\, t=1,\ldots,n\right)\\
&= F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}(x_1,0,\ldots,x_n,0).
\end{aligned}$$

Further, as $P(X_t < 0,\, t=1,\ldots,n) = [F_Z(0)]^n$ and Z_i is independent of $\underline{X}_{i-1}, i \in \{t+1,\ldots,n\}$, then

$$\begin{aligned}
&P(0 \le X_i \le x_i,\, i=1,\ldots,t, X_i < 0,\, i=t+1,\ldots,n)\\
&= F_{\left(X_1^+,-X_1^-,\ldots,X_t^+,-X_t^-\right)}(x_1,0,\ldots,x_t,0)\,[F_Z(0)]^{n-t}.
\end{aligned}$$

In this way we obtain

$$\begin{aligned}
&F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) \ge F_{\left(X_1^+,-X_1^+,\ldots,X_n^+,-X_n^+\right)}(x_1,0,\ldots,x_n,0)\\
&\quad + [F_Z(0)]^n + \sum_{t=1}^{n-1} F_{\left(X_1^+,-X_1^-,\ldots,X_t^+,-X_t^-\right)}(x_1,0,\ldots,x_t,0)\,[F_Z(0)]^{n-t}.
\end{aligned}$$

□

Remark 3.1. A lower bound of $F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)$ depending only on the coefficients of the model and on the distribution function of Z may be obtained taking into account the results of the previous section. In fact, if we consider, for

example, that X follows a TARCH(q) model (with $q \leq n$) we get for $1 \leq t \leq n$ and under the conditions of Theorem 2.1,

$$F_{\left(X_1^+,-X_1^-,\ldots,X_t^+,-X_t^-\right)}(x_1,0,\ldots,x_t,0) \geq$$

$$\geq \begin{cases} \prod_{j=1}^{q}\left[F_Z\left(\frac{x_j}{u_j}\right)-F_Z(0)\right]\prod_{j=q+1}^{t}\left[F_Z\left(\frac{x_j}{v_j}\right)-F_Z(0)\right], & 1\leq q<t, \\ \prod_{j=1}^{t}\left[F_Z\left(\frac{x_j}{u_j}\right)-F_Z(0)\right], & t\leq q, \end{cases}$$

where $u_j = \alpha_0 + \sum_{i=1}^{j-1}\alpha_i x_{j-i} + E\left(X_t^+\right)\sum_{i=j}^{q}(\alpha_i+\beta_i)$ and $v_j = \alpha_0 + \sum_{i=1}^{q}\alpha_i x_{j-i}$. So, for a process X following a TARCH(q) model we have, if $(x_1,\ldots,x_n) \in]0,+\infty[^n$,

$$\begin{aligned} &F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) \\ &\geq [F_Z(0)]^n + \sum_{t=1}^{q}\left\{[F_Z(0)]^{n-t}\prod_{j=1}^{t}\left[F_Z\left(\frac{x_j}{u_j}\right)-F_Z(0)\right]\right\} \\ &\quad + \sum_{t=q+1}^{n}\left\{[F_Z(0)]^{n-t}\prod_{j=1}^{q}\left[F_Z\left(\frac{x_j}{u_j}\right)-F_Z(0)\right]\prod_{j=q+1}^{t}\left[F_Z\left(\frac{x_j}{v_j}\right)-F_Z(0)\right]\right\}. \end{aligned}$$

□

Let us now consider $(x_1,\ldots,x_n) \in]-\infty,0[^n$. We get

$$F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) \geq P\left(Z_1 \leq \frac{x_1}{\theta},\ldots,Z_n \leq \frac{x_n}{\theta}\right) = \prod_{t=1}^{n} P\left(Z_t \leq \frac{x_t}{\theta}\right) = \prod_{t=1}^{n} F_Z\left(\frac{x_t}{\theta}\right).$$

Otherwise, for $(x_1,\ldots,x_n) \in [0,+\infty[^n$, an upper bound of $F_{(X_1,\ldots,X_n)}(-x_1,\ldots,-x_n)$ can be obtained in terms of $F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)$ and of $F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)$, as stated in the next theorem.

Theorem 3.2. *Let $X = (X_t, t \in \mathbb{Z})$ be a* TGARCH(p,q) *process such that the law of Z_t is diffuse. Then, for every $(x_1,\ldots,x_n) \in [0,+\infty[^n$,*

$$F_{(X_1,\ldots,X_n)}(-x_1,\ldots,-x_n) \leq F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n) - F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n).$$

Proof. As

$$
\begin{aligned}
&F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)\\
&\quad = P_{(X_1,\ldots,X_n)}\left(\prod_{t=1}^{n}[-x_t,x_t]\right)\\
&\quad = P_{(X_1,\ldots,X_n)}\left(\prod_{i=1}^{n}]-\infty,x_i]\cap\left(\overline{\bigcup_{t=1}^{n}\left(\mathbb{R}^{t-1}\times]-\infty,-x_t[\times\mathbb{R}^{n-t}\right)}\right)\right),
\end{aligned}
$$

then

$$
\begin{aligned}
&F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)\\
&= F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)-P_{(X_1,\ldots,X_n)}\left(\bigcup_{t=1}^{n}\left(\prod_{i=1}^{t-1}]-\infty,x_i]\times]-\infty,-x_t[\times\prod_{i=t+1}^{n}]-\infty,x_i]\right)\right).
\end{aligned}
$$

As, for every $t=1,\ldots,n$,

$$
\left(\prod_{i=1}^{n}]-\infty,-x_i]\right)\setminus\{(-x_1,\ldots,-x_n)\}\subset\bigcup_{t=1}^{n}\left(\prod_{i=1}^{t-1}]-\infty,x_i]\times]-\infty,-x_t[\times\prod_{i=t+1}^{n}]-\infty,x_i]\right)
$$

and $P(X_1=-x_1,\ldots,X_n=-x_n)=0$, then

$$
\begin{aligned}
&P_{(X_1,\ldots,X_n)}\left(\bigcup_{t=1}^{n}\left(\prod_{i=1}^{t-1}]-\infty,x_i]\times]-\infty,-x_t[\times\prod_{i=t+1}^{n}]-\infty,x_i]\right)\right)\\
&\geq P_{(X_1,\ldots,X_n)}\left(\left(\prod_{i=1}^{n}]-\infty,-x_i]\right)\setminus\{(-x_1,\ldots,-x_n)\}\right)=F_{(X_1,\ldots,X_n)}(-x_1,\ldots,-x_n).
\end{aligned}
$$

So,

$$
F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)\leq F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)-F_{(X_1,\ldots,X_n)}(-x_1,\ldots,-x_n).
$$

□

We recall the relation

$$
F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)\leq\prod_{t=1}^{n}\left[F_{|Z|}\left(\frac{x_t}{\alpha_0}\right)\right],\quad (x_1,\ldots,x_n)\in]0,+\infty[^n,
$$

valid for a general TGARCH(p,q) (Gonçalves and Mendes-Lopes [3]) which may contribute to build bounds useful in practice, that is, only expressed in terms of the coefficients of the model and characteristics of the process Z. We point out that the statement in Theorem 3.1 of these upper and lower bounds of $F_{(X_1,\ldots,X_n)}(x_1,\ldots,x_n)$ when $(x_1,\ldots,x_n)\in]0,+\infty[^n$ or $(x_1,\ldots,x_n)\in$

$]-\infty, 0[^n$ do not demand the weak stationarity of the process X. Nevertheless, to get some bounds useful in practice the weak stationarity of X may be necessary. This is the case with lower bounds of $F_{(|X_1|,\ldots,|X_n|)}(x_1,\ldots,x_n)$ for some subfamilies of these models deduced in Gonçalves and Mendes-Lopes [3].

4 Application to Control Charts

In problems related to assessing the performance of control charts for conditionally heteroskedastic processes with symmetric marginal distribution, it is important to evaluate the probabilities $P(|X_t| \leq x_t,\ t = 1,\ldots,n)$ [5], concretely when evaluating the distribution of the run length [6]. With this goal, upper and lower bounds for $P(|X_t| \leq x_t,\ t = 1,\ldots,n)$ were established in Gonçalves and Mendes-Lopes [3] when X is a TGARCH(p,q) process. In particular, if X is a TARCH(q) process and under general conditions, they conclude that the probability of no alarm until time n in the in-control state, namely $P\left(\max_{1\leq t\leq n} \frac{|X_t|}{\sigma_X} \leq x\right)$, can be evaluated, for certain values of x, using the independent generating process Z. From these results Gonçalves et al. [2] developed theoretical and simulated studies for control charts with symmetric limits. The lower and upper bounds obtained in Theorem 2.1 for the distribution function of $\left(X_1^+, -X_1^-, \ldots, X_n^+, -X_n^-\right)$ are crucial to generalize this approach to control charts with asymmetric control limits, in the sense that, for $x, y \in \mathbb{R}$, the probability of no alarm until time n in the in-control state is

$$\begin{aligned}
&P\left(\min_{1\leq t\leq n} X_t^- \geq x,\ \max_{1\leq t\leq n} X_t^+ \leq y\right)\\
&= P\left(X_t^- \geq x,\ t = 1,\ldots,n,\ X_t^+ \leq y,\ t = 1,\ldots,n\right)\\
&= P\left(-X_t^- \leq -x,\ t = 1,\ldots,n,\ X_t^+ \leq y,\ t = 1,\ldots,n\right)\\
&= F_{\left(X_1^+,-X_1^-,\ldots,X_n^+,-X_n^-\right)}(y,-x,\ldots,y,-x).
\end{aligned}$$

These asymmetric limits became relevant when the marginal distribution of the generating process is skewed, as a symmetric marginal distribution of the generating process necessarily leads a symmetric marginal distribution of the conditionally heteroskedastic process.

For example, if $p = q = 1$ and $x < 0$ and $y > 0$ we obtain the following bounds for the probability of no alarm until time n in the in-control state

$$\begin{aligned}
&\prod_{t=1}^{n}\left[F_Z\left(\frac{y}{w_t}\right) - F_Z\left(\frac{x}{w_t}\right)\right]\\
&\leq P\left(\min_{1\leq t\leq n} X_t^- \geq x,\ \max_{1\leq t\leq n} X_t^+ \leq y\right) \leq \left[F_Z\left(\frac{y}{\theta}\right) - F_Z\left(\frac{x}{\theta}\right)\right]^n,
\end{aligned}$$

where $w_t = \alpha_0 (1 + \alpha_1 y - \beta_1 x) \frac{1-\gamma_1^{t-1}}{1-\gamma_1} + \gamma_1^{t-1} E(\sigma_1)$ and $\theta = \alpha_0 (1 + \gamma_1)$. We note that, if X becomes independent $(\alpha_1 = \beta_1 = \gamma_1 = 0)$, the equality holds for both bounds as we have $P\left(\min_{1\le t\le n} X_t^- \ge x, \max_{1\le t\le n} X_t^+ \le y\right) = \left[F_Z\left(\frac{y}{\alpha_0}\right) - F_Z\left(\frac{x}{\alpha_0}\right)\right]^n$.

5 Conclusions

The results here presented, namely the evaluation of the finite dimensional laws of the process $(X^+, -X^-)$, lead us to the theoretical analysis of no alarm until time n in the in-control state in a way that is not necessarily symmetrical.

Moreover, the results established in Sect. 2 were also used to bound the laws of the process taking into mind its application to control charts with unilateral control limits. The theoretical development of this study, particularly in the context of control charts, as well as its evaluation with simulated and real data are planned for future work. The generalization of these results to other models, for example with several thresholds or random thresholds, is also an open question with relevance in financial time series.

Acknowledgements We thank the referees for their constructive suggestions.

References

1. Gonçalves, E., Leite, J., Mendes-Lopes, N.: Sobre as leis marginais de processos GARCH com limiares. In: Ferrão, M., Nunes, C., Braumann, C. (eds.) Estatística – Ciência Interdisciplinar, Actas do XIV Congresso da SPE, pp. 445-454. Edições SPE (2007)
2. Gonçalves, E., Leite, J., Mendes-Lopes, N.: The ARL of modified Shewhart control charts for conditionally heteroskedastic models. Stat. Papers (2011) doi: 10.1007/s00362-011-0408-z
3. Gonçalves, E., Mendes-Lopes, N.: On the distribution of generalized threshold ARCH stochastic processes. Int. J. Pure Appl. Math. **35**, 397–419 (2007)
4. Malmsten, H., Terasvirta, T.: Stylized facts of financial time series and three popular models of volatility. Eur. J. Pure Appl. Math. **3**, 443–477 (2010)
5. Pawlak, M., Schmid, W.: On the distributional properties of GARCH processes. J. Time Ser. Anal. **22**, 339–352 (2001)
6. Severin, T., Schmid, W.: Monitoring changes in GARCH processes. Allg. Stat. Arch. **83**, 281–307 (1999)
7. Taylor, S.J.: Asset Price Dynamics, Volatility and Prediction. Princeton University Press, Princeton (2007)

Modelling the Duration of Multihop Paths in Mobile Ad Hoc Networks

Gonçalo Jacinto, Nelson Antunes, and António Pacheco

Abstract

Mobile ad hoc networks are characterized by having nodes that are cooperative and communicate without any kind of infrastructure. The mobility and multihop capability of these networks lead the network topology to change rapidly and unpredictably; this aspect must be incorporated in effective models to describe the dynamics of multihop paths.

When modeling the duration of multihop paths, a great part of the literature assumes that the links of multihop paths behave independently. This simplifies the modeling and reduces the complexity of computations. However, each link shares a common node with each of its neighbor links, turning the independent link assumption generally not valid. In this chapter, we use a piecewise deterministic Markov model that characterizes the random behaviour of a multihop path not assuming independent links. We obtain the mean path duration of multihop paths and compare the results for the used model with the ones obtained by assuming independent links. Numerical results illustrate that independent link approximation results underestimate the mean path duration, with the most significant differences being observed with low node mobility and higher path durations.

G. Jacinto (✉)
DMAT/CIMA-UE and ECT of University of Évora, Portugal
e-mail: gjcj@uevora.pt

N. Antunes
CEMAT and FCT of University of Algarve, Portugal
e-mail: nantunes@ualg.pt

A. Pacheco
CEMAT and Department of Mathematics, Instituto Superior Técnico - TULisbon, Portugal
e-mail: apacheco@math.ist.utl.pt

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_25, © Springer-Verlag Berlin Heidelberg 2013

1 Introduction

The demand for wireless communications is experiencing a steady growth. In this respect, the integration of Mobile Ad Hoc Networks (MANETs) multihop capability into wireless networks is one of the most promising architectural upgrade to envisage area coverage without significant additional infrastructure cost. The rapid deployment and low configuration profile of MANETs make them suitable to be used in emergency and military scenarios, as well in sensor networks and vehicular networks, among others.

In MANETs the nodes can dynamically form a network in a self-organizing manner without the need of an existing fixed infrastructure. Nodes are expected to act cooperatively in order to route traffic and to allow the network to adapt to the highly dynamic status of its links and node mobility patterns. In view of the limited transmission range of nodes, when the source and destination nodes are at a distance greater than the transmission range, the communication between them is made by a multiple hop path, using the neighbour nodes to forward the traffic towards the destination node.

Once a multihop path is active, the mobility of nodes causes the frequent failure of the path and activation of a new link. Thus, node mobility affects the performance of a MANET (cf., e.g., [2, 7]). Therefore, the development of models integrating mobility and the connectivity demands of MANETs are essential to characterize the reliability of these networks. The functionality of the network depends on the reliability of communications paths, and these are dependent on the stability of the links along the multihop path and their dependence structure, with the state of the links limited by power constraints and channel effects.

Few studies in the literature address the reliability of multihop paths, with an exact analysis of this issue seeming to be limited to [1]. Moreover, most analytical studies that focus on link stability extend the analysis for multihop paths assuming that the links of a multihop path behave independently of each other (e.g., [4, 7, 8]).

One of the earliest analysis that includes mobility was done by McDonald and Znati [7], who addressed link and path availability assuming independent links and that nodes move according to a variation of the random walk mobility model. Xu et al. [8] used a Markov chain mobility model to derive several path metrics assuming that links behave independently. Han et al. [4] proved that when the link count is large, the distribution of path duration converges to an exponential distribution. La and Han [6] relaxed the independence assumption for the links in [4] requiring that the dependence between links goes away asymptotically with increasing link count. In addition, Bai et al. [2] investigated path durations under different mobility models and routing protocols based on simulations.

An effective piecewise deterministic Markov model that describes the random behaviour of a multihop path in a MANET is proposed in [1]. We will use this model, which assumes that the links of a multihop path do not behave independently, to compute the mean path duration and compare the results obtained from this model with those obtained assuming that links behave independently.

The path is characterized through a Piecewise Deterministic Markov Processes (PDMP, see [3]) where, for simplicity, the mobility of each node along the path is given by the random walk mobility model. A PDMP is a Markov process that follows deterministic trajectories between random jumps, which occur either spontaneously, in a Poisson-like fashion, or when the process hits the boundary of its state space. The usage of a PDMP to model a multihop path arises naturally, since PDMPs are a mix of deterministic motion and random events, just like the multihop path dynamics.

The mean path duration is obtained as the unique solution of a set of ordinary differential equations (ODEs), that calculated by a recursive method allow to obtain numerical solutions of the mean path duration. The numerical results obtained for the mean path duration are compared with those obtained when assuming independent links. We show that the independent links assumption can lead to a large underestimation of the mean path duration, especially in scenarios where there is a small number of links and a low node mobility, which originates larger mean path durations.

The chapter is organized as follows. Section 2 describes the multihop path model. Section 3 presents a recursive method for the computation of the mean path duration. Numerical results are presented in Sect. 4. Finally, Sect. 5 concludes the chapter.

2 Multihop Path Model

We assume that a multihop path is set up (or already active) at time 0 with $N-1$ links and extends from node 1 along nodes $2, 3, \ldots$, until it reaches terminal node N. We consider a transmission range R equal for all nodes in the multihop path and, given two consecutive nodes in the path, $i-1$ and i, with locations in the plane l^{i-1} and l^i, respectively, they can communicate if $\|l^{i-1} - l^i\| < R$.

Each node in the path moves across the plane independently of other nodes according to a variation of the random walk mobility model. In this model it is assumed that a node alternates between two phases: *pause* (0) and *move* (1). If at a transition instant a node goes into phase i, the amount of time it stays in phase i is drawn independently of the past according to a continuous distribution function F_i with support on the set $\mathbb{R}^+$. We assume that the hazard rate function of F_i, denoted by $\lambda_i(t) = \mathrm{d}F_i(t)/(1 - F_i(t))$, is bounded on the positive reals.

Denote by p^j the phase of node j and by m^j its mobility vector if $p^j = 1$ (i.e. the node is in the move phase). When the phase of node j changes to move, the node picks a mobility vector according to a distribution function F_M on an open set S_M. Choosing a mobility vector m corresponds to choosing independently a direction θ and a velocity v through $m = (v\cos\theta, v\sin\theta)$. The node travels from the current location in the direction and with the velocity drawn for the mobility vector during the entire phase duration, with distribution F_1. Once this time expires, independently of the past, the node pauses for a random time period with distribution F_0 before starting to move again.

To obtain the duration of a link between two nodes, we need to define the relative location and relative mobility vector of node i with respect to node $i-1$, which are given, respectively, by $l_r^i = l^i - l^{i-1}$ and $m_r^i = m^i - m^{i-1}$, where m^i (resp. m^{i-1}) is omitted in the expression if $p^i = 0$ (resp. $p^{i-1} = 0$), and if both nodes are in pause phases $m_r^i = \mathbf{0}$ with $\mathbf{0} = (0,0)$. Then, the duration of the link i, is given by $d_{link}(l_r^i, m_r^i) = Z(l_r^i, m_r^i)/\|m_r^i\|$, for $m_r^i \neq \mathbf{0}$, where $Z(l_r^i, m_r^i)$ denotes the distance that node i needs to travel to move out of the range of node $i-1$.

The multihop path model is characterized by a vector of phase states governed by an alternating Markov renewal process, and by a vector of phase attributes. The phase states describe the state of each relay node (moving or paused) while the phase attributes describe the sojourn time in the current state, the relative location between two consecutive nodes and its movement characteristics: its velocity and direction. Thus, we obtain a process $\mathbf{X} = (\mathbf{P}, \mathbf{A})$ with $\mathbf{P} = (P^i)_{1 \leq i \leq N}$ where P^i is the phase process of node i, and $\mathbf{A} = (\mathbf{E}, \mathbf{M}, \mathbf{L}_r)$ denotes the joint attribute process, where $\mathbf{E} = (E^i)_{1 \leq i \leq N}$ denotes the elapsed time since the previous phase transition of node i, $\mathbf{M} = (M^i)_{1 \leq i \leq N}$ denotes the mobility vector of node i, if node i is in move phase, and is omitted if node i is in pause phase, and $\mathbf{L}_r = (L_r^i)_{2 \leq i \leq N}$ denotes the relative location process of node i with respect to node $i-1$. From the definition of $\mathbf{X}$, a state will be denoted by $\mathbf{x} = (\mathbf{p}, \mathbf{a})$ where $\mathbf{a} = (\mathbf{e}, \mathbf{m}, \mathbf{l}_r)$ with the vector $\mathbf{p} = (p^1, \ldots, p^N)$ containing the phases of nodes, $\mathbf{e} = (e^1, \ldots, e^N)$ the elapsed times of the nodes in their current phases, $\mathbf{m} = (m^1, \ldots, m^N)$ including the mobility vectors of the nodes, having dimension N when all $p^j = 1$ and with m^j omitted if $p^j = 0$, and $\mathbf{l}_r = (l_r^2, \ldots, l_r^N)$ the relative locations of nodes $2, 3, \ldots, N$ relative to nodes $1, 2, \ldots, N-1$, respectively.

When the process departs from a state $\mathbf{x}$, the flow of the process describes the deterministic trajectory of $\mathbf{X}$ until the next jump, and is characterized by $\phi(t, \mathbf{x}) = (\mathbf{p}, \phi_{\mathbf{p}}(t, \mathbf{a}))$ with

$$\phi_{\mathbf{p}}(t, \mathbf{a}) = (\mathbf{e} + t\mathbf{1}, \mathbf{m}, \mathbf{l}_r + t\mathbf{m}_r), \qquad t \in \mathbb{R}$$

representing the evolution of the component $\mathbf{a}$ over time, where $\mathbf{1}$ denotes a vector of $1's$ with dimension N and the vector $\mathbf{m}_r = (m_r^2, \ldots, m_r^N)$ contains the relative mobility vectors of nodes $2, 3, \ldots, N$ relative to nodes $1, 2, \ldots, N-1$, respectively.

For a given phase vector $\mathbf{p}$, denote the set

$$S_{\mathbf{p}} =]0, \infty[^N \times {S_M}^{\sum p_i} \times {S_L}^{N-1}$$

where $S_L = \{x \in \mathbb{R}^2 : \|x\| < R\}$, as the set of the possible values of the attribute process $\mathbf{A}$, that is, the set where all the links of the multihop path are active. Let $\partial S_{\mathbf{p}}$ denote the boundary of the set $S_{\mathbf{p}}$, $\partial^- S_{\mathbf{p}}$ denote the disjoint union of the set of boundary points that take the process into $S_{\mathbf{p}}$, and B denote the set of boundary points at which the multihop path process exits from $S_{\mathbf{p}}$. When the process hits a state in the boundary B, which represents the set of states through which the multihop path disconnects, it means that the path breaks and $\mathbf{X}$ jumps to

an absorbing state which is denoted by Δ. The state space of the joint process $\mathbf{X}$ is denoted by

$$S_{\mathbf{X}}^{\Delta} = S_{\mathbf{X}} \cup \{\Delta\},$$

where the set $S_{\mathbf{X}}$ denotes the disjoint union of the sets $S_{\mathbf{p}}^{-} = S_{\mathbf{p}} \cup \partial^{-} S_{\mathbf{p}}$.

For $\mathbf{x} \in S_{\mathbf{X}}$, define $d_{path}(\mathbf{x})$ as the path duration constrained to no phase transitions of the nodes when departing from state $\mathbf{x}$ (in another words, if the mobile nodes maintain their states, is the time to hit a state in B, the set of states at which the multihop path is disconnected). This time is equal to infinity if all nodes are in pause phase or all nodes have the same mobility vector.

The function $\boldsymbol{\lambda} : S_{\mathbf{X}}^{\Delta} \to \mathbb{R}_0^+$ is a measurable function that characterizes the jump rate (or transition rate) from each state of the process. For $\mathbf{x} \in S_{\mathbf{X}}$ the jump rate depends only on the phase and the time since the last phase transition of each node, and is given by the sum of the hazard rate functions $\lambda_i(t) = \mathrm{d}F_i(t)/(1 - F_i(t))$ of the phase duration distributions F_i, $\boldsymbol{\lambda}(\mathbf{x}) = \sum_{i=1}^{n} \lambda_{p^i}(e^i)$. Since $\boldsymbol{\lambda}(\mathbf{x})$ denotes the rate at which the process will leave from a given state $\mathbf{x}$, and we are considering that the state Δ (the state of the process when at least one of the links is disconnected) is an absorbing state, $\boldsymbol{\lambda}(\Delta) = 0$.

The evolution of $\mathbf{X}$ starting from state $\mathbf{x} \in S_{\mathbf{X}}$ can be constructed as follows. The survival function of the first jump time T_1 is defined by

$$G_{\mathbf{x}}(t) = \begin{cases} \exp\left(-\int_0^t \boldsymbol{\lambda}(\phi(s,\mathbf{x}))\mathrm{d}s\right) & t < d_{\text{path}}(\mathbf{x}) \\ 0 & t \geq d_{\text{path}}(\mathbf{x}) \end{cases} \tag{1}$$

and the state at an instant of time before the first jump is given by the deterministic evolution of the process,

$$\mathbf{X}(t) = \phi(t, \mathbf{x}), \qquad t < T_1.$$

If $T_1 = d_{\text{path}}(\mathbf{x})$ the path breaks since the process hits a state in B and the next state $\mathbf{X}(T_1)$ is Δ with probability 1; the process then stays in Δ forever since the jump rate out of Δ is zero. Otherwise, $T_1 < d_{\text{path}}(\mathbf{x})$, which means that one of the nodes in the path changes phase and the next state of the multihop process $\mathbf{X}(T_1)$ has distribution $Q(\phi(T_1, \mathbf{x}), \cdot)$, next defined. The function $Q : (S_{\mathbf{X}} \cup B) \times \mathscr{E} \to [0, 1]$ is the transition measure where $\mathscr{E}$ denotes the event space of $S_{\mathbf{X}}$, and is such that for $\mathbf{x} \in S_{\mathbf{X}}$, $Q(\mathbf{x}, \cdot)$ is a probability measure defined by

$$\begin{aligned} Q(\mathbf{x}, \{\mathbf{x}^{(j)}\}) &= \lambda_{p^j}(e^j)/\boldsymbol{\lambda}(\mathbf{x}); \quad p^j = 1 \\ Q(\mathbf{x}, \mathbf{d}\mathbf{x}^{(j)}) &= \lambda_{p^j}(e^j)/\boldsymbol{\lambda}(\mathbf{x}) F_M(\mathrm{d}m); \quad p^j = 0 \end{aligned} \tag{2}$$

where the new state of the process, $\mathbf{x}^{(j)}$, is next defined, and for $\mathbf{x} \in B$ we have $Q(\mathbf{x}, \{\Delta\}) = 1$. The interpretation of $\mathbf{P}$ and $\mathbf{A}$ makes it clear that from any $\mathbf{x} = (\mathbf{p}, \mathbf{a}) \in S_{\mathbf{X}}$ it is only possible to jump to a state where a node changes its phase

characteristics (phase, elapsed time in the phase, and mobility vector) and all the other values of the components remain the same. Thus, when $p^j = 1$, $\mathbf{x}^{(j)}$ coincides with $\mathbf{x}$ except for the fact that the phase of node j is a pause phase, $p^j = 0$, and the corresponding elapsed time in the phase is null, $e^j = 0$ (and its mobility vector will be omitted). Conversely, when $p^j = 0$, the new state of the process, $\mathbf{x}^{(j)}$, coincides with $\mathbf{x}$ except on the part concerning to node j whose phase becomes a move phase, $p^j = 1$, the corresponding elapsed time in the phase is null, $e^j = 0$, and the mobility vector is m. When the new state of the process is chosen, the process restarts from $\mathbf{X}(T_1)$ in a similar way. For a detailed description of the multihop path model, please see [1].

3 Mean Path Duration

Given the state of the multihop path process $\mathbf{x} \in S_{\mathbf{X}}^{\Delta}$, the mean path duration is denoted by

$$D(\mathbf{x}) = \mathbb{E}_{\mathbf{x}} \left(\int_0^{\infty} \mathbf{I}_{S_{\mathbf{X}}}(\mathbf{X}(s)) \mathrm{d}s \right) \tag{3}$$

where $\mathbf{I}_A$ is the indicator function of a set A. Note that since $\mathbf{I}_{S_{\mathbf{X}}}(\mathbf{X}(s))$ denotes the indicator function that the process $\mathbf{X}$ belonging to a state where all the links of the multihop path are connected, the integral in (3) denotes the amount of time that the process $\mathbf{X}$ remains connected, when departing from a state where all links are connected. Thus, the expected value, $D(\mathbf{x})$, is in fact the mean path duration of the multihop path process.

In [1] it is proved that the mean path duration written as the expectation of a functional of the multihop path process is the unique solution of a system of integro-differential equations. However, any direct method to solve them is quite problematic and depends very much on the specific characterization of the multihop path process (number of nodes, deterministic motion, jump rate, transition measure). To provide a way to calculate numerically $D(\mathbf{x})$, a recursive scheme is proposed for a feasible computation of the mean path duration.

Let D^0 be a function such that $D^0(\mathbf{x}) = 0$ for all $\mathbf{x} \in S_{\mathbf{X}}^{\Delta}$ and let $\mathscr{O}$ be an operator mapping the set of bounded measurable functions on $S_{\mathbf{X}}^{\Delta} \cup B$ into itself. The action of the operator $\mathscr{O}$ on D^0 originates the function $D^1 \equiv \mathscr{O} D^0$ defined by

$$D^1(\mathbf{x}) = \mathbb{E}_{\mathbf{x}} \left(\int_0^{T_1} \mathbf{I}_{S_{\mathbf{X}}}(X(s))\, \mathrm{d}s + D^0(\mathbf{X}(T_1)) \right), \mathbf{x} \in S_{\mathbf{X}}^{\Delta}.$$

Iterating $k(\geq 1)$ times the operator $\mathscr{O}$ on D^0 results into the function $D^k \equiv \mathscr{O}^k D^0$ given by

$$D^k(\mathbf{x}) = \mathscr{O} D^{k-1}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}} \left(\int_0^{T_1} \mathbf{I}_{S_{\mathbf{X}}}(X(s))\, \mathrm{d}s + D^{k-1}(\mathbf{X}(T_1)) \right)$$

for $\mathbf{x} \in S_{\mathbf{X}}^{\Delta}$.

The equation above signifies that, if the state of the process after $k-1$ phase transitions is known, the state of the process after k phase transitions is just given by the evolution of the process until the next phase change. The metric $D^k(\mathbf{x})$ denotes the mean path duration constrained to at most k jumps of the multihop process $\mathbf{X}$, when departing from state $\mathbf{x}$. As the number of jumps increases, in the limit, the mean path duration $D(\mathbf{x})$ defined by (3), is obtained

$$\lim_{k\to\infty} D^k(\mathbf{x}) = D(\mathbf{x}). \tag{4}$$

Let $f : S_{\mathbf{X}}^{\Delta} \cup B \to \mathbb{R}^+$ be a bounded measurable function and for $\mathbf{x} \in B$ define $f(\mathbf{x}) \equiv \lim_{t\downarrow 0} f(\phi(-t,\mathbf{x}))$. Denote the expected value of $\mathbf{X}$ just after a jump from $\mathbf{x}$ by

$$\mathscr{Q}f(\mathbf{x}) = \int_{S_{\mathbf{X}}^{\Delta}} f(\mathbf{y})Q(\mathbf{x},\mathbf{dy}),$$

where for $\mathbf{x} \in B$, $\mathscr{Q}f(\mathbf{x}) = f(\Delta)$.

For $\mathbf{x} \in S_{\mathbf{X}}$, the state of $\mathbf{X}$ after a short time t is, roughly, $\phi(t,\mathbf{x})$ with probability $1-\boldsymbol{\lambda}(\mathbf{x})t+o(t)$, while with probability $\boldsymbol{\lambda}(\mathbf{x})t+o(t)$ the process jumps to another state $\mathbf{X}(t)$ with transition measure Q and all other events have probability $o(t)$. Thus, we have

$$\mathbb{E}_{\mathbf{x}}(f(\mathbf{X}(t))) = (1-\boldsymbol{\lambda}(\mathbf{x})\boldsymbol{t})f(\phi(t,\mathbf{x})) + \boldsymbol{\lambda}(\mathbf{x})\boldsymbol{t}\mathscr{Q}f(\phi(t,\mathbf{x})) + o(t) \tag{5}$$

so that

$$\begin{aligned}&\frac{1}{t}\mathbb{E}_{\mathbf{x}}(f(\mathbf{X}(t)) - f(\mathbf{x}))\\ &\quad = \frac{1}{t}(f(\phi(t,\mathbf{x})) - f(\mathbf{x})) + \boldsymbol{\lambda}(\mathbf{x})\,(\mathscr{Q}f(\phi(t,\mathbf{x})) - f(\phi(t,\mathbf{x}))) + o(1).\end{aligned} \tag{6}$$

Denote by $\mathscr{A}f(\mathbf{x})$ the derivative of (5) where, in order to define the derivative of f with respect to the flow $\phi(t,\mathbf{x})$ in a rigorous way, we need to define its *phase function* $f_{\mathbf{p}} : S_{\mathbf{p}}^{-} \to \mathbb{R}^+$ by $f_{\mathbf{p}}(\mathbf{a}) = f(\mathbf{x})$. Therefore, as $t \to 0$ in (6) we obtain

$$\mathscr{A}f(\mathbf{x}) = \mathscr{V}f(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})(\mathscr{Q}f(\mathbf{x}) - f(\mathbf{x})),$$

for $\mathbf{x} \in S_{\mathbf{X}}$, where to simplify the notation we write $\mathscr{V}f(\mathbf{x})$ instead of the more accurate $\mathscr{V}_{\mathbf{p}}f_{\mathbf{p}}(\mathbf{a})$. Also any reference to a function $t \to f(\phi(t,\mathbf{x}))$ should be read as $t \to f_{\mathbf{p}}(\phi_{\mathbf{p}}(t,\mathbf{a}))$. The function $\mathscr{V}_{\mathbf{p}}$ is the vector field and $\phi_{\mathbf{p}}(t,\mathbf{a})$ is the unique integral curve of $\mathscr{V}_{\mathbf{p}}$ such that for almost all t,

$$\frac{\mathrm{d}}{\mathrm{d}t}f_{\mathbf{p}}(\phi_{\mathbf{p}}(t,\mathbf{a})) = \mathscr{V}_{\mathbf{p}}f_{\mathbf{p}}(\phi_{\mathbf{p}}(t,\mathbf{a})), \qquad \phi_{\mathbf{p}}(0,\mathbf{a}) = \mathbf{a}$$

is satisfied, since the function $t \to f_{\mathbf{p}}(\phi_{\mathbf{p}}(t,\mathbf{a}))$ is differentiable almost everywhere on $[0, d_{path}(\mathbf{x})[$.

The next result follows from Theorems 32.2 and 32.10 of Davis [3] conveniently applied to the expectation functional D^k of the PDMP $\mathbf{X}$ with finite time horizon and taking into account the specific boundary conditions.

Proposition 3.1. *Suppose that the function D^{k-1} is given. For each $\mathbf{x} \in S_{\mathbf{X}}$, $t \to D^k(\phi(t, \mathbf{x}))$ is an absolutely continuous function on $[0, d(\mathbf{x})[$ and D^k is the unique bounded solution of the equations*

$$\mathscr{V} f(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x}) \left(\mathscr{Q} D^{k-1}(\mathbf{x}) - f(\mathbf{x}) \right) = -1, \quad \mathbf{x} \in \mathbf{S}_{\mathbf{X}}, \tag{7}$$

and at a boundary state $\mathbf{x} \in B$, $f(\mathbf{x}) = f(\Delta) = 0$.

Note that in (7) the operator $\mathscr{Q}$ acts only on the given function D^{k-1}, so that the respective equations are ODEs. Combining this result with (4) provides a recursive way for computing the mean path duration D. For a complete proof of the results presented in this section, please see [5].

Computing D^k requires only to solve first order ordinary differential equations. The results of these calculations are then used to compute the next iteration $k + 1$. Since they are independent ODEs, they can be computed using parallel computation. The convergence of the solution depends on how large k has to be before D^k is close to D. Any direct implementation of these equations requires a discretization of the state space and solving at each grid point an independent ODE, providing the data for calculating the next iteration. It is unrealistic to hope that numerical solutions are possible for a medium size number of links in a single workstation due to the great number of computer processing cycles and the need of storing large amounts of data. However, it is possible to solve the equations in a single workstation in the case of one or two links and in one-dimensional ad hoc networks.

4 Numerical Results

In this section we illustrate an application of the preceding results and study the effect of the independence link assumption. The scenario proposed intends to model a scenario where nodes move with relatively low velocities and pause times. We consider that the phase durations are exponentially distributed with means of 30 s and 60 s in move and pause phase, respectively. The transmission range of a node is set up to 250 m. The mobility vector is obtained choosing a velocity (m/s) and direction of nodes uniformly distributed in $]1, 20[$ and $]0, 2\pi[$, respectively. For a multihop path with N nodes, initially each node i ($2 \leq i \leq N$) is deployed inside node $i - 1$'s radio coverage with an angle uniformly distributed in $]0, 2\pi[$ and with a distance following a triangular distribution in the interval (0,250) with mode 62.5. The initial phase of a node is picked randomly with probabilities proportional to the mean time spent in the phase.

Figure 1 shows the results of the mean path duration after each iteration for different link count. The departure states of the multihop path were sampled

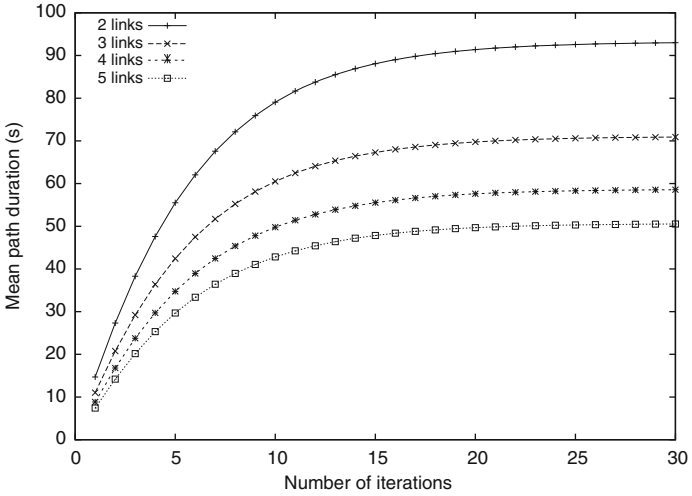

Fig. 1 Mean path duration after each iteration

according to the initial distribution, and their respective mean path durations were estimated in each iteration using Monte Carlo methods (in a single workstation). The results were averaged out in the final of each iteration. The difference between iterations gets smaller as the number of links increases since it gets more likely that a path failure occurs after a small number of phase transitions. However, all curves have converged before iteration 25.

In Fig. 2 we investigate the impact of neglecting the dependency between links in the mean path duration and study the impact of the mean time in pause phase on the mean path duration. Numerical routines were developed for independent links. We conclude that the mean path duration increases with the increase in the mean time in pause phase, and the independent link assumption leads always to an underestimation of the mean path duration. The observed percentage errors from assuming independent links were always higher than 3.5 % and achieved values higher than 25 %. The results are rather sensitive to the mean time in pause phase and getting an estimate for the mean path duration using the link independence assumption may in fact lead to a large bias when the inactive time of a node is large. Smaller values of the percentage error occur with paths with a higher number of hops and in scenarios where mobile nodes have higher mobility (no pause phase).

Figure 3 shows the impact of the mean time in move phase on the mean path duration. We conclude that with the increase in the mean time in move phase, the mean path duration decreases and also the percentage error arising from assuming independent links. We can again observe that higher percentage errors occur in scenarios where nodes have low mobility (inclusion of pause phases) and in multihop paths with a small number of hops.

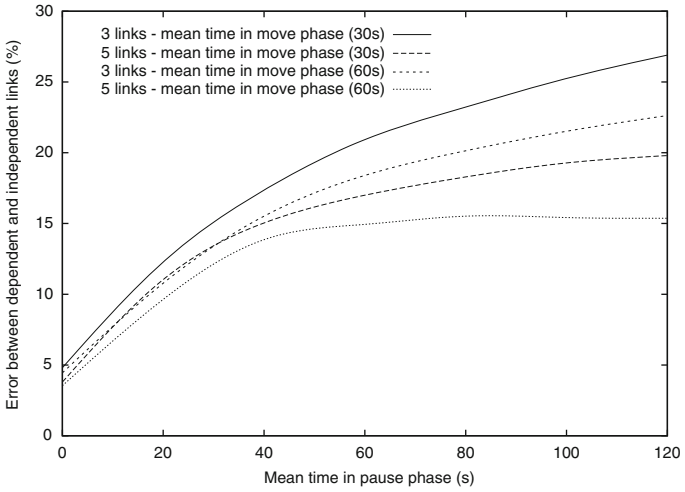

Fig. 2 Error (%) between mean path duration using dependent and independent links vs mean time in pause phase

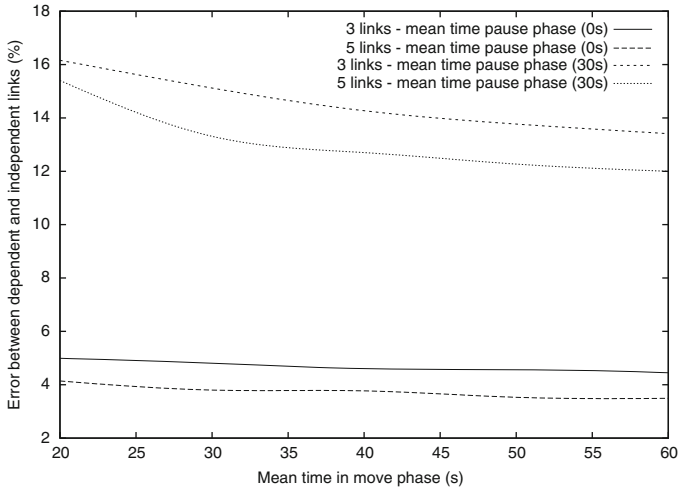

Fig. 3 Error (%) between mean path duration using dependent and independent links vs mean time in move phase

In Fig. 4 we observe the impact of the mobile nodes velocity on the mean path duration. We conclude that with the increase in the mean node velocity, leading to a decrease in mean path duration, the percentage error by assuming independent links increases, so higher node velocity associated with low node mobility (long

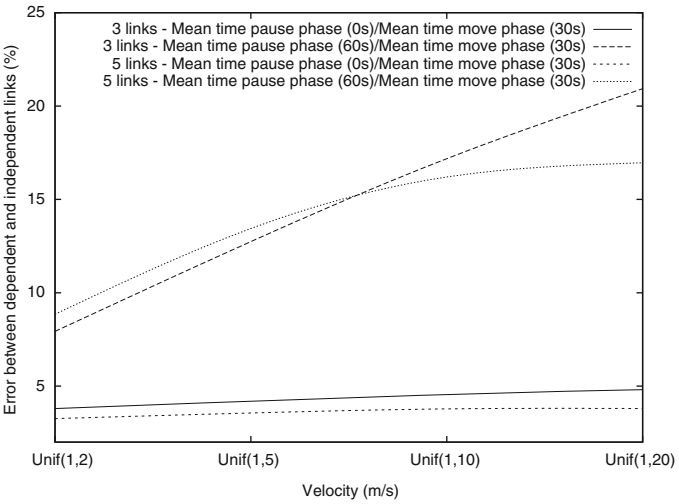

Fig. 4 Error (%) between mean path duration using dependent and independent links vs mobile node velocity

pause phases) increases the underestimation error of mean path durations when the independent links assumption is used.

5 Conclusion

This chapter uses an analytical framework to characterize the random behaviour of a multihop path under a PDMP that allows to describe the mean path duration through a set of ordinary differential equations and a recursive scheme for its computation. The results obtained using this model were compared with the corresponding results obtained assuming that links behave independently. We concluded that in scenarios with a small number of links, high velocities, and low node mobility (inclusion of long pause phases), the independent links assumption can lead to large underestimation of the mean path duration. In these scenarios, the percentage error can achieve values higher than 25 %. In the best case scenario, that is, with no pause phases, low velocities, and high number of links, the observed percentage errors were always higher than 3.0 %. These results can be used to improve the performance of routing algorithms in MANETs.

References

1. Antunes, N., Jacinto, G., Pacheco, A.: An analytical framework to infer multihop path reliability in MANETs. SIGMETRICS '10: Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 323–332, New York (2010)

2. Bai, F., Sadagopan, N., Krishnamachari, B., Helmy, A.: Modeling path duration distributions in MANETs and their impact on reactive routing protocols. IEEE J. Select. Areas Comm. **22**, 1357–1373 (2004)
3. Davis, M.: Markov Models and Optimization. Chapman and Hall, London/New York (1993)
4. Han, Y., La, R.J., Makowski, A.M., Lee, S.: Distribution of path durations in mobile ad-hoc networks: Palm's theorem to the rescue. Comput. Networks **50**, 1887–1900 (2006)
5. Jacinto, G.: Modeling and performance evaluation of mobile ad hoc networks. PhD thesis, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal (2011)
6. La, R.J., Han, Y.: Distribution of path durations in mobile ad hoc networks and path selection. IEEE/ACM Trans. Networks **15**, 993–1006 (2007)
7. McDonald, A.B., Znati, T.: A path availability model for wireless ad-hoc networks. Proceedings IEEE Wireless Communication Network Conference, vol. 1, pp. 35–40, Orleans, LA (1999)
8. Xu, S., Blackmore, K.L., Jones, H.M.: An analysis framework for mobility metrics in mobile ad hoc networks. EURASIP J. Wirel. Comm. Network. **1**, 26–26 (2007)

Risk Measures and Stochastic Orders Using Integrals of Distorted Quantile Functions

Miguel Mendes and Ignacio Cascos

Abstract

We construct new risk functionals and stochastic orderings based on integrals using discontinuous distortion functions. We prove a result on the subadditivity of the risk functional. As to the stochastic order we give a characterization of this order which allows us to construct spectral risk measures that are consistent with that order.

1 Introduction

In this chapter we call *integral of distorted quantile function* to all integrals of the form

$$\int_0^t q_X\left(\varphi\left(s\right)\right)\,\mathrm{d}s \tag{1}$$

where q_X is the pseudo-inverse (commonly known as quantile function) of the distribution function $F_X(x) := P(X \leq x)$ for a given probability measure P on a measure space $(\Omega, \mathscr{F})$ and $\varphi : [0,1] \to [0,1]$ is a non-decreasing function satisfying $\varphi(0) = 0$ and $\varphi(1) = 1$. This function is interpreted as a pseudo-inverse of a given *distortion function* $\psi : [0,1] \to [0,1]$ which generates a *distorted probability* $P^* := \psi \circ P$. The distortion function will be assumed to possess discontinuities and to the best of our knowledge this had not been considered previously.

M. Mendes (✉)
FEUP and CMUP, Universidade do Porto, Portugal
e-mail: mmendes@fc.up.pt

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

P.E. Oliveira et al. (eds.), *Recent Developments in Modeling and Applications in Statistics*, Studies in Theoretical and Applied Statistics,
DOI 10.1007/978-3-642-32419-2_26, © Springer-Verlag Berlin Heidelberg 2013

Integrals of this form appear naturally in the following classical construction. Let $X_1, \ldots, X_n$ be i.i.d. random variables. It can be easily seen that the expectation of $X_{1:n} := \min\{X_1, \ldots, X_n\}$ can be written as

$$\mathbf{E}(X_{1:n}) = \int_0^1 q_X\left(1-(1-s)^{1/n}\right) \mathrm{d}s$$

where $\varphi(s) := 1-(1-s)^{1/n}$ is the inverse function of $\psi(u) := 1-(1-u)^n$ and ψ is such that $P(X_{1:n} \le x) = \psi(P(X \le x))$ for all $x \in \mathbb{R}$.

Its relation with risk measures is also straightforward. If one sees X as a random variable representing the net worth after discounting of a certain portfolio and interprets $X_{1:n}$ as a worst case scenario over n simulations, the risk of $X_{1:n}$ calculated by means of the Expected Shortfall is given by the following integral:

$$ES_t(X_{1:n}) = -\frac{1}{t}\int_0^t q_X\left(1-(1-s)^{1/n}\right) ds\,.$$

In Sect. 2 we will make use of integrals of form (1) in a more general setting in order to obtain new risk functionals. These functionals will have an interesting decomposition formula resulting from a change of variables which is proved in the Appendix.

As far as stochastic orderings are concerned, we recall that the random variable X is dominated by Y *uniformly* if and only if for every $0 < t \le 1$

$$\int_0^t q_X(s)\,\mathrm{d}s \le \int_0^t q_Y(s)\,\mathrm{d}s\,.$$

Again, if we replace X by $X_{1:n}$ we obtain a new stochastic order which can be included in the generalization discussed in Sect. 3 (see also Cascos and Mendes [5]).

2 Risk Measures

In measuring the risk of a given portfolio modeled by a random variable X, risk functionals are employed with the purpose of calculating a number which represents a capital requirement that must be met so that X is considered acceptable (see the seminal paper by Artzner et al. [2]). These functionals typically satisfy a certain set of properties and according to several factors some may be preferred over others. A list of concepts which are usually considered important are the following:

1. *Monotonicity*: if $X \le Y$ then $\rho(X) \ge \rho(Y)$.
2. *Positive homogeneity*: $\rho(\lambda X) = \lambda\rho(X)$ for $\lambda > 0$.
3. *Law invariance*: if X and Y are identically distributed then $\rho(X) = \rho(Y)$.
4. *Cash-invariance*: $\rho(X+m) = \rho(X) - m$ for every $m \in \mathbb{R}$.
5. *Sub-additivity*: $\rho(X+Y) \le \rho(X) + \rho(Y)$.

6. *Comonotonic additivity*: if X and Y are comonotonic then

$$\rho(X+Y)=\rho(X)+\rho(Y).$$

We now introduce a risk functional based on the use of integrals of distorted quantile functions as aforementioned.

Definition 2.1. The φ-distorted Expected Shortfall is the functional

$$ES_t^{\varphi}(X):=-\frac{1}{t}\int_0^t q_X(\varphi(s))\,\mathrm{d}s$$

where φ is a pseudo-inverse of a given distortion function ψ.

At this moment, one issue must be clarified. If ψ is continuous then integrals of the form (1) which correspond to ES_1^{φ} coincide with the Choquet integral under the corresponding distorted probability P^*. And for $t<1$, the φ-distorted Expected Shortfall is nothing more than a truncated Choquet integral.

Let us now apply formula (3) in Appendix to the case when $f=q_X$ defined on an interval $[0,t]$ for $t\leq 1$ and $\psi:[0,1]\to[0,1]$ represents a distortion function with pseudo-inverse φ. Then

$$\int_0^{\psi(u^+)} q_X(\varphi(s))\,\mathrm{d}s=\int_0^u q_X(x)\,\psi'(x)\,\mathrm{d}x+\sum_{x\in[0,u]} q_X(x)\left\{\psi\left(x^+\right)-\psi\left(x^-\right)\right\}.$$

In order to replace $\psi(u^+)$ by any real number $t\in[0,1]$, we note that $u:=\varphi(t)$ can be such that $\psi(u^-)<t<\psi(u^+)$. Therefore

$$\begin{aligned}\int_0^t q_X(\varphi(s))\,\mathrm{d}s&=\int_0^{\psi(u^+)} q_X(\varphi(s))\,\mathrm{d}s-\int_t^{\psi(u^+)} q_X(\varphi(s))\,\mathrm{d}s\\&=\int_0^{\psi(u^+)} q_X(\varphi(s))\,\mathrm{d}s-q_X(u)\left\{\psi\left(u^+\right)-t\right\}\end{aligned}$$

Consequently

$$\int_0^t q_X(\varphi(s))\,\mathrm{d}s=\int_0^u q_X(x)\,\psi'(x)\,\mathrm{d}x+S_t(X)$$

where $u=\varphi(t)$ and $S_t(X):=\sum_{x\in[0,u)} q_X(x)\{\psi(x^+)-\psi(x^-)\}+q_X(u)\{t-\psi(u^-)\}$. Note that if X is bounded then $\sum_{x\in[0,u)} q_X(x)\{\psi(x^+)-\psi(x^-)\}<\infty$ because $\sum_{x\in[0,u)}\{\psi(x^+)-\psi(x^-)\}\leq 1$.

In terms of risk functionals we rewrite (2) as

$$ES_t^{\varphi}(X) = \Psi_t(X) + \Gamma_t(X)$$

where $\Psi_t(X) := -\frac{1}{t}\int_0^{\varphi(t)} q_X(u)\,\psi'(u)\,\mathrm{d}u$ and $\Gamma_t(X) := -\frac{1}{t}S_t(X)$.

We note without proof that properties 1, 2, 3, and 6 are satisfied by ES_t^{φ}, Ψ_t and Γ_t whereas 4 holds for ES_t^{φ} and Γ_t but not necessarily for Ψ_t. In fact, we have the following simple fact.

Proposition 2.1. *For a given $t \in (0,1]$, the risk functional Ψ_t is cash-invariant if and only if ψ is continuous in $[0, \varphi(t)]$.*

Proof. Clearly, for $m \in \mathbb{R}$, we have that

$$\Psi_t(X+m) = \Psi_t(X) - m \Leftrightarrow \int_0^{\varphi(t)} \psi'(u)\,\mathrm{d}u = t$$

which is equivalent to saying that

$$\psi\left(\varphi(t)^-\right) - \sum_{x\in[0,\varphi(t))} \left\{\psi\left(x^+\right) - \psi\left(x^-\right)\right\} = t\,. \tag{2}$$

By definition of pseudo-inverse functions we have that $\psi\left(\varphi(t)^-\right) \le t$. However, if $\psi\left(\varphi(t)^-\right) < t$ then (2) cannot hold since $\sum_{x\in[0,\varphi(t))}\{\psi(x^+) - \psi(x^-)\} \ge 0$. Therefore, $\psi\left(\varphi(t)^-\right) = t$ and hence $\sum_{x\in[0,\varphi(t))}\{\psi(x^+) - \psi(x^-)\} = 0$. Conversely, if ψ is continuous then, of course, $\int_0^{\varphi(t)} \psi'(u)\,\mathrm{d}u = t$. □

This result implies that the property of cash-invariance of Ψ_t is not of interest here because it would mean that $ES_t^{\varphi} \equiv \Psi_t$ as a consequence of the continuity of ψ. Obviously, ES_t^{φ} is a novelty as long as ψ is discontinuous.

As far as subadditivity is concerned we note that Γ_t is basically a sum of VaR measures rescaled by the jumps at the corresponding discontinuities of ψ. Recently, it was proved by Daníelsson et al. [6] that if X and Y have a jointly regularly varying nondegenerate tails with tail index $\alpha > 1$, then there is a sufficiently low probability level p such that

$$\mathrm{VaR}_p(X+Y) \le \mathrm{VaR}_p(X) + \mathrm{VaR}_p(Y)\,.$$

Therefore, assuming the same conditions on the distributions of random variables and choosing such a low level p as in Daníelsson et al. [6] the subadditivity of ES_t^{φ} is valid in the context of the following simple result.

Proposition 2.2. *If ψ has at most a finite number of discontinuities in $[0, \varphi(t)]$, then Ψ_t is sub-additive if and only if $\psi'_-(a) \ge \psi'_+(a)$ for all $a \in (0, \varphi(t))$.*

Proof. First note that Ψ_t can be written in the form

$$\Psi_t(X) = -\int_0^1 q_X(u)\,\frac{\psi'(u)}{t}\mathbf{I}_{\{0\le u\le\varphi(t)\}}\,\mathrm{d}u$$

which is a spectral representation of Ψ_t. By Acerbi [1] we know that Ψ_t is sub-additive if and only if

$$\int_{a-\varepsilon}^{a}\psi'(u)\,\mathrm{d}u \ge \int_{a}^{a+\varepsilon}\psi'(u)\,\mathrm{d}u$$

for all $a \in (0,\varphi(t))$ and all $\varepsilon > 0$ such that $[a-\varepsilon, a+\varepsilon] \subset [0,\varphi(t)]$. This is equivalent to saying that

$$\psi\left(a^-\right) - \psi\left[(a-\varepsilon)^+\right] - \sum_{x\in(a-\varepsilon,a)}\left\{\psi\left(x^+\right)-\psi\left(x^-\right)\right\} \ge$$
$$\ge \psi\left[(a+\varepsilon)^-\right] - \psi\left(a^+\right) - \sum_{x\in(a,a+\varepsilon)}\left\{\psi\left(x^+\right)-\psi\left(x^-\right)\right\}.$$

Since ε is arbitrary we can take ε sufficiently small so that a is the only possible point of discontinuity (here we have used the fact that points of discontinuity do not accumulate). Consequently, our previous condition reduces to

$$\psi\left(a^-\right) - \psi(a-\varepsilon) \ge \psi(a+\varepsilon) - \psi\left(a^+\right) \Leftrightarrow \psi'_-(a) \ge \psi'_+(a)$$

by letting $\varepsilon \downarrow 0$. Of course, if $\psi'_-(a) \ge \psi'_+(a)$ for all $a \in (0,\varphi(t))$, then condition (3) is satisfied. □

The result on the subadditivity of VaR mentioned above and the latter proposition yield the following corollary.

Corollary 2.1. *Let X and Y be random variables with jointly regularly varying nondegenerate tails with tail index $\alpha > 1$ and let p be sufficiently small so that $VaR_p(X+Y) \le VaR_p(X) + VaR_p(Y)$. Then, assuming that ψ has finitely many discontinuities $\{x_i\}_{i=1}^k$ such that $x_i < p$ for all $i \in \{1,\dots,k\}$ and $\psi'_-(a) \ge \psi'_+(a)$ for all $a \in (0,\varphi(t))$, we have that for all $0 < t \le 1$*

$$ES_t^{\varphi}(X+Y) \le ES_t^{\varphi}(X) + ES_t^{\varphi}(Y)$$

provided that all integrals are well defined.

Example 2.1. Market regulators impose capital requirements involving the VaR_λ measure usually for $\lambda = 0.01$ or $\lambda = 0.05$. Let us suppose that a conservative market player wants to mix VaR with a sub-additive measure. He/she can choose a percentage $p \in]0,1[$ of how much of the VaR measure will be used and devise

a distortion function ψ that has a discontinuity at $x = \lambda$ with jump equal to p and is such that $\psi'_{-}(x) \geq \psi'_{+}(x)$ for all $x \in [0,1]$. Moreover, in order to give more relevance to low quantiles ψ can me made to exhibit a higher derivative close to the origin. If in addition he/she chooses $t = 1$ the whole range of the portfolio is then covered in any risk measure of the type $ES_1^{\varphi}(X) = \Psi_1(X) + p.\mathrm{VaR}_{\lambda}(X)$. □

3 Stochastic Orders

We now define a stochastic order for random variables in the following manner.

Definition 3.1. Given two random variables X and Y, we say that $X \preceq_{\varphi} Y$ if and only if for all $t \in (0,1]$

$$\int_0^t q_X(\varphi(s))\,\mathrm{d}s \leq \int_0^t q_Y(\varphi(s))\,\mathrm{d}s\,.$$

where φ is again a pseudo-inverse of a distortion function ψ.

Our next result will be used in obtaining conditions for spectral measures to be consistent with $\preceq_{\varphi}$.

Proposition 3.1. *Given X and Y we have that $X \preceq_{\varphi} Y$ if and only if for all nonincreasing functions $f : (0,1) \to [0,\infty)$,*

$$\int_0^1 f(s)\,q_X(\varphi(s))\,\mathrm{d}s \leq \int_0^1 f(s)\,q_Y(\varphi(s))\,\mathrm{d}s\,.$$

Proof. Here we reproduce the argument in the proof of Lemma 3.45 in Föllmer and Schied [7]. Let us assume that

$$\int_0^t q_X(\varphi(s))\,\mathrm{d}s \leq \int_0^t q_Y(\varphi(s))\,\mathrm{d}s$$

for all $t \in (0,1]$. Without loss of generality we assume that f is left-continuous. Therefore, there exists a Radon measure μ such that $f(t) = \mu([t,1])$. By Fubini's theorem we have that

$$\begin{aligned}
&\int_0^1 f(s)\,q_X(\varphi(s))\,\mathrm{d}s = \int_0^1 \int_s^1 q_X(\varphi(s))\,\mu(\mathrm{d}t)\,\mathrm{d}s \\
&= \int_0^1 \int_0^t q_X(\varphi(s))\,\mathrm{d}s\,\mu(\mathrm{d}t) \leq \int_0^1 \int_0^t q_Y(\varphi(s))\,\mathrm{d}s\,\mu(\mathrm{d}t) \\
&= \int_0^1 f(s)\,q_Y(\varphi(s))\,\mathrm{d}s\,.
\end{aligned}$$

The converse implication follows by taking $f(s) = \mathbb{I}_{(0,t]}$. □

Let us consider now the problem of finding spectral measures([1])

$$M_\phi(X) := -\int_0^1 q_X(u)\,\phi(u)\,\mathrm{d}u$$

which are consistent with a given stochastic order $\preceq_\varphi$([2]). By this we mean that we want to find conditions on the risk aversion function ϕ such that $X \preceq_\varphi Y$ implies $M_\phi(X) \leq M_\phi(Y)$.

Suppose φ is a continuous and a.e. differentiable function (note that φ is also nondecreasing by assumption). Then, by the change of variables $u = \varphi(s)$ we have that $\int_0^1 q_X(u)\,\phi(u)\,\mathrm{d}u = \int_0^1 q_X(\varphi(s))\,\phi(\varphi(s))\,\varphi'(s)\,\mathrm{d}s$. Consequently, we conclude from Proposition 3.1 that if f, given by $f(s) := -\phi(\varphi(s))\,\varphi'(s)$, is nonincreasing, then M_ϕ must be consistent with $\preceq_\varphi$.

Example 3.1. Let us consider the case of the function $\varphi(s) = 1-(1-s)^{1/n}$ which corresponds to the stochastic order $\preceq_{icv_n}$ introduced in Cascos and Mendes [5]. Let $\mathscr{R}(\varphi)$ be the set of all risk-aversion functions ϕ such that $\phi(\varphi(\cdot))\,\varphi'(\cdot)$ is a nondecreasing function.

First, we see that $\phi(\varphi(s))\,\varphi'(s) = \frac{1}{n}\phi\left(1-(1-s)^{1/n}\right)(1-s)^{1/n-1}$ and that for $u \leq v$ we have that $\overline{u} := 1-(1-u)^{1/n} \leq 1-(1-v)^{1/n} := \overline{v}$. Consequently, $\phi(\varphi(\cdot))\,\varphi'(\cdot)$ being nondecreasing is equivalent to

$$\overline{u} \leq \overline{v} \Rightarrow \phi(\overline{u})\,(1-\overline{u})^{1-n} \leq \phi(\overline{v})\,(1-\overline{v})^{1-n}\ .$$

by the change of variable $\overline{s} := 1-(1-s)^{1/n}$. Therefore, $\phi \in \mathscr{R}(\varphi)$ if and only if ϕ is a risk aversion function and $\phi(x)\,(1-x)^{1-n}$ is nondecreasing in $[0,1)$, that is to say that $\phi(x) = g(x)\,(1-x)^{n-1}, x \in [0,1)$, for some nondecreasing function $g : [0,1) \to \mathbb{R}_0^+$. □

Appendix

Let $f : [a,b] \to \mathbb{R}$ be an a.e. differentiable function such that for every x both $f(x^+)$ and $f(x^-)$ exist. We define the *jump differential* of f, $\mathbf{d}f$, by its Lebesgue decomposition, i.e.

$$\mathbf{d}f := \mathrm{d}f + \delta f; \quad \mathrm{d}f \ll \lambda,\ \delta f \perp \lambda$$

[1] See Acerbi [1] for more details on the properties of spectral risk measures.

[2] For more details on consistency issues cf. Bäuerle and Müller [3].

where, as usual, $\mathrm{d}f = f'(x)\,\mathrm{d}\lambda$ and λ represents Lebesgue measure. As to the singular component, we define δf as the cumulative jump measure, that is, for every Borel set A, $\delta f(A) := \sum_{x\in A}\{f(x^+) - f(x^-)\}$.

In the following let $\psi : [c_1, c_2] \to [d_1, d_2]$ be an a.e. c^1 increasing function and let φ represent any of its pseudo-inverses. Since ψ is increasing we have that both $\psi(x^+)$ and $\psi(x^-)$ are well defined for all x. For convenience and ease of notation, we write $\psi\left(c_1^-\right) := \psi(c_1) = d_1$ and $\psi\left(c_2^+\right) := \psi(c_2) = d_2$.

Proposition 3.2. *Let $f : [d_1, d_2] \to \mathbf{R}$ be a Lebesgue integrable function. Then for every $\alpha, \beta \in [c_1, c_2]$ with $\alpha < \beta$ we have that*

$$\int_{\psi(\alpha^-)}^{\psi\left(\beta^+\right)} f \circ \varphi \,\mathrm{d}\lambda = \int_{\alpha}^{\beta} f \,\mathbf{d}\psi\,. \tag{3}$$

Proof. By a change of variables (see Billingsley [4] Theorem 16.13) we have that

$$\int_{\varphi^{-1}(A)} f \circ \varphi(x)\ \lambda(\mathrm{d}x) = \int_A f(x)\ \lambda\varphi^{-1}(\mathrm{d}x)$$

where, $\lambda\varphi^{-1}(S) := \lambda\left(\varphi^{-1}(S)\right)$ for every Lebesgue measurable set S. Since ψ is a pseudo-inverse of φ we do not necessarily have $\varphi^{-1}(S) = \psi(S)$. Nevertheless, $\varphi^{-1}([a,b])$ is an interval with endpoints $\psi(a^-)$ and $\psi(b^+)$. Therefore

$$\lambda\left(\varphi^{-1}([a,b])\right) = \int_{\varphi^{-1}([a,b])} \lambda(\mathrm{d}x) = \psi\left(b^+\right) - \psi\left(a^-\right)\ .$$

First we note that if A is an interval of the form $[a,b] = \cup_{i=1}^n [a_{i-1}, a_i]$ where ψ' is continuous in the interior of each $[a_{i-1}, a_i]$, then

$$\begin{aligned}\int_A \psi'(x)\ \mathrm{d}x &= \sum_{i=1}^{n}\left\{\psi\left(a_i^-\right) - \psi\left(a_{i-1}^+\right)\right\}\\ &= \psi\left(b^-\right) - \psi\left(a^+\right) - \sum_{i=1}^{n-1}\left\{\psi\left(a_i^+\right) - \psi\left(a_i^-\right)\right\}\ .\end{aligned}$$

And by writing $\psi(x^-) = \psi(x^+) - \{\psi(x^+) - \psi(x^-)\}$ for $x \in \{a, b\}$ we get

$$\int_A \psi'(x)\ \mathrm{d}x = \psi\left(b^+\right) - \psi\left(a^-\right) - \sum_{i=0}^{n}\left\{\psi\left(a_i^+\right) - \psi\left(a_i^-\right)\right\}$$

which implies that $\lambda\left(\varphi^{-1}([a,b])\right) = \int_A \psi'(x)\ dx + \sum_{i=0}^{n}\left\{\psi\left(a_i^+\right) - \psi\left(a_i^-\right)\right\}$. □

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

References

1. Acerbi, C.: Spectral measures of risk: A coherent representation of subjective risk aversion. J. Banking Finance **26**, 1505–1518 (2002)
2. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
3. Bäuerle, N., Müller, A.: Stochastic orders and risk measures: consistency and bounds. Insur. Math. Econ. **38**, 132–148 (2006)
4. Billingsley, P.: Probability and Measure, 3rd edn. Wiley Series in Prob. and Math. Stats. John Wiley & Sons, New York (1995)
5. Cascos, I., Mendes, M.: Lorenz curves of extrema. In: Borgelt, C., et al. (eds.) Combining Soft Computing and Statistical Methods in Data Analysis, pp. 81–88. Springer, Berlin (2010)
6. Daníelsson, J., Jorgensen, B., Samorodnitsky, G., Sarma, M., de Vries, C.: Fat tails, VaR and subadditivity. http://www.riskresearch.org/files/JD-Cd-BJ-SM-GS-35.pdf (2011). Accessed in March 2010
7. Föllmer, H., Schied, A.: Stochastic Finance, 2nd edn. de Gruyter Studies in Mathematics, Walter de Gruyter, Berlin, New York, p. 27 (2004)