

Carolina Marchant and Víctor Leiva

Chilean Journal of Statistics: A forum for the Americas and the World in COVID-19 pandemic

1

Ruth Burkhalter and Yuhlong Lio

Bootstrap control charts for the generalized Pareto distribution percentiles

3

Carlos López-Vázquez, Andrómaca Tasistro, and Esther Hochsztain

Exact tables for the Friedman rank test: Case with ties

23

Lucas de Oliveira Ferreira de Sales, André Luís Santos de Pinho, Francisco Moisés Cândido de Medeiros, and Marcelo Bourguignon

Control chart for monitoring the mean in symmetric data

37

Agatha S. Rodrigues, Vinicius F. Calsavara, Eduardo Bertolli, Stela V. Peres, and Vera L. D. Tomazella

Bayesian long-term survival model including a frailty term: Application to melanoma data

53

Abraão D.C. Nascimento, Kássio F. Silva, and Alejandro C. Frery

Distance-based edge detection on synthetic aperture radar imagery

71

Malinda Coa and Ernesto Ponsot

Alternatives to the logit model in the situation of factor levels aggregation in binomial responses

83

Francisco J. Ariza-Hernandez and Eduardo Gutiérrez-Peña

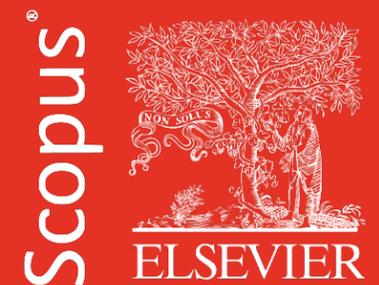
Bayesian analysis of an item response model with an AEP-based link function

103

CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

A free open access journal indexed by



Volume 12 Number 1
April 2021

ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society

SOCHÉ 
SOCIEDAD CHILENA DE ESTADÍSTICA

AIMS

The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (www.soche.cl). The ChJS takes the place of *Revista de la Sociedad Chilena de Estadística*, which was published from 1984 to 2000.

The ChJS covers a broad range of topics in statistics, as well as in artificial intelligence, big data, data science, and machine learning, focused mainly on research articles. However, review, survey, and teaching papers, as well as material for statistical discussion, could be also published exceptionally. Each paper published in the ChJS must consider, in addition to its theoretical and/or methodological novelty, simulations for validating its novel theoretical and/or methodological proposal, as well as an illustration/application with real data.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

EDITORS-IN-CHIEF

Víctor Leiva *Pontificia Universidad Católica de Valparaíso, Chile*
Carolina Marchant *Universidad Católica del Maule, Chile*

EDITORS

Héctor Allende Cid *Pontificia Universidad Católica de Valparaíso, Chile*
Danilo Alvares *Pontificia Universidad Católica de Chile*
José M. Angulo *Universidad de Granada, Spain*
Robert G. Aykroyd *University of Leeds, UK*
Narayanawamy Balakrishnan *McMaster University, Canada*
Michelli Barros *Universidade Federal de Campina Grande, Brazil*
Carmen Batanero *Universidad de Granada, Spain*
Ionut Bebu *The George Washington University, US*
Marcelo Bourguignon *Universidade Federal do Rio Grande do Norte, Brazil*
Márcia Branco *Universidade de São Paulo, Brazil*
Oscar Bustos *Universidad Nacional de Córdoba, Argentina*
Luis M. Castro *Pontificia Universidad Católica de Chile*
George Christakos *San Diego State University, US*
Enrico Colosimo *Universidade Federal de Minas Gerais, Brazil*
Gauss Cordeiro *Universidade Federal de Pernambuco, Brazil*
Francisco Cribari-Neto *Universidade Federal de Pernambuco, Brazil*
Francisco Cysneiros *Universidade Federal de Pernambuco, Brazil*
Mário de Castro *Universidade de São Paulo, São Carlos, Brazil*
José A. Díaz-García *Universidad Autónoma Agraria Antonio Narro, Mexico*
Raul Fierro *Universidad de Valparaíso, Chile*
Jorge Figueroa-Zúñiga *Universidad de Concepción, Chile*
Isabel Fraga *Universidade de Lisboa, Portugal*
Manuel Galea *Pontificia Universidad Católica de Chile*
Diego Gallardo *Universidad de Atacama, Chile*
Christian Genest *McGill University, Canada*
Viviana Giampaoli *Universidade de São Paulo, Brazil*
Marc G. Genton *King Abdullah University of Science and Technology, Saudi Arabia*
Patricia Giménez *Universidad Nacional de Mar del Plata, Argentina*
Hector Gómez *Universidad de Antofagasta, Chile*
Yolanda Gómez *Universidad de Atacama, Chile*
Emilio Gómez-Déniz *Universidad de Las Palmas de Gran Canaria, Spain*
Daniel Griffith *University of Texas at Dallas, US*
Eduardo Gutiérrez-Peña *Universidad Nacional Autónoma de México*
Nikolai Kolev *Universidade de São Paulo, Brazil*
Eduardo Lalla *University of Twente, Netherlands*
Shuangzhe Liu *University of Canberra, Australia*
Jesús López-Fidalgo *Universidad de Navarra, Spain*
Liliana López-Kleine *Universidad Nacional de Colombia*
Rosângela H. Loschi *Universidade Federal de Minas Gerais, Brazil*
Manuel Mendoza *Instituto Tecnológico Autónomo de México*
Orietta Nicolis *Universidad Andrés Bello, Chile*
Ana B. Nieto *Universidad de Salamanca, Spain*
Teresa Oliveira *Universidade Aberta, Portugal*
Felipe Osorio *Universidad Técnica Federico Santa María, Chile*
Carlos D. Paulino *Instituto Superior Técnico, Portugal*
Fernando Quintana *Pontificia Universidad Católica de Chile*
Nalini Ravishanker *University of Connecticut, US*
Fabrizio Ruggeri *Consiglio Nazionale delle Ricerche, Italy*
José M. Sarabia *Universidad de Cantabria, Spain*
Helton Saulo *Universidade de Brasília, Brazil*
Pranab K. Sen *University of North Carolina at Chapel Hill, US*
Giovani Silva *Universidade de Lisboa, Portugal*
Julio Singer *Universidade de São Paulo, Brazil*
Milan Stehlik *Johannes Kepler University, Austria*
Alejandra Tapia *Universidad Católica del Maule, Chile*
M. Dolores Ugarte *Universidad Pública de Navarra, Spain*

Chilean Journal of Statistics

VOLUME 12, NUMBER 1

APRIL 2021

CONTENTS

Carolina Marchant and Víctor Leiva <i>Chilean Journal of Statistics: A forum for the Americas and the World in COVID-19 pandemic</i>	1
Ruth Burkhalter and Yuhlong Lio <i>Bootstrap control charts for the generalized Pareto distribution percentiles</i>	3
Carlos López-Vázquez, Andr�maca Tasistro, and Esther Hochsztain <i>Exact tables for the Friedman rank test: Case with ties</i>	23
Lucas de Oliveira Ferreira de Sales, Andr� Lu�s Santos de Pinho, Francisco Mois�s C�ndido de Medeiros, and Marcelo Bourguignon <i>Control chart for monitoring the mean in symmetric data</i>	37
Agatha S. Rodrigues, Vinicius F. Calsavara, Eduardo Bertolli, Stela V. Peres, and Vera L. D. Tomazella <i>Bayesian long-term survival model including a frailty term: Application to melanoma datas</i>	53
Abra�o D.C. Nascimento, K�ssio F. Silva, and Alejandro C. Frery <i>Distance-based edge detection on synthetic aperture radar imagery</i>	71
Malinda Coa and Ernesto Ponsot <i>Alternatives to the logit model in the situation of factor levels aggregation in binomial responses</i>	83
Francisco J. Ariza-Hernandez and Eduardo Guti�rrez-Pe�a <i>Bayesian analysis of an item response model with an AEP-based link function</i>	103

TWELFTH VOLUME – FIRST NUMBER
EDITORIAL PAPER

Chilean Journal of Statistics: A forum for the Americas and the World in COVID-19 pandemic

CAROLINA MARCHANT¹ and VÍCTOR LEIVA²

¹Faculty of Basic Sciences, Universidad Católica del Maule, Talca, Chile

²School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

The first issue of the twelfth volume of the Chilean Journal of Statistics (ChJS) was published on 30 April 2021. The ChJS has more than a decade of life in its current version published in English and almost four decades from its origins in the *Revista de la Sociedad Chilena de Estadística*. We have published three issues in times of trial due to the COVID-19 pandemic. This pandemic has been very relevant for statistics because its use has allowed different governments to establish regulations to stop its spread. The World Health Organization declared a pandemic in March 2020. On 30 April 2021, more than 150 million cases have been confirmed, with more than 3.16 million deaths attributed to COVID-19, making it one of the deadliest pandemics in history. Medicine, science, statistics, and the generation of new knowledge have played a fundamental role, with scientific journals having a preponderant role in the publication of quality research. We believe that the world will overcome this situation, but we are sure that the new customs acquired during this period, such as interconnectivity, teleworking, teleconferencing, and virtuality, will remain with us forever. Although its development began long before this pandemic, the areas of big data, data science, machine learning, and statistics rose in prominence in 2020-2021. These areas play an important role in artificial intelligence, science, and engineering; indeed, in practically all areas of knowledge.

The scientific and editorial production of this volume would not have been achieved without the valuable contributions of many people. We are pleased to inform the international community that outstanding researchers, this time from the all America, have honored us by publishing their exciting work in our journal so the we acknowledge their relevant contributions. We are publishing articles written by colleagues from Argentina, Brazil, Chile, Ecuador, Mexico, Uruguay, Venezuela, and the United States (USA). We also thank all the anonymous reviewers who have contributed to maintaining ChJS' high-quality standards. Furthermore, we feel obliged and pleased to thank our prestigious editorial board, made up of colleagues from the five continents and listed in <http://chjs.mat.utfsm.cl/board.html>, who have collaborated from their positions to increase our visibility and quality of the works published by the ChJS. Of course, we must also thank the President and the Board of Directors of the Chilean Statistics Society (listed in <https://soche.cl/quienes-somos>) and the entire Chilean statistical community for placing on us, the Editors-In-Chief of The ChJS, their confidence in our work.

The first issue of the twelfth volume of the ChJS comprises seven articles written by, as mentioned, researchers from across the Americas. Details of these papers are as follows:

- (i) Our first paper is based on statistical process control and authored by Ruth Burkhalter and Yuhlong Lio from US, who designed control charts by using bootstrap methods for monitoring percentiles of the generalized Pareto distribution, and applied them to engineering data.
- (ii) In the second paper, considering non-parametric inference, Carlos López-Vázquez, Andrómaca Tasistro, and Esther Hochsztain, from Uruguay, presented exact tables for the Friedman rank test considering the case with ties, and applied them to geoportals navigation analysis, a web-based solution to provide open spatial data.
- (iii) The third paper is authored by Lucas de Oliveira Ferreira de Sales, André Luís Santos de Pinho, Francisco Moisés Cândido de Medeiros, and Marcelo Bourguignon, from Brazil, who introduced control charts used in the monitoring of the mean when the data are symmetrically distributed, and applied them to wine production data.
- (iv) Based on Bayesian survival analysis, Agatha S. Rodrigues, Vinicius F. Calsavara, Eduardo Bertolli, Stela V. Peres, and Vera L.D. Tomazella, also from Brazil, presented our fourth paper utilizing a Bayesian long-term survival model including a frailty term, which was applied to melanoma real data.
- (v) The fifth paper is authored by Abraão D.C. Nascimento, Kássio F. Silva, and Alejandro C. Frery, from Argentina and Brazil, who studied distance-based edge detection on synthetic aperture radar imagery.
- (vi) In the sixth paper, Malinda Coa and Ernesto Ponsot, from Venezuela, investigated alternatives to the logit model in the situation of factor levels aggregation in binomial distributed response variables.
- (vii) This issue of the ChJS closes with a paper authored by Francisco J. Ariza-Hernandez and Eduardo Gutiérrez-Peña, from Mexico, who performed a Bayesian analysis of an item response model with a link function based on the asymmetric exponential power distribution, and applied it to education data.

As the Chilean Statistics Society, we are proud because we continue to be an open-access journal, publishing works free of any article processing charges (APC). In addition, we are happy and pleased to be indexed to the Elsevier Scopus and Clarivate ISI WoS systems. We are very motivated because, at the beginning of 2021, we have already received 24 submissions from different countries.

Finally, we would like the international statistical and data-science communities, our editorial board, and our collaborators to champion the ChJS as a twelve-year, international, free of charges, and open-access journal, with fair and high-quality reviews, that cares about gender equality. Indeed, close to 50% of the papers published in this issue were authored by women, filling us with pride. We encourage the international scientific community to submit their works to the ChJS.

Carolina Marchant and Víctor Leiva
Editors-in-Chief
Chilean Journal of Statistics
<http://soche.cl/chjs>

STATISTICAL PROCESS CONTROL
RESEARCH PAPER

Bootstrap control charts for the generalized Pareto distribution percentiles

RUTH BURKHALTER* and YUHLONG LIO

Department of Mathematical Sciences, University of South Dakota, Vermillion, USA

(Received: 24 November 2020 · Accepted in final form: 06 January 2021)

Abstract

Lifetime percentile is an important indicator of product reliability. Recently, numerous quality control charts have been built for the quantiles of different distributions. Because of the positive support and flexibility, the Pareto distribution is one of the useful distributions to model lifetime. But the statistical quality control for the Pareto percentiles has not been considered. The current work aims to establish quality control charts for the Pareto distribution percentiles. The least squared error, maximum likelihood and a modified moment method estimators are proposed for monitoring the Pareto distribution percentiles. However, the sampling distributions of percentile estimators are neither known nor bell shape. As a result, the well-known Shewhart-type control chart may not be appropriately applied to monitor the Pareto distribution percentiles. The bootstrap procedure and normality approximations are proposed to establish control charts. An intensive Monte Carlo simulation study is conducted to compare the performance among the proposed bootstrap and Shewhart-type control charts. The simulation study shows that the bootstrap control chart based on the maximum likelihood estimator outperforms the rest control charts considered. Finally, a numerical example is utilized to illustrate the application of the bootstrap control chart based on maximum likelihood estimator.

Keywords: Average run length · False alarm rate · Quality control chart · Parametric bootstrap · Percentile.

Mathematics Subject Classification: Primary 62F40 · secondary 62P30.

1. INTRODUCTION

As product lifetime is a key aspect metric in industry, certain standards for the quality of a product lifetime are often required to prevent faulty or inferior products from reaching the consumer (Aykroyd et al., 2019). The statistical quality control charts have been very useful tools to improve product lifetime quality as well as reliability. Therefore, researchers have developed percentile control charts for many different lifetime distributions recently. For example, Lio and Park (2008) studied control charts for Birnbaum-Saunders percentiles, Lio and Park (2010) explored control charts for the inverse Gaussian percentiles, Lio et al. (2014) developed quality control charts for the Burr type-X percentiles, Rezac et al. (2015) developed percentile control charts for the Burr type-XII distribution that has been

*Corresponding author. Email: Ruth.Burkhalter@usd.edu

published in the ChJS and [Chiang et al. \(2017\)](#) investigated the percentile control charts for the generalized exponential distributions.

Since [Pickands \(1975\)](#) introduced the Pareto distribution, many authors have studied the properties of the Pareto distribution and eventually developed the two-parameter generalized Pareto distribution that has the rate and shape parameters. Some of these authors include [Hosking and Wallis \(1987\)](#), [Hüsler et al. \(2011\)](#), [Chen et al. \(2017\)](#) and [Salmasi and Yari \(2017\)](#). Due to the flexibility of the two-parameter generalized Pareto distribution with positive support, the generalized Pareto distribution would have been useful for lifetime modeling. However, based on our best knowledge, no any research work has addressed the generalized Pareto distribution percentiles. Therefore, the goal of this study is to investigate the quality control of the generalized Pareto distribution percentiles.

When the exact sampling distribution of the parameter estimator is not available, the approximated sampling distributions, such as asymptotic normal or bootstrap sample distribution, would be used to inference the parameter concerned. However, the asymptotic normal distribution is usually for large sample size case or the bell-shape sampling distribution. For the quality control chart established based on small sample size, the commonly used Shewhart-type control chart that is based on normal distribution may not be appropriate because the sampling distribution of a percentile estimator is usually not either known nor near a bell shape one. Hence, the parametric bootstrap procedure has been proposed to approximate the sampling distribution of the percentile estimator such that the control chart could be built. For more information, readers may refer to the aforementioned works on percentile quality control charts. For a thorough introduction to the bootstrap method, see [Gunter \(1992\)](#), [Efron and Tibshirani \(1993\)](#) and [Young \(1994\)](#). One distinct advantage of the bootstrap method is it allows the establishment of control chart limits when the sampling distribution of an estimator is unknown. This paper uses this fact extensively while using the least square error (LSE), maximum likelihood (ML) and modified moment method (MMM) estimations, respectively. A minor disadvantage could be the computational time of the bootstrap method. Recently, with access to more powerful computers, the runtime can be reduced to a reasonable amount.

While studying in different areas, many authors confirmed the superiority of the bootstrap method to the Shewhart chart and shown significant characteristics of bootstrapping. To name a few, [Nichols and Padgett \(2005\)](#) found that a parametric bootstrap chart could detect an out of control process faster than a Shewhart-type chart. [Lio and Park \(2008\)](#), [Lio and Park \(2010\)](#), [Lio et al. \(2014\)](#), [Rezac et al. \(2015\)](#) and [Chiang et al. \(2017\)](#) showed that bootstrap charts based on the maximum likelihood estimate or the moment method estimate performed better than the Shewhart-type chart when monitoring the lifetime percentiles. The above discussions motivate the current investigation of the parametric bootstrap control charts based on the ML, MMM and LSE estimators, respectively, for the generalized Pareto percentiles. Then, all the proposed parametric bootstrap control charts and the Shewhart-type chart are compared using the computer simulation.

In order to create the control charts, Section 2 presents the three different estimation methods, which include the ML, MMM and LSE methods, for the unknown distribution parameters. The procedures of the Shewhart-type and parametric bootstrap charts are addressed in Section 3. After the control charts are developed based on the aforementioned four different estimation methods, the average of run lengths (ARLs), standard error of the ARL (SEARL) as well as the average of upper control limits (UCLs) and lower control limits (LCLs), and their respective standard deviations, are obtained through a simulation study and used to compare and determine which method is the best for monitoring the generalized Pareto percentiles in Section 4. In this same section, a numerical example is given for the illustration purpose. Finally, some remarks and suggestions is addressed in Section 5.

2. PARAMETER ESTIMATION

In this section, we introduce the two-parameter generalized Pareto distribution and three different estimation methods.

2.1 THE GENERALIZED PARETO DISTRIBUTION

Let X be the random variable of the two-parameter generalized Pareto distribution that has the probability density function (PDF), cumulative distribution function (CDF) and percentile function respectively given as,

$$\begin{aligned} f(x; \alpha, \lambda) &= \alpha\lambda(1 + x\lambda)^{-(\alpha+1)}, \quad x > 0, \\ F(x; \alpha, \lambda) &= 1 - (1 + x\lambda)^{-\alpha}, \quad x > 0 \\ Q(p; \alpha, \lambda) &= \frac{1}{\lambda}((1 - p)^{-1/\alpha} - 1), \quad 0 < p < 1, \end{aligned} \quad (1)$$

where $\lambda > 0$ is the rate parameter and $\alpha > 0$ is the shape parameter. Three estimation procedures for the unknown distribution parameters and percentiles is presented next.

2.2 THE ML ESTIMATORS

Let X_1, \dots, X_n be a random sample of size n from the generalized Pareto distribution given in Equation (1). The corresponding log-likelihood function is given as

$$l(\alpha, \lambda) = n \log(\alpha) + n \log(\lambda) - (\alpha + 1) \sum_{i=1}^n \log(1 + X_i\lambda).$$

Setting the partial derivative of $l(\alpha, \lambda)$ with respect to α and λ equal to zero, respectively, two normal equations are obtained as

$$\begin{aligned} \frac{n}{\alpha} &= \sum_{i=1}^n \log(1 + X_i\lambda) \\ \frac{n}{\lambda} &= (\alpha + 1) \sum_{i=1}^n \frac{X_i}{1 + X_i\lambda}. \end{aligned} \quad (2)$$

The system of Equation (2) produces

$$\alpha = \frac{n}{\sum_{i=1}^n \log(1 + X_i\lambda)} \quad (3)$$

and

$$\frac{n}{\lambda} = \left(\frac{n}{\sum_{i=1}^n \log(1 + X_i\lambda)} + 1 \right) \sum_{i=1}^n \frac{X_i}{1 + X_i\lambda}. \quad (4)$$

The solution of λ to Equation (4) could be obtained by the unit-root function, unit-root, in R and labeled by $\hat{\lambda}_n$. Plugging $\hat{\lambda}_n$ into Equation (3), the solution $\hat{\alpha}_n$ is obtained. The solutions, $\hat{\alpha}_n$ and $\hat{\lambda}_n$, are called the ML estimates of α and λ , respectively. $\hat{\alpha}_n$ and $\hat{\lambda}_n$ may also be simultaneously obtained by optimization function `optim` of R. The ML estimate of the p th quantile can be stated as

$$\hat{Q}_n(p; \hat{\alpha}_n, \hat{\lambda}_n) = \frac{1}{\hat{\lambda}_n} \left((1-p)^{-1/\hat{\alpha}_n - 1} \right), \quad 0 < p < 1.$$

However, the exact sampling distributions of $\hat{\alpha}_n$, $\hat{\lambda}_n$ and $\hat{Q}_n(p; \hat{\alpha}_n, \hat{\lambda}_n)$ are unknown. Therefore, the exact quality control chart for $Q(p; \alpha, \lambda)$ cannot be established through $\hat{Q}_n(p; \hat{\alpha}_n, \hat{\lambda}_n)$. It can be shown that $\sqrt{n}((\hat{\alpha}_n, \hat{\lambda}_n) - (\alpha, \lambda)) \rightarrow N_2(\mathbf{0}, \mathbf{I}^{-1}(\alpha, \lambda))$ where N_2 is the bivariate normal distribution with mean vector as the two-dimension zero vector, $\mathbf{0}$ and two by two variance covariance matrix as the inverse of the Fisher information matrix, $\mathbf{I}(\alpha, \lambda)$, given as

$$\mathbf{I}(\alpha, \lambda) = -\frac{1}{n} \begin{bmatrix} \mathbb{E}\left(\frac{\partial^2 l(\alpha, \lambda)}{\partial \alpha^2}\right) & \mathbb{E}\left(\frac{\partial l(\alpha, \lambda)}{\partial \alpha \partial \lambda}\right) \\ \mathbb{E}\left(\frac{\partial l(\alpha, \lambda)}{\partial \lambda \partial \alpha}\right) & \mathbb{E}\left(\frac{\partial^2 l(\alpha, \lambda)}{\partial \lambda^2}\right) \end{bmatrix} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha^2} & \frac{1}{\lambda(\alpha+1)} \\ \frac{1}{\lambda(\alpha+1)} & \frac{1}{\lambda^2} - \frac{2}{\lambda^2(\alpha+2)} \end{bmatrix}.$$

More detail calculation procedures for the four entries of $\mathbf{I}(\alpha, \lambda)$ are as follows. The Fisher information matrix is presented as

$$\mathbf{I}(\alpha, \lambda) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix},$$

where $I_{11}, I_{12}, I_{21}, I_{22}$ can be obtained through

$$\begin{aligned} I_{11} &= -\frac{1}{n} \mathbb{E} \left(\frac{\partial^2 l(\alpha, \lambda)}{\partial \alpha^2} \right) = \frac{1}{\alpha^2} \\ I_{12} = I_{21} &= -\frac{1}{n} \mathbb{E} \left(\frac{\partial l(\alpha, \lambda)}{\partial \alpha \partial \lambda} \right) = -\frac{1}{n} \mathbb{E} \left(-\sum_{i=1}^n \frac{x_i}{1+x_i \lambda} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \int_0^\infty \frac{x}{1+x\lambda} f(x) dx = \alpha \lambda \int_0^\infty x(1+x\lambda)^{-\alpha-2} dx = \frac{1}{\lambda(\alpha+1)}, \\ I_{22} &= -\frac{1}{n} \mathbb{E} \left(\frac{\partial^2 l(\alpha, \lambda)}{\partial \lambda^2} \right) = -\frac{1}{n} \mathbb{E} \left(-\frac{n}{\lambda^2} + (\alpha+1) \sum_{i=1}^n \frac{X_i^2}{(1+x_i \lambda)^2} \right) = \frac{1}{\lambda^2} - \frac{2}{\lambda^2(\alpha+2)} \end{aligned}$$

It can be shown that

$$\frac{\hat{Q}_n(p; \hat{\alpha}_n, \hat{\lambda}_n) - Q(p; \alpha, \lambda)}{\sigma_{p,n}} \rightarrow N(0, 1), \quad 0 < p < 1,$$

where $\sigma_{p,n}^2 = (1/n) \nabla Q(p; \alpha, \lambda)^\top \mathbf{I}^{-1}(\alpha, \lambda) \nabla Q(p; \alpha, \lambda)$, for $0 < p < 1$, and $\nabla Q(p; \alpha, \lambda)$ is the gradient of $Q(p; \alpha, \lambda)$ with respect to α and λ . Thus, a Shewhart chart can be constructed using the asymptotic normal distribution to monitor the generalized Pareto percentile.

2.3 THE MMM ESTIMATORS

Given n sample observations, x_1, \dots, x_n , from the generalized Pareto distribution. In order to find the moment method estimates of α and λ , let the first order sample moment about zero be equal to the population mean and the second order sample moment about zero be equal to the population second moment about zero. Then, two required equations for moment method estimates can be expressed as

$$E(X) = \frac{1}{n} \sum_{i=1}^n x_i, \quad E(X^2) = \frac{1}{n} \sum_{i=1}^n x_i^2, \quad (5)$$

where X is the generalized Pareto distribution random variable. However, the solutions to the system of Equation (5) are difficult to obtain. Also the α solution is restricted to $\alpha > 2$ to ensure $E(X^2)$ finite. Hence, a modified moment method is needed so that the solutions are easier to solve and there is no restriction on α .

Let $U = \alpha \log(1 + \lambda X)$, then it can be easily shown that U has exponential distribution with mean equal 1. Let $X_{(1)} < \dots < X_{(n)}$ be the ordered statistic of X_1, \dots, X_n . Then, $U_{(1)} < \dots < U_{(n)}$ is the ordered statistic of $U_i = \alpha \log(1 + \lambda X_i)$ for $i = 1, \dots, n$. Denote $Y_1 = nU_{(1)}, \dots, Y_n = U_{(n)} - U_{(n-1)}$ or alternatively, $Y_1 = n\alpha \log(1 + \lambda X_{(1)}), Y_2 = (n-1)\alpha(\log(1 + \lambda X_{(2)}) - \log(1 + \lambda X_{(1)})), \dots, Y_n = \alpha(\log(1 + \lambda X_{(n)}) - \log(1 + \lambda X_{(n-1)}))$. It can be shown that Y_1, \dots, Y_n are random sample from the exponential distribution with mean equal to one. Let $g(\lambda) = 2 \sum_{i=1}^{n-1} (\log(T_n) - \log(T_i))$, where $T_i = \sum_{j=1}^i Y_j / \alpha$. It can be shown that $g(\lambda)$ has a chi-square distribution with degree of freedom $2n-2$ and αT_n has a gamma distribution $G(1, n)$. Following Wang (2008) and Rezac et al. (2015), λ can be estimated by the unique solution of $g(\tilde{\lambda}) = 2(n-2)$ and α can be estimated by $\tilde{\alpha} = (n-1)/T_n$. The estimates, $\tilde{\lambda}_n$ and $\tilde{\alpha}_n$, are called MMM estimations of λ and α , respectively. Then, the MMM estimator of the generalized Pareto percentile $Q(p, \alpha, \lambda)$ based on the MMM estimators, $\tilde{\alpha}_n$ and $\tilde{\lambda}_n$, is defined as $\tilde{Q}_n(p; \tilde{\alpha}_n, \tilde{\lambda}_n) = (1/\tilde{\lambda}_n) ((1-p)^{-1/\tilde{\alpha}_n} - 1)$, for $0 < p < 1$. However, the exact sampling distributions of $\tilde{\alpha}_n$, $\tilde{\lambda}_n$ and $\tilde{Q}_n(p; \tilde{\alpha}_n, \tilde{\lambda}_n)$ are unknown.

2.4 THE LSE ESTIMATORS

The LSE estimators of the generalized Pareto distribution parameters are obtained by minimizing the following sum of squares with respect to α and λ ,

$$\sum_{i=1}^n \left(F(X_{(i)}; \alpha, \lambda) - \frac{i}{n+1} \right)^2,$$

where $F(X_{(i)}; \alpha, \lambda) = 1 - (1 + X_{(i)}\lambda)^{-\alpha}$ for $i = 1, \dots, n$. The solutions of α and λ can be obtained simultaneously by optimization function, `optim`, in R and are labeled by $\bar{\alpha}_n$ and $\bar{\lambda}_n$, respectively. Then, the LSE estimator of $Q(p; \alpha, \lambda)$ is defined as $\bar{Q}_n(p; \bar{\alpha}_n, \bar{\lambda}_n) = (1/\bar{\lambda}_n) ((1-p)^{-1/\bar{\alpha}_n} - 1)$, for $0 < p < 1$. Again, the exact sampling distributions of $\bar{\alpha}_n, \bar{\lambda}_n$ and $\bar{Q}_n(p; \bar{\alpha}_n, \bar{\lambda}_n)$, respectively, are unknown.

3. STATISTICAL CONTROL CHARTS

In this section, we perform the Shewhart-type and parametric bootstrap charts.

3.1 ASSUMPTIONS

In Phase I, there are several assumptions, which are that the k in-control subgroup samples of size m are randomly collected from the generalized Pareto PDF of Equation (1) for the control chart setting. Let $n = m \times k$ denote the total sample size used in Phase I and $\hat{\alpha}_n$ and $\hat{\lambda}_n$ be the ML estimates of α and λ , respectively. The process for creating the Shewhart-type and parametric bootstrap charts is illustrated in the following sections.

3.2 THE SHEWHART-TYPE CHARTS

Using the ML estimation procedure described in Section 2.2, the ML estimate of the 100 p th percentile for $0 < p < 1$ based on each subgroup sample of size m from the Phase I process is $\hat{Q}_m(p; \hat{\alpha}_m, \hat{\lambda}_m) = (1/\hat{\lambda}_m)((1-p)^{-1/\hat{\alpha}_m} - 1)$ where $(\hat{\alpha}_m, \hat{\lambda}_m)$ is the ML estimates of (α, λ) . Then, the Shewhart-type chart for monitoring the 100 p th percentile, $Q(p; \alpha, \lambda)$, by using $\hat{Q}_m(p; \hat{\alpha}_m, \hat{\lambda}_m)$ for $0 < p < 1$ can be constructed with the steps:

- (1) Using all n sample observations from Phase I in-control process, the ML estimates of α and λ were obtained above. Then the asymptotic standard error of $\hat{Q}_m(p; \hat{\alpha}_m, \hat{\lambda}_m)$ can be estimated by

$$\hat{\sigma}_{\hat{Q}_m} = \sqrt{\frac{1}{m} \nabla Q^\top(p; \hat{\alpha}_n, \hat{\lambda}_n) \hat{\mathbf{I}}_n^{-1}(\hat{\alpha}_n, \hat{\lambda}_n) \nabla Q(p; \hat{\alpha}_n, \hat{\lambda}_n)}.$$

- (2) For the j th subgroup sample of size m , the ML estimates of α , λ and $Q(p; \alpha, \lambda)$ are found by using the procedure of Section 2.2 and denoted by $\hat{\alpha}_m^j$, $\hat{\lambda}_m^j$ and $\hat{Q}_m^j(p; \hat{\alpha}_m^j, \hat{\lambda}_m^j)$, respectively, for $j = 1, \dots, k$. The sample mean, $\tilde{Q}_m(p)$, of $\hat{Q}_m^j(p; \hat{\alpha}_m^j, \hat{\lambda}_m^j)$ for $j = 1, \dots, k$ is obtained as

$$\tilde{Q}_m(p) = \frac{1}{k} \sum_{j=1}^k \hat{Q}_m^j(p; \hat{\alpha}_m^j, \hat{\lambda}_m^j).$$

- (3) The control limits of the Shewhart-type chart are given as

$$\text{LCL}_{\text{SH}} = \tilde{Q}_m(p) - z_{(1-\gamma/2)} \hat{\sigma}_{\tilde{Q}_m}, \quad \text{UCL}_{\text{SH}} = \tilde{Q}_m(p) + z_{(1-\gamma/2)} \hat{\sigma}_{\tilde{Q}_m},$$

where $\tilde{Q}_m(p)$ is the center line (CL), $z_{1-\gamma/2}$ satisfies $\Phi(z_{1-\gamma/2}) = 1 - \gamma/2$ with $0 < \gamma < 1$, Φ is the standard normal CDF and γ is the false alarm rate (FAR).

After the control limits of the Shewhart-type chart are determined, future samples of size m (Phase II samples) are drawn from the generalized Pareto process to compute the plot statistic $\hat{Q}_m(p; \hat{\alpha}_m, \hat{\lambda}_m)$. If $\hat{Q}_m(p; \hat{\alpha}_m, \hat{\lambda}_m)$ is between the control limits found above, then the process is assumed to be in control. If not, signal that the process is out-of-control.

3.3 PARAMETRIC BOOTSTRAP CHARTS

The parametric bootstrap chart based on the ML estimation method is constructed as

- (1) Using all n observations collected during the Phase I in-control process, the ML estimates, $\hat{\alpha}_n$ and $\hat{\lambda}_n$, of α and λ were obtained above.
- (2) Generate m parametric bootstrap observations from the generalized Pareto distribution given in Equation (1), with $\alpha = \hat{\alpha}_n$ and $\lambda = \hat{\lambda}_n$. Denote the parametric bootstrap observations by x_1^*, \dots, x_m^* .

- (3) Find the ML estimates of α and λ using x_1^*, \dots, x_m^* from Step 2. The obtained ML estimates of α and λ are labeled by $\hat{\alpha}_m^*$ and $\hat{\lambda}_m^*$, respectively.
- (4) Find the bootstrap estimate of the 100 p th percentile, denoted $\hat{Q}_m^*(p; \hat{\alpha}_m^*, \hat{\lambda}_m^*)$ by plugging $\hat{\alpha}_m^*$ and $\hat{\lambda}_m^*$ into the quantile function, $Q(p; \alpha, \lambda)$, that is,

$$\hat{Q}_m^*(p; \hat{\alpha}_m^*, \hat{\lambda}_m^*) = Q(p; \hat{\alpha}_m^*, \hat{\lambda}_m^*) = \frac{1}{\hat{\lambda}_m^*} ((1-p)^{-1/\hat{\alpha}_m^*} - 1).$$

- (5) Repeat Steps 2 through 4 M times to obtain a size M bootstrap sample, $\hat{Q}_{m,j}^*(p; \hat{\alpha}_{m,j}^*, \hat{\lambda}_{m,j}^*)$, $j = 1, \dots, M$, where M is a given large positive integer.
- (6) Given a FAR, γ , find the $(\gamma/2)$ th and $(1-\gamma/2)$ th empirical quantiles of the bootstrap sample from Step 5 as the LCL and UCL, respectively, where the empirical quantiles can be obtained by using R quantile function. The CL is given as

$$\bar{Q}_m^*(p) = \frac{1}{M} \sum_{j=1}^M \hat{Q}_{m,j}^*(p; \hat{\alpha}_{m,j}^*, \hat{\lambda}_{m,j}^*).$$

The LCL and UCL developed above and the plot statistic, $\hat{Q}_m^*(p; \hat{\alpha}_m^*, \hat{\lambda}_m^*)$ is called the ML bootstrap chart. Following the same steps established in this section and replacing $\hat{\alpha}(\hat{\alpha}^*)$ and $\hat{\lambda}(\hat{\lambda}^*)$ by $\tilde{\alpha}(\tilde{\alpha}^*)$ and $\tilde{\lambda}(\tilde{\lambda}^*)$, respectively. Then, the corresponding MMM bootstrap chart is obtained. Similarly, replacing $\hat{\alpha}(\hat{\alpha}^*)$ and $\hat{\lambda}(\hat{\lambda}^*)$ by $\bar{\alpha}(\bar{\alpha}^*)$ and $\bar{\lambda}(\bar{\lambda}^*)$, respectively. Then, the corresponding LSE bootstrap chart is developed.

4. NUMERICAL STUDIES

In this section, based on the aforementioned four different estimation methods, we obtain the ARL, SEARL, UCL, LCL, and their respective standard deviations, through a simulation study and used to compare and determine which method is the best for monitoring the generalized Pareto percentiles. We close this section with a numerical example to show potential applications.

4.1 SIMULATION SCENARIO

To compare the performance among the proposed generalized Pareto distribution quantile control charts for monitoring lower quantiles below median, a Monte Carlo simulation study was executed using R, a programming language and environment originally developed by [Ihaka and Gentleman \(1996\)](#). The R code is available from the authors on request.

The performance quality of the control charts was based on the ARL and its SEARL. The average LCL and average UCL and their corresponding standard errors were also recorded for each method discussed in Section 2. From a practical standpoint, very few samples are available for lifetime testing in industry for quality control because the lifetime test is destructive and expensive. Hence, in the simulation, only sample sizes of $m = 4, 5$ and 6 with $k = 20$ subgroups were collected randomly. This simulation also considered a variety of false alarm rates (FARs), specifically, $0.1, 0.01, 0.0027$, and 0.002 . The control limits of 100 p th percentiles, where $p = 0.01, 0.05, 0.10$ and 0.25 , were found using the empirical distribution of $M = 10,000$ bootstrap observations for bootstrap control charts. The simulation process was repeated 10,000 times to find an accurate estimation of the ARL, SEARL, average of LCL, average of UCL and their respective standard errors running a self developed R program through the hp laptop with window 10. It took about 16.5

hours to run each one submission of R program for monitoring one percentile to product 10,000 LCLs, UCLs and run lengths for four control charts with FAR = 0.1, 0.01, 0.0027 and 0.002, respectively. The ARL, average LCL and average UCL are the average of 10,000 run lengths, LCLs and UCLs, respectively. The SEARL and standard deviations of the average of LCL and UCL are calculated by using standard deviations of 10,000 run lengths, LCLs and UCLs divided by squared root of 10,000, respectively.

4.2 SIMULATION RESULTS

In Tables 1 through 4, the ARLs and SEARLs are compared. An appropriate control chart has ARL near $1/\text{FAR}$ that is also known as the nominal ARL. The simulated ARLs and SEARLs for the Shewhart-type chart are shown in Table 1. The Shewhart-type chart overestimates the nominal ARL for FAR=0.1 and underestimates for FAR=0.01, 0.0027, and 0.002. That indicates overall narrow control limits except the case of FAR = 0.1. Table 2 shows the simulated ARLs and SEARLs for the LSE bootstrap chart. This process highly overestimates the nominal ARL. Table 3 shows the simulated ARLs and standard deviations for the MMM bootstrap chart. This chart is simply inconsistent. For FAR=0.1, it does fairly well across the percentiles tested. However, for smaller FAR, it underestimates the nominal ARL for smaller percentiles and overestimates for larger percentiles. Finally, Table 4 shows the simulated ARLs and SEARLs for the ML bootstrap chart. The ARLs in this chart stay close to the nominal ARL with small SEARL relative to the corresponding ARL across all percentiles and sample sizes used. All SEARLs shown in Tables 1 through 4 are very small compared with their respective ARLs. In Tables 5 through 8, the averages of LCLs and UCLs of each chart are compared. In Table 5, notice that the Shewhart-type chart has a negative average lower bound. Since the charts are used to monitor lifetime data, a negative lower bound implies that the normal approximation is not appropriate and the Shewhart-type chart is not appropriate for detecting a low percentile deteriorate. Also note that some of the average LCLs shown in Table 7 for the MMM bootstrap charts are set at 0+. Actually those numbers are very small positive. Again, the MMM bootstrap charts are not appropriate to use particularly monitoring a low percentile deteriorate. The calculated standard errors of average LCLs and UCLs for each control chart from 10,000 simulation runs are displayed in Tables 9 through 12. Some calculated standard errors for the LCLs and UCLs of the MMM bootstrap chart shown in Table 11 are 0+ that are actually very small positive numbers. Tables 9 through 12 show all standard errors are very small. As the Shewhart-type chart, the LSE bootstrap chart, and the MMM bootstrap chart have all been eliminated from consideration, the only chart left is the ML bootstrap chart. This chart's ARL stays close to the nominal ARL (Table 4) with small SEARL relative to the corresponding ARL, its LCLs and UCLs are reliable (Table 8), and the standard error for the averages of LCLs and UCLs are small. As a result, the ML bootstrap chart is assessed for monitoring the out of control.

Out of control testing analyzes how quickly the ML bootstrap chart detects a downward shift in the distribution percentiles. This type of downward shift indicates a product's lifetime is shortening. Looking at the Pareto quantile function, it is clear that as the α and (or) λ increase, the quantile function decreases. As a result, when running the out of control testing on ML bootstrap chart, the parameters α and (or) λ were shifted upwards. The results of this test are mainly based on the ARL and SEARL. To calculate ARL and SEARL for each out of control setting, the simulation study were conducted 10,000 runs and each run with 10,000 bootstrap sample observations. Table 13 through Table 15 display the simulation results. In Table 13, both α and λ shifted upwards. Let α_0 and λ_0 be the in-control parameter inputs and increase the values of α_0 and λ_0 to α_1 and λ_1 . In Table 14, λ_0 from the in-control process is fixed and α_0 from the in-control process

is increased to α_1 . In Table 15, α_0 from the in-control process is fixed and λ_0 from the in-control process is increased to λ_1 . In viewing of Table 13 through Table 15, it can be seen that all ARLs are relatively small compared to the nominal ARL and all SEARLs are very small, too. These results confirm that the ML bootstrap control chart is reliable for monitoring generalized Pareto percentiles.

Table 1. Shewhart-type in-control ARL estimates and corresponding standard deviations for generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	15.4038	0.3053	16.1360	0.2922	15.7627	0.2674
$p = 0.05$	15.2520	0.3026	15.9680	0.2886	15.6154	0.2619
$p = 0.10$	15.2592	0.3036	15.6528	0.2785	15.3232	0.2564
$p = 0.25$	14.6227	0.2501	15.1762	0.2680	14.6056	0.2441
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	37.9077	0.8082	43.0204	0.8511	45.0765	0.8593
$p = 0.05$	37.4736	0.7973	42.5070	0.8302	44.4339	0.8347
$p = 0.10$	37.0809	0.7489	41.5448	0.7988	43.4403	0.8132
$p = 0.25$	35.0311	0.6345	39.4974	0.7754	40.9858	0.7740
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	55.3873	1.1429	65.0647	1.3018	70.2976	1.3643
$p = 0.05$	54.7378	1.1404	64.7746	1.3200	70.2678	1.3980
$p = 0.10$	53.9932	1.1237	63.4313	1.2958	68.8031	1.3463
$p = 0.25$	50.3986	0.9611	59.6773	1.2202	64.3715	1.2834
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	59.9459	1.2528	70.5459	1.4171	78.2119	1.5724
$p = 0.05$	59.0701	1.2362	70.0550	1.4683	77.8436	1.5785
$p = 0.10$	58.2528	1.2071	68.6909	1.4198	76.4253	1.5344
$p = 0.25$	54.0512	1.0232	64.5977	1.3066	71.1732	1.4163

Table 2. LSE bootstrap in-control ARLs estimates and corresponding standard deviations for the generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	12.740	0.1918	12.7904	0.1954	12.7874	0.1900
$p = 0.05$	12.7783	0.1901	12.7938	0.1982	12.7571	0.1914
$p = 0.10$	12.7063	0.1947	12.9092	0.1978	12.8133	0.1892
$p = 0.25$	13.1900	0.2027	13.1257	0.2076	13.2774	0.2036
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	179.819	3.4544	186.3573	3.6381	189.9276	3.8459
$p = 0.05$	180.4755	3.4162	185.7970	3.7856	195.4479	4.1130
$p = 0.10$	180.9999	3.4464	187.6843	3.7808	190.5604	3.9170
$p = 0.25$	184.2501	3.6159	188.6343	3.7525	194.9903	4.0870
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	788.459	17.7730	812.0935	18.4439	844.5480	19.1808
$p = 0.05$	788.6242	17.1870	821.3853	18.4914	870.5871	19.7251
$p = 0.10$	771.9857	16.2963	819.5059	17.9659	852.3299	19.8192
$p = 0.25$	796.4193	17.0323	807.9862	17.5517	872.3674	21.6409
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	1120.381	24.9256	1165.7726	29.7523	1234.9286	30.7898
$p = 0.05$	1127.6833	24.9995	1172.5420	28.1306	1245.6503	29.3309
$p = 0.10$	1101.3921	23.7495	1193.0949	29.3921	1221.7993	28.5823
$p = 0.25$	1118.6962	24.9399	1148.6458	27.4038	1272.1061	31.8478

Table 3. MMM bootstrap in-control ARL estimates and corresponding standard deviations for the generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	9.6570	0.1335	9.7628	0.1373	9.6871	0.1332
$p = 0.05$	9.3388	0.1352	11.6867	0.1609	10.9639	0.1515
$p = 0.10$	8.7744	0.1272	13.2327	0.2123	10.3124	0.1400
$p = 0.25$	9.6432	0.1406	12.4179	0.2023	10.7797	0.1503
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	49.7678	0.7441	55.1232	0.8122	58.5411	0.8671
$p = 0.05$	89.8256	1.2590	116.3314	1.5610	88.5108	1.2978
$p = 0.10$	109.3266	1.9285	102.2867	1.2401	114.7982	1.6952
$p = 0.25$	102.24451	1.8508	113.6496	1.4729	137.2954	2.4680
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	60.9588	0.8813	68.8012	0.9745	74.45920	1.0695
$p = 0.05$	147.0226	2.2652	191.4860	3.0285	186.7040	2.4625
$p = 0.10$	445.3300	8.0744	383.9819	7.0894	427.8512	6.3043
$p = 0.25$	377.1021	7.2660	452.4614	7.5417	467.0340	7.4687
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	62.1741	0.8982	70.4454	0.9950	76.7008	1.1057
$p = 0.05$	156.1286	2.3482	201.7414	2.9981	202.7630	2.5971
$p = 0.10$	561.8492	10.0323	720.5037	14.6875	564.9076	9.8149
$p = 0.25$	589.4910	9.4193	606.7636	9.4168	576.9963	10.5925

Table 4. ML bootstrap in-control ARL estimates and corresponding standard deviations for generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	9.3209	0.1309	9.0851	0.1250	9.2029	0.1266
$p = 0.05$	9.2294	0.1313	9.2691	0.1294	9.2289	0.1295
$p = 0.10$	9.2964	0.1316	9.1935	0.1302	9.3212	0.1292
$p = 0.25$	9.1746	0.1280	9.3039	0.1285	9.2309	0.1278
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	91.6366	1.6801	89.7949	1.5565	88.8241	1.4242
$p = 0.05$	90.2183	1.6010	90.1718	1.5683	88.2176	1.4181
$p = 0.10$	90.3458	1.6343	91.3064	1.5865	90.5653	1.4959
$p = 0.25$	93.2211	1.6978	90.6854	1.5839	87.6807	1.4678
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	343.0466	7.0766	336.3652	6.4181	335.8107	6.2239
$p = 0.05$	339.5141	6.7849	333.6685	6.5899	328.6546	6.2053
$p = 0.10$	337.7009	6.8637	339.3910	6.4750	340.7111	6.4211
$p = 0.25$	355.4715	7.5188	340.8538	6.7502	330.1604	6.1004
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	472.3910	10.1697	456.5624	9.1939	449.1549	8.4255
$p = 0.05$	470.6944	10.0895	452.7038	9.4391	445.0275	8.8036
$p = 0.10$	459.5464	9.7480	464.8904	9.3621	458.1780	8.9725
$p = 0.25$	488.0070	10.7023	464.8904	10.4099	450.7093	8.8848

Table 5. Shewhart-type in-control LCL and UCL for generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	-0.0008	0.0104	-0.0003	0.0097	0.0001	0.0092
$p = 0.05$	-0.0037	0.0533	-0.0013	0.0495	0.0005	0.0469
$p = 0.10$	-0.0073	0.1096	-0.0022	0.1019	0.0016	0.0966
$p = 0.25$	-0.0152	0.3002	-0.0018	0.2800	0.0085	0.2658
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	-0.0039	0.0136	-0.0031	0.0125	-0.0025	0.0118
$p = 0.05$	-0.0199	0.0694	-0.0157	0.0639	-0.0126	0.0601
$p = 0.10$	-0.0403	0.1427	-0.0316	0.1314	-0.0253	0.1235
$p = 0.25$	-0.1044	0.3894	-0.0814	0.3597	-0.0643	0.3386
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	-0.0054	0.0151	-0.0044	0.0138	-0.0037	0.0129
$p = 0.05$	-0.0272	0.0768	-0.0222	0.0705	-0.0186	0.0660
$p = 0.10$	-0.0554	0.1578	-0.0451	0.1448	-0.0376	0.1357
$p = 0.25$	-0.1451	0.4301	-0.1177	0.3961	-0.0974	0.3718
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	-0.0057	0.0154	-0.0047	0.0141	-0.0039	0.0132
$p = 0.05$	-0.0288	0.0784	-0.0236	0.0719	-0.0199	0.0673
$p = 0.10$	-0.0586	0.1610	-0.0479	0.1477	-0.0402	0.1383
$p = 0.25$	-0.1538	0.4387	-0.1254	0.4038	-0.1045	0.3788

Table 6. LSE bootstrap in-control LCL and UCL for the generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	0.0011	0.0104	0.0012	0.0094	0.0014	0.0088
$p = 0.05$	0.0057	0.0538	0.0064	0.0486	0.0070	0.0454
$p = 0.10$	0.0117	0.1127	0.0133	0.1021	0.0146	0.0948
$p = 0.25$	0.0313	0.3344	0.0378	0.3008	0.0417	0.2779
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	0.0005	0.0221	0.0006	0.0183	0.0007	0.0160
$p = 0.05$	0.0025	0.1142	0.0031	0.0940	0.0038	0.0826
$p = 0.10$	0.0051	0.2376	0.0065	0.1970	0.0078	0.1714
$p = 0.25$	0.0142	0.6872	0.0181	0.5679	0.0219	0.4937
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	0.0003	0.0329	0.0004	0.0257	0.0005	0.0215
$p = 0.05$	0.0016	0.1697	0.0022	0.1314	0.0027	0.1115
$p = 0.10$	0.0033	0.3523	0.0044	0.2756	0.0056	0.2307
$p = 0.25$	0.0091	1.0105	0.0124	0.7889	0.0158	0.6608
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	0.0003	0.0354	0.0004	0.0274	0.0005	0.0227
$p = 0.05$	0.0014	0.1825	0.0019	0.1399	0.0025	0.1178
$p = 0.10$	0.0029	0.3791	0.0040	0.2932	0.0052	0.2437
$p = 0.25$	0.0081	1.0857	0.0112	0.8393	0.0145	0.6969

Table 7. MMM bootstrap in-control LCL and UCL for the generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	0.0005	0.0140	0.0006	0.0122	0.0007	0.0112
$p = 0.05$	0.0044	0.0577	0.0036	0.0633	0.0044	0.0577
$p = 0.10$	0.0070	0.1491	0.0084	0.1307	0.0098	0.1186
$p = 0.25$	0.0248	0.4184	0.0284	0.3636	0.0323	0.3306
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	0.0000+	0.0254	0.0000+	0.0210	0.0000+	0.0183
$p = 0.05$	0.0004	0.0961	0.0002	0.1096	0.0004	0.0961
$p = 0.10$	0.0010	0.2700	0.0016	0.2247	0.0022	0.1943
$p = 0.25$	0.0051	0.7561	0.0069	0.6235	0.0091	0.5396
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	0.0000+	0.0336	0.0000+	0.0269	0.0000+	0.0230
$p = 0.05$	0.0000+	0.1195	0.0000+	0.1391	0.0000+	0.1195
$p = 0.10$	0.0002	0.3583	0.0005	0.2878	0.0008	0.2443
$p = 0.25$	0.0019	1.0049	0.0029	0.8000	0.0042	0.6782
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	0.0000+	0.0353	0.0000+	0.0282	0.0000+	0.0239
$p = 0.05$	0.0000+	0.1240	0.0000+	0.1456	0.0000+	0.1241
$p = 0.10$	0.0001	0.3773	0.0004	0.3005	0.0006	0.2541
$p = 0.25$	0.0015	1.0573	0.0023	0.8373	0.0034	0.7046

Table 8. ML bootstrap in-control upper and lower control limits for generalized Pareto percentiles with $\alpha = 2.5$ and $\lambda = 1.0$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	0.0011	0.0116	0.0012	0.0106	0.0013	0.0099
$p = 0.05$	0.0057	0.0595	0.0064	0.0546	0.0070	0.0510
$p = 0.10$	0.0120	0.1225	0.0134	0.1127	0.0146	0.1055
$p = 0.25$	0.0346	0.3408	0.0388	0.3166	0.0424	0.2918
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	0.0006	0.0210	0.0007	0.0181	0.0008	0.0161
$p = 0.05$	0.0030	0.1072	0.0036	0.926	0.0041	0.0828
$p = 0.10$	0.0062	0.2207	0.0075	0.1913	0.0085	0.1713
$p = 0.25$	0.0176	0.6149	0.0214	0.5284	0.0245	0.4729
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	0.0004	0.0278	0.0005	0.0232	0.0006	0.0202
$p = 0.05$	0.0022	0.1418	0.0028	0.1185	0.0032	0.1039
$p = 0.10$	0.0045	0.2918	0.0058	0.2452	0.0068	0.2148
$p = 0.25$	0.0126	0.8144	0.0163	0.6761	0.0192	0.5923
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	0.0004	0.0292	0.0005	0.0242	0.0006	0.0211
$p = 0.05$	0.0020	0.1492	0.0026	0.1239	0.0031	0.1081
$p = 0.10$	0.0041	0.3068	0.0054	0.2564	0.0064	0.2238
$p = 0.25$	0.0116	0.8567	0.0151	0.7071	0.0181	0.6168

Table 9. The Shewhart-type chart standard errors for the UCL and LCL.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	7.39×10^{-6}	2.33×10^{-5}	5.55×10^{-6}	1.97×10^{-5}	4.40×10^{-6}	1.69×10^{-5}
$p = 0.05$	3.69×10^{-5}	1.18×10^{-4}	2.74×10^{-5}	1.02×10^{-4}	2.22×10^{-5}	8.63×10^{-5}
$p = 0.10$	7.31×10^{-5}	2.42×10^{-4}	5.52×10^{-5}	2.07×10^{-4}	4.50×10^{-5}	1.77×10^{-4}
$p = 0.25$	1.82×10^{-4}	6.58×10^{-4}	1.40×10^{-4}	5.55×10^{-4}	1.16×10^{-4}	4.76×10^{-4}
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	1.23×10^{-5}	3.02×10^{-5}	9.09×10^{-6}	2.54×10^{-5}	6.84×10^{-6}	2.15×10^{-5}
$p = 0.05$	6.15×10^{-5}	1.53×10^{-4}	4.51×10^{-5}	1.31×10^{-4}	3.40×10^{-5}	1.10×10^{-4}
$p = 0.10$	1.22×10^{-4}	3.14×10^{-4}	9.02×10^{-5}	2.66×10^{-4}	6.80×10^{-5}	2.25×10^{-4}
$p = 0.25$	3.03×10^{-4}	8.47×10^{-4}	2.20×10^{-4}	7.08×10^{-4}	1.67×10^{-4}	6.01×10^{-4}
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	1.52×10^{-5}	3.34×10^{-5}	1.13×10^{-5}	2.80×10^{-5}	8.61×10^{-6}	2.36×10^{-5}
$p = 0.05$	7.59×10^{-5}	1.70×10^{-4}	5.67×10^{-5}	1.44×10^{-4}	4.30×10^{-5}	1.21×10^{-4}
$p = 0.10$	1.52×10^{-4}	3.47×10^{-4}	1.13×10^{-4}	2.92×10^{-4}	8.60×10^{-5}	2.46×10^{-4}
$p = 0.25$	3.79×10^{-4}	9.34×10^{-4}	2.79×10^{-4}	7.78×10^{-4}	2.13×10^{-4}	6.59×10^{-4}
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	1.58×10^{-5}	3.41×10^{-5}	1.18×10^{-5}	2.85×10^{-5}	9.01×10^{-6}	2.41×10^{-5}
$p = 0.05$	7.91×10^{-5}	1.73×10^{-4}	5.92×10^{-5}	1.47×10^{-4}	4.50×10^{-5}	1.23×10^{-4}
$p = 0.10$	1.58×10^{-4}	3.54×10^{-4}	1.19×10^{-4}	2.98×10^{-4}	9.01×10^{-5}	2.51×10^{-4}
$p = 0.25$	3.96×10^{-4}	9.52×10^{-4}	2.92×10^{-4}	7.93×10^{-4}	2.24×10^{-4}	6.71×10^{-4}

Table 10. The LSE bootstrap chart standard errors for the UCL and LCL.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	2.29×10^{-6}	2.98×10^{-5}	2.31×10^{-6}	2.37×10^{-5}	2.28×10^{-6}	1.96×10^{-5}
$p = 0.05$	1.18×10^{-5}	1.56×10^{-4}	1.21×10^{-5}	1.22×10^{-4}	1.19×10^{-5}	1.03×10^{-4}
$p = 0.10$	2.48×10^{-5}	3.29×10^{-4}	2.51×10^{-5}	2.64×10^{-4}	2.52×10^{-5}	2.17×10^{-4}
$p = 0.25$	7.17×10^{-5}	1.01×10^{-3}	7.35×10^{-5}	7.94×10^{-4}	7.37×10^{-5}	6.50×10^{-4}
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	1.16×10^{-6}	1.15×10^{-4}	1.33×10^{-6}	7.85×10^{-5}	1.41×10^{-6}	5.86×10^{-5}
$p = 0.05$	5.93×10^{-6}	5.97×10^{-4}	6.85×10^{-6}	3.99×10^{-4}	7.32×10^{-6}	3.08×10^{-4}
$p = 0.10$	1.24×10^{-5}	1.24×10^{-3}	1.41×10^{-5}	8.55×10^{-4}	1.53×10^{-5}	6.37×10^{-4}
$p = 0.25$	3.47×10^{-5}	3.61×10^{-3}	4.06×10^{-5}	2.48×10^{-3}	4.39×10^{-5}	1.84×10^{-3}
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	8.45×10^{-7}	2.36×10^{-4}	1.04×10^{-6}	1.46×10^{-4}	1.16×10^{-6}	1.03×10^{-4}
$p = 0.05$	4.33×10^{-6}	1.22×10^{-3}	5.30×10^{-6}	7.36×10^{-4}	6.00×10^{-6}	5.40×10^{-4}
$p = 0.10$	8.92×10^{-6}	2.54×10^{-3}	1.09×10^{-5}	1.57×10^{-3}	1.25×10^{-5}	1.12×10^{-3}
$p = 0.25$	2.49×10^{-5}	7.25×10^{-3}	3.09×10^{-5}	4.56×10^{-3}	3.54×10^{-5}	3.17×10^{-3}
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	7.90×10^{-7}	2.69×10^{-4}	9.79×10^{-7}	1.64×10^{-4}	1.11×10^{-6}	1.13×10^{-4}
$p = 0.05$	4.05×10^{-6}	1.39×10^{-3}	5.01×10^{-6}	8.22×10^{-4}	5.80×10^{-6}	6.01×10^{-4}
$p = 0.10$	8.33×10^{-6}	2.90×10^{-3}	1.03×10^{-5}	1.76×10^{-3}	1.20×10^{-5}	1.24×10^{-3}
$p = 0.25$	2.32×10^{-5}	8.23×10^{-3}	2.92×10^{-5}	5.10×10^{-3}	3.40×10^{-5}	3.50×10^{-3}

Table 11. The MMM bootstrap chart standard errors for the UCL and LCL.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	1.51×10^{-6}	3.12×10^{-5}	1.71×10^{-6}	2.42×10^{-5}	1.92×10^{-6}	2.03×10^{-5}
$p = 0.05$	1.03×10^{-5}	2.02×10^{-4}	1.05×10^{-5}	1.47×10^{-4}	1.21×10^{-5}	1.20×10^{-4}
$p = 0.10$	2.43×10^{-5}	4.08×10^{-4}	2.41×10^{-5}	2.98×10^{-4}	2.69×10^{-5}	2.35×10^{-4}
$p = 0.25$	8.02×10^{-5}	1.15×10^{-3}	7.74×10^{-5}	8.30×10^{-4}	8.39×10^{-5}	6.58×10^{-4}
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	0.00+	6.96×10^{-5}	0.00+	4.69×10^{-5}	0.00+	3.58×10^{-5}
$p = 0.05$	8.13×10^{-7}	3.70×10^{-4}	1.47×10^{-6}	2.53×10^{-4}	2.50×10^{-6}	2.07×10^{-4}
$p = 0.10$	3.97×10^{-6}	7.83×10^{-4}	4.81×10^{-6}	5.46×10^{-4}	6.99×10^{-6}	4.12×10^{-4}
$p = 0.25$	1.71×10^{-5}	2.27×10^{-3}	1.89×10^{-5}	1.54×10^{-3}	2.59×10^{-5}	1.14×10^{-3}
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	0.00+	1.10×10^{-4}	0.00+	7.13×10^{-5}	0.00+	5.24×10^{-5}
$p = 0.05$	0.00+	5.36×10^{-4}	0.00+	3.72×10^{-6}	0.00+	2.45×10^{-4}
$p = 0.10$	1.64×10^{-6}	1.11×10^{-3}	2.35×10^{-6}	7.89×10^{-6}	3.49×10^{-6}	5.40×10^{-4}
$p = 0.25$	7.87×10^{-6}	3.22×10^{-3}	1.08×10^{-5}	2.22×10^{-6}	1.36×10^{-5}	1.55×10^{-3}
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	0.00+	1.19×10^{-4}	0.00+	7.71×10^{-5}	0.00+	5.63×10^{-5}
$p = 0.05$	0.00+	5.76×10^{-4}	0.00+	3.98×10^{-4}	0.00+	2.69×10^{-4}
$p = 0.10$	1.25×10^{-6}	1.20×10^{-3}	1.71×10^{-6}	8.37×10^{-4}	3.01×10^{-6}	5.78×10^{-4}
$p = 0.25$	6.53×10^{-6}	3.40×10^{-3}	9.26×10^{-6}	2.34×10^{-3}	1.13×10^{-5}	1.63×10^{-3}

Table 12. The ML bootstrap chart standard errors for the UCL and LCL.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	LCL	UCL	LCL	UCL	LCL	UCL
$\gamma_0 = 0.1$ (FAR)						
$p = 0.01$	2.10×10^{-6}	2.70×10^{-5}	2.23×10^{-6}	2.15×10^{-5}	2.39×10^{-6}	1.79×10^{-5}
$p = 0.05$	1.09×10^{-5}	1.38×10^{-4}	1.15×10^{-5}	1.10×10^{-4}	1.24×10^{-5}	9.28×10^{-5}
$p = 0.10$	2.30×10^{-5}	2.85×10^{-4}	2.44×10^{-5}	2.30×10^{-4}	2.60×10^{-5}	1.92×10^{-4}
$p = 0.25$	6.98×10^{-5}	7.98×10^{-4}	7.25×10^{-5}	6.33×10^{-4}	7.49×10^{-5}	5.36×10^{-4}
$\gamma_0 = 0.01$ (FAR)						
$p = 0.01$	1.19×10^{-6}	6.53×10^{-5}	1.17×10^{-6}	4.63×10^{-5}	1.15×10^{-6}	3.50×10^{-5}
$p = 0.05$	6.17×10^{-6}	3.33×10^{-4}	6.04×10^{-6}	2.37×10^{-4}	6.07×10^{-6}	1.82×10^{-4}
$p = 0.10$	1.29×10^{-5}	4.95×10^{-4}	1.29×10^{-5}	4.95×10^{-4}	1.29×10^{-5}	3.78×10^{-4}
$p = 0.25$	3.82×10^{-5}	1.95×10^{-3}	3.90×10^{-5}	1.37×10^{-3}	3.95×10^{-5}	1.06×10^{-3}
$\gamma_0 = 0.0027$ (FAR)						
$p = 0.01$	1.00×10^{-6}	1.04×10^{-4}	1.01×10^{-6}	7.03×10^{-5}	1.01×10^{-6}	5.07×10^{-5}
$p = 0.05$	5.18×10^{-6}	5.32×10^{-4}	5.22×10^{-6}	3.57×10^{-4}	5.31×10^{-6}	2.64×10^{-4}
$p = 0.10$	1.08×10^{-5}	1.10×10^{-3}	1.11×10^{-5}	7.45×10^{-4}	1.12×10^{-5}	5.46×10^{-4}
$p = 0.25$	3.08×10^{-5}	3.12×10^{-3}	3.27×10^{-5}	2.05×10^{-3}	3.36×10^{-5}	1.52×10^{-4}
$\gamma_0 = 0.002$ (FAR)						
$p = 0.01$	9.67×10^{-7}	1.14×10^{-4}	9.95×10^{-7}	7.60×10^{-5}	9.90×10^{-7}	5.46×10^{-5}
$p = 0.05$	4.97×10^{-6}	5.80×10^{-4}	5.11×10^{-6}	3.86×10^{-4}	5.21×10^{-6}	2.84×10^{-4}
$p = 0.10$	1.03×10^{-5}	1.20×10^{-3}	1.08×10^{-5}	8.04×10^{-4}	1.10×10^{-5}	5.87×10^{-4}
$p = 0.25$	2.93×10^{-5}	3.40×10^{-3}	3.18×10^{-5}	2.22×10^{-3}	3.28×10^{-5}	1.63×10^{-3}

Table 13. Out of control ML estimate chart for the generalized Pareto distribution with out of control parameters $\alpha = 5.0$ and $\lambda = 2.5$.

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	1.3464	0.00706	1.2344	0.00562	1.1528	0.00435
$p = 0.05$	1.3357	0.00689	1.2183	0.00535	1.1515	0.00432
$p = 0.10$	1.3276	0.00686	1.2051	0.00505	1.1248	0.00383
$p = 0.25$	1.2960	0.00653	1.1751	0.00471	1.1023	0.00349
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	3.0699	0.02886	2.3627	0.01966	1.9373	0.01421
$p = 0.05$	3.0619	0.02866	2.3092	0.01887	1.9441	0.01416
$p = 0.10$	3.0602	0.02856	2.2908	0.01862	1.8689	0.01348
$p = 0.25$	2.9571	0.02707	2.1870	0.01750	1.1782	0.01260
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	6.3779	0.07389	4.0018	0.03924	2.9852	0.02778
$p = 0.05$	6.3924	0.07448	3.9766	0.03851	2.9838	0.02724
$p = 0.10$	6.4617	0.07726	3.9563	0.03936	3.69870	0.02615
$p = 0.25$	6.3339	0.07325	3.8642	0.03799	2.7893	0.02583
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	7.9494	0.09725	4.7668	0.04981	3.4133	0.03319
$p = 0.05$	7.9494	0.09697	4.7029	0.04788	4.99500	0.03261
$p = 0.10$	8.2089	0.10630	4.7029	0.04931	3.3173	0.03219
$p = 0.25$	7.9494	0.09476	4.6614	0.05051	4.99500	0.03142

Table 14. Out of control ML estimate charts for the generalized Pareto distribution with out of control parameters $\alpha = 5.0$ and $\lambda = 1.0$

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	4.5318	0.0438	4.0891	0.0391	3.8218	0.0349
$p = 0.05$	4.5271	0.0439	4.0515	0.0386	3.6853	0.0338
$p = 0.10$	4.4212	0.0429	3.9901	0.0375	3.6906	0.0344
$p = 0.25$	4.2499	0.0413	3.6662	0.0352	3.3230	0.0307
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	25.8143	0.3380	20.2286	0.2511	16.6361	0.1876
$p = 0.05$	25.5503	0.3208	20.1428	0.2450	16.5277	0.1892
$p = 0.10$	25.5291	0.3254	19.6476	0.2364	16.0774	0.1877
$p = 0.25$	25.1304	0.3175	19.1412	0.2393	15.3536	0.1731
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	82.2897	1.1592	58.3665	0.8167	43.1321	0.5541
$p = 0.05$	80.8592	1.0892	57.2253	0.7818	43.2578	0.5656
$p = 0.10$	82.1383	1.1715	57.2081	0.7799	43.2304	0.5705
$p = 0.25$	81.1916	1.1376	56.9895	0.8237	41.4291	0.5506
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	114.3089	1.7078	78.5059	1.1334	57.2333	0.8035
$p = 0.05$	112.5334	1.6530	78.3478	1.1540	57.5751	0.7900
$p = 0.10$	115.9670	1.7793	78.3478	1.1379	57.8656	0.8186
$p = 0.25$	112.5334	1.6395	78.3928	1.2019	55.6368	0.7737

Table 15. Out of control ML estimate charts for the generalized Pareto distribution with out of control parameters $\alpha = 2.5$ and $\lambda = 2.0$

Parameters	$n = 4$		$n = 5$		$n = 6$	
	ARL	SEARL	ARL	SEARL	ARL	SEARL
	$\gamma_0 = 0.1$ (FAR)		$1/\gamma_0 = 10$			
$p = 0.01$	4.8353	0.0463	4.4966	0.0442	4.1259	0.0390
$p = 0.05$	4.8245	0.0471	4.3706	0.0429	4.0561	0.0376
$p = 0.10$	4.7334	0.0447	4.3496	0.0420	3.9753	0.0373
$p = 0.25$	4.6432	1.7211	4.1408	0.0401	3.7211	0.0352
	$\gamma_0 = 0.01$ (FAR)		$1/\gamma_0 = 100$			
$p = 0.01$	28.5761	0.3498	23.0087	0.2683	19.8312	0.2316
$p = 0.05$	28.6142	0.3493	22.8181	0.2667	19.6298	0.2276
$p = 0.10$	28.1088	0.3444	23.4356	0.2844	19.3925	0.2204
$p = 0.25$	28.5066	0.3539	22.4157	0.2812	18.4769	0.2127
	$\gamma_0 = 0.0027$ (FAR)		$1/\gamma_0 = 370.37$			
$p = 0.01$	88.7289	0.3498	67.6586	0.9705	52.6793	0.6839
$p = 0.05$	89.0722	1.2382	66.6053	0.8885	52.3929	0.6743
$p = 0.10$	88.4970	1.2085	66.2546	0.8827	50.8437	0.6362
$p = 0.25$	88.8254	1.2125	65.7238	0.9371	50.2462	0.6486
	$\gamma_0 = 0.002$ (FAR)		$1/\gamma_0 = 500$			
$p = 0.01$	122.4647	1.9951	93.4767	1.4172	69.6476	0.9444
$p = 0.05$	122.9314	1.7426	90.2651	1.2825	69.6476	0.9432
$p = 0.10$	120.3574	1.6882	90.2651	1.2781	68.2267	0.8971
$p = 0.25$	122.3561	1.7211	89.0618	1.2991	67.4059	0.9072

4.3 ILLUSTRATIVE EXAMPLE

Assume certain machine parts have failure times in terms of years that have a generalized Pareto distribution with $\alpha = 2.5$ and $\lambda = 1.0$. Since no real-world data could be obtained during this study, an R program was created to generate twenty subgroups with six machine part lifetimes a piece. Since λ is a rate parameter, without the loss of generality λ can be selected as one with a reasonable measurement unit in lifetime measure. α is the shape parameter that should not be too large or too small. These subgroups were made independently from the in-control generalized Pareto distribution with $\alpha = 2.5$ and $\lambda = 1.0$. These twenty subgroups are reported in Table 16. The designer of the parts is concerned about the tenth percentile of the lifetime of his parts, $Q(0.10; \alpha_0, \lambda_0)$. After the first twenty subgroups, assume that the process was shifted to out of control where $\alpha_1 = 5.0$ and $\lambda_1 = 2.5$ and another twenty subgroups were generated with six machine part lifetimes a piece. These twenty subgroups are displayed in Table 17. The ML bootstrap chart was developed based on the twenty in-control subgroups in Table 16 where the FAR=0.0027 and B=10,000. The control limits were are LCL = 0.0129 and UCL = 0.1646. The center line is CL = 0.06007. Figure 1 (top) shows the control chart for the in-control percentiles and Figure 1 (bottom) shows the same control chart for monitoring the out of control tenth percentile. In Figure 1 (top) all of the tenth percentiles calculated based on twenty subgroups, respectively, are within the control limits and spread around the CL. In Figure 1 (bottom), notice that the first tenth percentile signals an out of control process. While not all of the tenth percentiles are outside of the control chart limits in this figure, they are grouped rather tightly and are all well below the CL. Thus, the ML bootstrap chart is successful in indicating that a process is out of control.

Table 16. Twenty subgroups of machine part lifetimes generated from the generalized Pareto distribution with $\alpha_0 = 2.5$ and $\lambda_0 = 1.0$.

Subgroup number	Lifetime observations						
1	0.5819	2.3860	0.1465	0.2941	0.3153	0.1461	
2	1.1770	0.0462	0.1149	0.7403	0.2062	1.8740	
3	0.0467	0.1225	0.0617	0.2224	0.9333	0.4332	
4	5.2410	0.8200	0.4782	0.5241	0.0439	1.3510	
5	0.6481	0.0607	0.1688	0.1478	0.8709	0.2992	
6	0.9731	0.0956	0.0493	1.6600	0.3200	0.3037	
7	0.1695	0.0998	1.2960	0.0525	0.0936	1.3860	
8	0.2028	0.0197	0.8517	0.7443	0.6432	0.1275	
9	0.0614	0.1126	0.1307	0.0167	0.5010	1.2790	
10	0.6115	0.5098	1.0260	0.9001	0.2065	0.0695	
11	1.7350	2.1580	1.1040	0.2383	0.3030	0.1099	
12	0.0758	0.4705	0.0119	0.1444	0.0568	0.8328	
13	1.2990	0.6935	0.2923	0.7409	0.4427	0.8387	
14	0.7045	0.1364	3.5530	0.0713	0.1115	0.5185	
15	0.2578	0.4141	0.2453	1.6400	0.2592	0.3155	
16	1.3130	0.0240	1.1280	0.0591	0.1310	0.0676	
17	0.5947	0.0189	0.4675	0.0356	1.4630	0.0643	
18	0.2588	0.1155	0.4547	1.2500	0.7298	0.1451	
19	0.1267	1.2390	0.0508	0.2061	0.2859	1.2500	
20	0.3479	0.0243	0.2715	0.0724	0.0877	1.3420	

Table 17. Twenty subgroups of machine part lifetimes generated from the generalized Pareto distribution with $\alpha_1 = 5.0$ and $\lambda_1 = 2.5$.

Subgroup number	Lifetime observations						
1	0.0342	0.0787	0.2626	0.0460	0.0135	0.0040	
2	0.2904	0.0627	0.0198	0.0252	0.0727	0.1682	
3	0.0049	0.2299	0.0117	0.0375	0.0613	0.0088	
4	0.0402	0.2255	0.0342	0.0958	0.1299	0.0905	
5	0.0450	0.0422	0.2306	0.1699	0.0893	0.0174	
6	0.0170	0.4364	0.2594	0.0518	0.2007	0.1366	
7	0.0547	0.0112	0.0004	0.0023	0.0761	0.0094	
8	0.0613	0.0013	0.7946	0.0365	0.1964	0.1364	
9	0.0018	0.1491	0.0472	0.1392	0.1302	0.0829	
10	0.0345	0.0032	0.0227	0.0420	0.0975	0.1786	
11	0.0199	0.0141	0.0103	0.0709	0.0095	0.2356	
12	0.0220	0.7481	0.0402	0.1396	0.0129	0.0989	
13	0.0166	0.0034	0.0148	0.1722	0.2251	0.0620	
14	0.0038	0.1211	0.2050	0.0061	0.2040	0.0528	
15	0.0093	0.0105	0.0855	0.0156	0.1116	0.0153	
16	0.0045	0.0648	0.2079	0.0912	0.0727	0.0258	
17	0.1148	0.1332	0.1420	0.2850	0.0859	0.0154	
18	0.1863	0.1126	0.2125	0.0102	0.0781	0.1808	
19	0.0356	0.2005	0.0333	0.2909	0.1731	0.6069	
20	0.0719	0.0411	0.1251	0.0564	0.4224	0.1394	

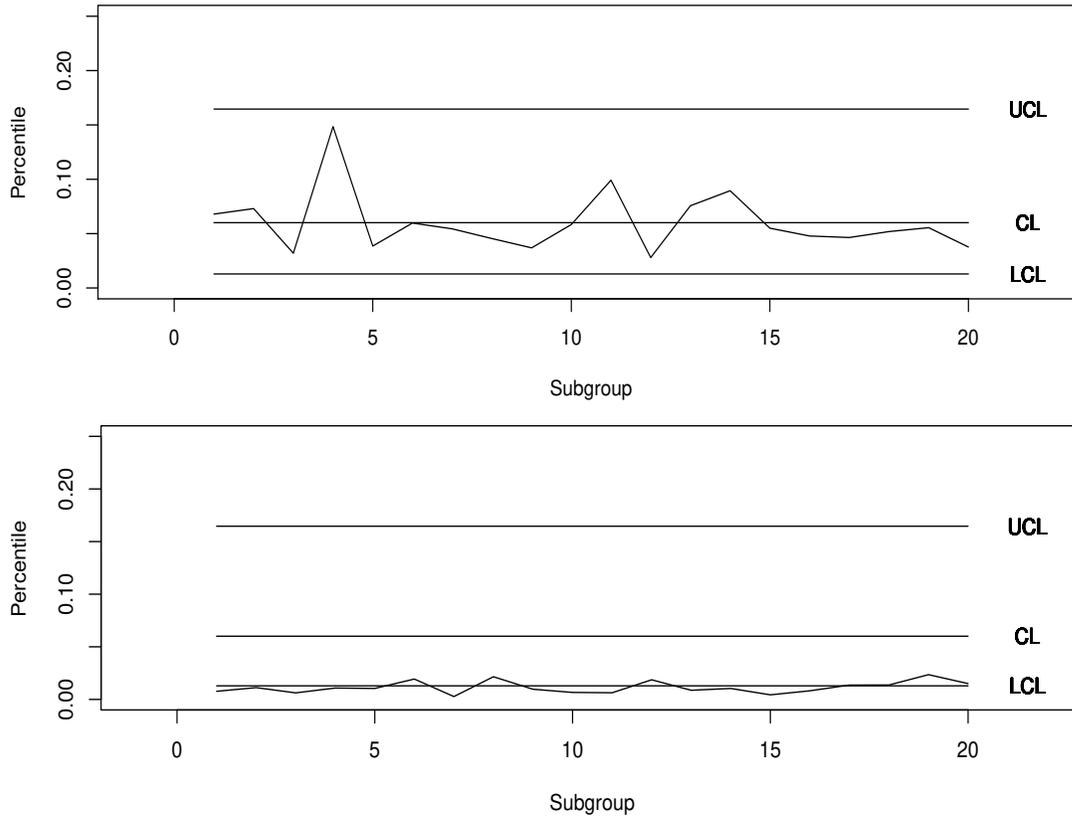


Figure 1. In control subgroups (top) and out control subgroups with FAR=0.0027.

5. CONCLUSIONS

In order to examine the Pareto percentiles, the Shewhart-type chart and three parametric bootstrap charts were constructed. As a result of the Monte Carlo simulation, it was discovered that the Shewhart-type chart was inadequate in providing appropriate control limits. Two of the parametric bootstrap charts were also shown to be unsuitable for providing appropriate control limits. The least squared error control chart consistently overestimated the nominal average run length and the modified moment method chart was inconsistent. However the maximum likelihood chart was shown to be acceptable choice for monitoring Pareto percentiles. As it also promptly detects an out of control process, as shown in the simulation and illustrative example, it is recommended for practical use.

It should be mentioned that the conclusions from the current research may not be applied to any other case with shape parameter, α , or rate parameter, λ , too far from the current values for generalized Pareto distribution. However, the current simulation procedures provides a guideline to run simulation study for different α and λ to make a selection of control chart method. When a real word application data are given, it is suggested to use Kolmogorov-Smirnov test with Akaike and Bayesian information criteria to select probability model. Then, the practitioners can follow the current research procedure to decide the control chart method after the lifetime distribution has been decided. Further research for multivariate control charts under non-normality can be explored (Marchant et al., 2019, 2018).

ACKNOWLEDGEMENTS

The authors would like to thank the Editors, Associate Editor and Referees for their suggestions and comments that led to a significant improvement of this manuscript.

REFERENCES

- Aykroyd, R.G., Leiva, V., and Ruggeri, F., 2019. Recent developments of control charts, identification of big data sources and future trends of current research. *Technological Forecasting and Social Change*, 144, 221–232.
- Chen, H., Cheng, W., Zhao, J., and Zhao, X., 2017. Parameter estimation for generalized Pareto distributions by generalized probability weighted moment-equations. *Communications in Statistics: Simulation and Computation*, 46, 7761–7776.
- Chiang, J., Jiang, N., Brown, T., Tsai, T., and Lio, Y., 2017. Control charts for generalized exponential distribution percentiles. *Communication in Statistics: Simulation and Computation*, 46, 7827–7843.
- Efron, B., Tibshirani, R., 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York.
- Gunter, B., 1992. Bootstrapping: How to make something from almost nothing and get statistically valid answers, part III. *Quality Progress*, 25, 119–122.
- Hosking, J. and Wallis, J., 1987. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29, 339–349.
- Hüsler, J., Li, D., and Raschke, M., 2011. Estimation for the generalized Pareto distribution using maximum likelihood and goodness of fit. *Communications in Statistics: Theory and Methods*, 40, 2500–2510.
- Ihaka, R. and Gentleman, R., 1996. A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.
- Lio, Y., Park, C., 2008. A bootstrap control chart for Birnbaum-Saunders percentiles. *Quality and Reliability Engineering International*, 24, 585–600.
- Lio, Y. and Park, C., 2010. A bootstrap control chart for inverse Gaussian percentiles. *Journal of Statistical Computation and Simulation*, 80, 287–299.
- Lio, Y., Tsai, T., Aslam, M., and Jiang, N., 2014. Control charts for monitoring Burr type-X percentiles. *Communication in Statistics: Simulation and Computation*, 43, 761–776.
- Marchant, C., Leiva, V., Christakos, G., and Cavieres, M.F., 2019. Monitoring urban environmental pollution by bivariate control charts: new methodology and case study in Santiago, Chile. *Environmetrics*, 30, e2551.
- Marchant, C., Leiva, V., Cysneiros, F.J.A., and Liu, S., 2018. Robust multivariate control charts based on Birnbaum-Saunders distributions. *Journal of Statistical Computation and Simulation*, 88, 182–202.
- Nichols, M. and Padgett, W., 2005. A bootstrap control chart for Weibull percentiles. *Quality and Reliability Engineering International*, 22, 141–151.
- Pickands, J., 1975. Statistical inference using extreme order statistics. *The Annals of Statistics*, 29, 119–131.
- Rezac, J., Lio, Y., and Jiang, N., 2015. Burr type-XII percentile control charts. *Chilean Journal of Statistics*, 6(1), 67–87.
- Salmasi, M. and Yari, G. 2017. On generalized order statistics from generalized Pareto distribution. *Communications in Statistics: Simulation and Computation*, 46, 5682–5697.
- Wang, B., 2008. Statistical inference for the Burr type-XII distribution. *ACTA Mathematica Scientia*, 28, 1103–1108.
- Young, G., 1994. Bootstrap: More than a stab in the dark. *Statistical Science*, 9, 382–415.

NON-PARAMETRIC INFERENCE
RESEARCH PAPER

Exact tables for the Friedman rank test: Case with ties

CARLOS LÓPEZ-VÁZQUEZ^{1,*}, ANDRÓMACA TASISTRO², and ESTHER HOCHSZTAIN³

¹LatinGEO Lab IGM+ORT, Universidad ORT, Uruguay

²Permanent Education Unit, Faculty of Economic Sciences and Management,
Universidad de la República, Uruguay

³Faculty of Economic Sciences and Management, Universidad de la República, Uruguay

(Received: 28 July 2020 · Accepted in final form: 01 April 2021)

Abstract

Exact tables for the case without ties of the Friedman statistic test proposed have been available since its inception. A modified statistic suitable for the case with ties has been derived 30 years later, and it appears in a text book nearly after 40 years. However, exact tables for the case of ties were never offered. Here we present for the first time a reduced set of exact tables for such a case, thus filling a gap. If a problem allows ties, the proper exact tables should be used thus disregarding other workarounds commonly suggested in the literature. The availability of exact tables for the case of ties is relevant for applied research because an hypothesis test decision when ties occur may be different if tables for the case without ties are used instead. We illustrate the effect of using the correct tables with both an example and a real data case study in the context of geoportals navigation analysis.

Keywords: Friedman test · Exact distribution · Non-parametric methods · R software.

Mathematics Subject Classification: Primary 62G10 · Secondary 62F05.

1. INTRODUCTION

The problem of analyzing the rankings resulting from a wine contest with k wines and N judges has been addressed by [Friedman \(1937\)](#). The null hypothesis is that there is no difference between the wines. In other situations, the wines might be medical treatments and the judges are patients. Original data might be of type ordinal, or it might be of continuous type (interval and ratio, as defined by [Stevens, 1946](#)). In that case, when ranking, he circumvents the normality requirements of other parametric tests like analysis of variance.

Tied ranks might appear with ordinal data, but also with continuous one. For example, if the values arise from a measurement device with finite accuracy, the uncertainty in their readings leads naturally to possible ties. In his seminal paper Friedman only considered the case without ties. Under such assumption he offered two asymptotic estimates valid for: a) large N irrespective of k and b) small N and moderate and large k . The asymptotic

*Corresponding author. Email: carloslopez@uni.ort.edu.uy

expressions were inaccurate for the case with both k and N small so he offered some exact tables. The set of (k, N) pairs covered was modest, mostly due to the limited computing facilities of the time. The availability of exact tables and asymptotic approximations for ranking problems is not unusual; see [Sen et al. \(2011\)](#) for another example.

Exact tables for the case of ties were never published. In this proposal we offer for the first time some exact tables for the case of both small k and N . Exact tables for the Friedman rank test in the case of ties are relevant to applied research. In particular, we will show that the conclusion of an hypothesis test when ties occur may differ if tables for the case without ties are used instead of the ones proposed in this paper.

The paper is organized as follows. Section 2 describes the generalized formulae valid with or without ties as presented by various authors. On Section 3 we will comment about how the problem with ties is dealt with in textbooks, tutorials and reference guide. On principle, ties might be up to couples, triplets, etc. or without restrictions, something to be discussed in Section 4, showing that the results might vary depending on how many identical elements are allowed in the ties. Afterwards, in Section 5 an illustrative numerical example taken from a book will be presented followed by a case study in Section 6 showing the effect of misusing the asymptotic estimate and/or the wrong tables. The computational procedure will be commented in Section 7, and the new tables are presented in Section 8. Finally, some conclusions are sketched in Section 9.

2. RELATED WORKS

2.1 THE CASE WITHOUT TIES

[Friedman \(1937\)](#) proposed a rank test to avoid normality assumptions in analysis of variance. The method of ranks involves two steps, creating a two-way table: 1) rank data in each row 2) test if the all columns come from the same universe. The null hypothesis is that there are no differences between the columns. It can be proved that the statistic given by

$$\chi_r^2 = \frac{12}{Nk(k+1)} \sum_{j=1}^k \left(\sum_{i=1}^N r_{ij} \right)^2 - 3N(k+1),$$

is asymptotically χ_{k-1}^2 distributed for large N , and to a normal distribution with mean $k-1$ and standard deviation $\sqrt{2(k-1)(N-1)/N}$ for small N and moderate to large k . The asymptotic estimate must be used if the (k, N) pair of the problem under consideration is not covered by the available tables. The jump between table values and asymptotic estimate might be large, so the set of exact tables for given (k, N) pairs steadily grew with time. [Friedman \(1937\)](#) considered only the cases of $k = 3$ for N up to 9, and $k = 4$ for N up to 4. [Kendall and Babington-Smith \(1939\)](#) extended the tables for $k = 3$ up to $N = 10$; $k = 4$ up to $N = 6$ and analyzed the case $k = 5$, $N = 3$. [Owen \(1962\)](#) published exact tables for $k = 3$ and N up to 15; $k = 4$ and N up to 8 without disclosing the computation procedure. [Hollander and Wolfe \(1999\)](#) provided tables for $k = 5$ and N up to 5. [Odeh \(1977\)](#) extended the tables considering the cases $k = 5$ for N up to 8, and $k = 6$ for N up to 6. [Martin et al. \(1993\)](#) extended the case $k = 4$ up to $N = 15$. A significant contribution was provided by [van de Wiel \(2000\)](#), who extended previous work considering the case $k = 3$ with N up to 25, $k = 4$ for N up to 20 and for $k = 5$ he offered tables for N up to 12. More recently [López-Vázquez and Hochsztain \(2019\)](#) drastically expanded the set of tables using a code in the R software implemented by [Schneider et al. \(2016\)](#). There, exact tables up to $N = 204$, 41, 13, 7, 4 and 2 for $k = 3, 4, 5, 6, 7$ and 8 were performed based on an algorithm proposed by [van de Wiel et al. \(1999\)](#). All of these efforts were for the case without ties.

2.2 THE CASE WITH TIES

In problems with discrete random variables ties are likely to appear, a fact acknowledged in [Friedman \(1937\)](#). It is difficult to explain why a correction for ties only was mentioned for the first time in [Marascuilo and McSweeney \(1967\)](#), and appeared significantly later in a textbook in [Conover \(1980\)](#). Other equivalent expressions appeared even later, like the one by [Siegel and Castellan \(1988\)](#) or the one proposed by [Corder and Foreman \(2009\)](#). Apparently they were derived independently, even though they produced exactly the same value. Unlike the simple case without ties, the rank for the case with ties has more than one alternative. Most of the literature used the mid-rank method, which assures that the sum of ranks for each judge is constant. According to [Sprent and Smeeton \(2007\)](#) the generalized statistic (now valid either with or without ties) still has the same asymptotic estimate as the original one proposed by [Friedman \(1937\)](#). However, despite they acknowledge that for low k and N the asymptotic estimate is not accurate, no exact tables were provided.

The correction for ties proposed by [Marascuilo and McSweeney \(1967\)](#) is stated as

$$\chi_r^2 = \frac{\frac{12}{k(k+1)} \sum_{j=1}^k \frac{R_j^2}{N} - 3N(k+1)}{1 - \frac{\sum_{s=1}^d (t_s^3 - t_s)}{N(k^3 - k)}}.$$

The numerator is the familiar statistic for the case without ties. For the correction term, d is the number of set of ties and t_i is the number of tied scores in the i -th set of ties. If there are no ties the denominator is 1. The expression was not widely cited in the literature. [Conover \(1980\)](#) proposed an expression defined as

$$\chi_r^2 = \frac{(k-1) \sum_{j=1}^k [R_j - N(k+1)/2]^2}{\sum_{i=1}^N \sum_{j=1}^k r_{ij}^2 - Nk(k+1)^2/4},$$

where R_j is the sum of the ranks r_{ij} for treatment j . Apparently, a number of alternative formulations for the same statistic were derived independently. [Siegel and Castellan \(1988\)](#) proposed a slightly more complicated expression given by

$$\chi_r^2 = \frac{12 \sum_{j=1}^k R_j^2 - 3N^2k(k+1)^2}{Nk(k+1) + \frac{\left(Nk - \sum_{i=1}^N \sum_{j=1}^{g_i} t_{i,j}^3 \right)}{k-1}},$$

where g_i is the number of sets of tied ranks in the i -th group and $t_{i,j}$ is the size of the j -th set of tied ranks in the i -th group.

Using the same definition for g_i and $t_{i,j}$, [Hollander and Wolfe \(1999\)](#) proposed a different expression, given by

$$\chi_r^2 = \frac{12 \sum_{j=1}^k R_j^2 - 3N^2k(k+1)^2}{Nk(k+1) - \frac{\sum_{i=1}^N \left\{ \left(\sum_{j=1}^{g_i} t_{i,j}^3 \right) - k \right\}}{k-1}}.$$

Gibbons and Chakraborti (2010) suggested another expression to be written as

$$S = \sum_{i=1}^N \left[\sum_{j=1}^k \left(r_{ij} - \frac{(N+1)}{2} \right) \right]^2; \quad \chi_r^2 = \frac{12(N-1)S}{Nk(N^2-1) - \sum_{i=1}^N \sum_{j=1}^{g_i} (t_i^3 - t_i)}$$

Buskirk et al. (2013) included other notation established as

$$\chi_r^2 = \frac{12 \sum_{j=1}^k R_j^2 - 3N^2k(k+1)^2}{Nk(k+1) + \frac{\sum_{i=1}^N t_i^3 - t_i}{k-1}},$$

with t_i being the number of observations involved in a tie for the i -th case.

Boos and Stefanski (2013) proposed a compact expression, now without the need to count the number of ties explicitly, given by

$$\chi_r^2 = \frac{(k-1)N^2 \sum_{j=1}^k \left[\frac{1}{N} \sum_{i=1}^N r_{ij} - \frac{(k+1)}{2} \right]^2}{\sum_{i=1}^N \sum_{j=1}^k r_{ij}^2 - Nk(k+1)^2/4}.$$

In our computations, we use the expression from Corder and Foreman (2009), which is equivalent to the earlier ones, stated by

$$\chi_r^2 = \frac{N(k-1) \left[\sum_{j=1}^k \frac{R_j^2}{N} - C_F \right]}{\sum_{i=1}^N \sum_{j=1}^k r_{ij}^2 - C_F},$$

where

$$C_F = \frac{Nk(k+1)^2}{4}.$$

3. RECOMMENDED STRATEGIES FOR THE PROBLEM WITH TIES

Without going as deep as Richardson (2019), who compared many aspects of non-parametric statistics textbooks, it is fit to mention how they treated the case of the Friedman test with ties. Our search included some of the books considered by Richardson (2019), all of them intended for a statistical audience, but also some others designed with other communities in mind. An example for the food industry might be Granato et al. (2014) and we have included some others in the list. The frontline of science is usually found at papers, not books. However, papers are typically known only to a very small community. Also, the textbooks offer guidelines to a variety of users, not necessarily experts in the field. Thus, it is important to assess how the case of the Friedman test with ties is considered in the literature intended to reach a large audience.

According to the literature, the alternatives at hand for a problem with ties are:

- Use the generalized statistic, and compare it against the asymptotic estimate based upon the χ^2 approximation thus ignoring the need of exact tables (Plichta and Garzon, 2009; Alvo and Yu, 2010; Sheskin, 2011; Vidakovic, 1999; Hettmansperger and McKean, 2011; Buskirk et al., 2013; Boos and Stefanski, 2013; Granato et al., 2014).
- Use the generalized statistic, and acknowledge that the χ^2 approximation will not be accurate for low k and N . Use as a surrogate the exact table without ties (van Belle et al., 2004; Zar, 2010; Gibbons and Chakraborti, 2010; Linebach et al., 2014; Chechile, 2020).
- Same as before, but noticing that they lack an exact table for the problem with ties (Sprent and Smeeton, 2007; Hollander et al., 2014).
- Ignore the effects of ties, and use both the traditional Friedman statistic as well as its associated exact tables (Greene and D’Oliveira, 2005; Daniel and Cross, 2013; Corder and Foreman, 2009, 2014).
- Break the ties, assigning random ranks through a Monte Carlo experiment, and then use the standard Friedman statistic (Rayner et al., 2005).
- Assume that the problem with ties can be handled just by using midranks (Lehman, 1975; Canavos, 1988; Derrac et al., 2011; Liu et al., 2017).

There is no good reason to keep using the traditional statistic as proposed by Friedman, because the generalized one considers both situations. However, neglecting the fact that the asymptotic estimate is only valid for mid to large k and N (López-Vázquez and Hochsztain, 2019), or that the exact tables are not valid for cases with ties might have a devastating effect on the conclusions. We will illustrate it with some cases, but before let us analyze a somewhat subtle detail.

4. TYPES OF TIES ALLOWED

In general, ties might involve $2, \dots, k$ wines, denoted here as p -tuples. The case without ties is equivalent to set $p = 1$. If, for some reason the problem of interest just allows ties of pairs but not triplets, we should use $p = 2$. The general case “with ties” is equivalent to set $p = k$. We have yet to find practical examples where p is not equal to either 1 or k , but if they exist the distinction might be important because the exact tables are different. We illustrate it in Table 1, which corresponds to the case $k = 5$ and $N = 4$. The possibilities range from $p = 1$ (denoted as “no ties”) to $p = k = 5$ (denoted as “with ties”).

Table 1. Effect on the critical values when ties are restricted to p -tuples. Case of $k = 5$ and $N = 4$.

	10	5	2.5	1.0	0.5	0.1
with ties	7.471	8.675	9.699	10.873	11.671	13.105
4-tuples	7.474	8.675	9.699	10.886	11.676	13.111
3-tuples	7.478	8.675	9.699	10.889	11.684	13.143
2-tuples	7.474	8.684	9.707	10.880	11.730	13.143
no ties	7.600	8.800	9.800	11.200	12.000	13.200

This has consequences not considered before by the literature. If the problem under consideration allows tied ranks, the corresponding exact tables must be used even if in the data under analysis there are no cases with ties. Thus, the strategy of “breaking the ties” with any procedure, thus reducing the problem to the case without ties, is flawed. The differences will be evident only for low k and N , when exact tables are needed. Otherwise, since the asymptotic estimate is the same for the case with or without ties there will be no difference.

5. EFFECT OF MISUSING THE STANDARD TABLES FOR THE PROBLEM WITH TIES

Because it appears in a textbook we will first consider an example from [Corder and Foreman \(2009\)](#). They presented a simple example which is summarized in [Table 2](#). It reports the ranks of the response of seven employees ($N = 7$) under three alternatives ($k = 3$) to deal with their tardiness: a) do nothing (denoted as baseline), b) one month with a monetary incentive of \$10, and c) another month with double incentive. They want to determine if either of the payback deduction strategies modified employee tardiness.

Table 2. Rank of tardiness after considering three incentive initiatives (from [Corder and Foreman, 2009](#)).

Employee	Rank of monthly tardiness		
	Baseline	Month 1	Month 2
1	2	3	1
2	3	2	1
3	2	3	1
4	3	2	1
5	3	1.5	1.5
6	2.5	2.5	1
7	3	2	1

In order to compute the statistic, we first find R_i summing table entries by columns as

$$\begin{aligned} R_1 &= 2 + 3 + 2 + 3 + 3 + 2.5 + 3 = 18.5 \\ R_2 &= 3 + 2 + 3 + 2 + 1.5 + 2.5 + 2 = 16 \\ R_3 &= 1 + 1 + 1 + 1 + 1.5 + 1 + 1 = 7.5 \end{aligned}$$

The denominator holds the sum of squares of the table entries as well as the C_F term

$$\begin{aligned} \sum_{i=1}^7 \sum_{j=1}^3 r_{ij}^2 &= 2^2 + 3^2 + 1^2 + 3^2 + 2^2 + 1^2 + 2^2 + 3^2 + 1^2 + 3^2 + 2^2 + 1^2 + 3^2 + 1.5^2 + \\ &+ 1.5^2 + 2^2 + 2.5^2 + 1^2 + 3^2 + 2^2 + 1^2 = 97, \end{aligned}$$

$$C_F = \frac{7 * 3 * (3 + 1)^2}{4} = 84.$$

Thus the χ_r^2 value computation is presented as

$$\chi_r^2 = 7 * 2 \left[\frac{18.5^2 + 16^2 + 7.5^2}{7} - 84 \right] / (97 - 84) = \frac{133}{13} = 10.23.$$

For $\alpha = 0.05$ ([Corder and Foreman, 2009](#)) stated that the critical value is 7.140. After a quick check it is possible to confirm that the critical value presented corresponds to the “without ties” problem (see for example the tables from [Martin et al., 1993](#)). If we choose not to use exact tables, the χ^2 asymptotic approximation provides a critical value of $\chi_{0.05,2}^2 = 4.605$. The correct value for the “with ties” problem should have been 5.769, taken from our [Table 5](#). In this case the null hypothesis would be rejected using either critical value, but the differences observed are not negligible. An interesting case arises for $\alpha = 0.005$; the “without ties” table offers 10.286 as the critical value. The critical value χ^2 is now $\chi_{0.005,2}^2 = 10.597$. Both are larger than the χ_r^2 so according to this the null hypothesis should be rejected. However, using our [Table 5](#) the exact critical value is 9.083, now lower

than the statistic value $\chi_r^2 = 10.23$. Thus, according to the correct table, we could not reject the null hypothesis.

6. CASE STUDY

We present a case study in the context of geoportals navigation analysis. As stated by [Jiang et al. \(2020\)](#) and [Bernabé-Poveda and González \(2014\)](#) a geoportal is a web-based solution to provide open spatial data sharing and online geo-information management. The concept of geoportals has become key for spatial data and geoinformation accessing and sharing. We perform geoportal navigation analysis based on geoportal web server logs (click-stream data) following the guidelines given by ([Markov and Larose, 2007](#); [Bhavani et al., 2017](#); [Bhuvaneswari and Muneeswaran, 2021](#)).

As indicated by [Srivastava et al. \(2019\)](#), whenever a user requests a particular web resource on a website, an entry is recorded into a log file which is automatically stored and maintained by the web server. The log file is named web server log or clickstream. We preprocessed web server logs and computed three variables to be used in this case study: pages per session, session duration and average time per page.

The double-entry table presented in [Table 3](#) shows the coefficient of variation (rounded to two digits) of pages per session (CVPPS), where rows represent four levels I to IV of session duration and columns represent three levels I to III of average time per page. Levels were defined by percentile groups. Rankings by row are presented in [Table 4](#), and we can observe that one tie occurs for Session duration level II. [Figure 1](#) shows the R output of the Friedman test.

We want to assess if there is some relationship between the CVPPS and the session duration levels.

Table 3. Coefficient of variation of pages per session by session duration and average time per page levels.

Session duration levels	Coefficient of variation		
	Average time per page levels		
	I	II	III
I	0.12	0.08	0.19
II	0.11	0.21	0.21
III	0.08	0.19	0.26
IV	0.13	0.22	0.27

Table 4. Coefficient of variation of pages per session by session duration and average time per page levels.

Session duration levels	Ranking of coefficient of variation of pages per session		
	Average time per page levels		
	I	II	III
I	2	1	3
II	1	2.5	2.5
III	1	2	3
IV	1	2	3

```

> rankeddata
      [,1] [,2] [,3]
[1,]    2  1.0  3.0
[2,]    1  2.5  2.5
[3,]    1  2.0  3.0
[4,]    1  2.0  3.0
> friedman.test(rankeddata)
      Friedman rank sum test
data: rankeddata
Friedman chi-squared = 5.7333, df = 2, p-value = 0.05689

```

Figure 1. Friedman test output using R 3.6.3.

Following the same steps as before, we compute intermediate values and the resulting statistic, considering correction by ties as

$$\begin{aligned}
 R_1 &= 2 + 1 + 1 + 1 = 5 \\
 R_2 &= 1 + 2.5 + 2 + 2 = 7.5 \\
 R_3 &= 3 + 2.5 + 3 + 3 = 11.5
 \end{aligned}$$

$$\sum_{i=1}^4 \sum_{j=1}^3 r_{ij}^2 = 2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 2.5^2 + 2^2 + 2^2 + 3^2 + 2.5^2 + 3^2 + 3^2 = 55.5,$$

$$C_F = \frac{4 * 3 * (3 + 1)^2}{4} = 48.$$

Thus, the χ_r^2 value is

$$\chi_r^2 = 4 * (3 - 1) \left[\frac{5^2 + 7.5^2 + 11.5^2}{4} - 48 \right] / (55.5 - 48) = \frac{43}{7.5} = 5.733,$$

The null hypothesis is that there are no differences between the columns, that is, for any given level of session duration, the coefficient of variation of pages per session (CVPPS) is the same at all levels of average time per page.

Friedman χ_r^2 statistic value is 5.733. At 5% significance level, if we consider exact tables without ties as provided by López-Vázquez and Hochsztain (2019) or Martin et al. (1993) the critical value for $k=3$ $N=4$ is 6.500, and therefore the decision is not to reject the null hypothesis. The same holds true if we consider the $\chi_{0.05,2}^2$ value for $k = 3$ (7.815), as the statistic value (5.733) is again less than the critical value. Considering the R-output shown in Figure 1, as p -value 0.05689 is larger than 0.05 we should conclude that the decision is not to reject the null hypothesis. However, when there are ties, as we can see in Table 3 it is necessary to use the exact tables presented in this paper. As we can see in Table 5, the correct critical value in this case at the 5% significance level is 5.571, and as a consequence the decision is to reject the null hypothesis. Therefore, using either the wrong exact table or the chi-square approximation results in a different decision than using the correct exact table presented in this paper. We can acknowledge the practical importance of using the proper tables for the Friedman Rank-Test in the case of ties, leading to a correct decision in the hypothesis test. And thus concluding that the different values of CVPPS have an effect over the Session duration levels.

7. COMPUTATIONAL PROCEDURE TO PRODUCE THE TABLES

The problem of computing the exact tables for this problem has been barely considered in the literature. To the best of our knowledge, only [Hollander and Wolfe \(1999\)](#) proposed a brute-force procedure to compute the exact conditional distribution of the Friedman statistic valid for each particular case. We used instead a two step procedure in order to compute the general, exact tables, when there are ties among the data values. Firstly, we built the set of possible cases considering ties (always allowing up to k -tuples), and as a second step we computed the statistic for all the valid combinations. Hence, this is thus a combinatorial problem. For very small k computing the first step posed no special requirements. The computation was carried using the R software version 4.0.0 in a personal computer. The computation time for the second step was manageable for very small k and N , but the runtime requirements quickly grow along k and N . For example, for the case of $k = 4$, $N = 7$ the computations using Octave 3.8.2 required over 90 days of wall time using up to 100 nodes in parallel. The valid combinations were arranged in sets and computed independently using a nearly embarrassingly parallel approach. It is worth mentioning that, due to the combinatorial nature of the problem, our approach is only capable of dealing with modest cases in reasonable time. The facility (described by [Nesmachnow and Iturriaga, 2019](#)) is a LINUX-based cluster equipped with Intel Xeon-Gold 6138 2.00GHz processors.

8. RESULTING TABLES

Tables 5, 6, 7, 8, and 9 are offered for the case of $k = 3, 4, 5, 6$ and 7, respectively. Following the style used by [Martin et al. \(1993\)](#), last row of each table holds the corresponding $\chi_{\alpha, k-1}^2$ asymptotic estimate. It illustrates the jump with respect to the exact table values for finite N .

Table 5. Critical values of the statistic χ_r^2 in the case of ties up to k -tuples for $k = 3$ and N up to 11.

	0.100	0.050	0.025	0.010	0.005	0.001
$N = 3$	4.667	5.000	5.600	5.636	6.000	
$N = 4$	4.667	5.571	6.000	6.857	7.429	7.600
$N = 5$	4.588	5.647	6.615	7.444	8.316	9.294
$N = 6$	4.571	5.727	6.778	7.913	8.667	10.174
$N = 7$	4.560	5.769	6.870	8.222	9.083	10.583
$N = 8$	4.526	5.793	6.909	8.296	9.250	11.143
$N = 9$	4.563	5.813	6.968	8.357	9.455	11.467
$N = 10$	4.563	5.846	7.000	8.471	9.548	11.730
$N = 11$	4.550	5.850	7.048	8.581	9.657	12.000
$\chi_{\alpha, 2}^2$	4.605	5.991	7.378	9.210	10.597	13.816

Table 6. Critical values of the statistic χ_r^2 in the case of ties up to k -tuples for $k = 4$ and N up to 6.

	0.100	0.050	0.025	0.010	0.005	0.001
$N = 2$	5.400	5.842	5.842	6.000		
$N = 3$	5.893	6.692	7.444	8.111	8.379	8.793
$N = 4$	6.081	7.184	8.100	9.079	9.750	10.784
$N = 5$	6.136	7.370	8.478	9.652	10.421	11.936
$N = 6$	6.161	7.446	8.660	10.019	10.964	12.765
$\chi_{\alpha, 3}^2$	6.251	7.815	9.348	11.345	12.838	16.266

Table 7. Critical values of the statistic χ_r^2 in the case of ties up to k -tuples for $k = 5$ and N up to 4.

	0.100	0.050	0.025	0.010	0.005	0.001
$N = 2$	6.703	7.263	7.568	7.789	7.897	8.000
$N = 3$	7.241	8.267	9.091	9.964	10.400	11.154
$N = 4$	7.471	8.675	9.699	10.873	11.671	13.105
$\chi_{\alpha,4}^2$	7.779	9.488	11.143	13.277	14.860	18.467

Table 8. Critical values of the statistic χ_r^2 in the case of ties up to k -tuples for $k = 6$ and N up to 3.

	0.100	0.050	0.025	0.010	0.005	0.001
$N = 2$	8.065	8.710	9.118	9.485	9.688	9.918
$N = 3$	8.580	9.697	10.637	11.634	12.255	13.284
$\chi_{\alpha,5}^2$	9.236	11.070	12.833	15.086	16.750	20.515

Table 9. Critical values of the statistic χ_r^2 in the case of ties up to k -tuples for $k = 7$ and $N = 2$.

	0.100	0.050	0.025	0.010	0.005	0.001
$N = 2$	9.303	10.073	10.624	11.094	11.345	11.725
$\chi_{\alpha,6}^2$	10.645	12.592	14.449	16.812	18.548	22.458

9. CONCLUSIONS

The Friedman rank test for the case without ties has been used for decades, but only until recently the case with ties was considered. Despite a correction for the original formulae is available, and that the asymptotic approximations are the same, no exact tables for low k and N have been published. Here we present the first ones, and illustrate its importance by showing that even a simple case published in a book suffers badly for using the wrong table in the computations. In addition, we consider that the tables for the case without ties are only applicable for problems which cannot accept ties, and not merely because the data do not show ties. This questioned some strategies that propose to break the ties reducing the problem to one without ties. In addition, the type of ties allowed (only pairs, only triplets, etc.) have a noticeable effect on the final decision, at least for small k and N . To build the exact tables we used a naive approach, which is combinatorial and can only deal with very small k and N . Further expansion of the exact tables will require using different algorithms, in the line of those of [van de Wiel \(2000\)](#) or [van de Wiel et al. \(1999\)](#).

Future works include expanding the exact tables and develop an R package to calculate Friedman rank test p -value based in the exact tables for the case of ties as described in this paper.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The experiments presented in this paper were carried out using ClusterUY (<https://cluster.uv>). We want to acknowledge Dr. Sergio Nesmachnow and Dr. Gabriel Usera for providing easy access to the computing facility. We also acknowledge Col. Norbertino Suárez and Lic. Sergio Marshall, both of the Military Geographic Institute, for providing access to the web server logs used in the case study.

REFERENCES

- Alvo, M. and Yu, P. L. H., 2014. *Statistical Methods for Ranking Data*. Springer, New York.
- Bernabé-Poveda, M.A. and González, M.E., 2014. Sobre la necesaria usabilidad de los geoportales como puertas de entrada a las IDE. *Revista Internacional de Ciencia y Tecnología de la Información Geográfica*, 1(14), 1–5.
- Bhavani, B and Sucharita, V. and Satyanarana, K.V.V., 2017. Review on techniques and applications involved in web usage mining. *International Journal of Applied Engineering Research*, 12, 15994–15998.
- Bhuvanewari, M.S. and Muneeswaran, K., 2021. User community detection from web server log using between user similarity metric. *International Journal of Computational Intelligence Systems*, 14, 266–281.
- Boos, D.D. and Stefanski, L.A., 2013. *Essential Statistical Inference: Theory and Methods*. Springer, New York.
- Buskirk, T. D.; Willoughby, L. M. and Tomazic, T.T., 2013. *Nonparametric Statistical Techniques*. Oxford Handbooks Online, Oxford, UK.
- Canavos, G.C., 1988. *Probability and Statistics: Applications and Methods*. McGraw-Hill, Mexico.
- Conover, W.J., 1980. *Practical Nonparametric Statistics*. Wiley, New York.
- Corder, G.W. and Foreman, D.I., 2009. *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley, NJ, US.
- Corder, G.W. and Foreman, D.I., 2014. *Nonparametric Statistics: A Step-by-Step Approach*. Wiley, NJ, US.
- Chechile, R.A., 2020. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution-Free Methods*. MIT Press, Massachusetts, US.
- Daniel, W.W. and Cross, C.L., 2013. *Biostatistics: A Foundation for Analysis in the Health Sciences*. Wiley, New York.
- Derrac, J., García, S., Molina, D., and Herrera, F., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1, 3–18.
- Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32, 675–701.
- Gibbons, J. D. and Chakraborti, S., 2010. *Nonparametric Statistical Inference*. CRC Press, Boca Raton, FL, US.
- Granato, D., Calado, V.M.A., and Jarvis, B., 2014. Observations on the use of statistical methods in food science and technology. *Food Research International*, 55, 137–149.
- Greene, J. and D'Oliveira, M., 2005. *Learning to Use Statistical Tests in Psychology*. Open University Press.
- Hettmansperger, T.P. and McKean, J.W., 2011. *Robust Nonparametric Statistical Methods*. CRC Press, Boca Raton, FL, US.
- Hollander, M. and Wolfe, D.A., 1999. *Nonparametric Statistical Methods*. Wiley, New York.
- Hollander, M., Wolfe, D.A. and Chicken, E., 2014. *Nonparametric Statistical Methods*. Wiley, New York.
- Jiang, H., van Genderen, J., Mazzetti, P., Koo, H. and Chen, M., 2020. Current status and future directions of geoportals. *International Journal of Digital Earth*, 13, 1093–1114.
- Kendall, M.G. and Babington-Smith, B., 1939. The problem of m rankings. *Annals of Mathematical Statistics*, 10, 275–287.
- Korosteleva, O., 2013. *Nonparametric Methods in Statistics with SAS Applications*. Chapman and Hall, New York.
- Lehman, E.L. and D'Abrera, H.J.M., 1975. *Nonparametrics. Statistical Methods Based on Ranks*. Holden-Day, New York.

- Linebach, J.A., Tesch, P.T., and Kovacsiss, L.M., 2014. *Nonparametric Statistics for Applied Research*. Springer, New York.
- Liu, Z., Blash, E., and John, V., 2017. Statistical comparison of image fusion algorithms: Recommendations. *Information Fusion*, 36, 251–260.
- López-Vázquez, C. and Hochsztain, E., 2019. Extended and updated tables for the Friedman rank test. *Communications in Statistics: Theory and Methods*, 48, 268–281.
- Lovie, P., 2005. Coefficient of Variation. *Encyclopedia of Statistics in Behavioral Science*. Wiley, Online Library.
- Marascuilo, L.A., and McSweeney, M., 1967. Nonparametric post-hoc comparisons for trend. *Psychological Bulletin*, 67, 401–412.
- Markov, Z. and Larose, D.T., 2007. *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*. Wiley, New York.
- Martin, L., Leblanc, R. and Toan, N., 1993. Tables for the Friedman rank test. *Canadian Journal of Statistics*, 21, 39–43.
- Nesmachnow S. and Iturriaga S., 2019. Cluster-UY: collaborative scientific high performance computing in Uruguay. In: Torres M., Klapp J. (eds) *Supercomputing. ISUM 2019. Communications in Computer and Information Science*, vol 1151. Springer, Cham
- Odeh, R.E., 1977. Extended tables of the distribution of Friedman's s-statistic in the two way layout. *Communications in Statistics: Simulation and Computation*, 6, 29–48.
- Owen, D.B., 1962. *Handbook of Statistical Tables*. Addison-Wesley Publishing Company, Reading, Massachusetts, US.
- Pélabon, C., Hilde, C., Einum, S. and Gamelon, M., 2020. On the use of the coefficient of variation to quantify and compare trait variation. *Evolution Letters*, 4, 180–188.
- Plichta, S.B. and Garzon, L.S., 2009. *Statistics for Nursing and Allied Health*. Wolters Kluwer Health /Lippincott, Williams and Wilkins.
- Rainer, J.C.W., Best, D.J., Brockhoff, P.B., and Rayner, G.D., 2005. *Nonparametrics for Sensory Science: A More Informative Approach*. Blackwell Publishing.
- Richardson, A., 2019. A comparative review of nonparametric statistics textbooks. *The American Statistician*, 73, 360–366.
- Schneider, G., Chicken, E., and Becvarik, R., 2016. NSM3 Package: Functions and Datasets to Accompany Hollander, Wolfe, and Chicken - Nonparametric Statistical Methods. R software v. 1.16
- Sen, P.K., Salama, I.A., and Quade, D., 2011. Spearman's footrule: Asymptotics in applications. *Chilean Journal of Statistics*, 2(1), 3–20.
- Sheskin, D.J., 2011. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman-Hall, Boca Raton, FL, US.
- Siegel, S. and Castellan, N.J.Jr., 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, US.
- Sprent, P. and Smeeton, N., 2007. *Applied Nonparametric Statistical Methods*. CRC, NY.
- Srivastava, M., Srivastava; A.K. and Garg, R., 2019. Data preprocessing techniques in web usage mining: A literature review. In *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM)*, pp. 466-476
- Stevens, S.S., 1946. On the theory of scales of measurement. *Science*, 103, 677–680.
- van Belle, G., Fisher, L.D., Heagerty, P.J., and Lumley, T., 2004. *Biostatistics: A Methodology for the Health Sciences*. Wiley, New Work, US.
- van de Wiel, M.A., 2000. Exact distributions of distribution-free test statistics. Ph.D. Thesis, Eindhoven University of Technology, The Netherlands.
- van de Wiel, M.A., Di Bucchianico, A. and van der Laan, P., 1999. Symbolic computation and exact distributions of nonparametric test statistics. *Journal of the Royal Statistical Society D*, 48, 4, 507–516.

- Vidakovic, B., 2011. *Statistics for Bioengineering Sciences with MATLAB and WinBUGS Support*. Springer, New York, US.
- Zar, J.H., 2010. *Biostatistical Analysis*. Pearson, India.

STATISTICAL PROCESS CONTROL
RESEARCH PAPER

Control chart for monitoring the mean in symmetric data

LUCAS DE OLIVEIRA FERREIRA DE SALES^{1,2}, ANDRÉ LUÍS SANTOS DE PINHO^{3,*},
FRANCISCO MOISÉS CÂNDIDO DE MEDEIROS³ and MARCELO BOURGUIGNON³

¹Graduate Program in Statistics, Universidade de São Paulo, São Paulo, Brazil

²Graduate Program in Applied Mathematics and Statistics,
Universidade Federal do Rio Grande do Norte, Natal, Brazil

³Department of Statistics and Graduate Program in Applied Mathematics and Statistics,
Universidade Federal do Rio Grande do Norte, Natal, Brazil

(Received: 03 November 2020 · Accepted in final form: 22 April 2021)

Abstract

The control charts are the main tools used for monitoring quality characteristic. Usually the monitored characteristic is the process mean and the most used control charts for such monitoring are Shewhart \bar{X} , CUSUM and EWMA, which are based on two assumptions: independence between monitored samples and that the monitored variable follows a normal distribution. However, deviations from any of these assumptions imply poor control chart performance. Considering this, the present work proposes a control chart to monitoring the mean, based on the bootstrap method, for data that follows a distribution that belongs to the symmetric class of distributions. Simulation studies are performed for the proposed method, in order to evaluate the in-control and the out-of-control average run length, to evaluate the behavior of the control limits and to compare the proposed method with the traditional Shewhart \bar{X} . The simulation study indicates that the proposed approach presents better in-control average run length than the usual Shewhart \bar{X} . Regarding the power of detection, the proposed method presents good performance, being comparable to Shewhart \bar{X} , but with the advantage of a better in-control average run length. Practical use of the proposed approach is illustrated with a real example of pH of red wines.

Keywords: Bootstrap · Heavy-tailed distribution · Light-tailed distribution
· Statistical Process control · Symmetric distributions.

Mathematics Subject Classification: 62P30 · 62F99

1. INTRODUCTION

Competition in manufacturing industries has been growing around the world to achieve ever higher quality standards. Naveed et al. (2020) mentioned that the main concern of the companies is to maintain a positive reputation in the market. The authors also state that a key aspect to enable this goal is through the quality of the products. In this context,

*Corresponding author. Email: pinho@ccet.ufrn.br

statistical process control (SPC) is a powerful set of techniques to meet this end. More specifically, control charts are the most common tools in SPC used to monitor processes. The control chart proposed by [Shewhart \(1931\)](#), called the \bar{X} chart, is the most known and used SPC technique. The standard assumptions of this technique are: (i) the collections of independent samples over time, (ii) the monitored control characteristic follows a normal distribution. This method has the purpose of detecting shifts in the mean of magnitude greater than 1.5 standard deviation of the mean.

[Chakraborti et al. \(2008\)](#) emphasized the need to ascertain precisely if the monitored data follow the assumptions of the employed chart. However, in many practical situations, when verifying whether the data follow the standard assumptions, it is considered that the data are normally distributed just because they have a symmetrical shape. [Schilling and Nelson \(1976\)](#), [Borrer et al. \(1999\)](#), [Calzada and Scariano \(2001\)](#), and [Noorossana et al. \(2011\)](#) commented that when the monitored data do not follow a normal distribution, the usual Shewhart \bar{X} chart shows low performance in the monitoring, by generating more false alarms or by not detecting deviations from the true mean with the usual precision ([da Silva et al., 2019](#)).

It is in this context of low performance of the usual charts, under non-normal symmetrical distribution ([Rezac et al., 2015](#)), that arise alternatives methods to monitoring symmetrical data. For example, correction factors in quantiles of the distribution or in the form of control limits of the usual method ([Bai and Choi, 1995](#); [Tadikamalla and Popescu, 2007](#); [Tadikamalla et al., 2008](#)). Other approaches considered to solve the problem of non-normality are the non-parametric techniques ([Haq and Khoo, 2019](#); [Willemain and Runger, 1996](#); [Chakraborti et al., 2001](#)), and data transformation ([Qiu and Zhang, 2015](#)). There are also alternative procedures based on quantiles of distributions with a heavier tail than the tail of the normal distribution ([Calzada and Scariano, 2001](#); [Tsai et al., 2005](#); [Zhang et al., 2009](#)) and control charts via intensive computational methods ([Bajgier, 1992](#); [Seppala et al., 1995](#); [Liu and Tang, 1996](#)).

More recently, [Ahmed et al. \(2019\)](#) proposed a technique based on a more comprehensive class of distribution, known as the long-tailed symmetric (LTS), but in the context of small and moderate mean deviations. Nonetheless, there is a wider class of distributions, in which the LTS is a particular case, the symmetrical class of distributions or univariate elliptical ([Berkane and Bentler, 1986](#); [Fang et al., 1990](#); [Rao, 1990](#)).

Considering this scenario, this work has the objective of proposing, via parametric bootstrap method, a monitoring chart for the process mean for changes greater than 1.5 standard deviations. The underlying feature of the data distribution is the symmetric one. Besides, we focus on a wide class of symmetrical distribution known as the symmetrical distribution class. The in-control and out-of-control average run length (ARL_0 and ARL_1 , respectively) of the proposed method are evaluated through simulations and compared with the ARL_0 and ARL_1 of the standard Shewhart \bar{X} chart in different scenarios of the symmetrical class. The practical application of the method is illustrated by monitoring the pH of red wines. Finally, we argue that having a proposed method that provides a general framework for any member of the symmetric distribution class, regardless of the tail weight, leads to a better decision making.

This paper is organized as follows: this introductory section. [Section 2](#) presents briefly the symmetrical class of distributions. [Section 3](#) presents the proposed approach for monitoring changes in the process mean in symmetric data. [Section 4](#) presents the Average Run Length (ARL) performance of the proposed charts under different combinations of the model parameters based on simulation studies. Application to a real data set is presented in [Section 5](#). Final considerations are reported in [Section 6](#).

2. SYMMETRICAL DISTRIBUTION CLASS

In this section, we presents background on the symmetrical class of distributions.

Let X be a random variable, with $X \in \mathbb{R}$. The distribution of X belongs to the class of symmetric distributions with location parameter $\mu \in \mathbb{R}$ and scale parameter $\phi > 0$, if its probability density function is of the form:

$$f(x; \mu, \phi) = \frac{1}{\sqrt{\phi}} h\left(\frac{(x - \mu)^2}{\phi}\right), \quad x \in \mathbb{R},$$

for some function $h(u) > 0$, for $u > 0$, such that $\int_0^\infty u^{-\frac{1}{2}} h(u) du = 1$. The conditions imposed on h , guarantee that $f(x; \mu, \phi)$ is, in fact, a probability density function. The function h is called the density-generating function and it may depend on other parameters than μ and ϕ , which is the case of the Student- t and power-exponential distributions, for example.

We denote $X \sim S(\mu, \phi)$, if X belongs to the symmetric distributions class of parameters μ and ϕ . Some examples of distributions that belong to this class are shown in Table 1, as presented in [Medeiros and Ferrari \(2017\)](#), there are distributions with heavier tails (for example, Student- t distributions and type II logistic) and lighter tails (for example, power-exponential distribution with $-1 < \kappa < 0$ and type I logistic) than the normal distribution. Moreover, the class of symmetric distributions considers also bimodal distributions such as the generalized Kotz distribution.

Let us assume that $E(X) = \mu$ and $\text{Var}(X) = \xi\phi$ exist, for some constant $\xi > 0$. Furthermore, if $X \sim S(\mu, \phi)$, then $a + bX \sim S(a + b\mu, b^2\phi)$ with $a \in \mathbb{R}$ and $b \in \mathbb{R} - \{0\}$, that is, the distribution of any linear transformation of a random variable, which its distribution belongs to the symmetrical class, also belongs to the symmetrical class. Particularly, if $Z = (X - \mu)/\sqrt{\phi}$, then $Z \sim S(0, 1)$ and the probability density function of Z is given by

$$f(z) = f(z; 0, 1) = h(z^2), \quad z \in \mathbb{R},$$

where h is the density generating function of X . In order to estimate the parameters of these models, we adopt the maximum likelihood method. For more details on properties, demonstrations and theoretical results for the symmetric distribution class ([Berkane and Bentler, 1986](#); [Fang et al., 1990](#); [Rao, 1990](#)).

3. PROPOSED APPROACH

In this section, we propose a control chart using a parametric bootstrap method for a class of symmetric distributions. Our method is based on the work of [Bajgier \(1992\)](#) and [Liu and Tang \(1996\)](#), who used a non-parametric approach. Here, we make some modifications to use on a parametric bootstrap, since we are establishing theoretical results. When the distribution of the data is correctly identified, we generate samples of the suitable distribution in order to capturing the real nature of the data. As commented in [Efron and Tibshirani \(1994\)](#) and [Davison and Hinkley \(1997\)](#), when we fit a suitable distribution, the parametric bootstrap provides better results to estimate the quantiles, in our case the control limits, than the non-parametric bootstrap.

Before establishing results for the proposed control charts, we need to specify some notation and quantities such as sample size (m), subsample size (n), the frequency (s) (assumed here as $s = 1$), statistic used in the monitoring (in this case, \bar{X}), and lower and upper control limits (LCL and UCL). When $\bar{X} > \text{UCL}$ or $\bar{X} < \text{LCL}$, an action to search

Table 1. Density-generating function and ξ values, for some symmetric distributions.

Distribution	$h(u), \quad u > 0$	ξ
Normal	$\frac{1}{\sqrt{2\pi}}e^{-u/2}$	1
Student- t	$\frac{\nu^{\nu/2}}{B(1/2, \nu/2)}(\nu + u)^{-\frac{\nu+1}{2}}, \nu > 0$	$\frac{\nu}{\nu-2}, \nu > 2$
Type I logistic	$c \frac{e^{-u}}{(1+e^{-u})^2}, c \approx 1.4843$	0.7957
Type II logistic	$\frac{e^{-\sqrt{u}}}{(1+e^{-\sqrt{u}})^2}$	$\frac{\pi^2}{3}$
Kotz	$\frac{r^{(2N-1)/2}}{\Gamma(\frac{2N-1}{2})}, r > 0, N \geq 1$	$\frac{2N-1}{2r}$
Power-exponential ¹	$\frac{1}{c(\kappa)}e\left\{-\frac{1}{2}u^{1/(1+\kappa)}\right\}, -1 < \kappa \leq 1$	$2^{1+\kappa} \frac{\Gamma(1, 5(1+\kappa))}{\Gamma(\frac{1+\kappa}{2})}$

where Γ is the gamma functions and $c(\kappa) = \Gamma(1 + \frac{1+\kappa}{2})2^{1+(1+\kappa)/2}$.

for special causes in the process must be taken. Thus, when the process is in-control, it is desirable to have few false alarms to reduce the number of unnecessary stops in the process. In SPC, the usual metric to measure the performance of a control chart is the ARL until an out-of-control point is detected. When the process is in control, a large ARL_0 is desirable. Let us choose α , the probability of a type I error such that

$$\alpha = P(\bar{X} > \text{UCL} \mid \mu = \mu_0) + P(\bar{X} < \text{LCL} \mid \mu = \mu_0) \quad (1)$$

with μ_0 the value of μ when the process is in-control.

On other hand, if the process is out-of-control, it is desirable that the control chart signals very soon, that is, a low ARL_1 . The power of a control chart expressed as $1 - \beta$ is

$$P(\bar{X} > \text{UCL} \mid \mu = \mu_1) + P(\bar{X} < \text{LCL} \mid \mu = \mu_1), \quad (2)$$

where $\mu_1 = \mu_0 + \delta \sigma / \sqrt{m}$ is the mean when the process is out-of-control, σ is the standard deviation of the characteristic of interest, δ is the shift size expressed in units of the standard deviation of the mean and β is the probability of a type II error. Moreover,

$$ARL_0 = \frac{1}{\alpha}$$

and

$$ARL_1 = \frac{1}{1 - \beta}.$$

Note that from Equations (1) and (2) the control limits can be seen as quantiles of the distribution of the statistic used to monitor the process that provide a certain probability (α or $1 - \beta$). Thus, by fixing either (α or $1 - \beta$) or (ARL_0 or ARL_1), we get the control limits as the quantiles of the mean distribution. That said, based on the bootstrap method for control charts, proposed by Bajgier (1992) and Gandy and Kvaløy (2013), we obtain the control limits to monitor the process mean according to Algorithm 1.

Algorithm 1 Control limits to monitor the process mean.

- 1: Generate a $n \times m$ observation matrix of the considered symmetrical distribution, where n is the size of the subsample. Calculate

$$\bar{X}_i = \frac{1}{n} \sum_{l=(i-1)n+1}^{in} X_l, \quad i = 1, \dots, m.$$

- 2: Using the empirical distribution of \bar{X} , obtained using the samples of the symmetrical distribution in Step 1, obtain the quantiles of order $\alpha/2$ and $1 - \alpha/2$, referred here, respectively, as $\hat{q}_{\frac{\alpha}{2}}$ and $\hat{q}_{(1-\frac{\alpha}{2})}$.
- 3: Repeat Step 1 and 2 B times using the quantities

$$\widehat{\text{LCL}} = \frac{1}{B} \sum_{i=1}^B \hat{q}_{\frac{\alpha}{2},i} \quad \text{and} \quad \widehat{\text{UCL}} = \frac{1}{B} \sum_{i=1}^B \hat{q}_{(1-\frac{\alpha}{2}),i}$$

as the control limits.

Algorithm 1, described previously, can be schematically depicted as

Replicate 1	Replicate 2	...	Replicate B	
\bar{x}_1	\bar{x}_1	...	\bar{x}_1	
\vdots	\vdots	\vdots	\vdots	
\bar{x}_m	\bar{x}_m	...	\bar{x}_m	
$\hat{q}_{\frac{\alpha}{2},1}$	$\hat{q}_{\frac{\alpha}{2},2}$...	$\hat{q}_{\frac{\alpha}{2},B}$	$\rightarrow \widehat{\text{LCL}} = \frac{1}{B} \sum_{i=1}^B \hat{q}_{\frac{\alpha}{2},i}$
$\hat{q}_{(1-\frac{\alpha}{2}),1}$	$\hat{q}_{(1-\frac{\alpha}{2}),2}$...	$\hat{q}_{(1-\frac{\alpha}{2}),B}$	$\rightarrow \widehat{\text{UCL}} = \frac{1}{B} \sum_{i=1}^B \hat{q}_{(1-\frac{\alpha}{2}),i}$

Since this is a computationally intensive method, in order to obtain the control limits, we recommend simulated samples of sizes $m \geq 2,000$ and $B \geq 5,000$, aiming to obtain more accurate results, without possible bootstrap quantile bias. These values of m and B are suggested based on previous studies; see for more on the suggested values of m and B in Davison and Hinkley (1997). In practical situations, where the parameters of the process distribution are most likely unknown, we recommend the following procedure to estimate the control limits:

- (1) Obtain observed values $\mathbf{x} = (x_1, \dots, x_k)^\top$ (from sample of size k , under statistical control, of your population of interest) and adjust possible models for the data. Use the AIC (Akaike information criterion) and BIC (Bayesian information criterion) for selected model and goodness-of-fit techniques in order to evaluate the chosen distribution;
- (2) If the most suitable model belongs to the symmetric class, use $\hat{\boldsymbol{\theta}}(\mathbf{x})$, the maximum likelihood estimate of the parameters of the distribution, identified in Step (1), to obtain the control limits using Algorithm 1.

For more details on the computationally intensive and resampling procedure, see Davison and Hinkley (1997).

4. CHART PERFORMANCE EVALUATION AND COMPARISON

In this section, a detailed simulation study is conducted in order to gain insight into the detection abilities in the proposed control charts when we compare with the usual Shewhart charts.

The simulation study considers two distributions, Student- t and power-exponential. The Student- t distribution is usually used as an alternative to the normal distribution when the behavior of the data suggests a symmetrical distribution, but with tails heavier than the normal distribution. Lange et al. (1989) commented that the Student- t model can be seen as a robust parametric extension of the normal model, since it allows to reduce the influence of aberrant observations. On top of that, the Student- t allows the adjustment of the kurtosis of the data distribution through the ν parameter, which represents its degrees of freedom. For the purpose of evaluating the performance of the proposed chart, values of $\nu = \{3, 5, 10\}$ and 20 are considered in the simulation study. Additionally, the power-exponential distribution is also used because its κ parameter allows it to have both lighter and heavier tails than the normal distribution, thus making it a good alternative for non-normal symmetric data. The simulation study has lighter tail ($\kappa = \{-0.45, -0.25\}$) and heavier tail ($\kappa = \{0.3, 0.4\}$) scenarios than the normal distribution.

Furthermore, the scenarios are evaluated considering the following subsample size: $n = \{1, 2, 3, 10, 100, 500\}$. Bearing in mind that the proposed method is compared to the usual Shewhart chart, and we consider deviations in the mean of $\delta = \{0.0, \mp 1.5, \mp 2.0, \mp 3.0\}$ standard deviations. The analyzes presented below are based on the assessments of the ARL_0 , ARL_1 and the asymptotic behavior of the control limits, represented by $n = 100$ and 500. The parameter settings μ and ϕ adopted are intended to simulate several practical situations such as small (1, 2, 5) and large (100 and 200) values of process mean, in addition to considering the data dispersion index, which is given by

$$I_d = \frac{\text{Var}(X)}{E(X)}$$

and represents the variability of the data in relation to the mean. Regarding the dispersion index, we considered four categories: very underdispersed ($I_d \approx 0.033$), moderate underdispersed ($I_d \approx 0.67$), moderate overdispersed ($I_d \approx 1.67$) and very overdispersed ($I_d \approx 3.33$). The target value set for ARL_0 is 370.40 samples, which is equivalent to $\alpha = 0.0027$ (reference value of the Shewhart \bar{X} chart).

The computational routines were developed using the R software (R Core Team, 2018) version 3.6.2 for Windows platform and are available at:

<https://github.com/lucasdofs/Control-Chart-Symmetrical>.

The Shewhart LCL and UCL considered in the simulation are given by

$$\text{LCL} = \mu_0 - \frac{\sigma_0}{\sqrt{n}} \quad \text{and} \quad \text{UCL} = \mu_0 + \frac{\sigma_0}{\sqrt{n}},$$

where μ_0 and σ_0 are the in-control mean and in-control standard deviation, respectively.

Algorithm 2 proposes a way for estimating ARL_0 and ARL_1 .

Algorithm 2 Procedure to estimate ARL_0 and ARL_1 .

- 1: Generate $\mathbf{x}_{(n \times 5000)} = (x_1, \dots, x_{n \times 5000})^T$, a column vector of size $n \times 5000$ of the distribution of interest, and calculate

$$\bar{X}_i = \frac{1}{n} \sum_{l=(i-1)n+1}^{in} X_l, \quad i = 1, \dots, 5000.$$

- 2: The control limits are compared with the 5000 sample-shifted mean (shifted factor = $\delta \times \sigma_0 / \sqrt{n}$) and store the position of the first out of control sample, in which the value of the sample-shifted mean is higher than UCL or lower than LCL.
- 3: Steps 1 and 2 are repeated 10000 times independently, and ARL_0 or ARL_1 is calculated based on the average of the positions obtained in Step 2.

The diagram below illustrates Algorithm 2:

Replicate 1	Replicate 2	...	Replicate 10000
$\bar{x}_1 + \delta \frac{\sigma_0}{\sqrt{n}}$	$\bar{x}_1 + \delta \frac{\sigma_0}{\sqrt{n}}$...	$\bar{x}_1 + \delta \frac{\sigma_0}{\sqrt{n}}$
⋮	⋮	⋮	⋮
$\bar{x}_{5000} + \delta \frac{\sigma_0}{\sqrt{n}}$	$\bar{x}_{5000} + \delta \frac{\sigma_0}{\sqrt{n}}$...	$\bar{x}_{5000} + \delta \frac{\sigma_0}{\sqrt{n}}$
a_1	a_2	...	a_{10000}

Then, we estimate ARL_0 and ARL_1 by means of

$$\widehat{ARL}_j = \frac{\sum_{i=1}^{10000} a_i}{10000}, \quad j = 0, 1,$$

where a_i , for $i = 1, \dots, 10000$, represents the position of the first sample in which the value of the sample mean, plus δ standard deviations, is higher than UCL or lower than LCL.

Tables 2 to 5 and Tables 6 to 9 present the results of the computational study carried out for the Student- t and power-exponential distributions, respectively. The estimated results are expressed in terms of the quantities \widehat{ARL}_0 , \widehat{ARL}_1 and the control limits, in addition, in parentheses are \overline{ARL}_0 , \overline{ARL}_1 for the usual Shewhart control limits.

Table 2. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the Student- t with $\nu = 3$ (in parentheses are the \overline{ARL}_0 and \overline{ARL}_1 for the usual Shewhart limits).

Parameters		n	δ								
			LCL	UCL	-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 100.00, \phi = 1.00$ $\sigma^2 = 3.00, I_d = 0.03$	1	90.91	109.07	64.95 (4.08)	164.1 (10.72)	228.45 (23.58)	361.92 (72.68)	231.61 (23.52)	163.78 (10.75)	64.99 (4.08)	
	2	94.05	105.95	4.89 (1.56)	49.31 (2.48)	114.55 (7.78)	349.38 (79.85)	110.93 (7.72)	48.71 (2.45)	4.87 (1.07)	
	3	95.36	104.63	1.31 (1.01)	12.18 (1.48)	49.36 (3.37)	356.23 (85.81)	49.54 (3.30)	11.92 (1.35)	1.31 (1.01)	
	10	97.78	102.22	1.00 (1.00)	1.02 (1.00)	1.28 (1.00)	360.63 (112.37)	1.28 (1.00)	1.01 (1.00)	1.00 (1.00)	
	500	99.41	100.58	1.00 (1.00)	1.00 (1.00)	1.03 (1.00)	375.41 (175.90)	1.03 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 9.00, \phi = 2.00$ $\sigma^2 = 6.00, I_d = 0.67$	1	-3.82	21.81	64.37 (2.25)	162.81 (10.67)	226.73 (23.37)	351.00 (75.69)	220.81 (24.18)	158.93 (11.05)	62.39 (2.12)	
	2	0.58	17.40	4.98 (1.06)	49.48 (2.47)	113.92 (7.77)	348.40 (79.97)	112.77 (7.76)	48.99 (2.46)	4.83 (1.07)	
	3	2.44	15.58	1.32 (1.00)	12.03 (1.48)	49.55 (4.43)	360.44 (85.84)	50.98 (4.43)	12.58 (1.50)	1.32 (1.00)	
	10	5.85	12.15	1.00 (1.00)	1.01 (1.00)	1.29 (1.00)	351.47 (110.05)	1.28 (1.00)	1.01 (1.00)	1.00 (1.00)	
	500	8.18	9.82	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	358.47 (188.19)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 2.00, \phi = 1.00$ $\sigma^2 = 3.00, I_d = 1.50$	1	-7.08	11.04	65.10 (2.02)	161.00 (11.04)	224.12 (23.95)	349.92 (72.56)	226.8 (23.83)	160.31 (10.86)	63.45 (1.98)	
	2	-3.94	7.95	4.88 (1.04)	49.01 (2.46)	113.9 (7.76)	349.15 (79.81)	113.53 (7.53)	48.74 (2.47)	4.92 (1.00)	
	3	-2.64	6.64	1.31 (1.01)	12.18 (1.37)	49.65 (3.32)	354.25 (84.79)	49.65 (3.31)	12.18 (1.37)	1.31 (1.00)	
	10	-0.23	4.23	1.00 (1.00)	1.01 (1.00)	1.27 (1.00)	358.13 (110.50)	1.30 (1.00)	1.01 (1.00)	1.00 (1.00)	
	500	1.42	2.58	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	361.87 (178.28)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 3.00, \phi = 3.00$ $\sigma^2 = 9.00, I_d = 3.00$	1	-12.72	18.70	64.45 (2.01)	162.61 (10.57)	226.07 (24.34)	354.34 (71.58)	224.67 (23.21)	161.73 (10.74)	63.65 (1.98)	
	2	-7.28	13.29	4.88 (1.00)	49.05(3.21)	113.15(8.12)	354.46 (79.59)	112.42 (7.98)	48.27 (2.59)	4.93 (1.08)	
	3	-5.04	11.04	1.32 (1.00)	12.23 (1.24)	49.34 (4.02)	358.24 (84.78)	49.85 (3.31)	12.54 (1.12)	1.32 (1.02)	
	10	-0.86	6.87	1.00 (1.00)	1.01 (1.00)	1.30 (1.00)	352.81 (108.88)	1.30 (1.00)	1.02 (1.00)	1.00 (1.00)	
	500	2.00	4.00	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	368.31 (191.45)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

Table 3. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the Student- t with $\nu = 5$ (in parentheses are the ARL_0 and ARL_1 for the usual Shewhart limits).

Parameters	n	\widehat{LCL}	\widehat{UCL}	δ						
				-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 100.00, \phi = 2.00$ $\sigma^2 = 3.33, I_d = 0.03$	1	92.31	107.69	11.22 (2.03)	55.72 (7.13)	112.49 (17.93)	352.56 (87.86)	112.53 (17.71)	54.76 (7.87)	10.98 (1.97)
	2	95.07	104.93	1.48 (1.00)	7.15 (1.08)	23.83 (5.97)	352.57 (108.37)	23.90 (6.01)	7.13 (2.10)	1.47 (1.00)
	3	96.17	103.83	1.06 (1.01)	2.38 (1.44)	7.42 (3.23)	348.02 (125.75)	7.37 (3.44)	2.32 (1.92)	1.05 (1.00)
	10	98.12	101.88	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)	349.26 (199.15)	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	99.45	100.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	342.54 (342.54)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 5.00, \phi = 2.00$ $\sigma^2 = 3.33, I_d = 0.67$	1	-2.67	12.68	11.05(2.00)	54.23 (7.98)	108.10 (17.77)	339.16 (85.71)	108.00 (17.83)	54.50 (7.93)	10.98 (2.02)
	2	0.07	9.93	1.46 (1.02)	7.10 (2.35)	23.98 (5.99)	344.71 (126.00)	23.82 (6.02)	7.18 (2.49)	1.44 (1.01)
	3	1.69	8.83	1.06 (1.00)	2.31 (1.00)	7.49 (3.18)	352.36 (125.04)	7.39 (3.00)	2.33 (1.00)	1.06 (1.00)
	10	3.12	6.88	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)	342.91 (193.58)	1.06 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	4.45	5.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	341.06 (341.44)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 2.00, \phi = 2.00$ $\sigma^2 = 3.33, I_d = 1.67$	1	-5.69	9.68	11.41 (2.03)	55.49 (7.95)	110.58 (17.58)	350.81 (84.80)	109.42 (17.05)	55.32 (7.76)	11.29 (1.99)
	2	-2.93	6.93	1.47 (1.01)	7.18 (2.37)	24.03 (5.98)	342.45 (108.11)	24.07 (5.96)	7.17 (2.34)	1.45 (1.02)
	3	-1.83	5.83	1.07 (1.00)	2.37 (1.44)	7.43 (2.99)	353.08 (125.57)	7.51 (3.04)	2.32 (1.43)	1.06 (1.00)
	10	0.12	3.88	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)	344.14 (195.11)	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	1.45	2.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	342.59 (342.65)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 2.00$ $\sigma^2 = 3.33, I_d = 3.33$	1	-6.68	8.69	11.21 (2.00)	54.34 (7.98)	111.25 (17.82)	354.89 (86.91)	112.53 (17.86)	55.20 (7.99)	10.98 (2.00)
	2	-3.93	5.93	1.21 (1.10)	3.82 (2.35)	11.45 (5.94)	342.45 (108.01)	23.73 (6.05)	7.07 (2.40)	1.47 (1.11)
	3	-2.83	4.83	1.06 (1.00)	2.37 (1.45)	7.52 (3.12)	353.08 (125.33)	7.53 (2.87)	2.42 (2.00)	1.06 (1.00)
	10	-0.87	2.87	1.07 (1.00)	1.00 (1.00)	1.00 (1.00)	350.42 (201.52)	1.07 (1.01)	1.00 (1.00)	1.00 (1.00)
	100	0.45	1.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	346.90 (346.90)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	0.75	1.25	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	363.02 (363.02)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

Table 4. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the Student- t with $\nu = 10$ (in parentheses are the ARL_0 and ARL_1 for the usual Shewhart limits).

Parameters	n	\widehat{LCL}	\widehat{UCL}	δ						
				-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 40.00, \phi = 1.00$ $\sigma^2 = 1.25, I_d = 0.03$	1	36.08	43.91	3.88 (2.03)	16.22 (6.93)	41.07 (16.02)	349.80 (137.94)	40.03 (17.02)	15.69 (6.83)	3.38 (2.00)
	2	37.41	42.59	1.98 (1.00)	3.13 (2.33)	8.44 (5.56)	344.73 (185.14)	8.51 (5.60)	3.09 (2.33)	1.20 (1.00)
	3	37.94	42.06	1.02 (1.00)	1.63 (1.12)	3.65 (2.94)	348.76 (211.11)	3.73 (2.93)	1.65 (1.21)	1.02 (1.00)
	10	38.92	41.08	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	368.43 (309.80)	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	39.66	40.33	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	360.57 (360.57)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 2.00, \phi = 1.00$ $\sigma^2 = 1.25, I_d = 0.63$	1	-1.91	5.92	3.39 (2.01)	16.02 (5.84)	40.76 (16.05)	350.40 (137.34)	40.56 (16.05)	16.05 (5.85)	3.41 (1.98)
	2	-0.60	4.59	1.19 (1.02)	3.14 (2.36)	8.67 (5.57)	353.52 (183.05)	8.68 (5.53)	3.13 (2.32)	1.19 (1.00)
	3	-0.06	4.06	1.02 (1.00)	1.66 (1.43)	3.73 (2.94)	347.79 (209.95)	3.74 (2.97)	1.64 (1.49)	1.02 (1.00)
	10	0.92	3.08	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	362.89 (308.87)	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	1.66	2.33	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	359.04 (359.04)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 3.00, \phi = 4.00$ $\sigma^2 = 5.00, I_d = 1.67$	1	-4.68	10.80	3.38 (2.08)	15.69 (6.88)	40.23 (17.23)	350.53 (138.37)	40.46 (16.98)	16.15 (6.22)	3.42 (2.42)
	2	-2.19	8.19	1.19 (1.02)	3.16 (1.98)	8.64 (5.45)	353.80 (183.80)	8.58 (5.39)	3.12 (2.43)	1.19 (1.03)
	3	-1.12	7.12	1.02 (1.00)	1.62 (1.00)	3.69 (2.81)	349.12 (215.11)	3.69 (2.90)	1.63 (1.00)	1.02 (1.00)
	10	0.84	5.16	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	348.58 (302.91)	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	2.33	3.67	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	379.51 (379.51)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 2.00, \phi = 5.00$ $\sigma^2 = 6.25, I_d = 3.13$	1	-6.76	10.76	3.46 (2.04)	16.28 (6.11)	41.22 (17.07)	351.44 (135.79)	40.95 (16.97)	16.02 (5.94)	3.49 (1.99)
	2	-3.80	7.81	1.19 (1.00)	3.22 (1.98)	8.60 (5.21)	351.04 (185.96)	8.60 (4.99)	3.17 (1.49)	1.18 (1.00)
	3	-2.61	6.60	1.03 (1.00)	1.65 (1.02)	3.71 (2.38)	340.71 (211.32)	3.65 (2.61)	1.63 (1.02)	1.03 (1.01)
	10	-0.41	4.41	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	354.67 (303.27)	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	1.25	2.75	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	379.97 (379.98)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	1.67	2.33	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	345.63 (345.63)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

For both distributions considered, as n increases, the control limits get closer to the true value of the mean (process under control), regardless of the ν parameter, for the Student- t distribution, and the κ parameter for the power-exponential distribution. Furthermore, as expected, for the Student- t distribution, when n and ν increase, the sample mean distribution approaches a normal distribution. Thus, the control limits by the proposed method tend to approach the Shewhart's usual control limits. This behavior is also seen for the power-exponential distribution when n increases and κ approaches 0. This fact is noticeable when we observe the proximity of the ARL_0 values and considering the control limits using the proposed method and the usual Shewhart method. Moreover, the observed behavior of the proposed control limits occurs independently of the process dispersion index, thus showing the robustness in relation to this index.

Based on the ARL_0 for heavy-tailed data distribution, see Tables 2 to 5 and Tables 8 and 9, regardless of the scenario, the proposed method presents ARL_0 around 340 to 380

Table 5. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the Student- t with $\nu = 20$ (in parentheses are the \widehat{ARL}_0 and \widehat{ARL}_1 for the usual Shewhart limits).

Parameters	n	\widehat{LCL}	\widehat{UCL}	δ						
				-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 100.00, \phi = 3.00$ $\sigma^2 = 3.33, I_d = 0.03$	1	94.12	105.88	2.47 (2.00)	9.35 (6.49)	23.63 (15.54)	370.34 (205.13)	23.33 (15.48)	9.33 (6.54)	2.43 (2.00)
	2	95.99	104.01	1.15 (1.12)	2.59 (2.30)	6.31 (5.30)	338.44 (255.64)	6.22 (5.39)	2.59 (2.34)	1.14 (1.12)
	3	96.77	103.23	1.00 (1.00)	1.10 (1.02)	1.48 (1.13)	350.19 (293.38)	3.13 (1.10)	1.15 (1.01)	1.01 (1.00)
	10	98.27	101.73	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	340.78 (336.88)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	99.45	100.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	363.70 (363.70)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 5.00, \phi = 3.00$ $\sigma^2 = 3.33, I_d = 0.67$	1	-0.87	10.87	2.43 (2.02)	9.21 (6.56)	23.22 (15.63)	345.35 (205.13)	23.60 (15.49)	9.33 (6.60)	2.43 (2.00)
	2	0.99	10.10	1.15 (1.12)	2.60 (3.25)	6.32 (5.37)	350.44 (260.74)	6.21 (5.44)	2.55 (2.30)	1.15 (1.12)
	3	3.27	7.73	1.02 (1.02)	1.48 (1.09)	3.04 (2.95)	352.36 (284.71)	3.39 (2.93)	2.33 (1.46)	1.06 (1.02)
	10	3.27	7.73	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)	341.78 (338.52)	1.03 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	4.45	5.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	363.48 (368.48)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 2.00, \phi = 3.00$ $\sigma^2 = 3.33, I_d = 1.67$	1	-3.87	7.87	2.43 (1.99)	9.30 (6.59)	23.22 (15.41)	349.33 (206.78)	23.35 (15.47)	9.19 (6.61)	2.42 (2.03)
	2	-2.00	6.00	1.14 (1.08)	2.55 (2.36)	6.29 (5.41)	342.47 (257.90)	6.25 (5.40)	2.56 (2.33)	1.14 (1.02)
	3	-1.23	5.23	1.02 (1.01)	1.52 (1.46)	3.17 (2.93)	345.04 (285.07)	3.15 (2.95)	1.52 (1.47)	1.01 (1.00)
	10	0.27	3.73	1.03 (1.00)	1.00 (1.00)	1.00 (1.00)	347.54 (343.02)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	1.45	2.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	347.78 (347.65)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 3.00$ $\sigma^2 = 3.33, I_d = 3.33$	1	-4.87	6.87	2.43 (2.00)	9.29 (6.74)	23.19 (15.76)	344.41 (205.14)	23.23 (15.17)	9.25 (6.58)	2.49 (2.00)
	2	-3.00	5.00	1.14 (1.00)	2.58 (2.34)	6.31 (5.32)	342.45 (259.21)	6.33 (5.44)	2.56 (2.38)	1.14 (1.01)
	3	-2.23	4.23	1.02 (1.00)	1.52 (1.34)	3.18 (2.83)	345.04 (285.07)	3.15 (2.90)	1.52 (1.39)	1.01 (1.00)
	10	-0.73	2.73	1.05 (1.00)	1.00 (1.00)	1.00 (1.00)	347.54 (343.02)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	0.45	1.55	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	371.68 (371.68)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	0.75	1.25	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	370.84 (370.84)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

Table 6. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the power-exponential with $\kappa = -0.45$ (in parentheses are the \widehat{ARL}_0 and \widehat{ARL}_1 for the usual Shewhart limits).

Parameters	n	\widehat{LCL}	\widehat{UCL}	δ						
				-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 50.00, \phi = 3.00$ $\sigma^2 = 1.52, I_d = 0.03$	1	47.03	52.97	1.43 (1.97)	2.69 (5.42)	4.72 (14.95)	331.28 (2453.04)	4.83 (14.94)	2.75 (5.53)	1.47 (2.09)
	2	47.68	52.32	1.06 (1.13)	1.79 (2.31)	3.32 (4.98)	334.62 (1674.89)	3.30 (5.05)	1.76 (2.29)	1.06 (1.12)
	3	48.02	51.98	1.00 (1.01)	1.34 (1.48)	2.35 (2.93)	347.29 (978.20)	2.30 (2.87)	1.34 (1.49)	1.00 (1.01)
	10	48.86	51.14	1.00 (1.00)	1.00 (1.00)	1.03 (1.04)	357.29 (475.48)	1.04 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	49.63	50.37	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	362.80 (429.02)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 3.00, \phi = 4.00$ $\sigma^2 = 2.03, I_d = 0.67$	1	-0.43	6.43	1.43 (1.98)	2.70 (5.44)	4.74 (15.15)	338.91 (2353.73)	4.76 (14.82)	2.69 (5.50)	1.45 (2.01)
	2	0.32	5.68	1.06 (1.13)	1.77 (2.30)	3.29 (4.96)	342.17 (1629.00)	3.47 (5.09)	1.77 (2.30)	1.06 (1.13)
	3	0.72	5.28	1.01 (1.01)	1.34 (1.48)	2.33 (2.84)	342.95 (975.43)	2.27 (2.86)	1.33 (1.50)	1.01 (1.01)
	10	1.68	4.32	1.00 (1.00)	1.00 (1.00)	1.03 (1.05)	366.09 (463.89)	1.04 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	2.58	3.43	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	360.78 (360.78)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 3.00$ $\sigma^2 = 1.52, I_d = 1.52$	1	-1.97	3.97	1.43 (2.00)	2.74 (5.42)	4.71 (14.83)	348.63 (2510.84)	4.84 (14.76)	2.79 (5.55)	1.45 (2.04)
	2	-1.32	3.32	1.07 (1.13)	1.80 (2.30)	3.29 (4.97)	353.29 (961.80)	3.28 (5.03)	1.77 (2.27)	1.06 (1.12)
	3	-0.98	2.89	1.00 (1.01)	1.36 (1.50)	2.32 (2.89)	353.29 (961.80)	2.27 (2.86)	1.34 (1.48)	1.00 (1.01)
	10	-0.14	2.14	1.00 (1.00)	1.00 (1.00)	1.04 (1.03)	358.79 (474.93)	1.03 (1.05)	1.00 (1.00)	1.00 (1.00)
	100	0.46	1.54	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	425.74 (425.74)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 6.00$ $\sigma^2 = 3.04, I_d = 3.04$	1	-3.20	5.20	1.45 (2.01)	2.75 (5.53)	4.89 (14.85)	357.02 (2453.02)	4.79 (14.76)	2.71 (5.48)	1.46 (2.00)
	2	-2.28	4.28	1.07 (1.14)	1.78 (2.31)	3.30 (5.02)	338.52 (1653.62)	3.81 (5.05)	1.79 (2.32)	1.06 (1.13)
	3	-1.79	3.79	1.00 (1.00)	1.33 (1.33)	2.34 (2.33)	342.48 (990.77)	2.31 (2.88)	1.32 (1.48)	1.00 (1.00)
	10	-0.61	2.61	1.00 (1.00)	1.00 (1.00)	1.04 (1.05)	347.92 (450.07)	1.05 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	0.48	1.52	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	393.67 (393.67)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	0.77	1.23	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	365.62 (365.62)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

samples. Considering subsample sizes $n = \{1, 2, 3\}$ (most used in practical situations) the proposed method presents ARL_0 closer to 370.40 (target value) than the usual Shewhart method (see column $\delta = 0$ of Tables 2 to 5 and Tables 8 and 9). This behavior occurs independently of the dispersion index, showing the flexibility of the method for different situations. Considering the power of detection of the proposed method, for heavy-tailed distributions, some patterns are observed for all scenarios, they are: (i) for a fixed deviation δ , as n increases, the ARL_1 decreases, approaching one sample; (ii) for a fixed subsample size n , as δ increases, ARL_1 decreases, also approaching one sample.

Table 7. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the power-exponential with $\kappa = -0.25$ (in parentheses are the \widehat{ARL}_0 and \widehat{ARL}_1 for the usual Shewhart limits).

Parameters	n	δ								
		\widehat{LCL}	\widehat{UCL}	-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 20.00, \phi = 1.00$ $\sigma^2 = 0.67, I_d = 0.03$	1	17.82	22.18	1.63 (2.04)	3.73 (5.83)	7.78 (14.85)	353.72 (1475.83)	7.61 (14.66)	3.68 (5.62)	1.61 (1.98)
	2	18.38	21.62	1.09 (1.13)	1.96 (2.30)	4.01 (5.11)	348.92 (785.06)	3.88 (5.00)	1.96 (2.28)	1.08 (1.12)
	3	18.65	21.35	1.00 (1.00)	1.37 (1.05)	2.56 (2.92)	354.29 (583.45)	2.51 (2.86)	1.37 (1.47)	1.00 (1.00)
	10	19.24	20.76	1.00 (1.00)	1.00 (1.00)	1.03 (1.04)	361.93 (414.45)	1.04 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	19.76	20.24	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	373.80 (373.80)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 1.00$ $\sigma^2 = 0.67, I_d = 0.67$	1	-1.18	3.18	1.61 (2.01)	3.73 (5.87)	7.74 (14.62)	356.99 (1468.41)	7.75 (14.84)	3.75 (5.85)	1.60 (1.99)
	2	-0.62	2.62	1.08 (1.13)	1.97 (2.30)	3.93 (5.04)	356.25 (783.09)	4.00 (5.11)	1.97 (2.30)	1.09 (1.26)
	3	-0.34	2.35	1.01 (1.01)	1.38 (1.48)	2.52 (2.84)	352.44 (580.43)	2.53 (2.86)	1.39 (1.50)	1.00 (1.01)
	10	0.24	0.76	1.00 (1.00)	1.00 (1.00)	1.03 (1.04)	362.04 (413.75)	1.05 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	0.76	0.24	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	377.50 (377.50)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 2.00, \phi = 5.00$ $\sigma^2 = 3.34, I_d = 1.67$	1	-2.87	6.87	1.62 (2.01)	3.72 (5.89)	7.72 (14.82)	347.26 (1474.67)	7.62 (14.95)	3.68 (5.90)	1.61 (2.03)
	2	-1.62	5.62	1.08 (1.12)	1.94 (2.30)	3.85 (5.04)	343.57 (765.80)	3.94 (5.11)	1.98 (2.34)	1.08 (1.29)
	3	-1.02	5.02	1.01 (1.01)	1.39 (1.48)	2.52 (2.52)	352.94 (588.16)	2.54 (2.91)	1.39 (1.49)	1.00 (1.00)
	10	0.30	3.70	1.00 (1.00)	1.00 (1.00)	1.04 (1.05)	346.49 (414.75)	1.04 (1.05)	1.00 (1.00)	1.00 (1.00)
	100	0.46	2.54	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	353.19 (353.19)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 3.00$ $\sigma^2 = 3.34, I_d = 3.34$	1	-3.87	5.87	1.61 (2.00)	3.64 (5.75)	7.55 (14.84)	344.61 (1447.16)	7.61 (14.76)	3.69 (5.80)	1.60 (2.00)
	2	-2.62	4.62	1.08 (1.12)	1.95 (2.34)	3.92 (5.17)	342.42 (774.74)	3.98 (5.08)	1.98 (2.94)	1.09 (1.12)
	3	-2.02	4.02	1.00 (1.00)	1.38 (1.49)	2.50 (2.88)	350.27 (589.64)	2.51 (2.91)	1.38 (1.50)	1.00 (1.00)
	10	-0.70	2.70	1.00 (1.00)	1.00 (1.00)	1.04 (1.05)	353.87 (422.70)	1.03 (1.04)	1.00 (1.00)	1.00 (1.00)
	100	1.46	2.54	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	347.99 (347.99)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	0.75	1.25	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	365.62 (365.62)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

Table 8. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the power-exponential with $\kappa = 0.30$ (in parentheses are the \widehat{ARL}_0 and \widehat{ARL}_1 for the usual Shewhart limits).

Parameters	n	δ								
		\widehat{LCL}	\widehat{UCL}	-3.0	-2.0	-1.5	0	1.5	2.0	3.0
$\mu = 100.00, \phi = 2.00$ $\sigma^2 = 3.48, I_d = 0.03$	1	93.77	106.23	2.84 (1.99)	11.75 (6.99)	27.49 (15.43)	342.60 (156.80)	27.98 (15.35)	11.64 (6.99)	2.85 (2.00)
	2	95.80	104.20	1.16 (1.00)	2.79 (2.31)	7.24 (5.57)	339.54 (211.36)	7.26 (5.63)	2.82 (2.33)	1.16 (1.00)
	3	96.63	103.37	1.02 (1.00)	1.57 (1.10)	3.37 (1.85)	345.02 (243.33)	2.94 (1.98)	1.46 (1.10)	1.04 (1.00)
	10	98.21	101.78	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	341.86 (328.73)	1.04 (1.01)	1.00 (1.00)	1.00 (1.00)
	100	99.44	100.56	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	379.83 (379.83)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 5.00, \phi = 2.00$ $\sigma^2 = 3.48, I_d = 0.69$	1	-1.23	11.23	2.86 (2.00)	11.42 (7.04)	27.82 (15.48)	344.63 (159.19)	27.85 (15.68)	11.73 (7.02)	2.87 (1.98)
	2	0.80	9.20	1.16 (1.00)	2.81 (2.01)	7.13 (4.88)	345.25 (213.12)	7.33 (4.65)	2.84 (1.66)	1.17 (1.00)
	3	1.64	8.36	1.03 (1.00)	1.56 (1.00)	3.41 (2.01)	337.44 (245.43)	3.37 (1.98)	1.56 (1.02)	1.02 (1.00)
	10	3.22	6.78	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	331.18 (317.76)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	4.43	5.57	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	360.24 (360.24)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 1.00$ $\sigma^2 = 1.74, I_d = 1.74$	1	-3.41	5.41	2.85 (2.00)	11.78 (6.21)	28.40 (15.38)	340.78 (154.31)	28.40 (16.98)	11.79 (6.82)	2.90 (1.99)
	2	-1.97	3.97	1.16 (1.00)	2.76 (2.33)	7.22 (5.49)	343.14 (210.14)	7.16 (4.44)	2.82 (2.01)	1.16 (1.00)
	3	-1.38	3.38	1.02 (1.00)	1.56 (1.12)	3.40 (2.88)	345.69 (245.38)	3.31 (2.53)	1.56 (1.24)	1.02 (1.00)
	10	-0.26	3.38	1.00 (1.00)	1.00 (1.00)	1.04 (1.00)	336.31 (317.73)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	0.61	1.39	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	365.30 (365.30)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
$\mu = 1.00, \phi = 2.00$ $\sigma^2 = 3.48, I_d = 3.48$	1	-5.23	7.23	2.84 (2.01)	11.85 (6.33)	28.50 (16.22)	341.77 (156.24)	27.70 (15.99)	11.64 (7.01)	2.86 (2.12)
	2	-3.20	5.20	1.16 (1.00)	2.82 (1.87)	7.23 (3.34)	342.88 (212.48)	7.30 (3.88)	2.81 (1.97)	1.16 (1.00)
	3	-2.36	4.37	1.02 (1.00)	1.57 (1.21)	3.43 (1.87)	346.42 (249.51)	3.43 (2.01)	1.56 (1.21)	1.02 (1.00)
	10	-0.79	2.78	1.00 (1.00)	1.00 (1.00)	1.04 (1.00)	345.99 (319.95)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)
	100	0.61	1.39	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	361.58 (362.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
500	0.75	1.25	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	363.47 (363.47)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	

For the Student- t distribution in the most extreme scenarios ($\nu = 3$ and 5 ; see Tables 2 and 3), due to the behavior of the distribution, using the subsample size $n = 1$ does not prove to be the most recommended in these situations. However, when $n = 2$ the \widehat{ARL}_0 reduces considerably (about 50% or more), mainly when $\delta = 3$. In view of this, in situations where the data distribution have very heavy tails, we recommend the use of $n \geq 3$, as these subsample sizes have excellent \widehat{ARL}_0 , closer to the target value 370.40 than the usual Shewhart method. Regarding the heavy-tailed power-exponential distribution (Tables 8 and 9), the method shows excellent detection power, compatible with the usual Shewhart chart. Still on \widehat{ARL}_1 , for the power-exponential distribution with $0 < \kappa < 1$, it takes, on average, 30 samples to detect a shift of 1.5 standard deviations in the process

Table 9. Control limits, \widehat{ARL}_0 and \widehat{ARL}_1 of the proposed method, considering the power-exponential with $\kappa = 0.40$ (in parentheses are the \widehat{ARL}_0 and \widehat{ARL}_1 for the usual Shewhart limits).

Parameters	n	\widehat{LCL}	\widehat{UCL}	δ							
				-3.0	-2.0	-1.5	0	1.5	2.0	3.0	
$\mu = 200.00, \phi = 3.00$ $\sigma^2 = 6.38, I_d = 0.03$	1	192.27	208.74	3.38 (2.09)	14.31 (8.01)	33.75 (15.95)	343.75 (128.71)	34.42 (16.01)	14.33 (7.87)	3.29 (1.99)	
	2	194.19	205.81	1.18 (1.00)	3.06 (2.04)	8.11 (5.98)	339.18 (183.49)	8.24 (5.78)	3.09 (2.09)	1.19 (1.00)	
	3	195.37	204.65	1.02 (1.00)	1.61 (1.21)	3.62 (2.44)	350.76 (220.57)	3.58 (2.68)	1.60 (1.10)	1.02 (1.00)	
	10	197.57	202.43	1.00 (1.00)	1.00 (1.00)	1.05 (1.00)	341.72 (301.71)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)	
	500	199.25	200.75	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	363.53 (363.53)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 6.00, \phi = 2.00$ $\sigma^2 = 4.25, I_d = 0.69$	1	-1.13	13.13	3.38 (1.89)	14.49 (7.55)	34.42 (15.92)	352.32 (131.78)	33.83 (15.09)	14.32 (7.82)	3.67 (2.41)	
	2	1.25	10.74	1.18 (1.00)	3.03 (1.89)	8.18 (5.87)	344.72 (186.21)	8.28 (5.91)	3.03 (2.01)	1.18 (1.00)	
	3	2.22	9.78	1.07 (1.00)	1.63 (1.12)	3.69 (2.56)	344.36 (217.27)	3.61 (2.57)	1.60 (1.10)	1.02 (1.00)	
	10	4.02	7.98	1.00 (1.00)	1.00 (1.00)	1.04 (1.00)	338.18 (299.91)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)	
	500	5.38	6.61	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	365.11 (365.11)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 4.00, \phi = 3.00$ $\sigma^2 = 6.38, I_d = 1.60$	1	-4.73	12.73	3.40 (2.02)	14.33 (7.20)	33.99 (15.50)	344.13 (128.67)	33.75 (16.01)	14.29 (7.92)	3.39 (2.11)	
	2	-1.81	9.81	1.17 (1.00)	3.05 (2.15)	8.23 (6.21)	340.18 (184.24)	8.17 (6.64)	3.08 (2.65)	1.18 (1.00)	
	3	-0.63	8.63	1.02 (1.00)	1.61 (1.04)	3.61 (2.44)	347.98 (218.82)	3.65 (2.44)	1.62 (1.13)	1.02 (1.00)	
	10	1.57	6.43	1.00 (1.00)	1.00 (1.00)	1.04 (1.00)	340.06 (298.09)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)	
	500	3.21	4.76	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	359.28 (359.28)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
$\mu = 2.00, \phi = 3.00$ $\sigma^2 = 6.38, I_d = 3.19$	1	-6.73	10.73	3.31 (2.00)	14.32 (7.11)	34.18 (15.34)	339.57 (126.84)	34.69 (15.42)	14.62 (7.12)	3.36 (2.00)	
	2	-3.81	7.81	1.18 (1.00)	3.09 (2.31)	8.30 (5.46)	345.06 (193.59)	8.25 (5.45)	3.09 (2.35)	1.18 (1.00)	
	3	-2.63	6.62	1.02 (1.00)	1.61 (1.10)	3.64 (2.65)	336.19 (215.73)	3.68 (2.67)	1.62 (1.00)	1.02 (1.00)	
	10	-0.43	4.43	1.00 (1.00)	1.00 (1.00)	1.04 (1.00)	344.24 (304.75)	1.04 (1.00)	1.00 (1.00)	1.00 (1.00)	
	500	1.24	2.74	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	360.82 (360.82)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	
500	1.66	2.34	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	368.28 (368.28)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)		

mean, when using $n = 1$. However, with $n = 2$, the ARL_0 is already reduced to about 8 to 10 samples. This fact only reinforces the excellent applicability of the method for data with heavy tails.

Considering our method for light-tailed distributions (Tables 6 and 7), which furnishes a ARL_0 around 332 to 370 samples, the ARL_0 provided by the usual Shewhart method has higher values (especially when $n < 100$), around 360 to 2400 samples. This peculiar fact of high ARL_0 occurs because the tail of the normal distribution is heavier than that of the power-exponential distribution. Besides, when the process is under control, a sample (from the power-exponential distribution) will rarely exceed the usual Shewhart limits (designed to cover 99.73% of the samples when the process is normally distributed and under control). On the other hand, despite these high ARL_0 , the usual procedure shows to detect well large deviations in the process mean, taking a maximum of 16, 6 and 4 samples to detect shifts of 1.5, 2.0 and 3.0 standard deviations, respectively. Moreover, these ARL_0 reduce as n increases. This fact is, at first sight, surprising, as it is expected that the higher the ARL_0 , the longer it will take to detect a true alarm. However, as we are in the situation of large shifts (greater than 1.5 standard deviations), these changes in the mean are sufficient for the usual Shewhart chart to indicate such changes satisfactorily. Still for light-tailed distributions, the proposed method meet expectations and provides the ARL_1 within the desired range, since the proposed approach is based on the true distribution of the data (in this case the power-exponential). In addition to that, with regard to the power of detection, the proposed method always presents ARL_1 less than or equal to the usual method. Furthermore, it is worth mentioning that for both types of tail weights, the usual method detects changes in the mean almost instantly ($ARL_1 = 1$) when $n \geq 10$.

In short, the proposed method presents excellent performance in term of ARL_0 and ARL_1 , being the most recommended in cases with distributions of heavier tail than the normal distribution. On the other hand, in the context of light-tailed distributions, the usual Shewhart method is recommended, because although the proposed method has excellent performance, the Shewhart \bar{X} presents ARL_0 equal to or greater than that of the proposed method and it is comparable to the detection power of the proposed approach.

5. A REAL EXAMPLE

This section illustrates the applicability of the method proposed in Section 3, when the data is symmetric, but not necessarily with normal distribution. The data refers to the result of the pH of 1599 red wines produced by the Portuguese company Vinho Verde, one of the largest wine producers in Portugal, from May 2004 to February 2007. Cortez et al. (2009) provided more details about the Vinho Verde company and the specifications of the data set. The data used are available at <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>. In this section we also consider a comparison of the proposed method to obtain the control limits with the usual Shewhart method.

Based on Cortez et al. (2009), there are strong indications that the 1599 observations come from a process under control. Therefore, the first thousand observations are considered for phase I (process of constructing a control chart). In phase I, we perform a visual graphic analysis (see Figure 1), the descriptive statistics (see Table 10) and a symmetry test using the `symmetry.test` function of the `lawstat` package of R, which is based on Miao et al. (2006). In the normal boxplot in Figure 1 we note some possible “atypical” points. However, as the data is under control, it is more likely that these points are just points in the tail of a heavier tailed distribution than the normal distribution. In Table 10, we can observe that the mean are close to the median, differing just around one standard deviation ($\sqrt{0.249} \approx 0.5$). Furthermore, the coefficient of skewness and kurtosis are both close to zero, which is a indication of symmetry. In addition to the strong suggestion of symmetry, observed in Figure 1, and a brief analysis of the descriptive statistics, the symmetry test provides a p -value of 0.74, that is, there is a strong statistical evidence to not reject the hypothesis of data symmetry.

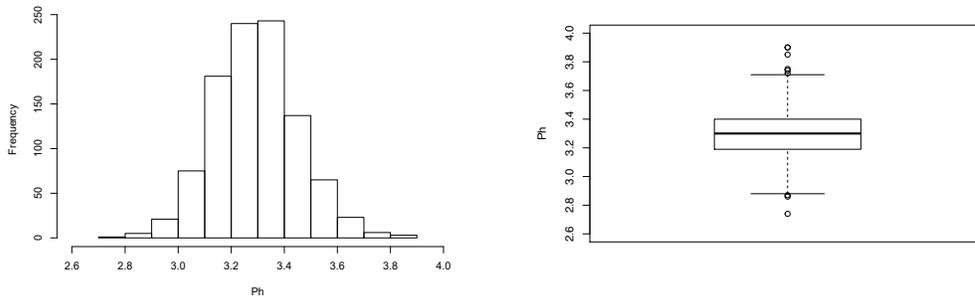


Figure 1. PH Histogram and boxplot of the first thousand wines.

Table 10. Descriptive statistics related to the first thousand pH observations (phase I) of the red wines.

Minimum	Mean	Median	Maximum	Variance	Dispersion index	Skewness	Kurtosis
2.740	3.299	3.300	3.900	0.249	0.008	0.185	0.417

After the assumption of symmetry is considered reasonable, we perform model adequacy tests to find out which symmetric distribution best fits the data. Table 11 shows the models considered, the estimated parameters, AIC and BIC. These measures and estimates are obtained using the `gamlss` package (Stasinopoulos et al., 2007) of the R software.

Based on Table 11, we see that the most suitable model for the data is the Student- t model with $\hat{\mu} = 3.299$, $\hat{\phi} = 0.007$, $\hat{\nu} = 2.841$ (which provides an estimated standard error $\hat{\sigma} = 0.1484$), with the lowest AIC and BIC among the concurrent models. In Figure 2, we present the quantile residuals (Dunn and Smyth, 1996) for the Student- t model, obtained using `gamlss` function. As expected, the quantile residual for the adjusted model are independent and normally distributed, which indicates that the postulated model is reasonable to the data.

Table 11. Parameter estimates, AIC and BIC for the models considered for the pH of red wines.

Distribution	Parameter estimates	AIC	BIC
Normal	$\hat{\mu} = 3.299, \hat{\phi} = 0.025$	-850.104	-840.288
Student- <i>t</i>	$\hat{\mu} = 3.299, \hat{\phi} = 0.007, \hat{\nu} = 2.841$	-853.585	-848.862
Power-exponential	$\hat{\mu} = 3.299, \hat{\phi} = 0.008, \hat{\kappa} = 0.558$	-852.388	-847.665
Type I logistic	$\hat{\mu} = 3.299, \hat{\phi} = 0.031$	-849.183	-839.367

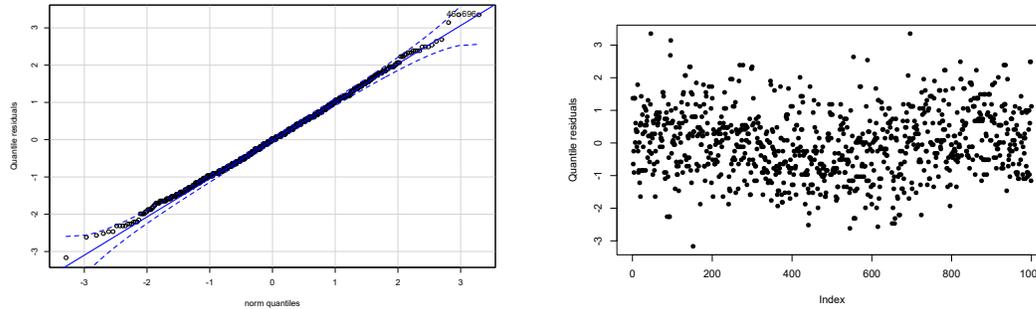


Figure 2. QQ-Plot and dispersion graphic of the quantile residuals for the Student-*t* model.

Thus, considering the Student-*t* model and the estimated parameters, we use the procedure described in Section 3 with $n = 1$ (chosen to preserve the original data monitoring scale) and a probability of false alarm equal to 0.0027. For this configuration, we obtain $LCL = 2.50$, $UCL = 4.10$ and an estimated ARL_0 of 373 samples. We see in Figure 3 (left) that the proposed method do not detect any change in the pH of the monitored wines. In contrast, in Figure 3 (right), we see that, even though the data is under control, the usual method of Shewhart, based on the normal distribution, detects changes in the average pH of the wines, thus generating false alarms. As expected, the proposed method performs better than the usual Shewhart method, in relation to the number of samples until a false alarm, when the data has a heavier tail than the normal distribution (in this case, Student-*t* distribution with 2.85 degrees of freedom).

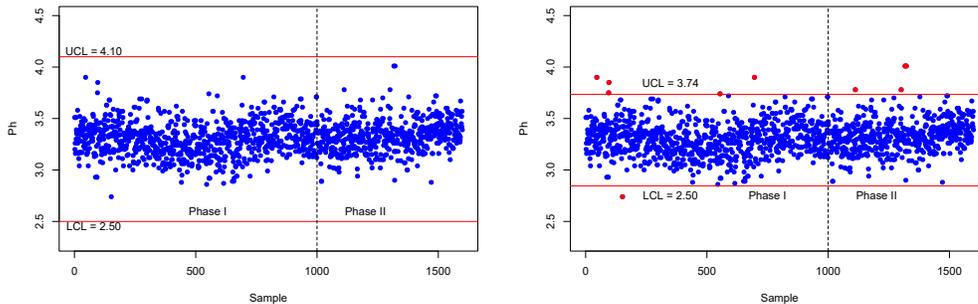


Figure 3. Control chart for monitoring the average pH of red wines from the Vinho Verde company produced from May 2004 to February 2007 with by the proposed (left) and he usual Shewhart (right) methods.

6. FINAL CONSIDERATIONS

In this work, a monitoring method via bootstrap is proposed and evaluated in order to monitor the mean of symmetric data whose distribution belongs to the symmetric distribution class. This method comes as an alternative to Shewhart \bar{X} chart when we want to monitor non-normal symmetric data, especially with heavy-tailed distribution data. The simulation study (illustrated with the Student- t and power-exponential distribution) shows that the proposed approach, for $\alpha = 0.0027$, provides in-control average run length between 340 and 380 samples and a good detection power, approaching one sample as n increases. Regarding the behavior of the control limits, for the proposed approach, they become closer to the mean when n increases.

In the context of light-tailed distribution, the proposed method presents good performance in-control average run length close to the desired value and good detection power). However, it is recommended the usual Shewhart \bar{X} chart, because in addition to presenting a detection power comparable to the proposed method, it has a false alarm rate lower than that of the proposed method. It is worth nothing that the great advantage of using the proposed method, instead of the usual Shewhart method, is in situations in which the data distribution has a heavier tail than normal distribution, since the proposed method has a lower rate of false alarm (being very close to the nominal value) and excellent detection power, as seen in the simulation study and illustrated in the monitoring of the average pH of red wines. Finally, the proposed method is robust to dispersion index variation. As future work we highlight: (i) to analyze the effect of the parameter estimation in the proposed method, (ii) to consider a joint monitoring of the mean and the standard deviation for symmetric class data and, (iii) to propose an EWMA and CUSUM charts for the symmetric class to monitor small deviations from the mean.

ACKNOWLEDGEMENT

To the reviewers for valuable contributions to this paper. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001 and CNPq-Brazil (421656/2018-2).

REFERENCES

- Ahmed, A., Sanaullah, A., and Hanif, M. 2019. A robust alternate to the HEWMA control chart under non-normality. *Quality Technology Quantitative Management*, 17, 423–47.
- Bai, D.S. and Choi, I. 1995. X and R control charts for skewed populations. *Journal of Quality Technology*, 27, 120–131.
- Bajgier, S., 1992. The use of bootstrap to construct limits on control charts. In *Proceedings of the decision Science Institute*, pp. 1611–1613.
- Berkane, M. and Bentler, P., 1986. Moments of elliptically distributed random variates. *Statistics and Probability Letters*, 4, 333–335.
- Borrer, C.M., Montgomery, D.C., and Runger, G.C., 1999. Robustness of the EWMA control chart to non-normality. *Journal of Quality Technology*, 31, 309–316.
- Calzada, M. and Scariano, S., 2001. The robustness of the synthetic control chart to non-normality. *Communications in Statistics: Simulation and Computation*, 30, 311–326.
- Chakraborti, S., Human, S., and Graham, M., 2008. Phase I statistical process control charts: an overview and some results. *Quality Engineering*, 21, 52–62.
- Chakraborti, S., Van der Laan, P., and Bakir, S., 2001. Nonparametric control charts: an overview and some results. *Journal of Quality Technology*, 33, 304–315.

- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47, 547–553.
- da Silva, G. Taconeli, C., Zeviani, W., and do Nascimento, I., 2019. Performance of Shewhart control charts based on neoteric ranked set sampling to monitor the process mean for normal and non-normal processes. *Chilean Journal of Statistics*, 10, 131–154.
- Davison, A.C. and Hinkley, D.V., 1997. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge, UK.
- Dunn, P.K. and Smyth, G.K., 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5, 236–244.
- Efron, B. and Tibshirani, R.J., 1994. *An Introduction to the Bootstrap*. CRC Press, Boca Raton, FL, US.
- Fang, K., Kotz, S., and Ng, K., 1990. *Symmetric Multivariate and Related Distribution*. Chapman Hall, London.
- Gandy, A. and Kvaløy, J.T., 2013. Guaranteed conditional performance of control charts via bootstrap methods. *Scandinavian Journal of Statistics*, 40, 647–668.
- Haq, A. and Khoo, M.B., 2019. A parameter-free adaptive EWMA mean chart. *Quality Technology and Quantitative Management*, 17:528–543.
- Lange, K.L., Little, R.J., and Taylor, J.M., 1989. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84, 881–896.
- Liu, R.Y. and Tang, J., 1996. Control charts for dependent and independent measurements based on bootstrap methods. *Journal of the American Statistical Association*, 91, 1694–1700.
- Medeiros, F.M. and Ferrari, S.L., 2017. Small-sample testing inference in symmetric and log-symmetric linear regression models. *Statistica Neerlandica*, 71, 200–224.
- Miao, W., Gel, Y.R., and Gastwirth, J.L., 2006. A new test of symmetry about an unknown median. In *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, pp. 199–214. World Scientific, Singapore.
- Naveed, M., Azam, M., Khan, N., and Aslam, M., 2020. Designing a control chart of extended EWMA statistic based on multiple dependent state sampling. *Journal of Applied Statistics*, 47, 1482–1492.
- Noorossana, R., Vaghefi, A., and Dorri, M., 2011. Effect of non-normality on the monitoring of simple linear profiles. *Quality and Reliability Engineering International*, 27, 425–436.
- Qiu, P. and Zhang, J., 2015. On phase II SPC in cases when normality is invalid. *Quality and Reliability Engineering International*, 31, 27–35.
- R Core Team, 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, B.P., 1990. Remarks on univariate elliptical distributions. *Statistics and Probability Letters*, 10, 307–315.
- Rezac, J., Lio, Y. L., and Jiang, N., 2015. Burr type-XII percentile control charts. *Chilean Journal of Statistics*, 6(1), 67–87.
- Schilling, E.G. and Nelson, P.R., 1976. The effect of non-normality on the control limits of \bar{X} charts. *Journal of Quality Technology*, 8, 183–188.
- Seppala, T., Moskowitz, H., Plante, R., and Tang, J., 1995. Statistical process control via the subgroup bootstrap. *Journal of Quality Technology*, 27, 139–153.
- Shewhart, W.A., 1931. *Economic Control of Quality of Manufactured Product*. D Van Nostrand Company, New York, US.
- Stasinopoulos, D.M. and Rigby, R.A., 2007. Generalized additive models for location scale and shape (gamlss) in R. *Journal of Statistical Software*, 23(7), 1–46.

- Tadikamalla, P., Banciu, M., and Popescu, D., 2008. An improved range chart for normal and long-tailed symmetrical distributions. *Naval Research Logistics*, 55, 91–99.
- Tadikamalla, P. and Popescu, D., 2007. Kurtosis correction method for \bar{X} and R control charts for long-tailed symmetrical distributions. *Naval Research Logistics*, 54, 371–383.
- Tsai, T.R., Lin, J.J., Wu, S.J., and Lin, H.C., 2005. On estimating control limits of \bar{X} chart when the number of subgroups is small. *The International Journal of Advanced Manufacturing Technology*, 26, 1312–1316.
- Willemain, T.R. and Runger, G.C., 1996. Designing control charts using an empirical reference distribution. *Journal of Quality Technology*, 28, 31–38.
- Zhang, L., Chen, G., and Castagliola, P., 2009. On t and EWMA t charts for monitoring changes in the process mean. *Quality and Reliability Engineering International*, 25, 933–945.

BAYESIAN SURVIVAL ANALYSIS
RESEARCH PAPER

Bayesian long-term survival model including a frailty term: Application to melanoma data

AGATHA S. RODRIGUES^{1,*}, VINICIUS F. CALSAVARA^{2,3}, EDUARDO BERTOLLI⁴,
STELA V. PERES⁵, and VERA L.D. TOMAZELLA⁶

¹Department of Statistics, Federal University of Espírito Santo, Vitória-ES, Brazil

²Department of Epidemiology and Statistics, A.C.Camargo Cancer Center, São Paulo-SP, Brazil

³Biostatistics and Bioinformatics Research Center, Cedars-Sinai Medical Center, Los Angeles-CA, USA

⁴Skin Cancer Department, A.C.Camargo Cancer Center, São Paulo-SP, Brazil

⁵Department of Information and Epidemiology, Fundação Oncocentro de São Paulo, São Paulo-SP, Brazil

⁶Department of Statistics, Federal University of São Carlos, São Carlos-SP, Brazil

(Received: 04 August 2020 · Accepted in final form: 28 January 2021)

Abstract

In this paper, we propose a flexible cure rate model including a frailty term, which was obtained by incorporating a random effect in the risk function of latent competing causes. The number of competing causes of the event of interest follows a negative binomial distribution, and the frailty variable follows a power variance function distribution, which includes other frailty models such as gamma, positive stable, and inverse Gaussian frailty models as special cases. The proposed model takes into account the presence of covariates and right-censored data, which are suitable for populations with a long-term survivors. Besides, it allows quantification of the degree of unobserved heterogeneity induced by unobservable risk factors, which is important to explain the lifetime. Once the posterior density function is not expressed in the closed form, Markov chain Monte Carlo algorithms are performed for the estimation procedure. Simulation studies were considered in order to evaluate the proposed model performance, and its practical relevance was illustrated in a real medical dataset from a population-based study of incident cases of melanoma diagnosed in the state of São Paulo, Brazil.

Keywords: Competing causes · Frailty models · Markov chain Monte Carlo · Negative binomial distribution · Power variance function

Mathematics Subject Classification: Primary 62N01 · Secondary 62P10.

1. INTRODUCTION

Clinical outcomes in oncology are fundamental for all healthcare providers. Information such as overall survival, disease-free survival, and cancer-specific survival can be obtained based on the cancer type and patient features, such as the clinical stage, sex, age, education level, type of treatment, and other information that is often available in medical records. The incidence of a tumor is not always related to its severity. For instance, carcinomas of the

*Corresponding author. Email: agatha.rodrigues@ufes.br

skin are very common worldwide, but their clinical outcomes are among the best in oncology. Melanoma is not the most common skin malignancy; however, it is one of the most dangerous ones due to its potential of metastatic dissemination. According to the Brazilian National Institute of Cancer (INCA), approximately 6,000 new cases were expected in 2018 (INCA, 2018); whereas, according to the International Agency for Research on Cancer (IARC), approximately 7,000 new cases were reported (IARC, 2021). The number of deaths in Brazil due to melanoma is estimated to be approximately 2,000 cases per year (INCA, 2018).

The staging system proposed by the American Joint Committee on Cancer (AJCC) is commonly used worldwide for melanoma. According to the latest edition (Gershenwald et al., 2017), clinical stage IV corresponds to metastatic disease, which carries the worst prognosis. Even though several new modalities of treatment have been reported recently, treating these patients is still challenging (Ascierto et al., 2018). Clinical stage III corresponds to the nodal spreading of the melanoma; in this scenario, surgery is routinely associated with radiotherapy and/or some modality of systemic treatment such as immunotherapy or targeted therapy (Eggermont and Dummer, 2017). Clinical stages I and II correspond to the melanoma being limited to the skin, which is associated with a better prognosis. These patients are normally treated with surgery, and the great majority will be alive after 10 years of follow-up (Gershenwald et al., 2017).

In the traditional survival analysis approach, it is assumed that all units under study are susceptible to the event of interest. However, such an assumption is violated in several situations, such as in melanoma cancer studies, when the event of interest is death by disease. In the literature, it is known that clinical stages I and II have a better prognosis, meaning that a proportion of patients will not die from the disease; these patients are termed as having “immune” elements, “cured”, or long-term survivors. Thus, a class of models, referred to as cure rate models consider this type of situation and have been studied by several authors. The Berkson-Gage model (Berkson and Gage, 1952) was probably the first model to propose the cured fraction, which is based on the assumption that only one cause is responsible for the occurrence of an event of interest (Cooner et al., 2007).

For melanoma, a patient death can be attributed to latent competing causes as the presence of cancer cells. These causes are based on the fact that each surviving carcinogenic cell can be characterized by an unknown time during which the cell could become a definitive tumor (Tsodikov et al., 2003). The books by Maller and Zhou (1996) and Ibrahim et al. (2001) as well as the articles by Tsodikov et al. (2003), Chen et al. (1999), Yin and Ibrahim (2005) and Rodrigues et al. (2009a) are key references.

Different distributions have been considered for the number of competing causes related to the occurrence of an event of interest. Chen et al. (1999) used Poisson distribution under a Bayesian approach, Rodrigues et al. (2009a) considered the negative binomial and geometric distributions, Rodrigues et al. (2009b) utilize the COM-Poisson distribution, Cancho et al. (2013) employed the power series distribution, Gallardo et al. (2017) considered the Yale-Simon distribution, Leão et al. (2018) assumed the Birnbaum-Saunders distribution, and Leão et al. (2020) used the zero-modified geometric distribution.

The promotion times are usually assumed to be independent and identically distributed, that is, the lifetimes of the carcinogenic cells follow a common distribution function, with the most common being exponential, piecewise exponential, and Weibull, among others (Cal-savara et al., 2017). Besides, the long-term survival models implicitly assume a homogeneous population for the susceptible units. Although covariates can be included in the model in order to explain some observable heterogeneity, there is an unobserved heterogeneity induced by unobservable risk factors that are not commonly considered in the model (Wienke, 2011).

The models that take into account the unobservable heterogeneity are known as frailty models (Vaupel et al., 1979). These models are characterized by the inclusion of a random effect, that is, an unobservable random variable that represents the information that cannot

be observed, such as unobservable risk factors. If an important covariate is not included in the model, this will increase the unobservable heterogeneity, thus affecting the inferences about the parameters in the model. Therefore, the inclusion of a frailty term can help to relieve this problem (Hougaard, 1991).

The frailty term can be included in an additive form in the model. However, a multiplicative effect on the baseline hazard function is often used. Multiplicative frailty models represent a generalization of the proportional hazards model introduced by Cox (1972), in which the frailty term acts multiplicatively on the baseline hazard function. This approach has been studied by several authors, notably Hougaard (1995), Sinha and Dey (1997) and Balakrishnan and Peng (2006). Other authors, such as Calsavara et al. (2013), Calsavara et al. (2017), Scudilio et al. (2019) and Calsavara et al. (2020) considered the frailty models in the presence of long-term survivors.

We propose a class of survival models including a frailty term in the risk function of latent competing causes (Cancho et al., 2011), where the distribution of the frailty is the power variance function (PVF) family suggested by Tweedie (1984) and derived independently by Hougaard (1986). This approach allows that the competitive causes (cancer cells) have different frailties and that the frailest will fail earlier than those that are less frail. In addition, we consider that the number of competing causes related to the occurrence of an event of interest is modeled by the negative binomial distribution. This class of models allows some well-known models, depending on the parameter values, to be used. Herein, we illustrate the applicability of the proposed model in a real medical dataset from a population-based study of incident cases of melanoma diagnosed in the state of São Paulo, Brazil.

The rest of the paper is organized as follows. In Section 2, we present cure rate models under latent competing causes and the frailty model following a PVF distribution for the random effect, and the proposed model. Bayesian inference and simulation studies are described in Section 3. The proposed methodology is illustrated with real melanoma data also in this section. Finally, some final remarks are considered in Section 4.

2. BACKGROUND AND PROPOSED MODEL

In this section, we provide preliminary notions of long-term survival models under the biologic perspective, considering a negative binomial distribution for latent causes. Also, notions of the frailty model with their respective unconditional survival and density functions, as well as the proposed model, are provided here.

2.1 CURE RATE MODELS AND FRAILITY MODELS

The time for the j th competing cause to produce the promotion time is denoted by Z_j , $j = 1, \dots, N$, where N represents the number of cancer cells. The variable N is unobservable with the probability mass function (PMF) $p_n = P(N = n|\Theta)$ for $n = 0, 1, \dots$. We assume that, conditional on N and on the parameter vector φ , Z_j s are independent and identically distributed with the cumulative distribution function $F(t|\varphi)$ and the survival function $S(t|\varphi) = 1 - F(t|\varphi)$. Also, we assume that Z_1, Z_2, \dots are independent from N . The observable time of the occurrence of the event of interest is defined as $T = \min\{Z_0, Z_1, \dots, Z_N\}$, where $P(Z_0 = \infty) = 1$, which leads to long-term survivors p_0 of the population not susceptible to the event occurrence. According to Rodrigues et al. (2009a), the survival function of the random variable T , conditional to parameter vector ϑ , is given by

$$S_{\text{pop}}(t|\vartheta) = P(T \geq t|\vartheta) = \sum_{n=0}^{\infty} P(N = n|\Theta)[S(t|\varphi)]^n = A_N[S(t|\varphi)],$$

where A_N is the probability generating function (PGF) of the random variable N , which converges when $s = S(t|\boldsymbol{\varphi}) \in [0, 1]$.

We suppose that the number of cancer cells (N), conditional to $\boldsymbol{\Theta} = (\eta, \theta)^\top$, follows a negative binomial distribution (Saha and Paul, 2005) with the PMF and PGF stated, respectively, as

$$p_n = P(N = n|\boldsymbol{\Theta}) = \frac{\Gamma(n + \eta^{-1})}{n!\Gamma(\eta^{-1})} \left(\frac{\eta\theta}{1 + \eta\theta} \right)^n (1 + \eta\theta)^{-1/\eta}$$

and

$$A_N(s) = \sum_{n=0}^{\infty} p_n s^n = [1 + \eta\theta(1 - s)]^{-1/\eta}, \quad 0 \leq s \leq 1,$$

for $n = 0, 1, \dots$, $\theta > 0$, $\eta \geq 0$ and $1 + \eta\theta > 0$, so that $E(N|\boldsymbol{\Theta}) = \theta$ and $\text{Var}(N|\boldsymbol{\Theta}) = \theta + \eta\theta^2$. As discussed by Tournoud and Ecochard (2008), the parameters of the negative binomial distribution have biological interpretations in which the mean number of competing causes is represented by θ , whereas η is the dispersion parameter.

Under this setup, the population survival is given by

$$S_{\text{pop}}(t|\boldsymbol{\vartheta}) = \{1 + \eta\theta[1 - S(t|\boldsymbol{\varphi})]\}^{-1/\eta}. \quad (1)$$

The long-term survivors is determined from Equation (1) as $p_0 = \lim_{t \rightarrow \infty} S_{\text{pop}}(t|\boldsymbol{\vartheta}) = (1 + \eta\theta)^{-1/\eta} > 0$.

Amico and Van Keilegom (2018) reviewed the literature on long-term survival models and it is a recommended reference about the subject.

The frailty model considers a proportional hazard structure conditional on the random effect V . The random effect, called frailty, is a nonnegative variable that indicates the fragility of the unit. According to proportional hazard approach described by Cox and Oakes (1984), the conditional hazard is expressed as $h(t|V) = Vh_0(t)$, where h_0 is the baseline hazard function.

The survival function of T conditional to $V = v$ is given by

$$S(t|V, \boldsymbol{\varphi}) = S_0(t|\boldsymbol{\varphi})^V, \quad (2)$$

where S_0 denotes the baseline survival function.

In this paper, we suppose that the frailty variable V in Equation (2) follows the family of PVF distributions with parameters μ , ψ , and γ , suggested by Tweedie (1984) and derived independently by Hougaard (1986).

Let V be a random variable following a PVF distribution with parameters μ , ψ , and γ so that the density function can be written as (Wienke, 2011)

$$f_v(v; \mu, \psi, \gamma) = \exp \left[-\psi(1 - \gamma) \left(\frac{v}{\mu} - \frac{1}{\gamma} \right) \right] \frac{1}{\pi} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{[\psi(1 - \gamma)]^{k(1-\gamma)} \mu^{k\gamma} \Gamma(k\gamma + 1)}{\gamma^k k!} v^{-k\gamma-1} \\ \times \sin(k\gamma\pi),$$

where $\mu > 0$, $\psi > 0$ and $0 < \gamma \leq 1$.

Following the historical definition of frailty originally introduced in the field of demography (Vaupel et al., 1979) and to make sure that the model is identifiable (Wienke, 2011), we consider the restriction $E(V|\mu, \psi, \gamma) = \mu = 1$. Consequently the $\text{Var}(V|\mu, \psi, \gamma) = \mu^2/\psi = \sigma^2$, where σ^2 is interpreted as the measure of unobserved heterogeneity.

In order to eliminate the unobserved quantities, the random effect can be integrated out. Thus, the marginal survival function is given by

$$S(t|\boldsymbol{\varphi}^*) = \mathbb{E}_V[S(t|v_j, \boldsymbol{\varphi})] = \int_0^\infty \exp[-H_0(t|\boldsymbol{\varphi})v_j] f_v(v_j|\gamma, \sigma^2) dv_j = L_v[H_0(t|\boldsymbol{\varphi})],$$

where $\boldsymbol{\varphi}^* = (\boldsymbol{\varphi}, \gamma, \sigma^2)^\top$ is the parameter vector, f_v is the density function of V conditional to γ and σ^2 , H_0 is the cumulative baseline hazard function and L_v denotes the Laplace transform of the frailty distribution.

The unconditional survival and density functions in the PVF frailty model are expressed, respectively, by

$$S(t|\boldsymbol{\varphi}^*) = \exp \left\{ \frac{1-\gamma}{\gamma\sigma^2} \left[1 - \left(1 + \frac{\sigma^2 H_0(t|\boldsymbol{\varphi})}{1-\gamma} \right)^\gamma \right] \right\} \quad (3)$$

and

$$f(t|\boldsymbol{\varphi}^*) = h_0(t|\boldsymbol{\varphi}) \left(1 + \frac{\sigma^2 H_0(t|\boldsymbol{\varphi})}{1-\gamma} \right)^{\gamma-1} \exp \left\{ \frac{1-\gamma}{\gamma\sigma^2} \left[1 - \left(1 + \frac{\sigma^2 H_0(t|\boldsymbol{\varphi})}{1-\gamma} \right)^\gamma \right] \right\}. \quad (4)$$

Besides providing an algebraic treatment of the closed form for the marginal survival, the PVF family is a flexible model in the sense that it includes many other frailty models as special cases. For instance, the gamma frailty model is obtained if $\gamma = 0$; and, in the case of $\gamma = 0.5$, the inverse Gaussian distribution is derived. The positive stable is a special case of the PVF distribution; however, to show this fact, some asymptotic considerations are necessary.

2.2 THE FRAILTY LONG-TERM SURVIVAL MODEL

Thus, as an alternative to the usual cure rate models given in Equation (1), we propose a new model that incorporates a frailty term for each competing cause and consider that, conditional on $N = n$ and on $\boldsymbol{\varphi}^*$, the latent times follow a survival function as in Equation (3). As the number of competing causes follows a negative binomial distribution, the population survival function with the PVF frailty is given by

$$S_{\text{pop}}(t|\boldsymbol{\vartheta}) = \left[1 + \eta\theta \left(1 - \exp \left\{ \frac{1-\gamma}{\gamma\sigma^2} \left[1 - \left(1 + \frac{\sigma^2 H_0(t|\boldsymbol{\varphi})}{1-\gamma} \right)^\gamma \right] \right\} \right) \right]^{-1/\eta}, \quad (5)$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\varphi}^*, \boldsymbol{\Theta})^\top$.

Usually, the most common choices for the promotion time distribution that specify the function $S(t|\boldsymbol{\varphi})$ have been exponential, piecewise exponential, or Weibull, among others. To capture the unobservable characteristics of each competing cause, we propose to incorporate a random effect (frailty term) on the baseline hazard function that acts multiplicatively in the promotion time. This approach allows that the competitive causes have different frailties and that the frailest will fail earlier than those that are less frail (Wienke, 2011).

We assume a Weibull distribution for the cumulative baseline hazard function, given by $H_0(t|\boldsymbol{\varphi}) = \exp(\alpha)t^\lambda$, where $\alpha \in \mathbb{R}$, $\lambda > 0$ and $\boldsymbol{\varphi} = (\alpha, \lambda)^\top$. Henceforward, we will refer to the model in which the survival function is as shown in Equation (5), by the PVF frailty cure rate model or simply the PVF cure rate model (PVFCR). Note that the usual cure rate model given in Equation (1) is obtained as $\sigma^2 \rightarrow 0$.

3. BAYESIAN INFERENCE AND SIMULATION STUDY

In this section, we provide the Bayesian inference and simulation studies in order to evaluate the performance of the Bayesian estimators of the proposed model under different sample sizes and degree of heterogeneity in the sample. Also, we provide here the real data application.

3.1 BAYESIAN INFERENCE

Let us consider the situation when the time to the event is not completely observed and is subject to right censoring. For a given sample of size m , the observed time for the i th unit is $W_i = \min\{T_i, C_i\}$, with $T_i = \min\{Z_{i0}, Z_{i1}, \dots, Z_{iN_i}\}$ and C_i is the censoring time, for $i = 1, \dots, m$. Let δ_i be an indicator variable, in which $\delta_i = 1$ if $W_i = T_i$ and $\delta_i = 0$ otherwise. We include the covariate through the expected number of competing causes by $E(N_i|\Theta) = \theta_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$, $i = 1, \dots, m$, where $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients. The observed data are represented by $\mathbf{D} = (m, \mathbf{w}, \boldsymbol{\delta}, \mathbf{X})$, $\mathbf{w} = (w_1, \dots, w_m)^\top$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_m)^\top$, and \mathbf{X} is an $m \times k$ matrix containing the covariates.

The likelihood function of parameter $\boldsymbol{\vartheta} = (\boldsymbol{\varphi}^*, \Theta)^\top = (\alpha, \lambda, \gamma, \sigma^2, \eta, \boldsymbol{\beta})^\top$ under noninformative censoring can be written as

$$\begin{aligned} L(\boldsymbol{\vartheta}|\mathbf{D}) &\propto \prod_{i=1}^m [f_{\text{pop}}(w_i|\boldsymbol{\vartheta})]^{\delta_i} [S_{\text{pop}}(w_i|\boldsymbol{\vartheta})]^{1-\delta_i} \\ &\propto \prod_{i=1}^m \left[\exp(\mathbf{x}_i^\top \boldsymbol{\beta}) f(w_i|\boldsymbol{\varphi}^*) \right]^{\delta_i} \left\{ 1 + \eta \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) [1 - S(w_i|\boldsymbol{\varphi}^*)] \right\}^{-\frac{1}{\eta} - \delta_i}, \end{aligned}$$

where $S(w_i|\boldsymbol{\varphi}^*)$ and $f(w_i|\boldsymbol{\varphi}^*)$ are given in Equations (3) and (4), respectively.

The posterior distribution of $\boldsymbol{\vartheta}$ comes out to be

$$\begin{aligned} \pi(\boldsymbol{\vartheta}|\mathbf{D}) &\propto \pi(\boldsymbol{\vartheta}) \lambda^r \exp \left[\sum_{i=1}^m \delta_i \mathbf{x}_i^\top \boldsymbol{\beta} + r \left(\alpha + \frac{1-\gamma}{\gamma \sigma^2} \right) \right] \prod_{i=1}^m \left[w_i^{\lambda-1} \left(1 + \frac{\sigma^2 \exp(\alpha) w_i^\lambda}{1-\gamma} \right)^{\gamma-1} \right]^{\delta_i} \\ &\times \prod_{i=1}^m \left[1 + \eta \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \left(1 - \exp \left\{ \frac{1-\gamma}{\gamma \sigma^2} \left[1 - \left(1 + \frac{\sigma^2 \exp(\alpha) w_i^\lambda}{1-\gamma} \right)^\gamma \right] \right\} \right) \right]^{-1/\eta - \delta_i} \\ &\times \prod_{i=1}^m \exp \left[- \left(\frac{1-\gamma}{\gamma \sigma^2} \right) \left(1 + \frac{\sigma^2 \exp(\alpha) w_i^\lambda}{1-\gamma} \right)^\gamma \right]^{\delta_i}, \end{aligned} \quad (6)$$

where $r = \sum_{i=1}^m \delta_i$ and $\pi(\boldsymbol{\vartheta})$ is the prior distribution of $\boldsymbol{\vartheta}$.

We consider independent prior distributions by defining them as $\boldsymbol{\beta} \sim \text{Normal}_{k+1}(\mathbf{0}, 100\mathbf{I})$, with \mathbf{I} being a $(k+1) \times (k+1)$ identity matrix, $\alpha \sim \text{Normal}(0, 100)$, $\gamma \sim \text{Uniform}(0, 1)$, and η , λ and σ^2 following a gamma distribution with mean value of 1 for all and variances of 1, 100 and 1, respectively. In this paper, no prior information about the parameters is available, which is the reason for the choice of non-informative prior distributions, besides of the assumption that the parameters are independent a priori. It is possible that the prior distributions can be postulated by expert knowledge and past experiences in situations they are available.

The posterior density of $\boldsymbol{\vartheta}$ in Equation (6) is analytically intractable because the integration of the joint density is not easy to perform. An alternative is to rely on Markov chain Monte Carlo (MCMC) simulations. Here, we consider the adaptive Metropolis-Hastings algorithm with a multivariate distribution as the proposed distribution (Haario et al., 2005)

implemented in the statistical package *LaplacesDemon* (Hall et al., 2020), which provides a friendly environment for Bayesian inference within the R program (R Core Team, 2020).

As a result, a sample of size n_p from the joint posterior distribution of $\boldsymbol{\vartheta}$ is obtained (eliminating burn-in and jump samples). The sample from the posterior can be expressed as $(\boldsymbol{\vartheta}_1, \boldsymbol{\vartheta}_2, \dots, \boldsymbol{\vartheta}_{n_p})$. The posterior mean of $\boldsymbol{\vartheta}$ can be approximated by

$$\widehat{\boldsymbol{\vartheta}} = \frac{1}{n_p} \sum_{k=1}^{n_p} \boldsymbol{\vartheta}_k, \quad (7)$$

and the posterior mean of the long-term survivors is approximated by

$$\widehat{p}_0 = \frac{1}{n_p} \sum_{k=1}^{n_p} (1 + \eta_k \theta_k)^{-1/\eta_k}. \quad (8)$$

Considering the function $Y_k(t) = S_{\text{pop}}(t|\boldsymbol{\vartheta}_k)$, where $S_{\text{pop}}(t|\boldsymbol{\vartheta}_k)$ is presented in Equation (5), conditional to $\boldsymbol{\vartheta}_k$, the posterior mean of the improper survival function is approximated by

$$\widehat{S_{\text{pop}}}(t|\boldsymbol{\vartheta}) = \frac{1}{n_p} \sum_{k=1}^{n_p} Y_k(t), \quad \text{for each } t > 0. \quad (9)$$

3.2 SIMULATION STUDY

For data generation in this simulation study, we consider the model in given in Equation (5) with the Weibull distribution for the cumulative baseline hazard function with $\alpha = 0$, $\lambda = 1$ (exponential distribution with a rate of $\exp(\alpha)$), and one binary covariate X drawn from a Bernoulli distribution with the parameter 0.5. The PVF frailty distribution parameters are $\gamma = 0.5$ and $\sigma^2 \in \{0.5, 1, 1.5, 2\}$. The data of failure times were simulated with $\eta = 0.5$, $\theta_l = \exp(\beta_0 + l\beta_1)$, and $l = 0, 1$, where $\beta_0 = -0.5$ and $\beta_1 = 0.7$. The attribution of the parameters' values is motivated by the estimates obtained from real dataset application in Section 3.3 when fitted the model with only sex as a covariate.

In this way, $p_{0l} = (1 + \eta\theta_l)^{-1/\eta}$, so that the long-term survivors for the two levels of X are $p_{00} = 0.59$ and $p_{01} = 0.39$. The censoring times were sampled from the exponential distribution with the parameter τ (rate), where τ was set in order to control the proportion of censored observations. The algorithm to generate the observed times and censoring indicators is presented in the Algorithm 1.

Algorithm 1 Data generation algorithm.

- 1: Draw $X_i \sim \text{Bernoulli}(0.5)$ and $u_i \sim \text{Uniform}(0, 1)$.
- 2: Let $X_i = l$. If $u_i < p_{0l}$, $t_i = \infty$, otherwise,

$$t_i = \frac{(1 - \gamma)}{\sigma^2 \exp(\alpha)} \left(\left\{ 1 - \frac{\gamma \sigma^2}{1 - \gamma} \log \left[1 - \left(\frac{u^{-\eta} - 1}{\eta \exp(\beta_0 + \beta_1 x_i)} \right) \right] \right\}^{1/\gamma} - 1 \right).$$

- 3: Draw $c_i \sim \text{Exponential}(\tau)$, which controls the proportion of censored observations.
 - 4: Let $w_i = \min\{t_i, c_i\}$.
 - 5: If $t_i < c_i$, set $\delta_i = 1$, otherwise, $\delta_i = 0$, for $i = 1, \dots, m$.
-

We consider four sample sizes, $m = 100, 300, 500$ and 1000 . For each combination of parameter values and sample sizes, we simulated $B = 1000$ random samples.

As mentioned previously, the Bayesian estimation procedures were performed using the adaptive Metropolis-Hastings algorithm such that the estimation of the covariance matrix is updated every 100 iterations. We generated 40,000 values for each parameter, disregarding the first 10,000 iterations to eliminate the effect of the initial values. In addition, jumps of size 30 were chosen to reduce the correlation effects between the samples. As a result, the final sample size of the parameters generated from the posterior distributions was $n_p = 1,000$. For good convergence results to be obtained, the convergence of the chains was monitored in all simulation scenarios, through monitoring graphics similar to what we did in the application (Section 3.3) and made available in the Appendix.

For each random sample, the estimates of $\boldsymbol{\vartheta}$ and the long-term survivors are obtained by Equation (7) and (8), respectively. We computed the average of B estimates of $\boldsymbol{\vartheta}$ (AE) and the root of the mean squared error (RMSE) of the estimators obtained from the PVFCR model. The results are summarized in Table 1.

According to the results, the average estimates of p_{00} and p_{01} were not affected by the increase of σ^2 value. Even for small sample sizes, the average estimates were close to the fixed values. The RMSE values appear reasonably close to zero as the sample size increases, except for the parameter σ^2 , which needs a large sample size close to zero. For a fixed sample size, the RMSE of the σ^2 estimation increases as the σ^2 also increases.

To discuss the computational time, we simulated 100 datasets of each configuration and summarize these times (in seconds) in Table 2. The computational time increases as the sample size increases. For example, when $m = 100$ we take about 20 seconds, on average, to fit the proposed model, while we need about 80 seconds on average when $m = 1000$, regardless σ^2 value. This simulation study was conducted in a computer with the following configuration: Intel(R) Core(TM) i7-core 1.80GHz[Cores 4] processor (logical processors 8), 8 GB RAM, and Microsoft Windows 10 Home Single Language operating.

3.3 APPLICATION

The melanoma dataset used in this study is part of a retrospective cohort of patients diagnosed with melanoma in the state of São Paulo, Brazil, between 2000 and 2014, with follow-up conducted until 2018. The records were provided by the Fundação Oncocentro de São Paulo (FOSP), which is responsible for coordinating the Hospital Cancer Registry of the State of São Paulo, and it can be downloaded in <http://www.fosp.saude.sp.gov.br>. The FOSP is a public institution connected to the State Health Secretariat, which assists the study and implementation of public policies in the field of Oncology.

The time to death due to cancer was defined as the period between the dates of melanoma diagnosis and death. Those patients who did not die due to melanoma during the follow-up period were characterized as right-censored observations. The sample size was $m = 5358$ patients and the percentage of censored observations was 71%. The explanatory variables measured at baseline were as follows: sex (male or female), age (≤ 45 years or > 45 years), education level (no formal education, primary school, high school, or college), and cancer clinical stage (I, II, III or IV).

This data was studied by Calsavara et al. (2020), where they evaluated only the effect of surgery in lifetime considering a non-proportional hazards model with a frailty term. Here, we also consider other relevant information available in the registry, such as gender, age at diagnosed, education level, and the clinical stage, as previously mentioned.

In the observations, 49.38% were male, and 79% were younger than 45 years old. For the education level, 58.3% had a primary school degree, 19.3% completed high school, 15% had a college degree and the remaining (7.4%) with no formal education. A total of 42.83% of the melanoma cases were classified as clinical stage I; II: 23.12%; III: 18.23%; and IV: 15.82%.

Table 1. The RMSE and the AE values for simulated data from the PVFCR model when $p_{00} = 0.59$, $p_{01} = 0.39$, $\beta_0 = -0.5$, $\beta_1 = 0.7$, $\alpha = 0$, $\lambda = 1$, $\eta = 0.5$, and $\gamma = 0.5$.

σ^2	Parameter	Sample size (m)							
		100		300		500		1000	
		RMSE	AE	RMSE	AE	RMSE	AE	RMSE	AE
0.5	p_{00}	0.071	0.587	0.039	0.588	0.032	0.587	0.022	0.588
	p_{01}	0.069	0.403	0.038	0.395	0.031	0.392	0.022	0.389
	β_0	0.545	-0.192	0.338	-0.299	0.294	-0.334	0.198	-0.393
	β_1	0.465	0.857	0.257	0.810	0.203	0.793	0.150	0.765
	η	1.055	1.394	0.825	1.116	0.727	1.008	0.531	0.835
	α	0.405	-0.210	0.289	-0.155	0.247	-0.133	0.188	-0.092
	λ	0.205	1.116	0.116	1.053	0.095	1.039	0.065	1.019
	γ	0.085	0.471	0.103	0.481	0.110	0.476	0.112	0.467
	σ^2	0.545	0.988	0.445	0.891	0.407	0.844	0.381	0.788
1	p_{00}	0.073	0.588	0.041	0.588	0.033	0.588	0.023	0.587
	p_{01}	0.072	0.404	0.038	0.396	0.031	0.394	0.021	0.390
	β_0	0.560	-0.182	0.392	-0.262	0.318	-0.307	0.232	-0.367
	β_1	0.481	0.871	0.288	0.837	0.218	0.810	0.159	0.775
	η	1.097	1.444	0.948	1.227	0.805	1.095	0.618	0.905
	α	0.538	-0.375	0.419	-0.310	0.360	-0.267	0.290	-0.192
	λ	0.168	1.064	0.108	1.015	0.086	1.001	0.069	0.996
	γ	0.094	0.455	0.121	0.444	0.124	0.436	0.127	0.422
	σ^2	0.293	1.056	0.278	0.990	0.273	0.956	0.325	0.960
1.5	p_{00}	0.068	0.585	0.042	0.585	0.033	0.586	0.023	0.586
	p_{01}	0.068	0.399	0.039	0.395	0.031	0.394	0.022	0.389
	β_0	0.584	-0.135	0.438	-0.220	0.333	-0.289	0.254	-0.344
	β_1	0.505	0.892	0.298	0.845	0.215	0.808	0.163	0.785
	η	1.167	1.516	1.048	1.309	0.836	1.130	0.673	0.961
	α	0.667	-0.534	0.536	-0.434	0.442	-0.359	0.369	-0.293
	λ	0.164	1.027	0.108	0.983	0.089	0.973	0.073	0.970
	γ	0.108	0.437	0.124	0.431	0.133	0.416	0.137	0.404
	σ^2	0.496	1.118	0.538	1.064	0.533	1.080	0.540	1.095
2	p_{00}	0.072	0.582	0.042	0.585	0.034	0.583	0.024	0.585
	p_{01}	0.066	0.395	0.038	0.393	0.030	0.392	0.021	0.390
	β_0	0.589	-0.128	0.463	-0.216	0.391	-0.245	0.269	-0.334
	β_1	0.483	0.894	0.289	0.850	0.239	0.823	0.162	0.784
	η	1.140	1.502	1.056	1.307	0.961	1.214	0.696	0.979
	α	0.751	-0.632	0.621	-0.530	0.572	-0.491	0.441	-0.370
	λ	0.154	0.988	0.110	0.957	0.099	0.950	0.083	0.950
	γ	0.105	0.438	0.130	0.419	0.133	0.415	0.132	0.407
	σ^2	0.920	1.147	0.948	1.112	0.946	1.117	0.890	1.211

Table 2. Minimum (Min.), first quartile (1qt), median, mean, third quartile (3qt), maximum (Max.) and standard deviation (SD) of the computational times (in seconds) to fit the proposed model for 100 simulated datasets when $p_{00} = 0.59$, $p_{01} = 0.39$, $\beta_0 = -0.5$, $\beta_1 = 0.7$, $\alpha = 0$, $\lambda = 1$, $\eta = 0.5$, and $\gamma = 0.5$.

σ^2	m	Min.	1qt	Median	Mean	3qt	Max.	SD
0.5	100	17.552	19.857	20.235	20.611	21.047	25.153	1.435
	300	32.851	33.305	33.680	34.034	34.615	37.338	1.000
	500	45.156	46.119	46.997	47.120	47.862	50.318	1.179
	1000	73.332	78.920	80.223	80.319	81.589	86.031	2.232
1	100	17.802	19.802	20.130	20.578	21.017	25.440	1.462
	300	32.669	33.108	33.414	33.756	34.104	36.512	0.908
	500	45.354	46.114	46.726	47.165	47.981	53.343	1.399
	1000	72.485	78.950	80.214	80.110	81.444	87.051	2.438
1.5	100	17.459	19.774	20.140	20.482	21.093	24.777	1.392
	300	32.601	33.218	33.464	33.786	34.084	36.288	0.894
	500	45.301	46.055	46.951	47.226	48.014	53.091	1.445
	1000	72.669	78.528	79.973	79.941	81.424	87.039	2.451
2	100	17.780	19.871	20.300	20.646	21.181	25.277	1.413
	300	32.703	33.261	33.598	34.031	34.622	37.533	1.062
	500	45.368	46.351	46.907	47.216	48.038	51.567	1.161
	1000	71.902	78.524	79.632	79.794	81.015	84.453	2.504

Figure 1 presents the Kaplan-Meier estimates for each explanatory variable. Of note, there was strong evidence that a fraction of the population had been long-term survivors. Among all of the variables considered in our study, those with clinical stage I melanoma had a better prognosis.

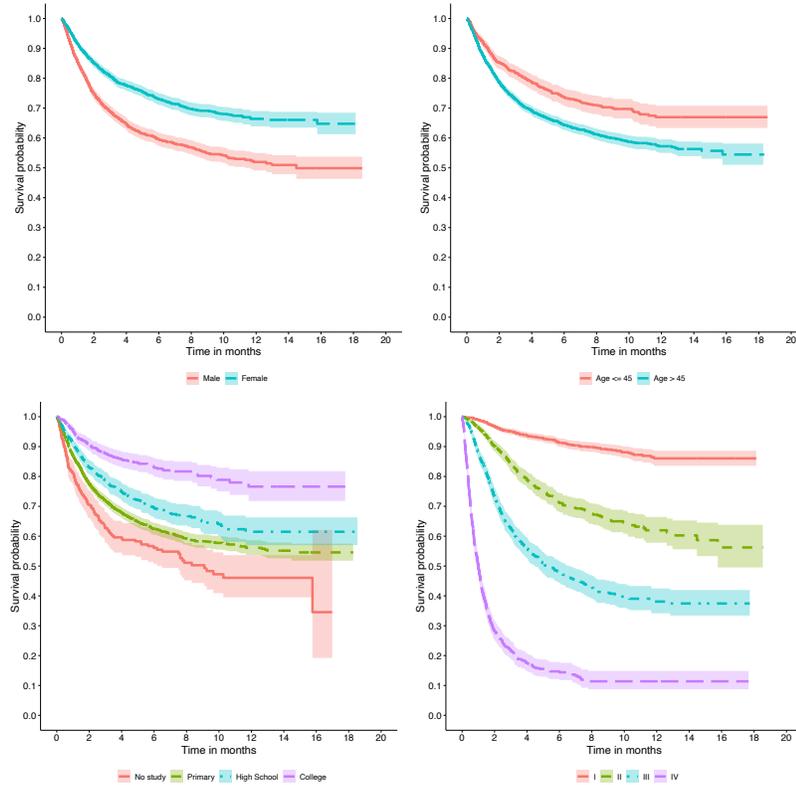


Figure 1. Kaplan-Meier estimates for the melanoma dataset grouped by sex, age, education level, and clinical stage, respectively.

To evaluate the effects of sex, age, education level, and clinical stage, the PVFCR model was fitted to the dataset. The adaptive Metropolis-Hastings algorithm was run, discarding the first 20,000 iterations as burn-in samples and using a jump of size 150 to avoid correlation problems, with a sample size of $n_p = 1,000$. The convergence of the chain was evaluated by multiple runs of the algorithm from different starting values and was monitored through graphical analysis. Good convergence results were obtained (see Appendix). The estimates of $\boldsymbol{\vartheta}$ and the long-term survivors were obtained by Equation (7) and (8), respectively, and the posterior mean of the improper survival function was given by Equation (9).

Table 3 lists the posterior mean, posterior standard deviation and 95% highest posterior density (95% HPD) intervals for all parameters from the PVFCR model. None of the parameters related to the explanatory variables have a 95% HPD value of zero.

The PVFCR model allows us to capture and to quantify the degree of unobservable heterogeneity, represented by σ^2 , obtaining a posterior mean of 1.159 (95% HPD: 0.018; 2.687), which indicates a reasonable degree of unobserved heterogeneity in the sample. It is of great importance in clinical practice, once important covariates were not observed and not available in the dataset, such as Breslow thickness, ulceration and Mitotic rate.

Breslow thickness is the single most important prognostic factor for clinically localized primary melanoma. It is measured from the top of the granular layer of the epidermis (or, if the surface is ulcerated, from the base of the ulcer) to the deepest invasive cell across the broad base of the tumor (dermal/subcutaneous). Ulceration is an integral component of the AJCC staging system and an independent predictor of outcome in patients with clinically localized primary cutaneous melanoma. Multiple studies indicate that mitotic count is an important prognostic factor for localized primary melanoma since it represents tumor cells division (Bertolli et al., 2019; Fonseca et al., 2020).

Table 3. The posterior mean, standard deviation (SD) and 95% HPD of the fitted PVFCR model parameters.

Parameter	Mean	SD	95% HPD	
			Lower	Upper
λ	1.585	0.060	1.465	1.705
α	-3.045	0.220	-3.500	-2.659
η	1.477	0.175	1.126	1.818
β_0	-2.516	0.249	-2.984	-2.010
β_{sex} (male)	0.572	0.078	0.402	0.715
β_{age} (>45 years)	0.311	0.097	0.115	0.492
$\beta_{\text{education}}$ (no formal study)	1.094	0.174	0.782	1.415
$\beta_{\text{education}}$ (primary school)	0.595	0.129	0.339	0.832
$\beta_{\text{education}}$ (high school)	0.432	0.149	0.156	0.738
β_{stage} (II)	1.338	0.117	1.123	1.565
β_{stage} (III)	2.492	0.132	2.259	2.760
β_{stage} (IV)	4.697	0.183	4.354	5.060
γ	0.380	0.246	0.001	0.813
σ^2	1.159	0.763	0.018	2.687

College is the baseline for education level and stage I is the baseline for the melanoma clinical stage.

All of the findings of this study are consistent with those observed in routine clinical practice. Sex and age have already been reported as prognostic factors, suggesting that young patients and women have a better prognosis (Sabel et al., 2005; Balch et al., 2014). The education level is very likely to be related to knowledge about diseases and the necessity of medical evaluation for an early diagnosis. Clinical staging is also used for prognosis stratification, and the curves shown in this paper are very similar to those presented in the three latest updates of the AJCC staging system for melanoma (Balch et al., 2001, 2009; Gershenwald et al., 2017). The long-term survivors' estimates and survival estimates for a specific patient can be seen in Figures 2 and 3, respectively. As expected, the patients with clinical stage IV melanoma had a worse prognosis, regardless of their sex and age. On the other hand, the patients in clinical stage I melanoma as well as females and those younger than 45 years old had a better prognosis.

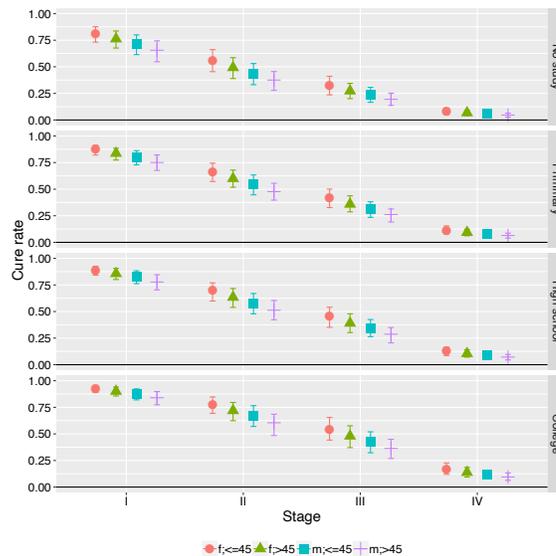


Figure 2. Long-term survivors' estimates (symbol) and 95% HPD intervals (bars) according to the fitted PVFCR model by considering sex (f, female and m, male), age (≤ 45 years and > 45 years), clinical stage (I, II, III and IV), and education level (no formal study, primary school, high school, and college).

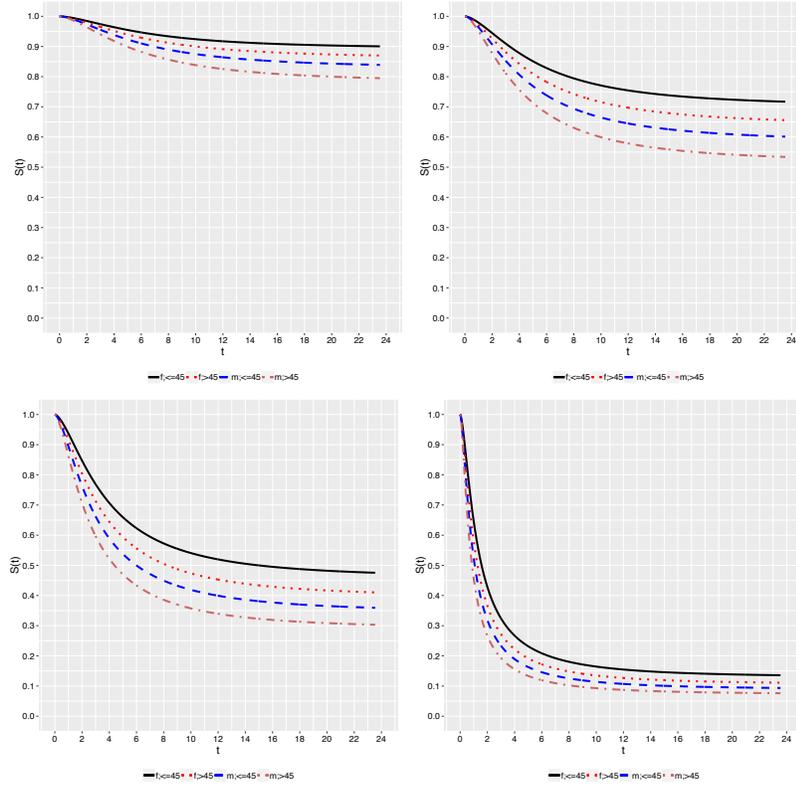


Figure 3. Survival functions estimated by the PVFCR model considering sex (f, female and m, male), age (≤ 45 years and > 45 years), and clinical stage I, II, III, and IV, respectively, for a fixed education level (high school category).

4. FINAL REMARKS

In this paper, we studied the cure rate model formulated by [Cancho et al. \(2011\)](#) in a different way, that is, we considered a random unobservable effect in promotion time of each competing cause, which allowed the unobserved heterogeneity to be quantified. The PVF frailty model was considered for the latent variables, and it included many other frailty models as special cases. A simulation study was conducted to illustrate the reliable performance of the Bayesian estimators of the proposed model, as the RMSE was reasonably close to zero as the sample size increased.

A point of attention is the fact that for large values of the parameter σ^2 , one needs a large sample size for RMSE goes close to zero. However, it is worth to note that we obtained satisfactory values of RMSE and average of the estimates when $\sigma^2 = 1$ that is the close value of the estimated σ^2 in the application to the real dataset.

The applicability of the proposed model was demonstrated with a real melanoma dataset, explaining the model fit results and discussing its relevance in the real world. We hope that this model can be generalized to wider applications in survival analysis.

ACKNOWLEDGEMENTS

The authors thank the Fundação Oncocentro de São Paulo for providing the melanoma dataset. The authors would like to thank the reviewers and the editors for their valuable contributions.

APPENDIX

MCMC CONVERGENCE MONITORING FOR PVFCR MODEL IN MELANOMA DATASET

A jump of size 150 was considered to reduce correlation effects between the samples, as one can see in the autocorrelation graphs in figures 4, 5 and 6. Thus, final samples are considered with a lag of 150. After burn-in (20000) and jump samples, a sample of 1000 size from the posterior distribution of the parameters is generated. The time series graphs in Figures 7, 8 and 6 were built from the final posterior distribution sample, in which a type of blur is observed in a small variability of sampled values.

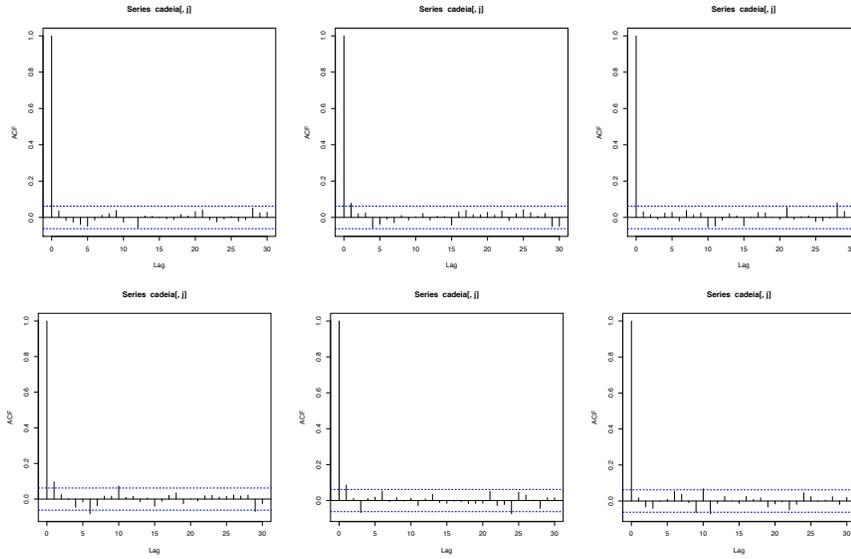


Figure 4. Autocorrelation graphs for λ , α , η , β_0 , β_{sex} and β_{age} parameters.

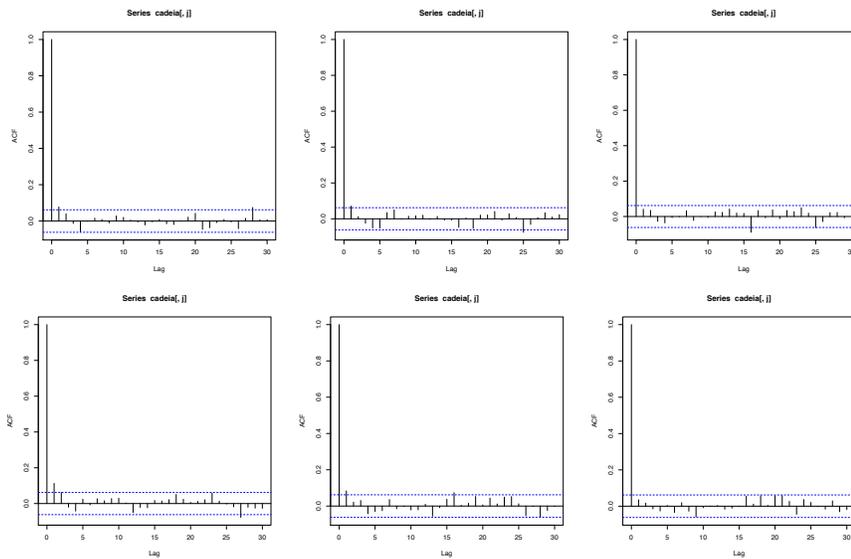


Figure 5. Autocorrelation graphs for β_{school} and β_{stage} parameters.

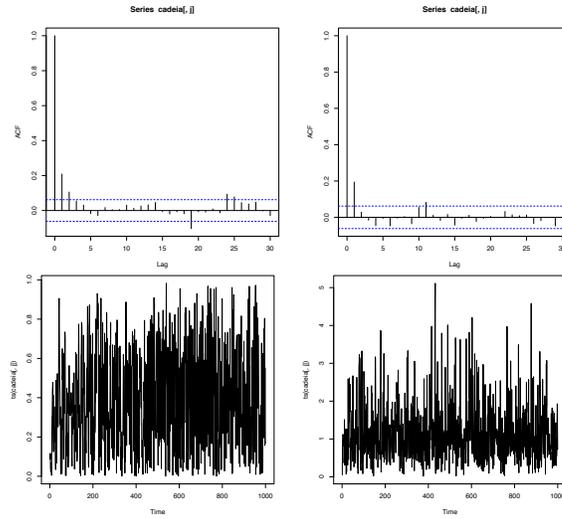


Figure 6. Autocorrelation (first panel) and time series (second panel) graphs for γ and σ^2 parameters.

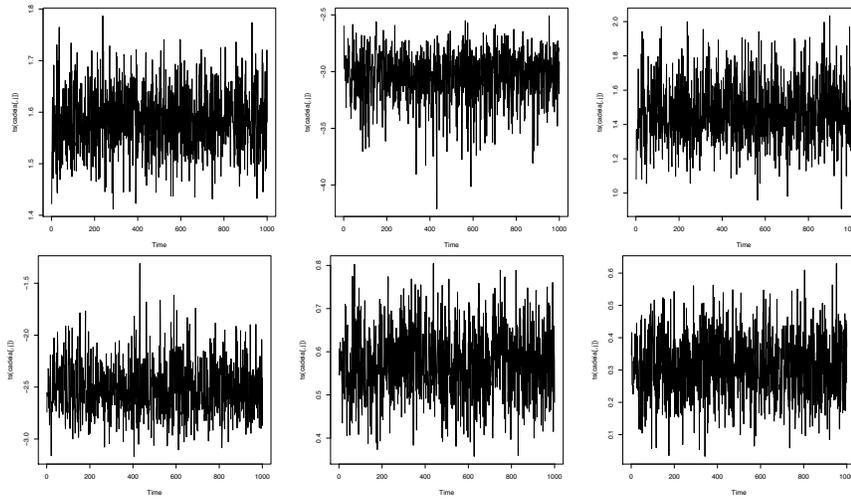


Figure 7. Time series graphs for λ , α , η , β_0 , β_{sex} and β_{age} parameters.

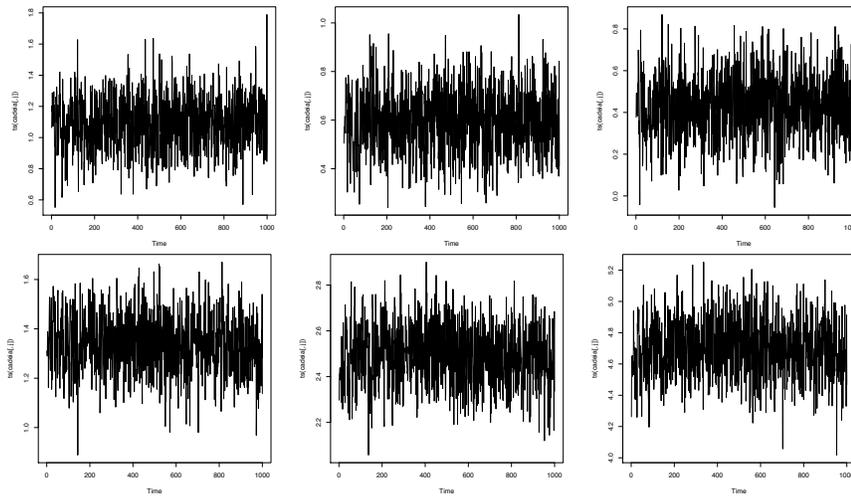


Figure 8. Time series graphs for β_{school} and β_{stage} parameters.

REFERENCES

- Amico, M. and Van Keilegom, I. 2018. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, 5, 311–342.
- Ascierto, P.A., Flaherty, K., and Goff, S., 2018. Emerging strategies in systemic therapy for the treatment of melanoma. *American Society of Clinical Oncology Educational Book*, 38, 751–758.
- Balakrishnan, N. and Peng, Y., 2006. Generalized gamma frailty model. *Statistics in Medicine*, 25, 2797–2816.
- Balch, C.M., Buzaid, A.C., Soong, S.J., et al., 2001. Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma. *Journal of Clinical Oncology*, 19, 3635–3648.
- Balch, C.M., Gershenwald, J.E., Soong, S.J. et al., 2009. Final version of 2009 AJCC melanoma staging and classification. *Journal of Clinical Oncology*, 27, 6199.
- Balch, C.M., Thompson, J.F., Gershenwald, et al., 2014. Age as a predictor of sentinel node metastasis among patients with localized melanoma: an inverse correlation of melanoma mortality and incidence of sentinel node metastasis among young and old patients. *Annals of Surgical Oncology*, 21, 1075–1081.
- Berkson, J. and Gage, R.P., 1952. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.
- Bertolli, E., Franke, V., Calsavara, V.F., de Macedo, M.P., Pinto, C.A.L., van Houdt, W.J., Wouters, M.W., Neto, J.P.D., and van Akkooi, A.C., 2019. Validation of a nomogram for non-sentinel node positivity in melanoma patients, and its clinical implications: a Brazilian–Dutch study. *Annals of Surgical Oncology*, 26, 395–405.
- Calsavara, V., Rodrigues, A., Tomazella, V., and de Castro, M., 2017. Frailty models power variance function with cure fraction and latent risk factors negative binomial. *Communications in Statistics: Theory and Methods*, 46, 9763–9776.
- Calsavara, V.F., Milani, E.A., Bertolli, E., and Tomazella, V., 2020. Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. *Statistical Methods in Medical Research*, 29, 2100–2118.
- Calsavara, V.F., Tomazella, V.L.D., and Fogo, J.C., 2013. The effect of frailty term in the standard mixture model. *Chilean Journal of Statistics*, 4, 95–109.
- Cancho, V.G., Louzada, F., and Ortega, E.M., 2013. The power series cure rate model: an application to a cutaneous melanoma data. *Communications in Statistics: Simulation and Computation*, 42, 586–602.
- Cancho, V.G., Rodrigues, J., and de Castro, M., 2011. A flexible model for survival data with a cure rate: a Bayesian approach. *Journal of Applied Statistics*, 38, 57–70.
- Chen, M.H., Ibrahim, J.G., and Sinha, D., 1999. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94, 909–919.
- Cooner, F., Banerjee, S., Carlin, B.P., and Sinha, D., 2007. Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, 102, 560–572.
- Cox, D.R., 1972. Regression models and life-tables. *Journal of the Royal Statistical Society B*, 34, 187–220.
- Cox, D.R. and Oakes, D., 1984. *Analysis of Survival Data*. Chapman and Hall, London.
- Eggermont, A.M. and Dummer, R., 2017. The 2017 complete overhaul of adjuvant therapies for high-risk melanoma and its consequences for staging and management of melanoma patients. *European Journal of Cancer*, 86, 101–105.

- Fonseca, I.B., Lindote, M.V.N., Monteiro, M.R., Doria Filho, E., Pinto, C. A.L., Jafelicci, A.S., de Melo Lôbo, M., Calsavara, V.F., Bertolli, E., and Neto, J.P.D., 2020. Sentinel node status is the most important prognostic information for clinical stage IIB and IIC melanoma patients. *Annals of Surgical Oncology*, 27, 4133–4140.
- Gallardo, D.I., Gómez, H.W., and Bolfarine, H., 2017. A new cure rate model based on the Yule–Simon distribution with application to a melanoma data set. *Journal of Applied Statistics*, 44, 1153–1164.
- Gershenwald, J.E., Scolyer, R.A., Hess, K.R., et al., 2017. Melanoma staging: evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA: A Cancer Journal for Clinicians*, 67, 472–492.
- Haario, H., Saksman, E., and Tamminen, J., 2005. Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20, 265–274.
- Hall, B., Hall, M., Statisticat, L.L.C., et al., 2020. Package ‘LaplacesDemon’. R package version 16.1.4.
- Hougaard, P., 1986. Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73, 387–396.
- Hougaard, P., 1991. Modelling heterogeneity in survival data. *Journal of Applied Probability*, 28, 695–701.
- Hougaard, P., 1995. Frailty models for survival data. *Lifetime Data Analysis*, 1, 255–273.
- IARC., 2021. Cancer Today. WHO. Available at <https://gco.iarc.fr/today/home> (accessed on 26-04-2021)
- Ibrahim, J.G., Chen, M., and Sinha, D., 2001. *Bayesian Survival Analysis*. Springer, New York.
- INCA, E., 2018. Incidência de câncer no Brasil [internet]. Instituto Nacional de Câncer José Alencar Gomes da Silva.
- Leão, J., Bourguignon, M., Gallardo, D. I., Rocha, R., and Tomazella, V., 2020. A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. *Statistics in Medicine*, 39, 3272–3284.
- Leão, J., Leiva, V., Saulo, H., and Tomazella, V., 2018. Incorporation of frailties into a cure rate regression model and its diagnostics and application to melanoma data. *Statistics in Medicine*, 37, 4421–4440.
- Maller, R.A. and Zhou, X., 1996. *Survival Analysis with Long-term Survivors*. Wiley New York.
- R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rodrigues, J., Cancho, V.G., de Castro, M., and Louzada-Neto, F., 2009a. On the unification of the long-term models. *Statistics and Probability Letters*, 79, 753–759.
- Rodrigues, J., de Castro, M., Cancho, V.G., and Balakrishnan, N., 2009b. Com–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, 139, 3605–3611.
- Sabel, M.S., Griffith, K., Sondak, V.K., Lowe, L., Schwartz, J.L., Cimmino, V.M., Chang, A.E., Rees, R.S., Bradford, C.R., and Johnson, T.M., 2005. Predictors of nonsentinel lymph node positivity in patients with a positive sentinel node for melanoma. *Journal of the American College of Surgeons*, 201, 37–47.
- Saha, K. and Paul, S., 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61, 179–185.
- Scudilio, J., Calsavara, V., Rocha, R., Louzada, F., Tomazella, V., and Rodrigues, A., 2019. Defective models induced by gamma frailty term for survival data with cured fraction. *Journal of Applied Statistics*, 46, 487–507.
- Sinha, D. and Dey, D., 1997. Semiparametric Bayesian analysis of survival data. *Journal of the American Statistical Association*, 92, 1195–1212.

- Tournoud, M. and Ecochard, R., 2008. Promotion time models with time-changing exposure and heterogeneity: Application to infectious diseases. *Biometrical Journal*, 50, 395–407.
- Tsodikov, A., Ibrahim, J., and Yakovlev, A., 2003. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98, 1063–1078.
- Tweedie, M.C.K., 1984. An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions: Proc. Indian Statistical Institute Golden Jubilee International Conference*, pp. 579–604.
- Vaupel, J.W., Manton, K.G., and Stallard, E., 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16, 439–454.
- Wienke, A., 2011. *Frailty Models in Survival Analysis*. Chapman and Hall, Boca Raton, FL, US.
- Yin, G. and Ibrahim, J., 2005. Cure rate models: A unified approach. *Canadian Journal of Statistics*, 33, 559–570.

SPATIAL STATISTICS AND IMAGE MODELING
RESEARCH PAPER

Distance-based edge detection on synthetic aperture radar imagery

ABRAÃO D.C. NASCIMENTO^{1,*}, KÁSSIO F. SILVA¹, AND ALEJANDRO C. FRERY²

¹DEPARTAMENTO DE ESTATÍSTICA, UNIVERSIDADE FEDERAL DE PERNAMBUCO, AV.
PROFESSOR MORAES REGO, 1235, RECIFE 50670-901, BRAZIL

²SCHOOL OF MATHEMATICS AND STATISTICS, VICTORIA UNIVERSITY OF WELLINGTON,
NEW ZEALAND

(Received: 01 March 2021 · Accepted in final form: 22 April 2021)

Abstract

Synthetic aperture radar is an efficient remote sensing tool by producing high spatial resolution images. But, synthetic aperture radar data suffer speckle noise effect that difficult their processing (for example, making boundary detection). We propose and assess edge detectors for synthetic aperture radar imagery based on stochastic distances between models. These edge detectors stem from generalized divergences with good asymptotic properties. Results reveal that divergence-based detectors can outperform the likelihood-based counterpart.

Keywords: Edge detection · \mathcal{G}_I^0 model · Synthetic aperture radar systems · Speckled data · Stochastic distance.

Mathematics Subject Classification: Primary 62B10 · Secondary 68U10.

1. INTRODUCTION

Because of its all-time, all-weather, high-penetration, and high-resolution imaging capability on a global scale, synthetic aperture radar (SAR) has become an essential tool for land survey, resource mapping, environmental monitoring, disaster rescue, and national security. SAR systems have progressed from low to high resolution, single polarization to full polarization, and single frequency to multifrequency. SAR images can be analyzed using a variety of techniques. There are primarily two types of methods depending on their theoretical foundations: electromagnetic (EM) physics methods based on Maxwell's equations (Kong, 1990), and statistical methods which focus on the image data.

Due to the high complexity of the EM approach, both theoretically and computationally, only simplified or empirical models for specific scenarios can be created. The statistical approach is focused on the relationships between pixel values and their distributional characteristics.

*Corresponding author. Email: abraao@de.ufpe.br

SAR image statistical analysis can be traced back to the 1950s. The first statistical study was focused on SAR clutter in the ocean. Since early radar images had poor resolution, the Rayleigh speckle model was developed under the assumption of the Central Limit Theorem, resulting in the Rayleigh distribution for the amplitude of radar echoes (Ward et al, 2006).

The Rayleigh model, however, became less reliable as the spatial resolution of SAR images improved: smaller areas comprise less elements in the summation, making the large sample assumption questionable in many cases.

Ward (1981) proposed the multiplicative model in 1981, a turning point in statistical SAR data description. Such an approach generalizes the Rayleigh model, and bridges the EM and statistical approaches (Yue et al, 2020, 2021)

The central problem to be described is the presence of an interference pattern, common to all images obtained with coherent illumination, called *speckle*. Although deterministic, the precise knowledge of speckle amounts to specifying the EM characteristics of each scattering element within a resolution cell. This is possible when there are a few well-known and simple backscatterers as, for instance, a single small sphere or a few dipoles (Sant'Anna et al, 2008). The multiplicative model is an adequate approach when the number of backscatterers or their properties are unknown.

The models that arise from the multiplicative description are neither Gaussian nor additive. Classical image processing techniques are, at best, sub-optimal in such scenario. Therefore, SAR imagery processing requires specialized models and methodologies. To that end, the use of Information Statistical Theory measures combined with the multiplicative modeling approach has been successfully adopted for treating SAR images.

Edge detection is one of the fundamental image processing techniques. Gambini et al (2006) proposed a method that relies on comparing two samples for estimating the position of the edge along a thin strip of data. Gambini et al (2008) compared five strategies based on SAR data or on estimates of the target roughness. Wei and Feng (2015), assuming a gamma model, derived a detector with low false alarm rate, but its performance strongly depends on the settings. Giron et al (2012) used a nonparametric approach with good results but, again, the performance is affected by the underlying distribution of the data.

In this paper, we assume that intensity SAR data follow the \mathcal{G}_I^0 model. This distribution is recognized in the literature as the universal model for this kind of observations (Mejail et al, 2003). Frey et al (2011) showed that using Information theory measures (as divergences and entropies) combined with statistical inference is a powerful methodology. Recently, Nikooravesh (2018) developed estimation procedures for the quantile function by means of Shannon and Tsallis entropies. We propose boundary detectors which are competitive with respect to those based on the joint likelihood, which are computationally demanding, as discussed by Nascimento et al (2014). We propose and discuss two boundary detection schemes based on the Kullback-Leibler (KL) and Rényi divergences. Additionally, we also investigate their performance at the limit case when intensities are gamma distributed. Results provide evidence in favor of the detector based on the Rényi divergence between \mathcal{G}_I^0 models.

The paper unfolds as follows. Section 2 recalls the \mathcal{G}_I^0 model. In Sections 3 and 4, we discuss the divergence measures and boundary detection procedures, respectively. Section 5 shows results of a simulation study and an application to an actual SAR image. Finally, Section 6 concludes the paper.

2. MODEL FOR SPECKLED DATA: THE \mathcal{G}_I^0 MODEL

The multiplicative model for the observation at position (i, j) of an intensity SAR image describes it as the outcome of the random variable $Z(i, j) = X(i, j)Y(i, j)$, where $X(i, j)$ and $Y(i, j)$ are independent random variables. The latter, which describes the speckle, follows

a unitary-mean gamma distribution with shape parameter $L \geq 1$; this parameter is known as “number of looks,” it is proportional to the signal-to-noise ratio, and it is often fixed for the whole image. The unobserved quantity of interest, $X(i, j)$, is called “backscatter.” The backscatter is positive, and contains all the relevant information about the target.

Assuming a Reciprocal Gamma law for the backscatter (Frery et al, 1997), we obtain that the density of Z is expressed by

$$f_Z(z; \alpha, \gamma, L) = \frac{L^L \Gamma(L - \alpha)}{\gamma^\alpha \Gamma(-\alpha) \Gamma(L)} z^{L-1} (\gamma + Lz)^{\alpha-L}, \quad z > 0,$$

where $\alpha < 0$ is the texture, and $\gamma > 0$ is the brightness. We denote this situation as $Z \sim \mathcal{G}^0(\alpha, \gamma, L)$, with $\Theta = \mathbb{R}_- \times \mathbb{R}_+ \times [1, \infty)$ the parameter space.

Frery et al (1997) proved the following result. Consider the sequence of random variables Z_1, Z_2, \dots in which $Z_i \sim \mathcal{G}_I^0(\alpha_i, \gamma_i, L)$. If $-\alpha_i, \gamma_i \rightarrow \infty$ such that $-\alpha_i/\gamma_i \rightarrow \beta_1$, then the following convergence in distribution holds:

$$Z_i \xrightarrow{\mathcal{D}} Z, \tag{1}$$

where Z follows a gamma distribution with mean β_1 and shape parameter L . In particular, if $L = 1$ then the convergence is towards an Exponential law.

Fig. 1 shows four single-look unitary-mean \mathcal{G}_I^0 densities with varying roughness: $\alpha \in \{-\infty, -10, -3, -1.5\}$. Since the \mathcal{G}_I^0 distribution is the Exponential law in the limit above, we plot this density in black to serve as a reference. The densities in linear scale might look like Exponential, but they are not, as revealed in the semi-logarithmic scale: \mathcal{G}_I^0 densities have heavier tails. The larger the texture parameter is, the heavier the tail is.

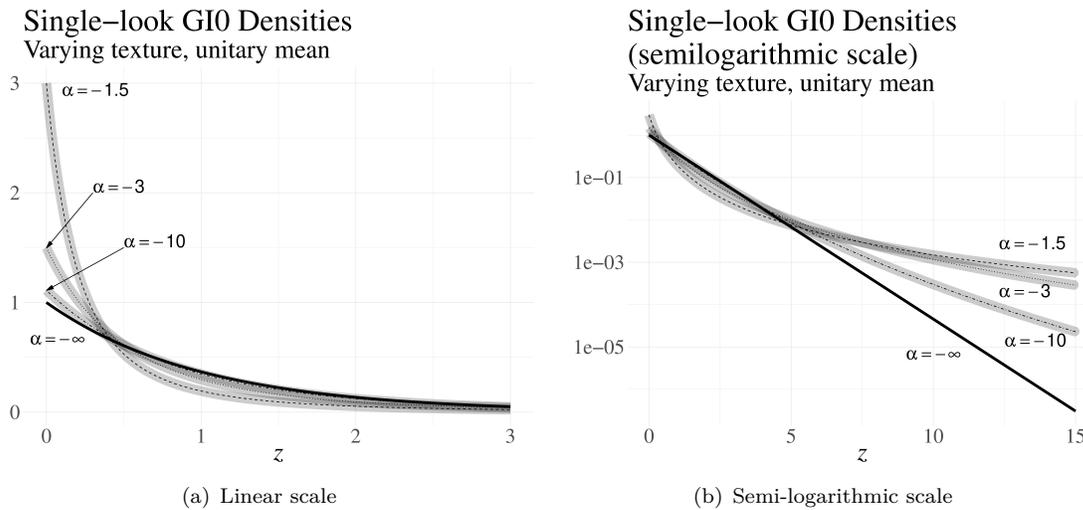


Figure 1. Single-look unitary-mean \mathcal{G}_I^0 densities with varying roughness.

A remarkable feature of this distribution is that it describes well extremely textured areas (urban centers), textured regions (forests), and areas with fully developed speckle and, thus, textureless (bare soil and crops, for instance). Figure 2 illustrates how α and γ can be interpreted.

Due to its desirable asymptotic properties (unbiasedness, normality, and efficiency), we use the maximum likelihood (ML) estimator for obtaining the parameters α and γ from data. Let $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$ be a random sample drawn from $Z \sim \mathcal{G}_I^0(\alpha, \gamma, L)$. The

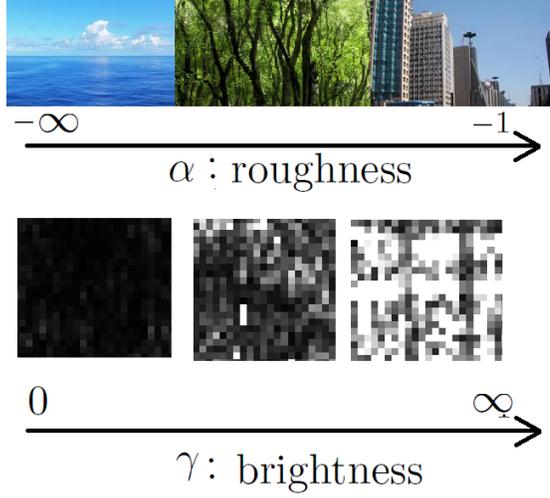


Figure 2. Properties of the \mathcal{G}_I^0 parameters.

likelihood function of the observed sample $\mathbf{z} = (z_1, z_2, \dots, z_n)$ is expressed as

$$\mathcal{L}(\alpha, \gamma; \mathbf{z}) = \left(\frac{L^L \Gamma(L - \alpha)}{\gamma^\alpha \Gamma(-\alpha) \Gamma(L)} \right)^n \prod_{i=1}^n z_i^{L-1} (\gamma + Lz_i)^{\alpha-L}.$$

Assuming L fixed, the ML estimates for α and γ , say $\hat{\alpha}$ and $\hat{\gamma}$, respectively, are the solution of the following system of nonlinear equations:

$$\begin{aligned} \psi^0(-\hat{\alpha}) - \psi^0(L - \hat{\alpha}) - \log(\hat{\gamma}) + \frac{1}{n} \sum_{i=1}^n \log(\hat{\gamma} + Lz_i) &= 0, \\ -\frac{\hat{\alpha}}{\hat{\gamma}} + \frac{\hat{\alpha} - L}{n} \sum_{i=1}^n (\hat{\gamma} + Lz_i)^{-1} &= 0, \end{aligned}$$

where ψ^0 is the digamma function. This nonlinear system does not have a closed-form solution, then, we rely on numerical optimization methods.

3. CONTRAST BASED ON INFORMATION AND DIVERGENCE

The KL divergence (or relative entropy) is a well-known way of comparing two distributions. Divergence measures are submitted to a systematic and comprehensive treatment and, as a result, [Salicrú et al \(1994\)](#) proposed the class of (h, ϕ) -divergences.

Let Z_1 and Z_2 be two random variables equipped with densities f_{Z_1} and f_{Z_2} , respectively, with common support $I \subseteq \mathbb{R}$. The (h, ϕ) -divergence between Z_1 and Z_2 is defined by

$$D_\phi^h(Z_1 \| Z_2) = h \left(\int_I \phi \left(\frac{f_{Z_1}(z; \boldsymbol{\theta}_1)}{f_{Z_2}(z; \boldsymbol{\theta}_2)} \right) f_{Z_2}(z; \boldsymbol{\theta}_2) dz \right),$$

where $\phi: (0, \infty) \rightarrow [0, \infty)$ is a convex function, $h: (0, \infty) \rightarrow [0, \infty)$ is a strictly increasing function with $h(0) = 0$, and indeterminate forms are assigned value zero. [Table 1](#) shows three choices of h and ϕ functions, and the resulting divergences.

Table 1. ϕ and h functions, and related divergences.

(h, ϕ) -distance/notation	$h(y)$	$\phi(x)$
Kullback-Leibler/KL	$y/2$	$(x - 1) \log(x)$
Rényi (order $0 < \beta < 1$)/RD: β	$\frac{1}{\beta-1} \log((\beta - 1)y + 1), 0 \leq y < \frac{1}{1-\beta}$	$\frac{x^{1-\beta} + x^\beta - \beta(x-1) - 2}{2(\beta-1)}$
Bhattacharyya/BA	$-\log(-y + 1), 0 \leq y < 1$	$-\sqrt{x} + \frac{x+1}{2}$

In particular, consider Z_1 and Z_2 be random variables following the \mathcal{G}_I^0 model and indexed by parameters $\boldsymbol{\theta}_1 = [\alpha_1, \gamma_1, L_1]^\top$ and $\boldsymbol{\theta}_2 = [\alpha_2, \gamma_2, L_2]^\top$, respectively. The KL divergence between Z_1 and Z_2 can be computed setting $A(\alpha, \gamma, L) \equiv L^L \Gamma(L - \alpha) [\gamma^\alpha \Gamma(-\alpha) \Gamma(L)]^{-1}$. With this

$$\begin{aligned}
 D_{\text{KL}}^{\mathcal{G}_I^0}(Z_1 \| Z_2) &= \int_0^\infty f_{Z_1}(z) \log \left(\frac{f_{Z_1}(z)}{f_{Z_2}(z)} \right) dz \\
 &= \int_0^\infty A(\alpha_1, \gamma_1, L_1) z^{(L_1-1)} (\gamma_1 + L_1 z)^{\alpha_1 - L_1} \left[\log \left(\frac{A(\alpha_1, \gamma_1, L_1)}{A(\alpha_2, \gamma_2, L_2)} \right) \right. \\
 &\quad \left. + \log \left(\frac{z^{L_1-1} (\gamma_1 + L_1 z)^{\alpha_1 - L_1}}{z^{L_2-1} (\gamma_2 + L_2 z)^{\alpha_2 - L_2}} \right) \right] dz \\
 &= \log \left(\frac{A(\alpha_1, \gamma_1, L_1)}{A(\alpha_2, \gamma_2, L_2)} \right) + (L_1 + L_2 - 2) \text{E}[\log(Z_1)] + (\alpha_1 - L_1) \text{E}[\log(\gamma_1 + L_1 Z_1)] \\
 &\quad - (\alpha_2 - L_2) \text{E}[\log(\gamma_2 + L_2 Z_1)],
 \end{aligned}$$

where E denotes the expected value, $\text{E}[\log(\gamma_1 + L_1 Z_1)] = \log(\gamma_1) + \psi(L - \alpha_1) - \psi(-\alpha_1)$, $\text{E}[\log(Z_1)] = \log(\gamma_1) - \psi(-\alpha_1) + \psi(L) - \Gamma(L)^{-1} [\log(L) + 1] - 1$ and $\text{E}[\log(\gamma_2 + L_2 Z_1)]$ is a quantify which can be defined in terms of the following integral

$$\begin{aligned}
 &\int_0^\infty \log(\gamma_2 + L_2 z) z^{L_1-1} (\gamma_1 + L_1 z)^{\alpha_1 - L_1} dz = -\frac{L_2^{-\alpha_1} [L_1 (\gamma_2 L_1 - \gamma_1 L_2)]^{-L_1}}{\alpha_1 (\alpha_1^2 - 1) \gamma_2 L_1 \Gamma(L_1 - \alpha_1 + 1)} \\
 &\times \left\{ -\frac{(L_1 - \alpha_1) (\gamma_2 L_1 - \gamma_1 L_2)^{L_1} \Gamma(L_1 + 1) (\gamma_1 L_2)^{\alpha_1 + 1} {}_3F_2 \left(1, 1, L_1 + 1; 2, \alpha_1 + 2; \frac{L_2 \gamma_1}{L_1 \gamma_2} \right)}{\Gamma(\alpha_1 - 1)} \right. \\
 &\times \pi \csc(\pi \alpha_1) + (\alpha_1 + 1) (\alpha_1 - L_1) (\gamma_2 L_1 - \gamma_1 L_2)^{L_1} \left[\gamma_2 \Gamma(2 - \alpha_1) \log(\gamma_2) \Gamma(L_1 + 1) (\gamma_1 L_2)^{\alpha_1} \right. \\
 &\left. + \pi \gamma_1 L_2 \csc(\pi \alpha_1) (\gamma_2 L_1)^{\alpha_1} \Gamma(L_1 - \alpha_1 + 1) {}_2F_1 \left(1 - \alpha_1, L_1 - \alpha_1 + 1; 2 - \alpha_1; \frac{L_2 \gamma_1}{L_1 \gamma_2} \right) \right] \\
 &\left. - \pi (\alpha_1^2 - 1) \csc(\pi \alpha_1) (\gamma_2 L_1)^{L_1 + 1} (\gamma_2 L_1 - \gamma_1 L_2)^{\alpha_1} \Gamma(L_1 - \alpha_1 + 1) \right\},
 \end{aligned}$$

where ${}_pF_q$ is the generalized hypergeometric function (Gradshteyn and Ryzhik, 1980, Sec. 9.18). Under the conditions of Equation (1), Equation (2) collapsed to

$$\begin{aligned}
 D_{\text{KL}}^\Gamma(Z_1 \| Z_2) &= L_1 \log(L_1/\beta_1) - L_2 \log(L_2/\beta_2) + \log(\Gamma(L_2)) - \log(\Gamma(L_1)) \\
 &\quad + L_1 \left(\frac{L_2/\beta_2}{L_1/\beta_1} - 1 \right) + [\psi(L_1) - \log(L_1/\beta_1)](L_1 - L_2).
 \end{aligned}$$

One can note that both $D_{\text{KL}}^{\mathcal{G}_I^0}(\cdot \| \cdot)$ and $D_{\text{KL}}^\Gamma(\cdot \| \cdot)$ are non-symmetric measurers. A simple solution for addressing the symmetry problem is the definition of a new measure d_ϕ^h expressed

by

$$d_\phi^h(Z_1, Z_2) = \frac{D_\phi^h(Z_1 \| Z_2) + D_\phi^h(Z_2 \| Z_1)}{2}. \quad (2)$$

In this paper, we work with the KL distances stated as

$$d_{\text{KL}}^{\mathcal{G}_I^0}(Z_1, Z_2) = \frac{D_{\text{KL}}^{\mathcal{G}_I^0}(Z_1 \| Z_2) + D_{\text{KL}}^{\mathcal{G}_I^0}(Z_2 \| Z_1)}{2}$$

and

$$d_{\text{KL}}^\Gamma(Z_1, Z_2) = \frac{D_{\text{KL}}^\Gamma(Z_1 \| Z_2) + D_{\text{KL}}^\Gamma(Z_2 \| Z_1)}{2}.$$

We also consider the Rényi distance (of order $0 < \beta < 1$) between \mathcal{G}_I^0 distributions given by

$$\begin{aligned} (\beta - 1) d_{\text{RE};\beta}^{\mathcal{G}_I^0}(Z_1, Z_2) &= \log \left(\int_0^\infty \frac{1}{2} [f_{Z_1}^\beta(z) f_{Z_2}^{1-\beta}(z) + f_{Z_2}^\beta(z) f_{Z_1}^{1-\beta}(z)] dz \right) \\ &= -\log(2) + \log \left[\int_0^\infty x^{\beta L_1 + (1-\beta)L_2 - 1} (\gamma_1 + L_1 x)^{\beta(\alpha_1 - L_1)} \right. \\ &\quad \times (\gamma_2 + L_2 x)^{(1-\beta)(\alpha_2 - L_2)} dx \times A(\alpha_1, \gamma_1, L_1)^\beta A(\alpha_2, \gamma_2, L_2)^{1-\beta} \\ &\quad \left. + \int_0^\infty x^{\beta L_2 + (1-\beta)L_1 - 1} \times (\gamma_2 + L_2 x)^{\beta(\alpha_2 - L_2)} (\gamma_1 + L_1 x)^{(1-\beta)(\alpha_1 - L_1)} dx \right. \\ &\quad \left. \times A(\alpha_2, \gamma_2, L_2)^\beta A(\alpha_1, \gamma_1, L_1)^{1-\beta} \right]. \end{aligned}$$

The expressions of the above integrations are suppressed for simplicity. Under the conditions of Equation (1), we have that

$$\begin{aligned} 2 \exp\{(\beta - 1) d_{\text{RE};\beta}^\Gamma(Z_1, Z_2)\} &= \left(\frac{\Gamma(L_2)}{(L_2/\beta_2)^{L_2}} \frac{(L_1/\beta_1)^{L_1}}{\Gamma(L_1)} \right)^{\beta-1} \\ &\quad \times \underbrace{\Gamma(\beta L_1 + (1-\beta)L_2)}_{L_{01}} \frac{[\beta(L_1/\beta_1) + (1-\beta)(L_2/\beta_2)]^{-L_{01}}}{(1/(\beta_1/L_1))^{L_1} \Gamma(L_1)} \\ &\quad + \left(\frac{\Gamma(L_1)}{(L_1/\beta_1)^{L_1}} \frac{(L_2/\beta_2)^{L_2}}{\Gamma(L_2)} \right)^{\beta-1} \Gamma(\underbrace{\beta L_2 + (1-\beta)L_1}_{L_{02}}) \\ &\quad \times \frac{[\beta(L_2/\beta_2) + (1-\beta)(L_1/\beta_1)]^{-L_{02}}}{(1/(\beta_2/L_2))^{L_2} \Gamma(L_2)}. \end{aligned}$$

In particular, one obtains the Bhattacharyya distance when $\beta = 1/2$ and the final expression is multiplied by $1/2$

$$\begin{aligned} d_{\text{BA}}^\Gamma(Z_1, Z_2) &= -\log \left(\int_0^\infty \sqrt{f_{Z_1}(z; L_1, L_1/\beta_1) f_{Z_2}(z; L_2, L_2/\beta_2)} dz \right) \\ &= -\frac{L_1 \log(L_1/\beta_1) + L_2 \log(L_2/\beta_2)}{2} + \left(\frac{L_1 + L_2}{2} \right) \log \left(\frac{(L_1/\beta_1) + (L_2/\beta_2)}{2} \right) \\ &\quad - \log \left(\Gamma \left(\frac{L_1 + L_2}{2} \right) \right) + \left[\frac{\log \Gamma(L_1) + \log \Gamma(L_2)}{2} \right]. \end{aligned}$$

The Hellinger (H) distance between gamma distributions can be derived from $d_{\text{BA}}^\Gamma(Z_1, Z_2)$ as

$$d_{\text{H}}^\Gamma(Z_1, Z_2) = 1 - e^{-d_{\text{BA}}^\Gamma(Z_1, Z_2)}.$$

We now employ these measures to propose new boundary detection tools. Although the previous contrast discussion considers that the number of looks are different and unknown, called “equivalent number of looks”, this parameter can be assumed common and known in the whole image. From now on, we assume it is a known constant for the \mathcal{G}_I^0 law and an estimable unknown constant for the Γ distribution as approached by [Anfinson et al \(2009\)](#). Under this setting, both models have the same parametric space dimension being, thus, comparable.

4. BOUNDARY DETECTORS

An edge detector seeks a point on a strip of data where the statistical properties change. The detection procedures used in this paper work in three stages: (i) identifying the centroid of the candidate area (in automatic, semiautomatic, or manual manner), (ii) detecting transition points which belong to the edge, and (iii) defining the contour using a imputation method among the transition points, such as B-Splines ([Gambini et al, 2006](#)). We focus our analysis on stages (ii) and (iii).

Assume that an initial region \mathcal{R} with centroid C is available. Rays are traced from C to points outside \mathcal{R} . They are of the form $\mathbf{s}^{(i)} = \overline{CP_i}$, where the angle between rays is $\angle(\mathbf{s}^{(i)}, \mathbf{s}^{(i+1)})$, for $i = 1, 2, \dots, S$, being S the number of rays. Finally, the data are collected in thin strips around these rays.

We assume that the data follow a \mathcal{G}_I^0 distribution, and that there are two populations: one inside the edge with $j^{(i)}$ observations, and another outside the edge with $N^{(i)} - j^{(i)}$ observations. We can then model the $N^{(i)}$ observations around segment $\mathbf{s}^{(i)}$, $1 \leq i \leq S$ as

$$\begin{cases} Z_k^{(i)} \sim \mathcal{G}_I^0(\alpha_A^{(i)}, \gamma_A^{(i)}, L), & \text{for } k = 1, \dots, j^{(i)}, \\ Z_k^{(i)} \sim \mathcal{G}_I^0(\alpha_B^{(i)}, \gamma_B^{(i)}, L), & \text{for } k = j^{(i)} + 1, \dots, N^{(i)}. \end{cases} \quad (3)$$

In the limit case, that is, under the conditions of Equation (1), Equation (3) becomes

$$\begin{cases} Z_k^{(i)} \sim \Gamma(L_A^{(i)}, L_A^{(i)}/\beta_A^{(i)}), & \text{for } k = 1, \dots, j^{(i)}, \\ Z_k^{(i)} \sim \Gamma(L_B^{(i)}, L_B^{(i)}/\beta_B^{(i)}), & \text{for } k = j^{(i)} + 1, \dots, N^{(i)}. \end{cases} \quad (4)$$

Note that Equation (3) collapses in Equation (4) if $L_A^{(i)} = L_B^{(i)} = L$, $-\alpha_A^{(i)}/\gamma_A^{(i)} \rightarrow \beta_A^{(i)}$ and $-\alpha_B^{(i)}/\gamma_B^{(i)} \rightarrow \beta_B^{(i)}$, but these laws are not nested and, therefore, are competitive.

The main idea is to find the edge $j^{(i)}$ th on the segment $\mathbf{s}^{(i)}$ as the point that provides the best configuration according with respect to a decision rule. In the following we present three different decision rules omitting, for the sake of brevity, the index (i) since only one strip is considered at each epoch.

The log-likelihood for the configuration stated in Equation (3) (or (4)) is given by

$$\ell(j) = \log(L(j)) = \sum_{k=1}^j \log(f_{Z_1}(z_k; \boldsymbol{\theta}_A)) + \sum_{k=j+1}^N \log(f_{Z_2}(z_k; \boldsymbol{\theta}_B)),$$

where $\boldsymbol{\theta}_A \in \{[\alpha_A, \gamma_A], [L_A, \beta_A]\}$ and $\boldsymbol{\theta}_B \in \{[\alpha_B, \gamma_B], [L_B, \beta_B]\}$.

Gambini et al (2006) showed that an efficient estimator, \widehat{j}_{ML} , for the index on the segment that corresponds to the transition point is stated as $\widehat{j}_{\text{ML}} = \arg \max_j \ell(j)$. However, this procedure is computationally demanding as it needs to evaluate two likelihood functions at each search step.

As discussed in Nascimento et al (2010), the distances derived in Section 3 can be scaled to be asymptotically distributed as chi-square statistics:

$$S_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}_1(j), \widehat{\boldsymbol{\theta}}_2(N-j)) = \frac{2j(N-j)v_{\mathcal{D}}}{N} d_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}_1(j), \widehat{\boldsymbol{\theta}}_2(N-j)),$$

where $v_{\mathcal{D}} = 1, \beta^{-1}, 4,$ and 4 for $\mathcal{D} = \text{KL}, \text{RD}:\beta, \text{BA},$ and H , respectively, and $\widehat{\boldsymbol{\theta}}_1(j) = [\widehat{\alpha}_A(j), \widehat{\gamma}_A(j)]$ and $\widehat{\boldsymbol{\theta}}_2(N-j) = [\widehat{\alpha}_B(N-j), \widehat{\gamma}_B(N-j)]$ are the maximum likelihood estimators for $\boldsymbol{\theta}_1 = (\alpha_A, \gamma_A)$ and $\boldsymbol{\theta}_2 = (\alpha_B, \gamma_B)$ using random samples of sizes j and $N-j$, respectively. Under mild conditions, $S_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}_1(j), \widehat{\boldsymbol{\theta}}_2(N-j))$ is asymptotically distributed as a χ^2_2 random variable under the null hypothesis $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

Thus, we propose novel detectors for finding edges on SAR intensities by seeking for the point that maximizes the test statistics between the two models, that is,

$$\widehat{j}_{\mathcal{D}} = \arg \max_j S_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}_1(j), \widehat{\boldsymbol{\theta}}_2(N-j)) = \arg \max_j S_{\mathcal{D}}(j),$$

where $\mathcal{D} = \{\text{KL}; \text{BA}; \text{H}; \text{RD}:\beta\}$.

5. NUMERICAL RESULTS

In the simulations, we study the performance of the two edge detectors here proposed. We use simulated data from two models, namely gamma and \mathcal{G}_I^0 and edge detectors based on the gamma and \mathcal{G}_I^0 distributions. With this, we verify the robustness of the detectors when fed with data that do not belong to the model they were originally devised. We utilize the absolute value of the difference between the mean detection and actual edge position as performance criterion, given by

$$D = |\bar{B} - 100|,$$

where \bar{B} is the sample mean of detected edge and 100 is the true edge position. The smaller this measure is, the better the performance is.

We performed a Monte Carlo simulation study with: i) 1000 replications for each situation; ii) \mathcal{G}_I^0 and Γ distributed data; iii) in each replicate we simulate a strip of data of size 1×200 . The first half from one distribution, and the second from another distribution.

The first set of experiments used edge detectors based on the gamma distribution, in which we set $\beta = 0.9$ in the Rényi distance. The observations are samples Z_1, Z_2, \dots, Z_{100} from $\Gamma(\nu_0, \nu_0/\beta_0)$, and $Z_{101}, Z_{102}, \dots, Z_{200}$ from $\Gamma(\nu_1, \nu_1/\beta_1)$, with $\beta_0 = \beta_1 = 1$ (unitary mean), $\nu_0 = 1$ and $\nu_1 = 4, 6, 8$.

The second set of experiments used edge detectors based on the \mathcal{G}_I^0 law and samples Z_1, Z_2, \dots, Z_{100} from $\mathcal{G}_I^0(-\nu_0, \nu_0, 4)$, and $Z_{101}, Z_{102}, \dots, Z_{200}$ from $\mathcal{G}_I^0(-\nu_1, \nu_1, 4)$, with $\nu_0 = 1.5, 3$ and $\nu_1 = 3, 5$. These parameters provide a small but representative set of values.

Table 2 shows the performance of the Gamma-based edge detectors. The best results (smallest errors) are highlighted in gray. In this table, finding an edge between Gamma samples seems a difficult task for Gamma-based detectors. The Hellinger-based detector performs best in call cases, but the mean errors are of the order of four pixels. The Rényi-based detector performs worst.

Table 2. Performance measures for gamma distances in the indicated model.

Distribution and parameters	gamma-based detectors				
	LR	KL	BA	H	RD, β
Gamma, $\nu_0 = 1, \nu_1 = 4$	26.51	6.41	11.35	5.05	60.50
Gamma, $\nu_0 = 1, \nu_1 = 6$	28.55	4.71	10.18	4.41	66.93
Gamma, $\nu_0 = 1, \nu_1 = 8$	29.64	4.68	10.88	3.96	68.94
$\mathcal{G}_0^I, \nu_0 = 1.5, \nu_1 = 3$	27.57	12.26	13.73	7.01	4.93
$\mathcal{G}_0^I, \nu_0 = 1.5, \nu_1 = 5$	18.19	1.25	20.80	0.14	1.70
$\mathcal{G}_0^I, \nu_0 = 3, \nu_1 = 5$	26.75	1.56	19.98	8.65	6.44

Table 3 shows the performance of the \mathcal{G}_I^0 -based edge detectors. In this table, overall, the two edge detectors based on the \mathcal{G}_I^0 distribution have similar performance. The Rényi-based detector produces consistently better results than the KL-based one when the data follow \mathcal{G}_I^0 laws. Although the latter detector is better in two out of three cases of Gamma-distributed samples, the differences are approximately 1% and 7%.

Table 3. Performance measures for \mathcal{G}_0^I distances.

Distribution and parameters	\mathcal{G}_0^I -based detectors	
	KL	RD, β
Gamma, $\nu_0 = 1, \nu_1 = 4$	6.30	6.68
Gamma, $\nu_0 = 1, \nu_1 = 6$	9.84	9.94
Gamma, $\nu_0 = 1, \nu_1 = 8$	10.70	10.20
$\mathcal{G}_0^I, \nu_0 = 1.5, \nu_1 = 3$	3.23	1.27
$\mathcal{G}_0^I, \nu_0 = 1.5, \nu_1 = 5$	5.12	1.33
$\mathcal{G}_0^I, \nu_0 = 3, \nu_1 = 5$	3.57	1.05

The results presented in Tables 2 and 3 led us to conclude that the safest and most versatile option for edge detection is the \mathcal{G}_I^0 -based detector that uses the Rényi distance with $\beta = 0.9$. We now are in position of submitting the detectors to real data. We present an application to an actual SAR image to assess the proposed detectors in practice.

Figure 3(a) displays a SAR image of crops in Foulum (Denmark) from the HH (horizontal-horizontal) polarization channel. This picture has been obtained by an EMISAR sensor with four nominal looks. Figure 3(b) shows the reference map, and Figure 3(c) identifies classes with shades of gray. According to the discussion about this image Foulum in Ferreira and Nascimento (2020), there is a centroid between the wheat and rapeseed areas. We use it to cast the rays on which the proposed detectors work. Figures 3(d) and 3(e) exhibit the result of detecting the edges that separate wheat from rapeseed.

The following analysis is made by visual inspection on the edges reconstructed from the estimated transition points and fourth-degree B-splines curves, as in Nascimento et al (2014).

The performance of \mathcal{G}_I^0 -based detectors is consistently better than those obtained from the Γ law. They provide the same good estimate of the edge. This result is in agreement with the simulation study once the \mathcal{G}_I^0 distribution is a better alternative to describe different SAR clutter. The Rényi distance furnishes the best detection for the Gamma-based detectors.

Simulation and real experiments were made in the R programming environment (Wickham, 2019); functions `integrate` and `maxLik(. , method=BFGS)` were used for numerical integration and obtaining ML estimates (equipped with moments method ones as initial point), respectively. All studies were performed in a Intel(R) Core(TM) i5-5200U processor at 2.20 GHz.

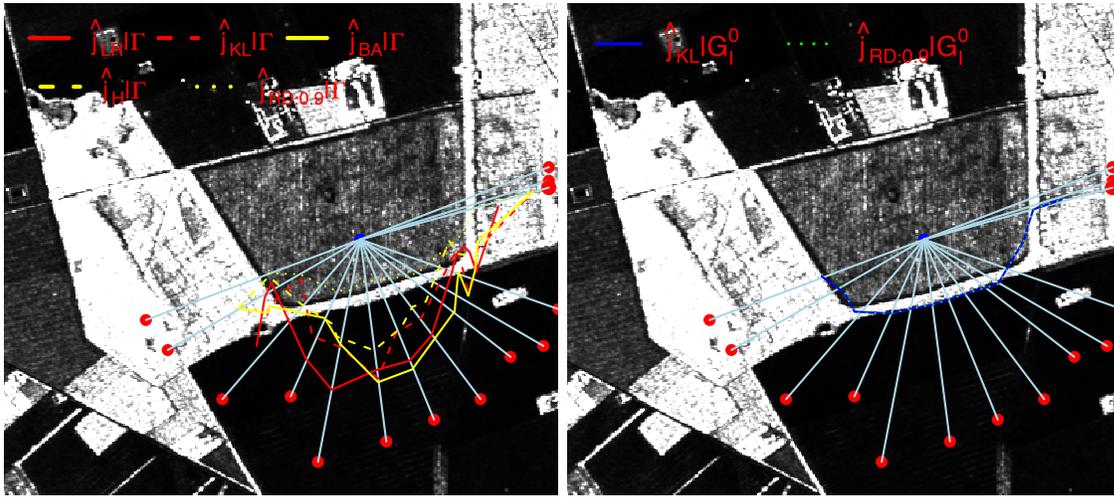
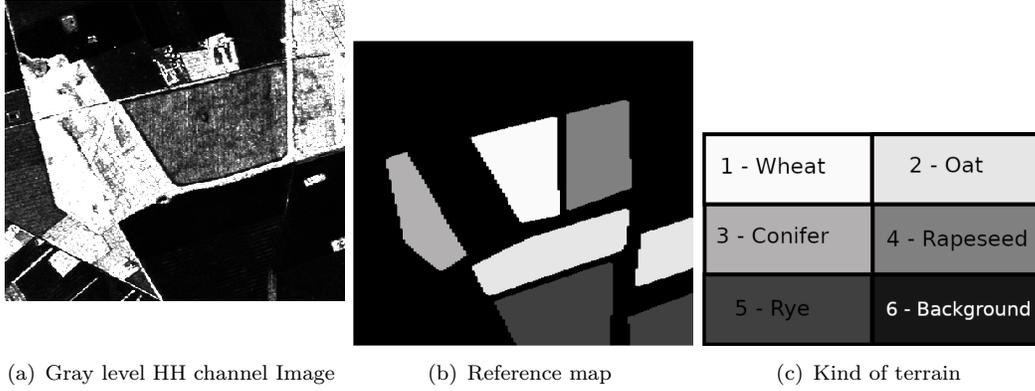


Figure 3. Application of edge detectors based on stochastic distances on an actual SAR image.

6. CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

In this paper, we proposed distance-based boundary detectors for synthetic aperture radar data modeled by the Γ and \mathcal{G}_I^0 laws. These proposals have wide applicability in practice, like the Monitoring of oil spill (Fan et al, 2015) and deforestation areas in forests (Bouvet et al, 2018). We quantified their performances with both simulated and actual data. Results provided evidence that detectors based on Kullbak-Leibler and Rényi distances for \mathcal{G}_I^0 models outperform ones based on distances between Γ limit cases and on the joint Γ likelihood, which have been employed in the synthetic aperture radar literature.

This paper has addressed only measures induced from the Γ and \mathcal{G}_I^0 distributions and approached the univariate aspect, that consist in two of its limitations.

In future works, we use the general distributions like complete \mathcal{G} (Frery et al, 1997) and KummerU (Deng et al, 2017).

REFERENCES

- Anfinsen, S.N., Doulgeris, A.P., and Eltoft, T., 2009. Estimation of the equivalent number of looks in polarimetric synthetic aperture radar imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 47, 3795–3809.
- Bouvet, A., Mermoz, S., Ballère, M., Koleck, T., and Toan, T.L., 2018. Use of the SAR Shadowing Effect for Deforestation Detection with Sentinel-1 Time Series. *Remote Sensing*, 10, 1–20.
- Deng, X., López-Martínez, C., Chen, J., and Han, P., 2017. Statistical modeling of polarimetric SAR data: A survey and challenges. *Remote Sensing*, 9, 1–34.
- Fan, J., Zhang, F., Zhao, D., and Wang, J., 2015. Oil Spill Monitoring Based on SAR Remote Sensing Imagery. *Aquatic Procedia*, 3, 112–118.
- Ferreira, J.A. and Nascimento, A.D.C., 2020. Shannon entropy for the \mathcal{G}_I^0 model: A new segmentation approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 2547–2553.
- Frery, A.C., Muller, H.J., Yanasse, C.C.F., and Sant’Anna, S.J.S., 1997. A model for extremely heterogeneous clutter. *IEEE Transactions on Geoscience and Remote Sensing*, 35, 648–659.
- Frery, A.C., Nascimento, A.D.C., and Cintra, R.J., 2011. Information theory and image understanding: An application to polarimetric SAR imagery. *Chilean Journal of Statistics*, 2, 81–100.
- Gambini, J., Mejail, M., Jacobo-Berlles, J., and Frery, A.C., 2006. Feature extraction in speckled imagery using dynamic B-spline deformable contours under the G_0 model. *International Journal of Remote Sensing*, 27, 5037–5059.
- Gambini, J., Mejail, M., Jacobo-Berlles, J., and Frery, A.C., 2008. Accuracy of edge detection methods with local information in speckled imagery. *Statistics and Computing*, 18, 15–26.
- Giron, E., Frery, A.C., and Cribari-Neto, F., 2012. Nonparametric edge detection in speckled imagery. *Mathematics and Computers in Simulation*, 82, 2182–2198.
- Gradshteyn, I.S. and Ryzhik, I.M., 1980. *Tables of Integrals, Series and Products*. Academic Press, New York.
- Kong, J., 1990. *Electromagnetic Wave Theory*. Wiley, New York.
- Mejail, M.E., Jacobo-Berlles, J., Frery, A.C., and Bustos, O.H., 2003. Classification of SAR images using a general and tractable multiplicative model. *International Journal of Remote Sensing*, 24, 3565–3582.
- Nascimento, A.D.C., Cintra, R.J., and Frery, A.C., 2010. Hypothesis testing in speckled data with stochastic distances. *IEEE Transactions on Geoscience and Remote Sensing*, 48, 373–385.
- Nascimento, A.D.C., Horta, M.M., and Frery, A.C., Cintra, R.J., 2014. Comparing edge detection methods based on stochastic entropies and distances for PolSAR imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7, 648–663.
- Nikooravesh, Z., 2018. Estimation of the probability function under special moments conditions using the maximum Shannon and Tsallis entropies. *Chilean Journal of Statistics*, 9, 55–64.
- Salicrú, M., Menéndez, M.L., Pardo, L., and Morales, D., 1994. On the applications of divergence type measures in testing statistical hypothesis. *Journal of Multivariate Analysis*, 51, 372–391.
- Sant’Anna, S.J.S., Lacava, J.C.S., and Fernandes, D., 2008. From Maxwell’s equations to polarimetric SAR images: A simulation approach. *Sensors*, 8, 7380–7409.

- Ward, K.D., 1981. Compound representation of high resolution sea clutter. *Electronic Letters*, 17, 561-563.
- Ward, K.D., Watts, S., and Tough, R.J., 2006. Sea clutter: scattering, the K distribution and radar performance. *IET*, 20.
- Wei, Q.R., Feng, D.Z. 2015. An efficient SAR edge detector with a lower false positive rate. *International Journal of Remote Sensing*, 36, 3773-3797.
- Wickham, H. 2019. *Advanced R*. CRC Press, Boca Raton, FL, US.
- Yue, D.X., Xu, F., Frery, A.C., and Jin, Y.Q. 2020. A generalized Gaussian coherent scatterer model for correlated SAR texture. *IEEE Transactions on Geoscience and Remote Sensing*, 58, 2947-2964.
- Yue, D.X., Xu, F., Frery, A.C., and Jin, Y.Q., 2021. SAR image statistical modeling Part I: Single-pixel statistical models. *IEEE Geoscience and Remote Sensing Magazine*, 9, 82-114.

STATISTICAL MODELING
RESEARCH PAPER

Alternatives to the logit model in the situation of factor levels aggregation in binomial responses

MALINDA COA¹ and ERNESTO PONSOT^{2,3,*}

¹Institute of Applied Statistics and Computing, University of Los Andes, Mérida, Venezuela

²Statistics Department, University of Los Andes, Mérida, Venezuela

³Topos: Data Science Consulting, Ibarra, Ecuador

(Received: 27 October 2020 · Accepted in final form: 26 April 2021)

Abstract

The detailed study of the logit, probit and cloglog link functions is presented for the generalized linear model with binomial response in the presence of the problem of explanatory factors levels aggregation. Expressions are deduced for the estimators of the parameters and their variances, in general terms, which allows for finding the particular results for any link function chosen. The impact of the link function on the estimates is illustrated, concluding that the use of the appropriate variance in the levels aggregation is preferable, regardless of the link function to be used.

Keywords: Binomial regression · Generalized linear models · Level sets · Link function

Mathematics Subject Classification: Primary 62J12 · Secondary 62J20.

1. INTRODUCTION

The binomial model pursues the same objectives as the classical regression model, however, they differ in their structure (Collet, 2002; Tutz, 2011). The crucial difference is that, in the binomial model, the dependent variable follows a distribution that takes only two possible values, zero and one, in contrast to the normally distributed response of the classical regression model, in which any real value can be observed. Thus, the categorization of the response variable in the binomial model leads to the second difference. This concerns the need for a link function between the explanatory variables and the mean of the response.

Within the context of binomial models, the logit model is the most widely studied and applied (Christensen, 1997; Hilbe, 2009; Hosmer and Lemeshow, 2000). It is a particular case of the generalized linear model (Nelder and Wedderburn, 1972), when logit is the link function between the random component and the systematic component of the model, with the probability of success in a Bernoulli trial being modeled. Factors or treatments, rather than variables, are postulated in the style of the conventional analysis of variance (McCullagh and Nelder, 1989; McCulloch and Searle, 2001).

In its simplest formulation, the logit model consists of a dichotomous response variable and a single explanatory factor. Additionally, it is assumed that the responses corresponding

* Corresponding author. Email: malindacoa@gmail.com, ernesto.pb@gmail.com

to the different levels of the explanatory factor are independent binomials. After adjusting and applying this model on data in a contingency table, suppose that the researcher decides to group some levels of the factor and reiterate the logit analysis in the usual way. With this procedure, [Ponsot et al. \(2009\)](#) demonstrated that a violation of the binomial assumption is incurred, with important implications for the variance. They suggest courses of action to correct the problem and at the same time improve the accuracy of the results. Specifically, based on the reference parameterization and the saturated model, the authors suggest a procedure that takes advantage of the computations of a first logit adjustment and corrects the distributional assumption about variance, producing more efficient estimates and with greater precision than those obtained if you decide to reiterate a logit adjustment. Through simulations, strong trends were shown in favor of the proposed method, even more, if the probabilities of success of the response variable associated are more dissimilar to each other.

Note that the aforementioned research is limited to the scope of the logit model, but what about the problem of factor levels aggregation in models that are usually competitors or alternatives to the logit model? The logit model is the most used for the advantages it offers, however, it does not always guarantee a good fit for all binomial response data, so the researcher may consider other alternatives ([Bonat et al., 2018](#); [Czado and Santner, 1992](#); [Czado and Munk, 2000](#)). Logit is the canonical link function for binomial response data, but probit is also popular ([McCullagh and Nelder, 1989](#); [Collet, 2002](#); [Hosmer and Lemeshow, 2000](#)). In fact, any differentiable monotonous function can serve as a link between the random and systematic components of the binomial model, so there are many other functions that could offer a better fit than the traditional logit model. Therefore, keeping the problem within the scope of generalized linear models, this work seeks to answer this question by generalizing the procedure suggested by [Ponsot et al. \(2009\)](#) so that it is applicable with any link function.

The paper has been organized as follows: Section 2 explains the problem of explanatory factors levels aggregation. In Section 3, the fit of binomial models under this situation is addressed through the usual method. In Section 4, we describe the method suggested by [Ponsot et al. \(2009\)](#). Section 5 develops the adjustment procedures of three of the best known binomial models (logit, probit and cloglog), using the methods proposed in Sections 3 and 4. An example of the potential application of these models is shown in Section 6. Finally, Section 7 presents the main conclusions derived from this work.

2. THE PROBLEM OF EXPLANATORY FACTORS LEVELS AGGREGATION

In Table 1, let y_i be the observed number of successes observed in the i -th level of the factor A and n_i the total number of observations for that level.

Table 1. Observed number of successes and total of the Y response versus the A factor levels.

		Y	
A	Number of successes	Total	
1	y_1	n_1	
2	y_2	n_2	
\vdots	\vdots	\vdots	
$k - 2$	y_{k-2}	n_{k-2}	
$k - 1$	y_{k-1}	n_{k-1}	
k	y_k	n_k	
Total	$y.$	$n.$	

The corresponding responses to the different levels of A are assumed independent of each other and from a binomial population in the number of successes ($Y = 1$). This is

$$Y_i \stackrel{\text{ind}}{\sim} \text{Bin}(y_i; n_i, p_i), \quad i = 1, \dots, k$$

where “ind” stands for independent, Y_i is the random variable that represents the number of successes in the i -th sample and p_i , considered constant, is the probability of success associated ($0 < p_i < 1$). Assuming a binomial distribution in the number of successes of the Y_i at each level of the explanatory factor, implies that $V[Y_i] = n_i p_i (1 - p_i)$ and $E[Y_i] = n_i p_i$.

Sometimes, after adjusting a model the researcher may decide to group levels of the A factor for various reasons. Suppose the levels k and $k - 1$ are added doing $y_{k-1}^* = y_{k-1} + y_k$ and $n_{k-1}^* = n_{k-1} + n_k$, obtaining an arrangement of the data as in Table 2. The situation could extend to more than two levels, simply by adding the last two, then these with the previous one, and so on.

Table 2. Number of successes and total of the Y response versus the A factor levels, after the aggregation of the k and $k - 1$ levels.

Y		
A	Number of successes	Total
1	y_1	n_1
2	y_2	n_2
\vdots	\vdots	\vdots
$k - 2$	y_{k-2}	n_{k-2}
$k - 1$	y_{k-1}^*	n_{k-1}^*
Total	$y.$	$n.$

Usually, by reiterating the model fit procedure, the researcher assumes that the new random variable $Y_{k-1}^* = Y_{k-1} + Y_k$, that arises from aggregation, still has a binomial distribution with variance

$$V_{\text{Bin}}[Y_{k-1}^*] = n_{k-1}^* p_{k-1}^* (1 - p_{k-1}^*), \tag{1}$$

where $n_{k-1}^* = n_{k-1} + n_k$ and $p_{k-1}^* = E[Y_{k-1}^*] / n_{k-1}^* = (n_{k-1} p_{k-1} + n_k p_k) / (n_{k-1} + n_k)$.

In this regard, Ponsot et al. (2009) demonstrated that with this proceeding, a violation of the original binomial assumption is incurred, with important implications for the estimated variances. In their work, the authors deduced the following:

- (1) If $p_{k-1} \neq p_k$, the binomial assumption is violated in the sample corresponding to the level of the response variable where aggregation arises: Y_{k-1}^* is actually distributed Poisson-binomial (and not binomial).
- (2) The right expression for the variance of Y_{k-1}^* is

$$V[Y_{k-1}^*] = n_{k-1} p_{k-1} (1 - p_{k-1}) + n_k p_k (1 - p_k). \tag{2}$$

Also, the authors argued that Equations (1) and (2) are not equivalent and that in general it has $V_{\text{Bin}}[Y_{k-1}^*] \geq V[Y_{k-1}^*]$. In this regard, let $\Delta V = V_{\text{Bin}}[Y_{k-1}^*] - V[Y_{k-1}^*]$. Then, we have

$$\Delta V = \frac{n_{k-1} n_k}{n_{k-1} + n_k} (p_{k-1} - p_k)^2. \tag{3}$$

From Equation 3, the authors deduced that:

- (1) If p_{k-1} and p_k are close ($p_{k-1} \approx p_k$), the difference $(p_{k-1} - p_k)^2 \rightarrow 0$, so that $\Delta V \rightarrow 0$ and $V_{\text{Bin}}[Y_{k-1}^*] \approx V[Y_{k-1}^*]$.
- (2) If p_{k-1} and p_k are distant from each other (which occurs, for example, when $p_{k-1} \approx 0$ and $p_k \approx 1$, or $p_{k-1} \approx 1$ and $p_k \approx 0$), then the difference $(p_{k-1} - p_k)^2 \rightarrow 1$, so that $\Delta V \rightarrow n_{k-1}n_k/(n_{k-1} + n_k)$ and $V_{\text{Bin}}[Y_{k-1}^*] \approx V[Y_{k-1}^*] + n_{k-1}n_k/(n_{k-1} + n_k)$. In this situation, the greatest difference between the variances occurs. This difference can be considerable depending on the n_{k-1} and n_k values.

Now, Figure 1(a) shows the behavior of the variance assumed by the researcher when the n_{k-1} and n_k parameters are fixed and vary the values of p_{k-1} and p_k . Clearly, it can be seen in the graph that as $p_{k-1} \approx 0$ and $p_k \approx 0$, or $p_{k-1} \approx 1$ and $p_k \approx 1$, then $V_{\text{Bin}}[Y_{k-1}^*] \rightarrow 0$. Meanwhile, relative maximums are obtained along the ordered pairs $(p_{k-1}, p_k) = (p_{k-1}, 0.5[1 + (n_{k-1}/n_k)(1 - 2p_{k-1})])$, where $V_{\text{Bin}}[Y_{k-1}^*]$ reaches the value $n_{k-1}^*/4$ in each one of them (see Appendix A).

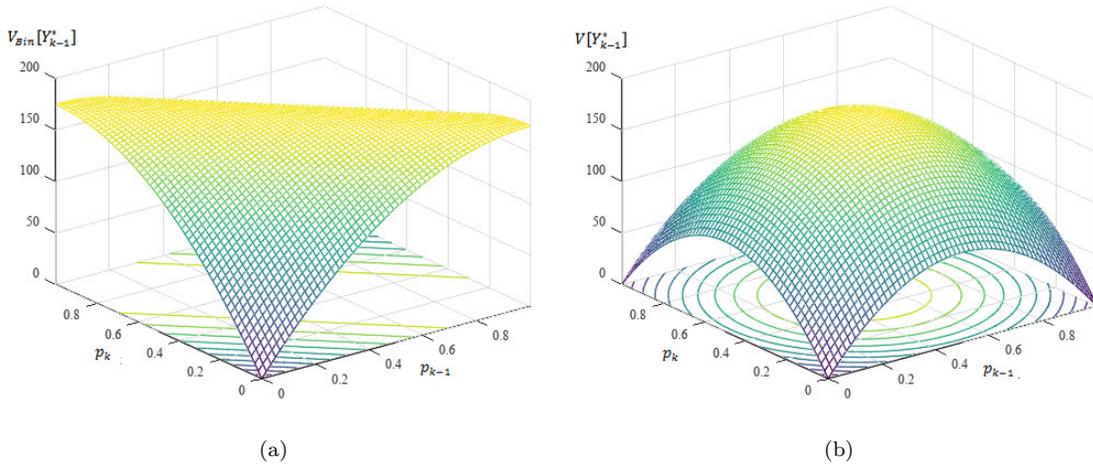


Figure 1. Y_{k-1}^* variances: (a) Binomial variance ($V_{\text{Bin}}[Y_{k-1}^*]$) assumed by the researcher; (b) True variance ($V[Y_{k-1}^*]$).

Note that the true variance (Figure 1(b)) shows similar behavior to the binomial variance, when p_{k-1} and p_k both tend to 0 or 1. In fact, whenever $p_{k-1} = p_k$. However, this behavior also occurs when one of them tends to 0 and the other to 1 (or viceversa). The minimum values of the variance occur for the cases mentioned, while the maximum occurs when both parameters take the value 0,5 ($p_{k-1} = p_k = 0.5$), being $V[Y_{k-1}^*] = n_{k-1}^*/4$ said maximum (see Appendix B).

In Figure 2, the difference between the variance assumed by the researcher and the true variance is shown. Indeed, this figure shows that the differences tend to 0 when $p_{k-1} \approx p_k$. These differences grow when p_{k-1} and p_k tend to opposite ends, reaching their maximum values in $p_{k-1} \approx 0$ and $p_k \approx 1$, and $p_{k-1} \approx 1$ and $p_k \approx 0$.

3. FITTING A BINOMIAL MODEL USING THE USUAL METHOD

Next, we describe the saturated binomial model before aggregation. It is possible to use the generalized linear model approach (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; McCulloch and Searle, 2001; Dobson, 2002; Agresti, 2007, 2015) to the data in Table 1, when the response variables Y_1, \dots, Y_k are supposed independent and follow a binomial

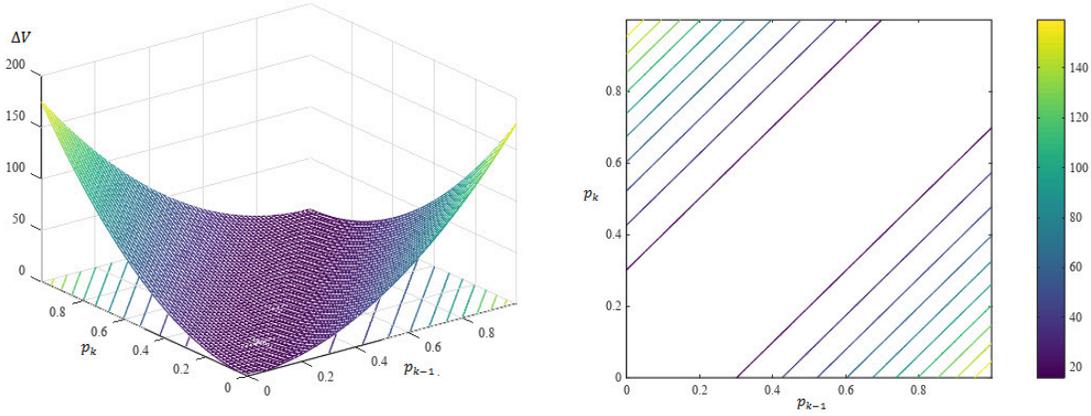


Figure 2. Difference between binomial and true variances ($\Delta V = V_{\text{Bin}}[Y_{k-1}^*] - V[Y_{k-1}^*]$).

distribution as this belongs to the exponential family of distributions. For the link, any monotonous and differentiable function can be used, however, the choice of it has given rise to the most important binomial models present in the literature (logit model, probit model, cloglog model, to mention some, among which the first one stands out).

In its simplest sense, the saturated model ($m = k$) and reference parameterization can be postulated, whereby the matrix \mathbf{X} is a square matrix (of order $k \times k$) and invertible (Ponsot, 2011). Being k the reference level, this parameterization leads to the model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, whose matrix representation is:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_{k-2} \\ \eta_{k-1} \\ \eta_k \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-2} \\ \beta_{k-1} \\ \beta_k \end{bmatrix}. \quad (4)$$

The \mathbf{X} matrix can be partitioned as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{j} & \mathbf{I} \\ \mathbf{1} & \mathbf{0}^\top \end{bmatrix} \Rightarrow \mathbf{X}^{-1} = \begin{bmatrix} \mathbf{0}^\top & \mathbf{1} \\ \mathbf{I} & -\mathbf{j} \end{bmatrix}. \quad (5)$$

Since $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \Rightarrow \boldsymbol{\beta} = \mathbf{X}^{-1}\boldsymbol{\eta}$. When saturated, the model raised in Equation (4) does not have sufficient degrees of freedom to calculate the *deviance* or Pearson statistics (Ponsot, 2011). However, you can still estimate its parameters ($\boldsymbol{\beta}$) and determine its statistical significance.

Let $\hat{\eta}_i = g(\hat{\mu}_i), i = 1, \dots, k$. From Equations (4) and (5), it follows in general terms that

$$\hat{\beta}_j = \begin{cases} \hat{\eta}_k & \text{if } j = 1 \\ \hat{\eta}_{j-1} - \hat{\eta}_k & \text{if } j = 2, \dots, k. \end{cases} \quad (6)$$

Since $\boldsymbol{\beta} \sim \text{AN}[\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}]$, the parameters variance is

$$V[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} = \mathbf{X}^{-1} \mathbf{W}^{-1} (\mathbf{X}^\top)^{-1}$$

where $\mathbf{W} = \text{diag} \{w_1, \dots, w_k\}$ with $w_i = (\partial \mu_i / \partial \eta_i)^2 V_{\text{Bin}}[Y_i]$, for $i = 1, \dots, k$.

Therefore, in general terms, the variance of $\widehat{\beta}$ is given by

$$V[\widehat{\beta}_j] = \begin{cases} \frac{V_{\text{Bin}}[Y_k]}{(\partial\mu_k/\partial\eta_k)^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{(\partial\mu_{j-1}/\partial\eta_{j-1})^2} + \frac{V_{\text{Bin}}[Y_k]}{(\partial\mu_k/\partial\eta_k)^2}, & \text{if } j = 2, \dots, k. \end{cases} \quad (7)$$

After fitting a saturated binomial model, assume that the last two levels $k-1$ and k of the A factor are added, leaving the data arranged as in Table 2. As already mentioned sometimes the researcher reiterates the usual adjustment process, as in the previous section, assuming that the new random variable $Y_{k-1}^* = Y_{k-1} + Y_k$ has variance $V_{\text{Bin}}[Y_{k-1}^*] = n_{k-1}^* p_{k-1}^* (1 - p_{k-1}^*)$, with $n_{k-1}^* = n_{k-1} + n_k$ and $p_{k-1}^* = (n_{k-1} p_{k-1} + n_k p_k) / n_{k-1}^*$.

In this new fit, the design matrix that arises from aggregation (call \mathbf{X}^*) now has dimensions $(k-1) \times (k-1)$ due to the elimination of the k -th row and k -th column. However, if the reference parameterization is maintained, then with respect to the aggregate level $k-1$ is possible to propose a model as in Equation (4) on the new data set.

As in Equation (6), the estimated parameter vector elements obtained by the usual method (denote the superscript h like $\widehat{\beta}^{*h}$) are expressed as

$$\widehat{\beta}_j^{*h} = \begin{cases} \widehat{\eta}_{k-1}^*, & \text{if } j = 1; \\ \widehat{\eta}_{j-1} - \widehat{\eta}_{k-1}^*, & \text{if } j = 2, \dots, k-1. \end{cases} \quad (8)$$

Observe that the new diagonal matrix \mathbf{W}^{*h} keeps the $k-2$ elements of \mathbf{W} , changing only the one that corresponds to the level $k-1$, that is, we have

$$\begin{aligned} \mathbf{W}^{*h} &= \text{diag} \{w_1, \dots, w_{k-1}^*\} \\ &= \text{diag} \left\{ \frac{(\partial\mu_1/\partial\eta_1)^2}{V_{\text{Bin}}[Y_1]}, \dots, \frac{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}{V_{\text{Bin}}[Y_{k-1}^*]} \right\}. \end{aligned}$$

Thus, the variances and covariances matrix is given by

$$V[\widehat{\beta}^{*h}] = [(\mathbf{X}^*)^\top \mathbf{W}^{*h} \mathbf{X}^*]^{-1} = (\mathbf{X}^*)^{-1} (\mathbf{W}^{*h})^{-1} [(\mathbf{X}^*)^\top]^{-1}$$

and as in Equation (7), the variance of $\widehat{\beta}_j^{*h}$ is

$$V[\widehat{\beta}_j^{*h}] = \begin{cases} \frac{V_{\text{Bin}}[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{(\partial\mu_{j-1}/\partial\eta_{j-1})^2} + \frac{V_{\text{Bin}}[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, & \text{if } j = 2, \dots, k-1. \end{cases} \quad (9)$$

Equations (8) and (9) are generalizations of the usual procedure for fitting binomial models when any differentiable monotone function is used as a link. Therefore, from now on, they are part of the context that we call the generalized usual method (GUM).

4. FITTING A BINOMIAL MODEL USING THE PONSOT METHOD

In the presence of the saturated model given in Equation (4), Ponsot et al. (2009) proposed a method to reiterate the adjustment of a binomial model, after the levels aggregation, when the first adjustment uses the logit link function.

Now, let $g(\mu_i)$ be any differentiable monotonous link function. Asymptotically, in the saturated model $V[\mathbf{X}\hat{\boldsymbol{\beta}}] = \mathbf{X}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X} \mathbf{X}^{-1} \mathbf{W}^{-1} (\mathbf{X}^\top)^{-1} \mathbf{X}^\top = \mathbf{W}^{-1}$. Then, under conditions of regularity and as a consequence of the usual central limit theorem, as well as properties of the maximum-likelihood estimators, we have that

$$g(\hat{\mu}_i) = x_i^\top \hat{\boldsymbol{\beta}} \sim \text{AN} \left(x_i^\top \boldsymbol{\beta}; w_i^{-1} = \left[\frac{(\partial \mu_i / \partial \eta_i)^2}{V_{\text{Bin}}[Y_i]} \right]^{-1} = \frac{V_{\text{Bin}}[Y_i]}{(\partial \mu_i / \partial \eta_i)^2} \right). \quad (10)$$

Applying the delta method in Equation (10) (Agresti, 2007), we get

$$\begin{aligned} \mu_i &= g^{-1}(\eta_i) \\ \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \eta_i}. \end{aligned}$$

Thus, we reach

$$\hat{\mu}_i \sim \text{AN} \left(\mu_i = g^{-1}(x_i^\top \boldsymbol{\beta}); V_{\text{Bin}}[Y_i] \right).$$

Now, when $k-1$ and k levels are added, the maximum-likelihood estimator of the new mean μ_{k-1}^* is given by

$$\hat{\mu}_{k-1}^* = \widehat{E}[Y_{k-1}^*] = \widehat{E}[Y_{k-1}] + \widehat{E}[Y_k] = \hat{\mu}_{k-1} + \hat{\mu}_k. \quad (11)$$

Due to $\hat{\mu}_{k-1}^*$ is the weighted sum of two linear functions of asymptotically independent normal random variables, their asymptotic distribution is also normal with

$$E[\hat{\mu}_{k-1}^*] = E[\hat{\mu}_{k-1}] + E[\hat{\mu}_k] = \mu_{k-1} + \mu_k$$

and

$$V[\hat{\mu}_{k-1}^*] = V[\hat{\mu}_{k-1}] + V[\hat{\mu}_k] = V_{\text{Bin}}[Y_{k-1}] + V_{\text{Bin}}[Y_k] = V[Y_{k-1}^*].$$

Again, using the delta method, the required distribution of $\hat{\eta}_{k-1}^* = g(\hat{\mu}_{k-1}^*)$ is asymptotically normal with expected value $E[\hat{\eta}_{k-1}^*] = g(\mu_{k-1}^*) = \eta_{k-1}^*$ and asymptotic variance stated as

$$(\sigma^2)_{k-1}^* = V[\hat{\mu}_{k-1}^*] \left[\frac{\partial g(\mu_{k-1}^*)}{\partial \mu_{k-1}^*} \right]^2 = V[Y_{k-1}^*] \left[\frac{\partial \eta_{k-1}^*}{\partial \mu_{k-1}^*} \right]^2. \quad (12)$$

From $(\sigma^2)_{k-1}^*$, Ponsot et al. (2009) suggested creating a matrix $\boldsymbol{\Sigma}$, equivalent to \mathbf{W}^{-1} from the original fit and from which its $k-2$ elements remains, but with the $k-1$ element added corrected for true variance $V[Y_{k-1}^*]$, as in Equation (12). This suggested matrix is of

the form expressed as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 & 0 \\ 0 & 1/w_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1/w_{k-2} & 0 \\ 0 & 0 & \cdots & 0 & (\sigma^2)_{k-1}^* \end{bmatrix}.$$

The variance and covariance matrix of the estimators using this suggested method (now denoted with the superscript s) is given by

$$V[\widehat{\boldsymbol{\beta}}^{*s}] = (\mathbf{X}^*)^{-1} \boldsymbol{\Sigma} [(\mathbf{X}^*)^\top]^{-1}$$

resulting in

$$V[\widehat{\beta}_j^{*s}] = \begin{cases} \frac{V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{(\partial\mu_{j-1}/\partial\eta_{j-1})^2} + \frac{V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, & \text{if } j = 2, \dots, k-1. \end{cases} \quad (13)$$

As in Equation (8), the elements of the estimated parameter vector using the suggested method ($\widehat{\boldsymbol{\beta}}^{*s}$) are defined as

$$\widehat{\beta}_j^{*s} = \begin{cases} \widehat{\eta}_{k-1}^* & \text{if } j = 1 \\ \widehat{\eta}_{j-1} - \widehat{\eta}_{k-1}^* & \text{if } j = 2, \dots, k-1. \end{cases} \quad (14)$$

The above results are a generalization of the method suggested by [Ponsot et al. \(2009\)](#), which we call the generalized suggested method (GSM). The GSM allows us to adjust a binomial model using the suggested method, but now using any link function.

Next, we describe differences between the variances obtained through GUM and GSM. Let $\Delta V(\widehat{\beta}_1^*) = V[\beta_1^{*h}] - V[\beta_1^{*s}]$ be the difference between the variances obtained by the GUM and the GSM, for the β_1^* parameter. Then, from Equations (9) and (13), we have

$$\begin{aligned} \Delta V(\widehat{\beta}_1^*) &= \frac{V_{\text{Bin}}[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2} - \frac{V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, \\ &= \frac{V_{\text{Bin}}[Y_{k-1}^*] - V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, \\ &= \frac{\Delta V}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, \end{aligned}$$

where ΔV is defined as in Equation (3).

For the parameters β_j^* , with $j = 2, \dots, k-1$, let $\Delta V(\hat{\beta}_j^*) = V(\hat{\beta}_j^{*h}) - V(\hat{\beta}_j^{*s})$ be the difference between the variances obtained by both methods. Again, from Equations (9) and (13), we get

$$\begin{aligned} \Delta V(\hat{\beta}_j^*) &= \left[\frac{V_{\text{Bin}}[Y_{j-1}]}{(\partial\mu_{j-1}/\partial\eta_{j-1})^2} + \frac{V_{\text{Bin}}[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2} \right] \\ &\quad - \left[\frac{V_{\text{Bin}}[Y_{j-1}]}{(\partial\mu_{j-1}/\partial\eta_{j-1})^2} + \frac{V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2} \right] \\ &= \frac{V_{\text{Bin}}[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2} - \frac{V[Y_{k-1}^*]}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2} \\ &= V[\beta_1^{*h}] - V[\beta_1^{*s}] = \Delta V(\hat{\beta}_1^*). \end{aligned}$$

This last result demonstrates that the difference in the variances obtained by both methods, for the first component of the vector $\hat{\beta}^*$ (that is, $\hat{\beta}_1^*$), is the same for the remaining components. Thus, we have that

$$\Delta V(\hat{\beta}_j^*) = \frac{\Delta V}{(\partial\mu_{k-1}^*/\partial\eta_{k-1}^*)^2}, \quad \text{for } j = 1, \dots, k-1. \quad (15)$$

In Equation (15), the denominator is dependent on the selected link function and the numerator is a function of ΔV . The same conditions established at the end of the Section 2 apply to p_{k-1} and p_k .

5. SOME BINOMIAL MODELS USING THE GUM AND GSM

Assume the k and $k-1$ level aggregation situation in Table 2, and let

$$\hat{p}_{k-1}^* = \frac{n_{k-1}\hat{p}_{k-1} + n_k\hat{p}_k}{n_{k-1} + n_k}, \quad (16)$$

as it follows from Equation (11). In this section, the adjustment procedure of three of the best known binomial models (logit, probit and cloglog) is shown, using the methods obtained in the previous sections.

The link function in the logit model is stated as

$$\eta_i = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\mu_i}{n-\mu_i}\right).$$

Then, we have

$$\frac{\partial\eta_i}{\partial\mu_i} = \frac{1}{n_i p_i (1-p_i)} = \frac{1}{V_{\text{Bin}}[Y_i]}.$$

Thus, we get

$$\frac{\partial \mu_i}{\partial \eta_i} = \text{V}_{\text{Bin}}[Y_i].$$

From Equations (8) and (14), the parameters estimated by both methods are identical and equal to

$$\widehat{\beta}_j^{*h} = \widehat{\beta}_j^{*s} = \begin{cases} \text{logit}(\widehat{p}_{k-1}^*), & \text{if } j = 1; \\ \text{logit}(\widehat{p}_{j-1}) - \text{logit}(\widehat{p}_{k-1}^*), & \text{if } j = 2, \dots, k-1. \end{cases}$$

The variances of the estimators obtained by GUM, according to Equation (9), are obtained as

$$\text{V}[\widehat{\beta}_j^{*h}] = \begin{cases} \frac{1}{\text{V}_{\text{Bin}}[Y_{k-1}^*]}, & \text{if } j = 1; \\ \frac{1}{\text{V}_{\text{Bin}}[Y_{j-1}]} + \frac{1}{\text{V}_{\text{Bin}}[Y_{k-1}^*]}, & \text{if } j = 2, \dots, k-1. \end{cases}$$

Meanwhile, the variances estimated by the GSM for these same estimators, according to Equation (13), are given by

$$\text{V}[\widehat{\beta}_j^{*s}] = \begin{cases} \frac{\text{V}[Y_{k-1}^*]}{(\text{V}_{\text{Bin}}[Y_{k-1}^*])^2}, & \text{if } j = 1; \\ \frac{1}{\text{V}_{\text{Bin}}[Y_{j-1}]} + \frac{\text{V}[Y_{k-1}^*]}{(\text{V}_{\text{Bin}}[Y_{k-1}^*])^2}, & \text{if } j = 2, \dots, k-1. \end{cases}$$

Regarding the differences between these variances, from Equation (15), we have that

$$\Delta \text{V}(\widehat{\beta}_j^*) = \Delta \text{V}(\widehat{\beta}_1^*) = \frac{\Delta \text{V}}{(\text{V}_{\text{Bin}}[Y_{k-1}^*])^2}.$$

Figure 3 graphically displays the behavior of $\Delta \text{V}(\widehat{\beta}_j^*)$ for the logit model. It is observed that its performance is very similar to that of ΔV shown in Figure 2.

The link function in the probit model is given by

$$\eta_i = \text{probit}(p_i) = \Phi^{-1}(p_i) = \Phi^{-1}\left(\frac{n_i p_i}{n_i}\right) = \Phi^{-1}\left(\frac{\mu_i}{n_i}\right).$$

Hence, we get

$$\mu_i = n_i \Phi(\eta_i)$$

and since $[\Phi]' = \phi$, then we have that

$$\frac{\partial \mu_i}{\partial \eta_i} = n_i \phi(\eta_i) = n_i \phi[\Phi^{-1}(p_i)].$$

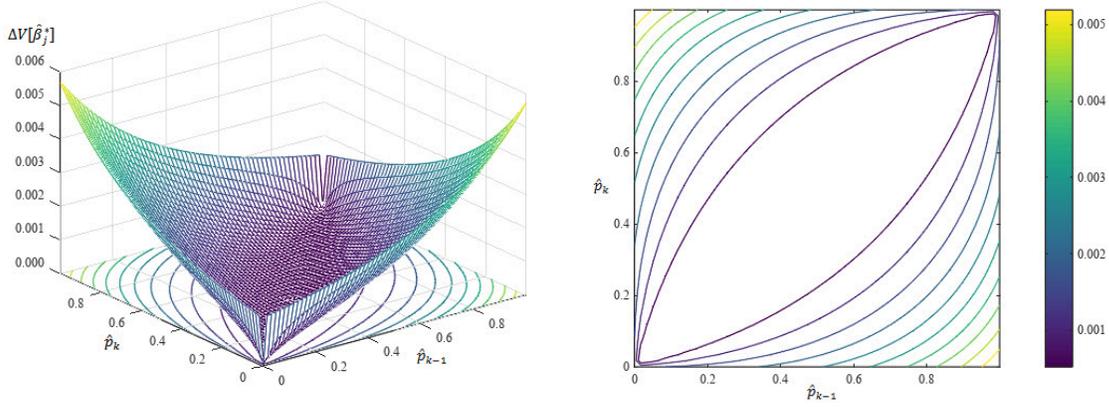


Figure 3. Differences between variances estimated using the GUM and GSM ($\Delta V(\hat{\beta}_j^*) = V(\hat{\beta}_j^{*h}) - V(\hat{\beta}_j^{*s})$) with logit link function

From Equation (8) and (14), the parameters estimated by the GUM and GSM are identical and equal to

$$\hat{\beta}_j^{*h} = \hat{\beta}_j^{*s} = \begin{cases} \text{probit}(\hat{p}_{k-1}^*), & \text{if } j = 1 \\ \text{probit}(\hat{p}_{j-1}) - \text{probit}(\hat{p}_{k-1}^*), & \text{if } j = 2, \dots, k - 1. \end{cases}$$

From Equation (9), the variances of the estimators obtained by the GUM are stated as

$$V[\hat{\beta}_j^{*h}] = \begin{cases} \frac{V_{\text{Bin}}[Y_{k-1}^*]}{\{n_{k-1}^* \phi[\Phi^{-1}(p_{k-1}^*)]\}^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{\{n_{j-1} \phi[\Phi^{-1}(p_{j-1})]\}^2} + \frac{V_{\text{Bin}}[Y_{k-1}^*]}{\{n_{k-1}^* \phi[\Phi^{-1}(p_{k-1}^*)]\}^2}, & \text{if } j = 2, \dots, k - 1. \end{cases}$$

Meanwhile, the variances estimated by the GSM, according to Equation (13), are expressed as

$$V[\hat{\beta}_j^{*s}] = \begin{cases} \frac{V[Y_{k-1}^*]}{\{n_{k-1}^* \phi[\Phi^{-1}(p_{k-1}^*)]\}^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{\{n_{j-1} \phi[\Phi^{-1}(p_{j-1})]\}^2} + \frac{V[Y_{k-1}^*]}{\{n_{k-1}^* \phi[\Phi^{-1}(p_{k-1}^*)]\}^2}, & \text{if } j = 2, \dots, k - 1. \end{cases}$$

From Equation (15), the differences between these estimated variances are defined by

$$\Delta V(\hat{\beta}_j^*) = \Delta V(\hat{\beta}_1^*) = \frac{\Delta V}{\{n_{k-1}^* \phi[\Phi^{-1}(p_{k-1}^*)]\}^2}.$$

Figure 4 graphically displays the behavior of $\Delta V(\hat{\beta}_j^*)$ for the probit model. Without considering the difference of scales inherent in each case, it is observed that it is similar to that of the logit model, differing mainly in the borders or neighborhoods where \hat{p}_{k-1} and \hat{p}_k both approach 0 or 1.

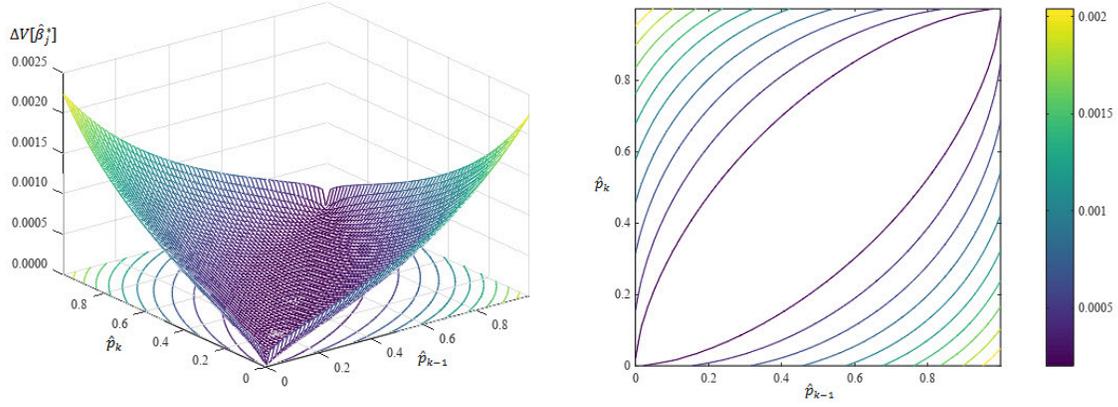


Figure 4. Differences between variances estimated using the GUM and GSM ($\Delta V(\widehat{\beta}_j^*) = V[\widehat{\beta}_j^{*h}] - V[\widehat{\beta}_j^{*s}]$) with probit link function.

The link function in the cloglog model is established by

$$\eta_i = \text{cloglog}(p_i) = \log[-\log(1 - p_i)] = \log[-\log(1 - \mu_i/n_i)] = g(\mu_i).$$

By differentiating, we obtain

$$\frac{\partial \eta_i}{\partial \mu_i} = -\frac{1}{n_i(1 - p_i) \log(1 - p_i)}.$$

Thus, we get

$$\frac{\partial \mu_i}{\partial \eta_i} = -n_i(1 - p_i) \log(1 - p_i).$$

From Equation (8) and (14), the parameters estimated by GUM and GSM are, once again, identical and equal to

$$\widehat{\beta}_j^{*h} = \widehat{\beta}_j^{*s} = \begin{cases} \text{cloglog}(\widehat{p}_{k-1}^*), & \text{if } j = 1; \\ \text{cloglog}(\widehat{p}_{j-1}^*) - \text{cloglog}(\widehat{p}_{k-1}^*), & \text{if } j = 2, \dots, k-1. \end{cases}$$

The variances of the estimators obtained by the GUM, according to Equation (9), are stated as

$$V[\widehat{\beta}_j^{*h}] = \begin{cases} \frac{V_{\text{Bin}}[Y_{k-1}^*]}{[n_{k-1}^*(1 - p_{k-1}^*) \log(1 - p_{k-1}^*)]^2}, & \text{if } j = 1 \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{[n_{j-1}(1 - p_{j-1}) \log(1 - p_{j-1})]^2} \\ + \frac{V_{\text{Bin}}[Y_{k-1}^*]}{[n_{k-1}^*(1 - p_{k-1}^*) \log(1 - p_{k-1}^*)]^2}, & \text{if } j = 2, \dots, k-1. \end{cases}$$

Meanwhile, the variances estimated by the GSM for these same parameters, according to Equation (13), are given by

$$V[\hat{\beta}_j^{*s}] = \begin{cases} \frac{V[Y_{k-1}^*]}{[n_{k-1}^*(1 - p_{k-1}^*) \log(1 - p_{k-1}^*)]^2}, & \text{if } j = 1; \\ \frac{V_{\text{Bin}}[Y_{j-1}]}{[n_{j-1}(1 - p_{j-1}) \log(1 - p_{j-1})]^2} \\ + \frac{V[Y_{k-1}^*]}{[n_{k-1}^*(1 - p_{k-1}^*) \log(1 - p_{k-1}^*)]^2}, & \text{if } j = 2, \dots, k - 1. \end{cases}$$

From Equation (15), we get

$$\Delta V(\hat{\beta}_j^*) = \Delta V(\hat{\beta}_1^*) = \frac{\Delta V}{[n_{k-1}^*(1 - p_{k-1}^*) \log(1 - p_{k-1}^*)]^2}.$$

Figure 5 shows the performance of $\Delta V(\hat{\beta}_j^*)$ for the cloglog model. Of course, it also constitutes a particularization of ΔV . However, it is observed that the region or border where $\Delta V(\hat{\beta}_j^*) \rightarrow 0$ is a little more extensive than the previous models.

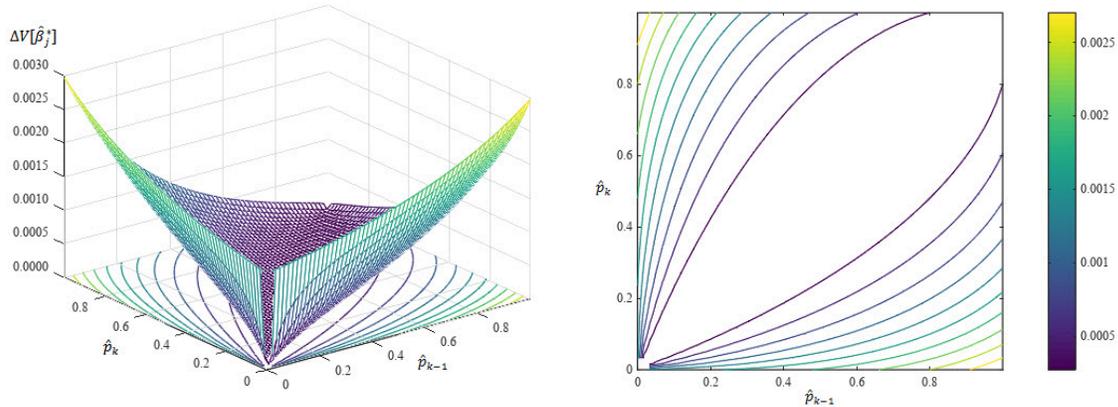


Figure 5. Differences between variances estimated using the GUM and GSM ($\Delta V(\hat{\beta}_j^*) = V[\hat{\beta}_j^{*h}] - V[\hat{\beta}_j^{*s}]$) with cloglog link function.

6. ILLUSTRATION OF THE PROCEDURES

Table 3 reproduces the example presented by Ponsot (2011). There, the situation of interest focused on studying the relationship between a Y response variable and an A explanatory factor with three levels.

Table 3. Example: $Y(0, 1)$ versus $A(1, 2, 3)$.

		Y		
A	0	1	Total	
1	189	161	350	
2	300	50	350	
3	32	318	350	
Total	521	529	1050	

From Equation (4), it follows the saturated model using the parameterization with the third level of the factor as reference stated as

$$\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix} = \begin{bmatrix} g(p_1) \\ g(p_2) \\ g(p_3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}.$$

Table 4 contains the estimates of the parameters of the linear predictor and its variances for the different binomial models, according to the link function used (for this case, logit, probit and cloglog links). This Table also contains the Wald χ^2 tests for $H_0: \beta_i = 0$ in order to check the statistical significance of the estimated parameters, as well as the 95% confidence intervals (CI) built for β_i . The predicted probabilities and their CI, following Agresti (2007), are shown in Table 5.

Table 4. Original model: $\hat{\beta}_i$ and Wald test ($H_0: \beta_i = 0$) according to the link function.

Link	i	Estimation of β_i				95% CI		
		$\hat{\beta}_i$	$V[\hat{\beta}_i]$	χ^2	p -value	Decision	LL	UL
logit	1	2.296	0.034	153.3	< 0.0001	Reject	1.933	2.660
	2	-2.457	0.046	131.5	< 0.0001	Reject	-2.877	-2.037
	3	-4.088	0.058	289.5	< 0.0001	Reject	-4.559	-3.617
probit	1	1.132	0.009	201.8	< 0.0001	Reject	1.148	1.516
	2	-1.432	0.013	154.3	< 0.0001	Reject	-1.658	-1.206
	3	-2.400	0.016	367.6	< 0.0001	Reject	-2.645	-2.154
cloglog	1	0.872	0.005	153.3	< 0.0001	Reject	0.734	1.010
	2	-1.356	0.011	161.8	< 0.0001	Reject	-1.565	-1.147
	3	-2.742	0.025	300.7	< 0.0001	Reject	-3.052	-2.432

where LL: lower limit and UL: upper limit.

Note that the three binomial models reject the null hypotheses $H_0: \beta_i = 0$, that is, their parameters are significant and fit the data well. Among them, the probit model is the one with the best fit for presenting higher values for the χ^2 statistic (which increases the power of the test). The predicted probabilities for each model are equal, regardless of the link function used, and their confidence intervals coincide in many cases, up to the order of thousandths. Now, suppose that after fitting any of these models, levels 2 and 3 of the A factor are added (see Table 3). Then, a new contingency table is obtained, such as the one shown in Table 6. In this case, the new model is given by

$$\begin{bmatrix} \eta_1^* \\ \eta_2^* \end{bmatrix} = \begin{bmatrix} g(p_1^*) \\ g(p_2^*) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \end{bmatrix}.$$

The new estimates obtained by using the GUM, that is, re-adjusted binomial models using logit, probit and cloglog link functions on the resulting contingency table (see Table 6) are shown in Table 7. The predicted probabilities for each of the models, without attention to parameters significance levels, are reproduced in Table 8.

Given the last two levels aggregation, the new parameter vector β^* is estimated differently than in the original model, as shown in Table 7. As for the new predicted probability vector (\hat{p}^*), without considering the significance levels of the parameters, the results are as expected: the first component is the same as the original model ($\hat{p}_1^* = \hat{p}_1$), while in the second component you get that $\hat{p}_2^* = 0.5257$, in accordance with Equation (16).

Table 5. Predicted probabilities and 95% CI according to link function.

Link	i	Estimation of p_i		95% CI	
		\hat{p}_i	LL	UL	
logit	1	0.4600	0.4084	0.5125	
	2	0.1429	0.1100	0.1836	
	3	0.9086	0.8736	0.9346	
probit	1	0.4600	0.4083	0.5124	
	2	0.1429	0.1093	0.1827	
	3	0.9086	0.8746	0.9352	
cloglog	1	0.4600	0.4094	0.5137	
	2	0.1429	0.1102	0.1841	
	3	0.9086	0.8755	0.9358	

where LL: lower limit, and UL: upper limit.

Table 6. Example: $Y(0, 1)$ versus $A(1, 2)$.

Y			
A	0	1	Total
1	189	161	350
2	332	368	700
Total	521	529	1050

Table 7. Usual procedure (GUM): $\hat{\beta}_i^*$ and Wald test ($H_0: \beta_i^* = 0$) according to the link function.

Link	i	Estimation of β_i^*				Decision	95% CI		
		$\hat{\beta}_i^*$	$V[\hat{\beta}_i^*]$	χ^2	p -value		LL	UL	L(CI)
logit	1	0.103	0.006	1.8	0.174	Not reject	-0.045	0.251	0.296
	2	-0.263	0.017	4.0	0.045	Reject	-0.521	-0.006	0.515
probit	1	0.065	0.002	1.9	0.174	Not reject	-0.028	0.157	0.185
	2	-0.165	0.007	4.0	0.045	Reject	-0.326	-0.004	0.322
cloglog	1	-0.293	0.003	30.2	< 0.000	Reject	-0.398	-0.189	0.209
	2	-0.191	0.009	3.9	0.047	Reject	-0.380	-0.003	0.377

where L(CI): length of CI, LL: lower limit, and UL: upper limit.

Table 8. Usual procedure (GUM): Predicted probabilities and 95% CI according to link function

Link	i	Estimation of p_i^*		95% CI		L(CI)
		\hat{p}_i^*	LL	UL		
logit	1	0.4600	0.4084	0.5125	0.1041	
	2	0.5257	0.4887	0.5625	0.0738	
probit	1	0.4600	0.4083	0.5124	0.1041	
	2	0.5257	0.4887	0.5625	0.0738	
cloglog	1	0.4600	0.4094	0.5137	0.1043	
	2	0.5257	0.4893	0.5631	0,0738	

where L(CI): length of CI, Ll: lower limit, and UL: upper limit.

Although these are the expected values, with $\alpha = 0.05$ and according to the results of Table 7, in the logit and probit models there is insufficient evidence to reject $H_0: \beta_1^* = 0$, in a strict statistical sense, and so the predicted probabilities in Table 8 for these models should not be considered valid. In fact, the correct predictions for the logit model would be

$$\hat{p}_1^* = \frac{\exp(\hat{\beta}_1^* + \hat{\beta}_2^*)}{1 + \exp(\hat{\beta}_1^* + \hat{\beta}_2^*)} = \frac{\exp(0 - 0.263)}{1 + \exp(0 - 0.263)} = \frac{\exp(-0.263)}{1 + \exp(-0.263)} = 0.4346 \quad \text{and}$$

$$\hat{p}_2^* = \frac{\exp(\hat{\beta}_1^*)}{1 + \exp(\hat{\beta}_1^*)} = \frac{\exp(0)}{1 + \exp(0)} = 0.5000$$

while those corresponding to the probit model would be given by

$$\hat{p}_1^* = \Phi(\hat{\beta}_1^* + \hat{\beta}_2^*) = \Phi(0 - 0.165) = \Phi(-0.165) = 0.4345 \quad \text{and}$$

$$\hat{p}_2^* = \Phi(\hat{\beta}_1^*) = \Phi(0) = 0.5000.$$

For that matter, only the cloglog model would remain valid since the nullity hypothesis for all its parameters is rejected.

In contrast, Table 9 and 10 present the estimates of the parameters and the predicted probabilities respectively, for the binomial models addressed, but now obtained through the GSM.

Table 9. Suggested procedure (GSM): $\hat{\beta}_i^*$ and Wald test ($H_0: \beta_i^* = 0$) according to the link function.

Link	i	Estimation of β_i^*				Decision	95% CI		
		$\hat{\beta}_i^*$	$V[\hat{\beta}_i^*]$	χ^2	p -value		LL	UL	L(CI)
logit	1	0.103	0.002	4.5	0.034	Reject	0.008	0.198	0.190
	2	-0.263	0.014	5.0	0.025	Reject	-0.494	-0.033	0.462
probit	1	0.065	0.001	4.5	0.034	Reject	0.005	0.124	0.119
	2	-0.165	0.005	5.0	0.025	Reject	-0.309	-0.020	0.289
cloglog	1	-0.293	0.001	73.3	< 0.000	Reject	-0.360	-0.226	0.134
	2	-0.191	0.008	4.8	0.028	Reject	-0.362	-0.020	0.341

where L(CI): length of CI, Ll: lower limit, and UL: upper limit.

Table 10. Suggested procedure (GSM): Predicted probabilities and 95% CI according to link function.

Link	i	Estimation of p_i^*		95% CI	
		\hat{p}_i^*	LL	UL	L(CI)
logit	1	0.4600	0.4084	0.5125	0.1041
	2	0.5257	0.5019	0.5494	0.0475
probit	1	0.4600	0.4083	0.5124	0.1041
	2	0.5257	0.5019	0.5494	0.0475
cloglog	1	0.4600	0.4094	0.5137	0.1043
	2	0.5257	0.5022	0.5497	0.0475

where L(CI): length of CI, Ll: lower limit, and UL: upper limit.

As expected by Equations (8) and (14), the estimates of β_i^* for each of the binomial models adjusted by using GUM and GSM are identical. However, the estimated variances are different in both procedures: those estimated by the GSM procedure are smaller. Therefore, it is preferable to those estimated by the GUM procedure. Of course, these decreases in the variances imply the reduction in the lengths of the confidence intervals, as it can be seen when comparing the respective tables. This applies both to the estimators of the linear predictor and to the second component of the predicted probability vector, regardless of the binomial model implemented.

Additionally, when the aggregate model data is adjusted by appealing to the suggested procedure, there is a change in the conclusions about the significance of β_1^* for the logit and probit models. Thus, estimates of the predicted probabilities in Table 10 are now statistically valid and better approximate the available data. The latter also applies to the cloglog model which, even though their estimates are considered valid when obtained by the usual GUM procedure, improve when the GSM is used.

7. CONCLUSIONS

This work constitutes a generalization of the method proposed by Ponsot et al. (2009) for fitting binomial logit models, in the situation of factor levels aggregation on a simple contingency table (with a factor and a dichotomous response variable). This generalization consists of an extension, both of the usual procedure and of the procedure suggested by the author, for the adjustment of binomial models by means of link functions not only logit, but also probit and cloglog.

The results showed that the problem of factor levels aggregation persists in models that are usually competitors or alternatives to the logit model. That is because, regardless of the link function used, the violation of the binomial assumption remains when the associated probabilities of success at aggregated levels are dissimilar. Then, the suggested procedure maintains its advantages with any of the link functions used, being it preferable to the usual adjustment procedure, as it offers the necessary correction and subsequent improvement of the results.

The link function that is selected does not favor the application of a particular method between the two presented. However, it was confirmed that, in any scenario, that is, when any link function is appealed to derive a binomial model, the estimates obtained by the suggested method improve when the correct distribution assumption is used. The choice between one method or another is based mainly on the probabilities of success associated with the aggregate levels. If they present slight differences, the estimates of the usual method are not very different from those of the suggested method. On the contrary, as these differences grow, it is better to rely on the suggested method.

In the future, it is expected to apply the comparisons made for the saturated model, using different link functions, to the general case of the aggregation of factor levels in the unsaturated model, as proposed by Ponsot et al. (2012). It is also expected to apply re-sampling techniques to study the behavior of the standard errors of the estimators in this more general case. Extensions to the so-called tobit model are also of interest (Barros et al., 2018; Desousa et al., 2018). Finally, the applications of the methodology is explored in the sense proposed by Da Silva et al. (2016), that is, for multinomial-ordinal models of a longitudinal nature.

ACKNOWLEDGEMENTS

The authors would like to thank the Editors, Associate Editor and Referees for their suggestions and comments that led to a significant improvement of this manuscript.

APPENDIX A. EXTREME VALUES OF $V_{\text{Bin}}[Y_{k-1}^*]$

Let $V_{\text{Bin}}[Y_{k-1}^*] = n_{k-1}^* p_{k-1}^* (1 - p_{k-1}^*)$, where $n_{k-1}^* = n_{k-1} + n_k$ and $p_{k-1}^* = (n_{k-1} p_{k-1} + n_k p_k) / n_{k-1}^*$. Then, the partial derivatives of $V_{\text{Bin}}[Y_{k-1}^*]$, with respect to p_{k-1} and p_k , are given by

$$\begin{aligned} \frac{\partial V_{\text{Bin}}[Y_{k-1}^*]}{\partial p_{k-1}} &= n_{k-1} \left(1 - \frac{n_{k-1} p_{k-1} + n_k p_k}{n_{k-1}^*} \right) + (n_{k-1} p_{k-1} + n_k p_k) \left(-\frac{n_{k-1}}{n_{k-1}^*} \right) \\ &= \frac{n_{k-1}}{n_{k-1}^*} (n_{k-1}^* - 2n_{k-1} p_{k-1} - 2n_k p_k), \\ \frac{\partial V_{\text{Bin}}[Y_{k-1}^*]}{\partial p_k} &= n_k \left(1 - \frac{n_{k-1} p_{k-1} + n_k p_k}{n_{k-1}^*} \right) + (n_{k-1} p_{k-1} + n_k p_k) \left(-\frac{n_k}{n_{k-1}^*} \right) \\ &= \frac{n_k}{n_{k-1}^*} (n_{k-1}^* - 2n_{k-1} p_{k-1} - 2n_k p_k). \end{aligned}$$

Equating to zero, the critical points are all those ordered pairs of the form

$$(p_{k-1}, p_k) = \left(p_{k-1}, 0.5 \left[1 + \frac{n_{k-1}}{n_k} (1 - 2p_{k-1}) \right] \right).$$

The second partial derivatives of $V_{\text{Bin}}[Y_{k-1}^*]$ are stated as

$$\begin{aligned} A &= \frac{\partial^2 V_{\text{Bin}}[Y_{k-1}^*]}{\partial p_{k-1}^2} = \frac{n_{k-1}}{n_{k-1}^*} (-2n_{k-1}) = -\frac{2n_{k-1}^2}{n_{k-1}^*} \\ B &= \frac{\partial^2 V_{\text{Bin}}[Y_{k-1}^*]}{\partial p_k \partial p_{k-1}} = \frac{n_{k-1}}{n_{k-1}^*} (-2n_k) = -\frac{2n_{k-1} n_k}{n_{k-1}^*} \\ C &= \frac{\partial^2 V_{\text{Bin}}[Y_{k-1}^*]}{\partial p_k^2} = \frac{n_k}{n_{k-1}^*} (-2n_k) = -\frac{2n_k^2}{n_{k-1}^*} \end{aligned}$$

while the Hessian is expressed by

$$H = AC - B^2 = \left(-\frac{2n_{k-1}^2}{n_{k-1}^*} \right) \left(-\frac{2n_k^2}{n_{k-1}^*} \right) - \left(-\frac{2n_{k-1} n_k}{n_{k-1}^*} \right)^2 = 0$$

Thus, that the Hessian criterion fails in the decision. Nevertheless, it is clear from the examination of Figure 1 that the set of points found are relative maximum values.

APPENDIX B. EXTREME VALUES OF $V[Y_{k-1}^*]$

Let $V[Y_{k-1}^*] = n_{k-1} p_{k-1} (1 - p_{k-1}) + n_k p_k (1 - p_k)$. Then, the partial derivatives of $V[Y_{k-1}^*]$, with respect to p_{k-1} and p_k , are given by

$$\begin{aligned} \frac{\partial V[Y_{k-1}^*]}{\partial p_{k-1}} &= n_{k-1} (1 - p_{k-1}) - n_{k-1} p_{k-1} = n_{k-1} (1 - 2p_{k-1}) \\ \frac{\partial V[Y_{k-1}^*]}{\partial p_k} &= n_k (1 - p_k) - n_k p_k = n_k (1 - 2p_k) \end{aligned}$$

which, when equated to zero, throw as a critical point $(p_{k-1}, p_k) = (0.5, 0.5)$.

As second partial derivatives of $V[Y_{k-1}^*]$, we have

$$A = \frac{\partial^2 V[Y_{k-1}^*]}{\partial p_{k-1}^2} = n_{k-1}(-2) = -2n_{k-1}$$

$$B = \frac{\partial^2 V[Y_{k-1}^*]}{\partial p_k \partial p_{k-1}} = 0$$

$$C = \frac{\partial^2 V[Y_{k-1}^*]}{\partial p_k^2} = n_k(-2) = -2n_k.$$

Consequently, the Hessian is given by $H = AC - B^2 = (-2n_{k-1})(-2n_k) - (0)^2 = 4n_{k-1}n_k$. Due to $H > 0$ and $A < 0$, then at the critical point $(p_{k-1}, p_k) = (0.5, 0.5)$, there is a relative maximum whose value is $V[Y_{k-1}^*] = (n_{k-1} + n_k)/4 = n_{k-1}^*/4$.

REFERENCES

- Agresti, A., 2007. *An Introduction to Categorical Data Analysis*. Wiley, NJ, US.
- Agresti, A., 2015. *Foundations of Linear and Generalized Linear Models*. Wiley, New York, US.
- Barros, M., Galea, M., Leiva, V., and Santos-Neto, M., 2018. Generalized tobit models: Diagnostics and application in econometrics. *Journal of Applied Statistics*, 45, 145–167.
- Bonat, W.H., Lopes, J.E., Shimakura, S.E., and Ribeiro Jr, P.J., 2018. Likelihood analysis for a class of simplex mixed models. *Chilean Journal of Statistics (ChJS)*, 9(2), 3–17.
- Christensen, R., 1997. *Log-Linear Models and Logistic Regression*. Springer, NY, US.
- Collet, D., 2002. *Modelling Binary Data*. Chapman and Hall, New York, US
- Czado, C. and Munk, A., 2000. Noncanonical links in generalized linear models -when is the effort justified? *Journal of Statistical Planning and Inference*, 87, 317–345.
- Czado, C. and Santner, T.J., 1992. The effect of link misspecification on binary regression inference. *Journal of Statistical Planning and Inference*, 33, 213–231.
- Da Silva, N.B., Amorim, L.D., Colosimo, E.A., and Heller, L., 2016. Modeling ordinal longitudinal outcomes: an applied perspective of marginal and conditional approaches. *Chilean Journal of Statistics*, 7(2), 51–68.
- Desousa, M., Saulo, H., Leiva, V., and Scalco, P., 2018. On a tobit-Birnbaum-Saunders model with an application to medical data. *Journal of Applied Statistics*, 45, 932–955.
- Dobson, A., 2002. *An Introduction to Generalized Linear Models*. Chapman and Hall, New York, US
- Hilbe, J.M., 2009. *Logistic Regression Models*. Chapman and Hall, New York, US
- Hosmer, D., and Lemeshow, S., 2000. *Applied Logistic Regression*. Wiley, New York, US.
- McCullagh, P., and Nelder, J., 1989. *Generalized Linear Models*. Chapman and Hall, New York, US.
- McCulloch, C., and Searle, S., 2001. *Generalized Linear, and Mixed Models*. Wiley, New York, US.
- Nelder, J.A., and Wedderburn, R.W.M., 1972. Generalized linear models. *Journal of the Royal Statistical Society A*, 135, 370–384.
- Ponsot, E., 2011. Estudio de la agregación de niveles de los factores en el modelo logit binomial. Tesis de Doctorado, Universidad de Los Andes, Mérida, Venezuela.
- Ponsot, E., Sinha, S., and Goitía, A., 2009. Sobre la agrupación de niveles del factor explicativo en el modelo logit binario. *Revista Colombiana de Estadística*, 32, 157–187.

- Ponsot, E., Sinha, S., and Goitía, A., 2012. Aggregation of explanatory factor levels in a binomial logit model: Generalization to the multifactorial unsaturated case. *Revista Colombiana de Estadística*, 35, 139–166.
- Tutz, G., 2011. *Regression for Categorical Data*. Cambridge University Press, Cambridge, UK.

BAYESIAN MODELING
RESEARCH PAPER

Bayesian analysis of an item response model with an AEP-based link function

FRANCISCO J. ARIZA-HERNANDEZ¹ and EDUARDO GUTIÉRREZ-PEÑA²

¹Facultad de Matemáticas, Universidad Autónoma de Guerrero, UAGro, Mexico,

²Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, IIMAS, UNAM, Mexico,

(Received: 17 February 2021 · Accepted in final form: 26 April 2021)

Abstract

We consider a robust Bayesian approach to the analysis of item response models, using the inverse of an asymmetric exponential power cumulative distribution function as a link function. This provides greater flexibility with respect to classic link functions such as the probit and the logit. We conduct a simulation study to evaluate the performance of our model. In order to draw samples from the posterior distribution of the parameters, we implement a Markov chain Monte Carlo scheme by means of the JAGS software. We also implement a posterior predictive model-checking method to assess the fit and relative performance of the various submodels. Finally, we provide a real-data example to illustrate the modeling approach proposed.

Keywords: Asymmetric exponential power (AEP) distribution · Generalized linear model · JAGS and R software · Rasch model · Sample-based inference.

Mathematics Subject Classification: Primary 62F15 · Secondary 62J12.

1. INTRODUCTION

Item response data come from applying a test to a set of individuals. The test is composed of a number of items. These tests are used extensively in schools, industry, and government, and for various purposes (see [Baker and Kim, 2004](#); [van der Linden and Hambleton, 1997](#); [Fox, 2010](#)). There is a very extensive literature about of the item response models, its development, description, and applications goes back to [Lord \(1952, 1980\)](#), who established the basis of item response theory (IRT), also called modern test theory.

Traditionally, frequentist analyses have been used in IRT. Recently, however, the Bayesian approach become very attractive for modeling item response data; (see [Ghosh et al., 2000](#); [Béguin and Glas, 2001](#); [Bazán et al., 2006](#); [Fox, 2010](#); [Azevedo et al., 2011, 2012](#); [Matteucci et al., 2012](#)). This approach allows one to incorporate additional information to the analysis and provides powerful estimation methods based on simulated samples from posterior distributions.

Although item response modeling can be employed in more general contexts (see [Reckase, 2009](#); [Fu et al., 2009](#); [Svetina, 2013](#); [Bacci et al., 2014](#)), and nonparametric settings (see [Karabatsos, 2016](#); [San Martin et al., 2011](#)). In this paper, we focus on dichotomous response

*Corresponding author. Email: arizahfj@uagro.mx

data and unidimensional models with a continuous latent trait. The Rasch model (see [Rasch, 1961](#)) is by far the most popular model in this latter case. It is basically a logistic model, so the probit model is commonly used as an alternative.

Both the logit and probit link functions (corresponding to the standard logistic and the standard normal distributions, respectively) have traditionally been utilized when modeling dichotomous response data. Both of these link functions are symmetric. However, when the proportion of ones in the observed sample is very different from the proportion of zeros, or vice versa, the symmetric links commonly used may not be appropriate, as they may lead to misspecified models (see [Chen et al., 1999a](#)). This situation is not uncommon with item response experimental data.

The fit of item response models can be improved significantly by using asymmetric links. Several authors have worked with asymmetric links. [Chen et al. \(1999a\)](#) proposed a class models with skewed link to analyze binary data with covariates, [Jiang et al. \(2013\)](#) derived a new class of symmetric power link functions to model binary data and applied it to the Protea co-occurrence data. More recently, [Durante \(2019\)](#) proved that in the case of probit regression models which have Gaussian priors for the coefficients, the posterior belongs to the class of unified skew-normal distributions. Also, [Naranjo et al. \(2015\)](#) employed an asymmetric exponential power (AEP) distribution for the error of a linear regression model, and the inverse of the AEP cumulative distribution function (CDF) as a link function in a regression model for binary data, but not in the context of IRT.

Models with asymmetrical link functions have also been proposed in IRT settings. [Samejima \(2000\)](#) proposed a family of models called the logistic positive exponent family, which provides asymmetric item characteristic curves (ICCs). [Bazán et al. \(2006\)](#) introduced a skew-probit IRT link function based on the skew normal distribution, while [Azevedo et al. \(2011\)](#) used skew-normal distributions to model latent traits in an IRT two-parameter probit model under centered parameterizations. However, these models are not as flexible as the AEP distribution, which not only allows one to handle symmetry/asymmetry but also light/heavy tails.

In this paper, we build on the work of [Zhu and Zinde-Walsh \(2009\)](#) and [Naranjo et al. \(2015\)](#) to propose a Bayesian item response model based on the AEP distribution.

The outline of the paper is as follows. In [Section 2](#), we briefly discuss the Rasch model and review the probability density and cumulative distribution functions of the AEP distribution. We describe the general model in [Section 3](#). Then, in [Section 4](#), we carry out Bayesian inferences on the parameters of interest via the just another Gibbs sampler (JAGS) software (see [Plummer, 2017](#)) within the R software (see [R Core Team, 2020](#)), and apply a posterior predictive model-checking method (see [Sinharay et al., 2006](#)) with the purpose of comparing various submodels. In [Section 5](#), we present a simulation study and conducted to assess the performance of the Bayesian estimates. Also, a real-data example is given in this section to illustrate the AEP-based IRT model. Finally, [Section 6](#) contains some concluding remarks.

2. PRELIMINARIES

2.1 THE RASCH MODEL

We model the probability of the correct answer, p_{ik} , corresponding to i -th individual in the k -th item, as $p_{ik} = P(Y_{ik} = 1 | \theta_i, a_k, b_k) = F(a_k \theta_i - b_k)$, for $i = 1, \dots, N$ and $k = 1, \dots, K$, where Y_{ik} is a random variable which takes the value of 1 if the i -th individual responds correctly to the k -th item and F is the CDF of a known parametric family. In the context of IRT, F is the ICC, $a_k > 0$ and $b_k \in \mathbb{R}$ are item parameters (called discrimination and difficulty parameters, respectively), and $\theta_i \in \mathbb{R}$ is the person parameter associated with the ability of individual i . The inverse of F is called the link function. The Rasch model is the

simplest and most traditional model for p_{ik} . It is given by

$$p_{ik} = P(Y_{ik} = 1 | \theta_i, b_k) = \frac{\exp(\theta_i - b_k)}{1 + \exp(\theta_i - b_k)}.$$

That is, the probability that the person i obtains a correct response to item k is a logistic function of the difference between the person's ability, θ_i , and difficulty of the item, b_k . Note that, if the person's ability is greater than the difficulty of the item, then the probability of success is higher in comparison with the probability of failure. A limitation of the Rasch model is that all items are assumed to discriminate between respondents in the same way (that is, $a_k = 1$ for all $k = 1, \dots, K$); as a result, items only differ in item difficulty (see [Fox, 2010](#)). The probit model is another popular model for p_{ik} ; it takes $p_{ik} = \Phi(\theta_i - b_k)$, where Φ is the standard normal CDF. The Rasch model can be approximated by a probit model by multiplying the parameter values by a scaling factor of 1.7.

2.2 THE AEP DISTRIBUTION

The probability density function (PDF) of the rescaled AEP distribution, proposed by [Zhu and Zinde-Walsh \(2009\)](#), is stated as

$$\tilde{f}_{\text{AEP}}(x | \mu, \tilde{\sigma}, \alpha, \delta_1, \delta_2) = \begin{cases} \frac{1}{\tilde{\sigma}} \exp \left\{ - \left| \frac{x - \mu}{\alpha \tilde{\sigma} / \Gamma(1 + 1/\delta_1)} \right|^{\delta_1} \right\}, & \text{if } x \leq \mu; \\ \frac{1}{\tilde{\sigma}} \exp \left\{ - \left| \frac{x - \mu}{(1 - \alpha) \tilde{\sigma} / \Gamma(1 + 1/\delta_2)} \right|^{\delta_2} \right\}, & \text{if } x > \mu; \end{cases} \quad (1)$$

where $\mu \in \mathbb{R}$ is the location parameter, $\tilde{\sigma} > 0$ is the scale parameter, $\alpha \in (0, 1)$ is the skewness parameter, and δ_1 and δ_2 are the left- and right-tail parameters, respectively ($\delta_1 > 0, \delta_2 > 0$). For convenience, we consider a reparametrization of the scale parameter in Equation (1), $\tilde{\sigma} = \sqrt{2\pi}\sigma$, so that

$$f_{\text{AEP}}(x | \mu, \sigma, \alpha, \delta_1, \delta_2) = \tilde{f}_{\text{AEP}}(x | \mu, \sqrt{2\pi}\sigma, \alpha, \delta_1, \delta_2).$$

With this parametrization, the density function of the AEP distribution is given by

$$f_{\text{AEP}}(x | \mu, \sigma, \alpha, \delta_1, \delta_2) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ - \left| \frac{x - \mu}{\sqrt{2\pi}\alpha\sigma / \Gamma(1 + 1/\delta_1)} \right|^{\delta_1} \right\}, & \text{if } x \leq \mu; \\ \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ - \left| \frac{x - \mu}{\sqrt{2\pi}(1 - \alpha)\sigma / \Gamma(1 + 1/\delta_2)} \right|^{\delta_2} \right\}, & \text{if } x > \mu. \end{cases} \quad (2)$$

We use $X \sim \text{AEP}(\mu, \sigma, \alpha, \delta_1, \delta_2)$ to denote Equation (2). Important properties of the AEP distribution have been discussed in the literature (see [Zhu and Zinde-Walsh, 2009](#); [Naranjo et al., 2015](#)). If $\alpha = 1/2$ and $\delta_1 = \delta_2$, the distribution is symmetric. An important special case is when $\delta_1 = \delta_2 = 2$ and $\alpha = 1/2$, in which case Equation (2) is the $N(\mu, \sigma^2)$ distribution. In [Figure 1](#), we show the PDF and CDF of the AEP distribution in Equation (2) for a range of parameter values.

For the standard version of the AEP distribution ($\mu = 0$, $\sigma = 1$), the CDF can be expressed as

$$F_{\text{AEP}}(x|\alpha, \delta_1, \delta_2) = \begin{cases} \alpha \left[1 - G \left(\left(\frac{x}{\sqrt{2\pi}\alpha/\Gamma(1+1/\delta_1)} \right)^{\delta_1}; \frac{1}{\delta_1}, 1 \right) \right], & \text{if } x \leq 0; \\ \alpha + (1-\alpha)G \left(\left(\frac{x}{\sqrt{2\pi}(1-\alpha)/\Gamma(1+1/\delta_2)} \right)^{\delta_2}; \frac{1}{\delta_2}, 1 \right), & \text{if } x > 0; \end{cases} \quad (3)$$

where $G(v; \gamma, \beta)$ is the gamma CDF given by

$$G(v; \gamma, \beta) = \frac{1}{\Gamma(\gamma)\beta^\gamma} \int_0^v t^{\gamma-1} \exp\{-t/\beta\} dt.$$

The proof of Equation (3) is given in Appendix A.

3. AN AEP-BASED GENERALIZED LINEAR MODEL FOR BINARY DATA

3.1 MODEL SPECIFICATION

The IRT model based on the AEP distribution is formally defined as follows. Let Y_{ik} be a random variable representing the response of the i -th individual to the k -th item. This response variable is discrete, taking only two possible values. We define $Y_{ik} = 1$ if the i -th individual's response to the k -th item is correct and $Y_{ik} = 0$ for an incorrect response. Then, we have

$$Y_{ik}|\theta_i, a_k, b_k, \alpha_k, \delta_{1k}, \delta_{2k} \sim \text{Bern}(p_{ik}), \quad (4)$$

where $\text{Bern}(p_{ik})$ denotes the Bernoulli distribution, for $i = 1, \dots, N$, $k = 1, \dots, K$, and p_{ik} is given by

$$\begin{aligned} p_{ik} &= P(Y_{ik} = 1|\theta_i, a_k, b_k, \alpha_k, \delta_{1k}, \delta_{2k}) \\ &= F_{\text{AEP}}(a_k\theta_i - b_k | \alpha_k, \delta_{1k}, \delta_{2k}). \end{aligned} \quad (5)$$

This represents the conditional probability that the i -th individual, with ability θ_i , responds correctly to the k -th item with discrimination parameter a_k and difficulty parameter b_k . The quantities α_k , δ_{1k} and δ_{2k} are the AEP parameters defined in Equation (2). This model assumes that a change in the probability of a specified response is described by the ICC in Equation (5), and that the responses to a pairs of items are statistically independent given the latent variable θ . The probability of success is modeled as a function of person, item and AEP parameters. Note that, for $\alpha = 0.5$ and $\delta_1 = \delta_2 = 2$, the Equation (5) reduces to the probit model.

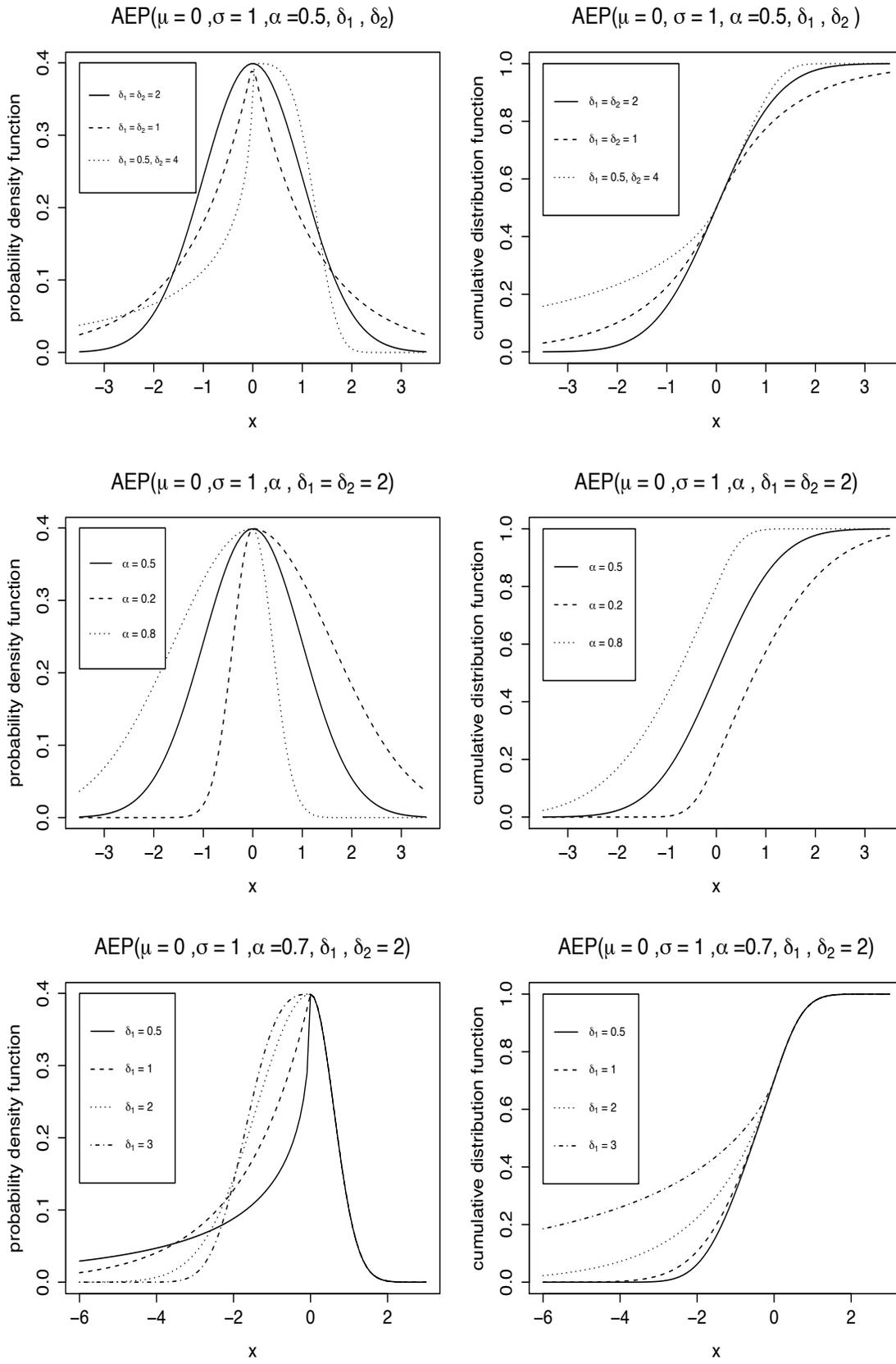


Figure 1. PDFs and CDFs of the AEP distribution for different values of the parameters.

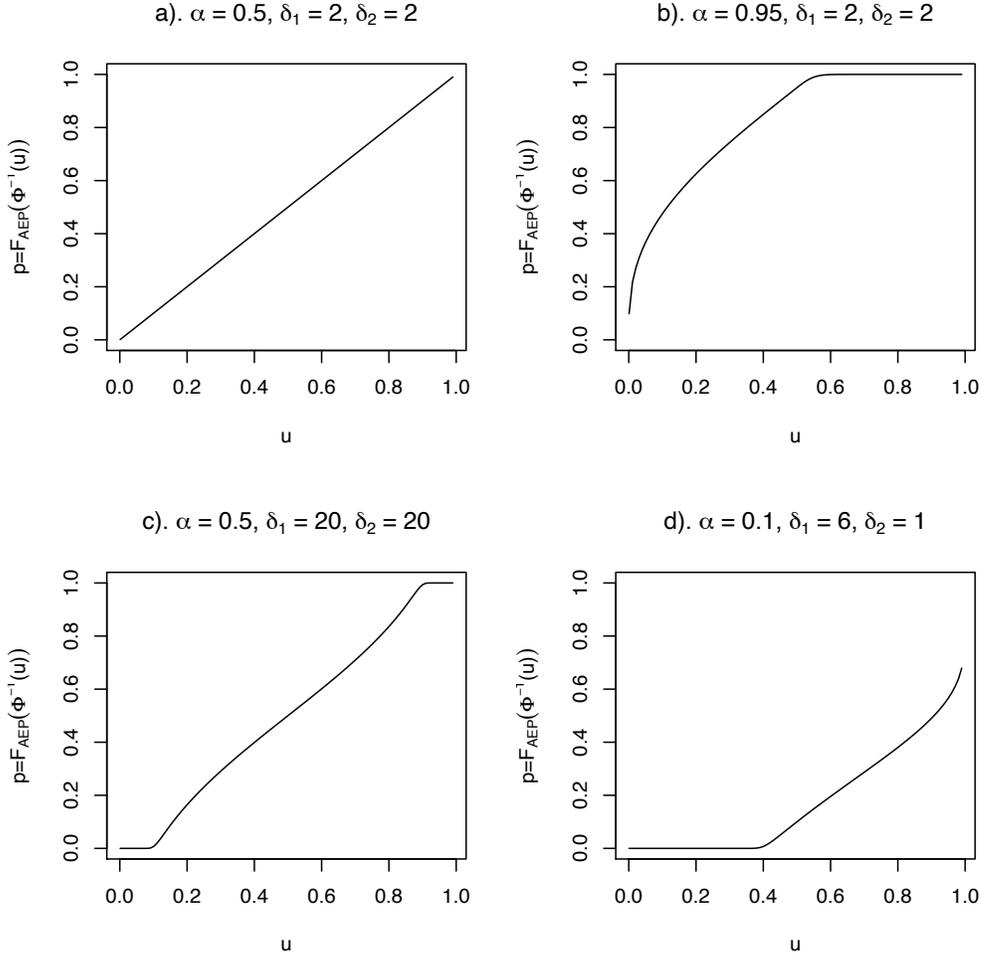


Figure 2. Comparison of AEP-based and probit link functions.

3.2 LIKELIHOOD FUNCTION

Let $\mathbf{y} = (y_{11}, \dots, y_{NK})^\top$ denote the observed item response data. Then, the likelihood function for the AEP-IRT model is stated as

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta}; \mathbf{y}) &= \prod_{i=1}^N \prod_{k=1}^K p_{ik}^{y_{ik}} (1 - p_{ik})^{1-y_{ik}} \\ &= \prod_{i=1}^N \prod_{k=1}^K [F_{\text{AEP}}(m_{ik} | \boldsymbol{\eta})]^{y_{ik}} [1 - F_{\text{AEP}}(m_{ik} | \boldsymbol{\eta})]^{1-y_{ik}}, \end{aligned}$$

where $m_{ik} = a_k \theta_i - b_k$, $i = 1, \dots, N$, $k = 1, \dots, K$; $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$, $\boldsymbol{\xi} = (\mathbf{a}, \mathbf{b})$ and $\boldsymbol{\eta} = (\boldsymbol{\alpha}, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$, with $\mathbf{a} = (a_1, \dots, a_K)$, $\mathbf{b} = (b_1, \dots, b_K)$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$, $\boldsymbol{\delta}_1 = (\delta_{1k}, \dots, \delta_{1k})$ and $\boldsymbol{\delta}_2 = (\delta_{2k}, \dots, \delta_{2k})$.

Note that the model proposed here is described in terms of N ability parameters, K discrimination parameters, K difficulty parameters, K skewness parameters and K pairs of tail parameters. Hence, it has a total of N person parameters and $5K$ unknown parameters. This model is overparameterized. In fact, for two different sets of parameter values the model may give the same success probabilities and so the model may be unidentifiable. For example, the linear predictor in Equation (5) can be written as $a_k \theta_i - b_k = a_k (10\theta_i -$

$50)/10 - (b_k - 50a_k/10) = a_k^*\theta_i^* - b_k^*$; that is, the model with a_k, b_k, θ_i is the same as with a_k^*, b_k^*, θ_i^* . Thus, the parameters cannot be uniquely estimated, unless certain constraints are imposed. From the Bayesian viewpoint, this problem may be solved by specifying suitable priors for the parameters of interest (see [Chen et al., 2003](#); [Matteucci et al., 2012](#); [Naranjo et al., 2015](#)).

As pointed out in Section 1, when the proportion of ones in the observed sample is very different from the proportion of zeros, or vice versa, the symmetric links commonly used may not be appropriate. To visualize the flexibility of the AEP-based link function with respect to the probit link function, in Figure 2 we plot $F_{\text{AEP}}(\Phi^{-1}(u)|\alpha, \delta_1, \delta_2)$ over the interval $(0, 1)$ for selected values of α, δ_1 and δ_2 .

4. BAYESIAN INFERENCE

4.1 PRIOR DISTRIBUTION

In this paper we use a Bayesian approach to make statistical inference about the parameters of interest. In this setting, the parameters are regarded as random variables and have prior distributions that reflect the uncertainty about their true values before observing the data. Several authors have suggested informative as well as noninformative prior distributions for the item parameters; for example, lognormal priors for the discrimination parameters and a normal prior for difficulty parameters (see [Albert, 1992](#); [Patz and Junker, 1999](#); [Rupp et al., 2004](#); [Fox and Glas, 2001](#); [Matteucci et al., 2012](#); [Bazán et al., 2006](#)). [Ghosh et al. \(2000\)](#) pointed out that, with noninformative priors, posterior distributions for item and person parameters may be improper when the sum of the binary responses for an item or person takes its minimum or maximum possible value. However, they prove that under certain conditions the joint posterior distribution is proper.

Here, we assume the item parameters to be exchangeable. We also assume monotonicity of the ICC, which is satisfied when the discrimination parameter is restricted to be positive. Thus, we assume the following prior distribution for the item parameters

$$(a_k, b_k) \sim N(\boldsymbol{\mu}_\xi, \boldsymbol{\Sigma}_\xi) I_A(a_k),$$

where $A = \{a \in \mathbb{R} : a > 0\}$, $k = 1, \dots, K$, and I_A is the indicator function of the set A . Note that this prior is not conjugate for the observed likelihood. A typical prior for person parameters assumes that the individual are chosen randomly from an unknown population, where each individual has the same probability of being chosen. Individuals are also assumed to be sampled independently, so we assume that

$$\theta_i \sim N(\mu_\theta, \sigma_\theta), \quad i = 1, \dots, N,$$

where $\mu_\theta, \sigma_\theta$ are known parameters. These priors have been suggested by others authors (see [Bazán et al., 2006](#); [Sinharay et al., 2006](#); [Fox, 2010](#); [Matteucci et al., 2012](#)). In generalized linear models, some authors have proposed an elicitation scheme for a class of informative prior distributions for the regression parameters based on historical data (see [Chen et al., 1999b, 2003](#)). [Naranjo et al. \(2015\)](#) proposed some alternative prior distributions for the AEP parameters, which have the advantage of allowing one to derive the full conditional distributions required for a Gibbs sampler. With some adjustments, the Jeffreys prior distribution can be computed from the Fisher information matrix given by [Zhu and Zinde-Walsh \(2009\)](#).

Assuming prior independence of the parameters, we can write the joint prior distribution as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta}) &= p(\boldsymbol{\theta}) p(\boldsymbol{\xi}) p(\boldsymbol{\eta}) \\ &= \prod_{i=1}^N p(\theta_i) \left\{ \prod_{k=1}^K p(a_k) p(b_k) p(\alpha_k) p(\delta_{1k}) p(\delta_{2k}) \right\}. \end{aligned}$$

4.2 POSTERIOR SAMPLING

By the Bayes theorem, the posterior distribution of the parameters of interest is established as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta} | \mathbf{y}) &= L(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta}; \mathbf{y}) p(\boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta}) / p(\mathbf{y}) \\ &\propto \prod_{i=1}^N \prod_{k=1}^K \left\{ [F_{\text{AEP}}(m_{ik} | \boldsymbol{\eta})]^{y_{ik}} [1 - F_{\text{AEP}}(m_{ik} | \boldsymbol{\eta})]^{1-y_{ik}} \right. \\ &\quad \times p(\theta_i) p(a_k) p(b_k) p(\alpha_k) p(\delta_{1k}) p(\delta_{2k}) \left. \right\}. \end{aligned} \quad (6)$$

Note that the joint posterior distribution is analytically intractable and thus obtaining the marginal posterior densities of the parameters is not an easy task; however, samples from Equation (6) can be obtained using Markov chain Monte Carlo (MCMC) techniques. The most common MCMC methods are the Gibbs sampling (see [Gelfand and Smith, 1990](#); [Casella and George, 1992](#)) and the Metropolis-Hastings (see [Metropolis et al., 1953](#); [Chib and Greenberg, 1995](#)). Currently, many of the MCMC algorithms have been already implemented in computer programs, such as, `WinBUGS` (see [Spiegelhalter et al., 2003](#)), `JAGS` (see [Plummer, 2017](#)) and `Stan` (see [Stan Development Team, 2014](#)). All of these software packages provide programs for Bayesian modeling through posterior simulation given a specified model and data. In particular, `JAGS` provides several samplers and attempts to use the most efficient one to update the parameters of the model at each iteration. The R packages named `R2WinBUGS`, `R2jags` and `rstan` allow one to run `WinBUGS`, `JAGS` and `Stan` from within R, respectively. There are several R packages for IRT. [Choi and Asilkalkan \(2019\)](#) presents a summary of the IRT package that have been developed over the last decade. In this paper, we utilize `JAGS` within R to obtain samples from the posterior distributions of interest (see [Appendix B](#)).

4.3 A POSTERIOR PREDICTIVE MODEL-CHECKING METHOD

The posterior predictive model-checking (PPMC) method is a popular Bayesian model-checking tool, has a strong theoretical basis, and can provide graphical or numerical summaries about the model fit (or lack thereof). For IRT models, [Sinharay et al. \(2006\)](#) presented an extensive explanation of the PPMC method and discuss different discrepancy measures to detect various violations to model assumptions. [Azevedo et al. \(2012\)](#) developed Bayesian methods for the multiple-group IRT model, including an estimation method based on MCMC and different posterior predictive assessment tools. The idea of PPMC is to generate replicate data sets by simulating from the posterior predictive distribution, and then compare these simulated samples with the observed data. If the replicated data and the observed data differ systematically, it is an indication of a potential model misfit.

The choice of discrepancy measure is crucial in the application of the PPMC method. In this paper, we used the Observed Score Distribution (OSD) as the discrepancy measure,

which has been employed by [Béguin and Glas \(2001\)](#). This discrepancy measure is given by

$$OSD = \sum_k \frac{[NC_k - E(NC_k)]^2}{E(NC_k)}, \tag{7}$$

where NC_k denotes the number of examinees getting exactly k correct items, and $E(NC_k)$ is the expected value of NC_k under the model, for $k = 0, 1, \dots, K$.

In addition, here we propose an alternative discrepancy measure based on the Kullback-Leibler divergence between the “true” model and an “approximate” model, stated as

$$D_{KL}(\pi||\tilde{\pi}) = \sum_k \pi_k \log \left(\frac{\pi_k}{\tilde{\pi}_k} \right), \tag{8}$$

where $\pi_k = E(NC_k)/N$ and $\tilde{\pi}_k = NC_k/N$.

In order to assess the fit of the IRT model to a given data set, we can repeat the following steps a large number of times:

- (1) Generate a draw of the parameters of interest from the posterior distribution given by Equation (6).
- (2) Obtain a data set from the model given in Equations (4)-(5), using the parameters drawn in the previous step.
- (3) Compute the values of the predictive and realized discrepancy measures given in Equations (7) or (8), utilizing the data set drawn in the previous step.

With the predictive and realized discrepancy measures, we can create plots to assess the fit of the IRT model.

5. NUMERICAL APPLICATIONS

5.1 SIMULATION STUDY

We carried out a simulation study to assess the performance of the Bayesian estimators of the parameters of interest. The procedure was applied to each of several combinations of data-generating and fitted models. Table 1 shows the cases we considered, that is:

- The AEP-III model is based on Equations (4)–(5). This model can describe both symmetry/asymmetry and light/heavy tails separately for each item.
- In the AEP-II model, the tails of the AEP distribution for each item are described by means of the parameters δ_{1k} and δ_{2k} , while α_k is held fixed at $\alpha_k = 0.5$.
- In the AEP-I model, the symmetry/asymmetry for each item is formulated by means of the skewness parameter α_k , while δ_{1k} and δ_{2k} are held fixed at $\delta_{1k} = \delta_{2k} = 2$.

Note that the AEP-I and AEP-II models are both particular cases of the AEP-III model.

Table 1. Cases examined in the simulation study.

Fitted model	Data-generating model			
	AEP-I	AEP-II	AEP-III	Probit
AEP-I	•			•
AEP-II		•		•
AEP-III			•	•
Probit	•	•	•	•

We now describe the simulation study:

- (1) We simulated $B = 100$ data sets from each data-generating model (see below for details).
- (2) For each simulated sample, we obtained Bayesian estimators of the parameters of interest, both for model given in Equations (4)–(5) and for the two-parameter probit model.
 - a) We calculated the Bayes estimators as the sample mean from Equation (6) using JAGS within R. We employed two chains, each with 26,000 iterations, with a burn-in of 1000 iterations and a thinning rate of 50, so we kept a total of 500 iterations to make inferences about the parameters of interest. The analysis of each sample took around 1.74 minutes on a computer with a 4 GHz Intel Core i7 processor and 32GB of RAM.
 - b) We calculated 95% credible intervals for each parameter; these intervals are based on the 2.5-th quantile and the 97.5-th quantile of the corresponding posterior sample.
- (3) From these B samples, we computed the mean squared error (MSE) of the estimators as

$$\widehat{\text{MSE}} = s^2(\theta) + \widehat{B}^2(\theta),$$

where $s^2(\theta) = \sum_{i=1}^B (\theta_i^B - \bar{\theta}_B)^2 / (B - 1)$ is the sample variance of the Bayes estimators, $\widehat{B}(\theta) = \bar{\theta}_B - \theta$ is the bias, $\bar{\theta}_B = \sum_{i=1}^B \theta_i^B / B$, and θ_i^B is the Bayes estimator corresponding to the i -th sample.

- (4) Finally, we computed the coverage of the corresponding credible intervals.

To perform this study, we used R together with the R2jags package (see Su and Yajima, 2020). Our simulated data sets consist of $N = 100$ individuals and $K = 3, 5, 10, 20$ items. The true values of the parameters utilized to generate the data sets were varying according to the Table 2.

Table 2. Parameter values for the simulation study.

Parameters	from	to
a_k	0.5	2.0
b_k	-2.0	2.0
α_k	0.1	0.9
δ_{1k}	0.5	4.0
δ_{2k}	0.5	4.0

We assumed the following priors for the discrimination and difficulty parameters:

$$a_k \sim N(1, 1)I(a_k > 0); \quad b_k \sim N(0, 1), \quad k = 1, \dots, K,$$

while, for the AEP parameters, we took the priors used by Naranjo et al. (2015), namely,

$$\alpha_k \sim \text{Beta}(1, 1); \quad \delta_{1k} \sim \text{Gamma}(1, 1) \quad \text{and} \quad \delta_{2k} \sim \text{Gamma}(1, 1), \quad k = 1, \dots, K.$$

In Tables 3 and 4, we show the estimated MSE and coverage for the cases considered in Table 1: AEP-I versus Probit, AEP-II versus Probit, AEP-III versus Probit, and Probit versus AEP- x , ($x = \text{I, II, III}$). Generally speaking, the AEP models outperform the probit model, especially AEP-III and AEP-II.

When we generated simulated data from an AEP- x model, and fitted both the corresponding AEP- x and the probit models, we observed that in all cases the coverage of the credible intervals for the item parameter was close to 100% for the AEP- x models. In contrast, the coverage obtained for the credible intervals of the discrimination and difficulty parameters of the probit model was much lower, even reaching zero in some cases. In general, the estimated MSE and bias are lower for the AEP- x models than for the probit model (see Table 3). Also, when we simulated data from the probit model, and fitted both the AEP- x and the probit models, we observed that both the coverage and the MSE are very similar for all models and close to 100%. As expected, all three AEP models fit the data generated from the probit model reasonably well (Table 4).

5.2 A REAL-DATA EXAMPLE

Next, an example is given to illustrate the Bayesian item response modeling approach proposed in this paper. We consider a data set previously analyzed by Fox (2010), which consists of 200 eighth-grade students that are subjected to a mathematics test with 5 items. The data set contains the responses of the examinees, where 1 indicates a correct answer and 0 an incorrect answer. We assume that the five items measure a unidimensional ability represented by θ , which is a continuous latent variable that takes values on the real line. We estimate the item parameters of both the probit IRT model and the AEP-IRT model using the MCMC methodology described above. This example was also implemented utilizing the JAGS package within R (see Appendix B for details).

The probability of a correct response by examinee i to item k is modeled by the following item response models:

$$\begin{aligned} \text{(Probit)} \quad & P(Y_{ik} = 1 | \theta_i, a_k, b_k) = \Phi(a_k \theta_i - b_k), \\ \text{(AEP-I)} \quad & P(Y_{ik} = 1 | \theta_i, a_k, b_k, \alpha_k) = F_{\text{AEP}}(a_k \theta_i - b_k | \alpha_k), \\ \text{(AEP-II)} \quad & P(Y_{ik} = 1 | \theta_i, a_k, b_k, \delta_{1k}, \delta_{2k}) = F_{\text{AEP}}(a_k \theta_i - b_k | \delta_{1k}, \delta_{2k}), \\ \text{(AEP-III)} \quad & P(Y_{ik} = 1 | \theta_i, a_k, b_k, \alpha_k, \delta_{1k}, \delta_{2k}) = F_{\text{AEP}}(a_k \theta_i - b_k | \alpha_k, \delta_{1k}, \delta_{2k}), \end{aligned}$$

for $i = 1, \dots, 200$ and $k = 1, \dots, 5$, where F_{AEP} is given in Equation (3), and Φ is the standard normal CDF.

The prior distributions used were as follows: for all models, we employed a $N(0, 1)$ distribution for the difficulty parameters, while a truncated normal distribution, $N(1, 1)I(a_k > 0)$, was utilized for the discrimination parameters. These values of the hyperparameters indicate a moderate level of discrimination and average level of difficulty. Assuming that the individuals are sampled independently from the population, we specified a $N(0, 1)$ for the ability parameters of all of the models. This restriction identifies the two-parameter item response model (see Fox, 2010). Finally, for the AEP parameters we took the priors used by Naranjo et al. (2015). That is, $\alpha_k \sim \text{Beta}(1, 1)$, $\delta_{1k} \sim \text{Gamma}(1, 1)$ and $\delta_{2k} \sim \text{Gamma}(1, 1)$, for $k = 1, \dots, 5$.

For each model, we employed two chains, each with 26000 iterations, and the first 1000 were discarded, taking a thinning rate of 50. Thus, 1000 posterior samples were used to obtain the summary statistics about the parameters of interest. Standard convergence diagnostics were carried out. To mention a few, the value of Gelman-Rubin \hat{R} was close to 1 for each parameter of interest and for all the models we considered. Also, the Geweke diagnostics were calculated and showed evidence of convergence.

Table 3. Estimated MSE | Bias | (coverage).

Parameters	AEP-I		Probit		AEP-II		Probit		AEP-III		Probit	
a_1	0.08	-0.005 (1.0)	0.08	0.02 (1.0)	0.12	-0.16 (1.0)	0.17	0.27 (0.70)	0.018	-0.10 (1.0)	0.148	0.36 (1.0)
a_2	0.17	-0.27 (0.96)	0.15	-0.23 (0.96)	0.54	-0.68 (0.92)	0.34	0.27 (0.90)	0.161	-0.33 (1.0)	0.220	-0.33 (1.0)
a_3	0.08	-0.05 (1.0)	0.09	0.07 (1.0)	0.04	-0.03 (1.0)	0.14	-0.46 (0.96)	0.043	-0.18 (1.0)	0.187	0.36 (0.96)
b_1	0.53	-0.72 (1.0)	0.30	-0.56 (0.33)	0.15	-0.09 (1.0)	0.08	0.23 (0.71)	0.011	0.09 (1.0)	0.944	0.95 (0.0)
b_2	0.15	0.38 (1.0)	0.54	0.73 (0.0)	0.05	-0.01 (1.0)	0.05	0.07 (0.60)	0.001	-0.01 (1.0)	0.031	0.11 (1.0)
b_3	1.25	1.11 (0.90)	1.59	-1.25 (0.0)	0.04	0.06 (1.0)	0.03	-0.04 (0.10)	0.012	-0.08 (1.0)	0.355	0.57 (0.41)
α_1	0.07	-0.27 (1.0)	-	-	-	-	-	-	0.095	-0.22 (1.0)	-	-
α_2	0.005	0.06 (1.0)	-	-	-	-	-	-	0.015	0.07 (1.0)	-	-
α_3	0.12	0.34 (1.0)	-	-	-	-	-	-	0.264	0.45 (1.0)	-	-
δ_{11}	-	-	-	-	0.46	0.21 (0.96)	-	-	0.354	-0.12 (1.0)	-	-
δ_{12}	-	-	-	-	2.04	0.26 (1.0)	-	-	0.794	-0.52 (1.0)	-	-
δ_{13}	-	-	-	-	4.84	-2.04 (0.84)	-	-	0.447	0.27 (0.84)	-	-
δ_{21}	-	-	-	-	4.48	-2.63 (0.68)	-	-	5.644	2.32 (0.68)	-	-
δ_{22}	-	-	-	-	1.61	-0.13 (1.0)	-	-	0.469	-0.38 (1.0)	-	-
δ_{23}	-	-	-	-	0.51	-0.48 (1.0)	-	-	0.324	0.38 (1.0)	-	-

Table 4. Probit versus AEP: Estimated MSE (coverage).

Parameters	Probit	AEP-I	AEP-II	AEP-III
a_1	0.16 (0.86)	0.16 (0.95)	0.23 (1.0)	0.14 (1.0)
a_2	0.15 (1.0)	0.15 (1.0)	0.27 (1.0)	0.03 (1.0)
a_3	0.18 (1.0)	0.17 (1.0)	0.06 (1.0)	0.16 (1.0)
b_1	0.09 (0.90)	0.11 (1.0)	0.06 (1.0)	0.47 (1.0)
b_2	0.05 (1.0)	0.01 (1.0)	0.03 (1.0)	0.02 (1.0)
b_3	0.08 (0.90)	0.39 (1.0)	0.16 (1.0)	0.51 (1.0)

Table 5 shows the parameter estimates for each of the fitted models. Posterior means are used as point estimates of the parameters of interest. The deviance information criterion –DIC– (see Spiegelhalter et al., 2002) for each model was obtained too, and was utilized to assess the fit of the various models.

Table 5. Parameter estimation for Fox (2010) data set.

Parameters	Probit	AEP-I	AEP-II	AEP-III
a_1	1.54	1.52	1.84	1.65
a_2	0.90	0.88	1.78	1.54
a_3	0.66	0.70	1.50	1.32
a_4	0.91	0.89	1.47	1.48
a_5	0.46	0.49	1.61	1.22
b_1	-0.27	-0.04	-0.26	-0.02
b_2	-0.79	-0.25	-0.52	-0.05
b_3	-0.11	0.05	0.31	0.55
b_4	-0.73	-0.35	-0.83	-0.19
b_5	-0.42	-0.18	-0.02	0.53
α_1	–	0.58	–	0.50
α_2	–	0.69	–	0.61
α_3	–	0.56	–	0.51
α_4	–	0.63	–	0.62
α_5	–	0.59	–	0.61
δ_{11}	–	–	3.80	1.36
δ_{12}	–	–	0.18	0.75
δ_{13}	–	–	0.76	0.56
δ_{14}	–	–	1.03	1.26
δ_{15}	–	–	0.10	0.67
δ_{21}	–	–	4.36	1.15
δ_{22}	–	–	5.68	1.61
δ_{23}	–	–	3.79	1.70
δ_{24}	–	–	3.50	1.38
δ_{25}	–	–	3.67	1.53
DIC	1332.9	1303.0	1463.6	2235.1

We observe that item five discriminates poorly in the AEP-I and Probit models, with values less than 1, except for item 1. Item 1 is the most discriminative item for all fitted models. The average estimated discrimination level is 0.89 with the AEP-I IRT model, which is slightly smaller than the prior mean. The posterior means for the skewness parameters show that the marginal posterior densities are non-symmetric and slightly positively skewed (see the AEP-III and AEP-I models). For the estimated difficulty parameter, we can observe that the item with higher values for difficulty is item 3, whereas the less difficult items are 2 and 4. The proportion of correct responses for each of the five items are estimated as 56%, 73%, 54%, 71%, and 65%, respectively. In Table 5, we also observe that, for the AEP-I IRT model, we have a lower value of DIC with respect to the probit, AEP-III and AEP-II IRT models. This criterion indicates that the AEP-I IRT model is preferable with respect to the other models considered.

We applied the PPMC method to assess the fit of each of the IRT models. We calculated the values of discrepancy measure OSD and compared the observed and predicted score distribution through plots of the discrepancy measures. Figure 3 and Figure 4 show the OSD and KL discrepancy measures, respectively. The PPMC method provides graphical evidence that the Probit and AEP-I models cannot adequately explain the observed score

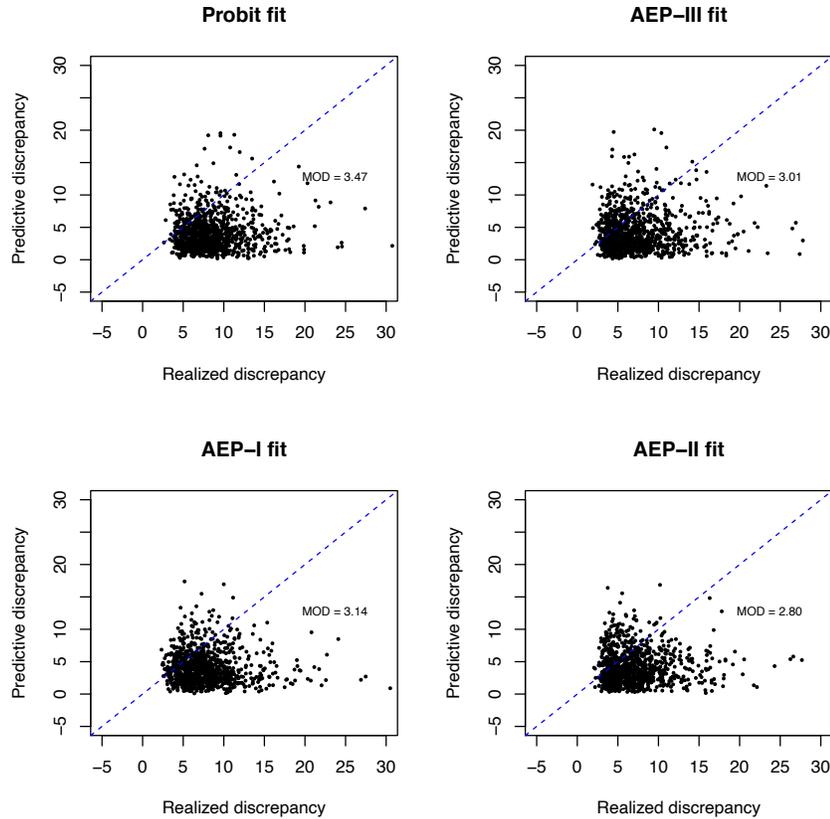


Figure 3. PPMC based on the OSD discrepancy.

distribution of the actual dataset, even though the AEP-I model seems preferable under the DIC criterion. As a quick numerical summary of the plots, we also calculated the average orthogonal distances from the 45-degree line to the points given by the realized and predictive OSDs. We used this mean orthogonal distance (MOD) to provide a quantitative measure of the fit (included in Figures 3 and 4). A large value of the MOD suggests that the model does not adequately capture the features of the data.

Neither the results in Figure 3 nor those in Figure 4 are in agreement with the results obtained using the DIC (see Table 5). The DIC has been criticized on several grounds (see Spiegelhalter et al., 2014). In this particular application, the PPMC procedure seems to yield better results.

6. CONCLUDING REMARKS

In this paper, we have proposed the use of link functions based on the asymmetric exponential power distribution to model item response data. These link functions provide great flexibility to model a wide range of item characteristic curve shapes and include the symmetric probit model as a special case. The resulting model can handle both symmetry/asymmetry and light/heavy tails at the same time.

In contrast with traditional approaches to IRT modeling, the Bayesian approach has a number of advantages. For one thing, inferences based on posterior simulations are both more flexible and relatively easy to implement in JAGS within the R software. Also, the possible lack of identifiability in the general IRT model may be tackled using suitable prior distributions.

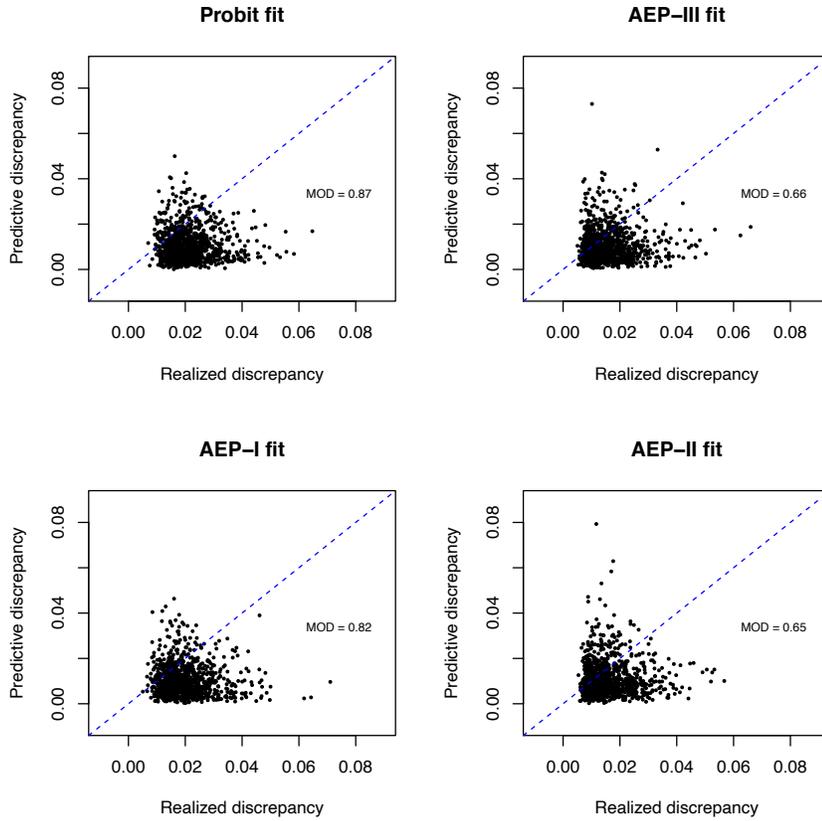


Figure 4. PPMC based on the D_{KL} discrepancy.

Our simulation study shows that the general IRT AEP-based model and the corresponding Bayesian estimates perform well. Our results also suggests that the DIC does not provide a good measure of model fit in our setting, perhaps because it is not based on a proper predictive criterion. By contrast, the posterior predictive model-checking procedure used here provides a nice graphical summary and, together with the mean orthogonal distance, provides a better way of comparing models. Moreover, in the real data example the Kullback-Leibler discrepancy proposed here seems to outperform the OSD discrepancy.

ACKNOWLEDGEMENTS

The authors would like to thank an Associate Editor and two anonymous reviewers for valuable comments which helped us improve the presentation of the paper. The first author was supported by a postdoctoral research fellowship from CONACYT, Mexico. The work of the second author was partially supported by the Sistema Nacional de Investigadores (CONACYT), Mexico.

APPENDIX A

Consider the standard version ($\mu = 0$, $\sigma = 1$) of the AEP PDF given in Equation (2) stated as

$$f_{\text{AEP}}(x|\alpha, \delta_1, \delta_2) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp \left\{ - \left| \frac{x}{\sqrt{2\pi}\alpha/\Gamma(1+1/\delta_1)} \right|^{\delta_1} \right\}, & \text{if } x \leq 0; \\ \frac{1}{\sqrt{2\pi}} \exp \left\{ - \left| \frac{x}{\sqrt{2\pi}(1-\alpha)/\Gamma(1+1/\delta_2)} \right|^{\delta_2} \right\}, & \text{if } x > 0. \end{cases}$$

The CDF of the AEP distribution is expressed by

- For $x \leq 0$,

$$\begin{aligned} F(x|\alpha, \delta_1, \delta_2) &= \int_{-\infty}^x f_{\text{AEP}}(z|\alpha, \delta_1, \delta_2) dz \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp \left\{ - \left| \frac{z}{\sqrt{2\pi}\alpha/\Gamma(1+1/\delta_1)} \right|^{\delta_1} \right\} dz. \end{aligned}$$

Now, making the change of variable

$$t = \left(\frac{|z|}{\sqrt{2\pi}\alpha/\Gamma(1+1/\delta_1)} \right)^{\delta_1}, \quad dt = \delta_1 |z|^{\delta_1-1} \left(\frac{\Gamma(1+1/\delta_1)}{\sqrt{2\pi}\alpha} \right)^{\delta_1} dz = -\delta_1 t z^{-1} dz,$$

we have

$$dz = -\frac{\sqrt{2\pi}\alpha t^{1/\delta_1-1}}{\Gamma(1/\delta_1)} dt.$$

Note also that $t \rightarrow +\infty$ as $z \rightarrow -\infty$, while $t \rightarrow t_1(x) \equiv \left(\frac{|x|}{\sqrt{2\pi}\alpha/\Gamma(1+1/\delta_1)} \right)^{\delta_1}$ as $z \rightarrow x$.

Hence,

$$\begin{aligned} F(x|\alpha, \delta_1, \delta_2) &= \alpha \int_{t_1(x)}^{\infty} \frac{t^{1/\delta_1-1} \exp(-t)}{\Gamma(1/\delta_1)} dt \\ &= \alpha \left(1 - \int_0^{t_1(x)} \frac{t^{1/\delta_1-1} \exp(-t)}{\Gamma(1/\delta_1)} dt \right) \\ &= \alpha \left[1 - G \left(t_1(x); \frac{1}{\delta_1}, 1 \right) \right]; \quad x \leq 0, \end{aligned} \tag{9}$$

where $G(\cdot)$ denotes the gamma CDF.

- For $x > 0$,

$$\begin{aligned} F(x|\alpha, \delta_1, \delta_2) &= \int_{-\infty}^x f_{\text{AEP}}(z|\alpha, \delta_1, \delta_2) dz \\ &= \int_{-\infty}^0 f_{\text{AEP}}(z|\alpha, \delta_1, \delta_2) dz + \int_0^x f_{\text{AEP}}(z|\alpha, \delta_1, \delta_2) dz \\ &= \alpha + \int_0^x f_{\text{AEP}}(z|\alpha, \delta_1, \delta_2) dz \quad \text{by Equation (9)} \end{aligned}$$

Similarly to the previous case, making the change of variable

$$t = \left(\frac{|z|}{\sqrt{2\pi}(1-\alpha)/\Gamma(1+1/\delta_2)} \right)^{\delta_2}, \quad dt = \delta_2 z^{\delta_2-1} \left(\frac{\Gamma(1+1/\delta_2)}{\sqrt{2\pi}(1-\alpha)} \right)^{\delta_2} dz,$$

we have

$$\begin{aligned} F(x|\alpha, \delta_1, \delta_2) &= \alpha + (1-\alpha) \int_0^{t_2(x)} \frac{t^{1/\delta_2-1} \exp(-t)}{\Gamma(1/\delta_2)} dt \\ &= \alpha + (1-\alpha) G\left(t_2(x); \frac{1}{\delta_2}, 1\right); \quad x > 0, \end{aligned}$$

where

$$t_2(x) = \left(\frac{|x|}{\sqrt{2\pi}(1-\alpha)/\Gamma(1+1/\delta_2)} \right)^{\delta_2}.$$

APPENDIX B. JAGS IMPLEMENTATION

The proposed models were all implemented in JAGS using the R2jags package to fit the models and to perform convergence diagnostics right within R. Here we use the data set of Section 5.2 to illustrate the implementation of our model in JAGS.

- (1) **Packages.** Load the required R packages:

```
library(R2jags)
library(coda)
library(lattice)
library(R2WinBUGS)
library(rjags)
```

- (2) **Data.** Read the data from the working directory:

```
setwd("my directory")
cito<-matrix(read.table(file="cito.txt", sep=",",),200,5,byrow=T)
N<-dim(cito)[1]; K<-dim(cito)[2]
cito.data <- list("cito","N","K")
```

- (3) **The model.** Write the model in BUGS code and save it as “cito.model.jags” in the working directory.

```
model{
  for (i in 1:N){
    for(k in 1:K){
      p[i,k]<-phi(a[k]*theta[i]-b[k])
      Y[i,k] ~ dbern(p[i,k])
    }
    theta[i] ~ dnorm(0,1)
  }
  for (i in 1:K){
    a[k] ~ dnorm(1,1)T(0,)
    b[k] ~ dnorm(0,1)
  }
}
```

- (4) **Parameters.** Define the parameters of interest:

```
cito.params<-c("a","b")
```

- (5) **Initial values.** Define the starting values for the MCMC runs:

```
cito.inits<-function()
{
  list("a"=c(0.5,0.5,0.5,0.5,0.5), "b"=c(0,0,0,0,0))
}
```

Alternatively, specify separate starting values for each chain:

```
units1<-list("a"=c(0.1,0.1,0.1,0.1,0.1), "b"=c(-4,-4,-4,-4,-4))
units2<-list("a"=c(3,3,3,3,3), "b"=c(4,4,4,4,4))
cito.inits2<-list(units1, units2)
```

- (6) **Fit.** Fit the model in JAGS:

```
set.seed(123)
fit.cito<-jags(data=cito.data, inits = cito.inits2, parameters.to.save=
  cito.params, n.chains =2, n.iter = 9000, n.burnin=1000,
  model.file="cito.model.jags")
print(fit.cito)
```

- (7) **Diagnostic.** Convert the model output into an MCMC object in order to have access to several convergence diagnostics:

```
cito.mcmc<-as.mcmc(cito,fit)
xyplot(cito.mcmc,layout=c(2,6), aspect="fill")
densityplot(cito.mcmc)
autocorr.plot(cito.mcmc)
elman.plot(cito.mcmc)
geweke.diag(cito.mcmc)
geweke.plot(cito.mcmc)
raftery.diag(cito.mcmc)
raftery.plot(cito.mcmc)
heidel.diag(cito.mcmc).
```

REFERENCES

- Albert, J.H., 1992. Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics*, 17, 251–269.
- Azevedo, C.L., Andrade, D.F., and Fox, J.P., 2012. A Bayesian generalized multiple group IRT model with model-fit assessment tools. *Computational Statistics and Data Analysis*, 56, 4399–4412.
- Azevedo, C.L., Bolfarine, H., and Andrade, D.F., 2011. Bayesian inference for skew-normal IRT model under the centred parametrization. *Computational Statistics and Data Analysis*, 55, 353–365.
- Bacci, S., Bartolucci, F., and Gnaldi, M., 2014. A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Communications in Statistics: Theory and Methods*, 43, 4:787–800.
- Baker, F. and Kim, S., 2004. *Item Response Theory: Parameter Estimation Techniques*. Statistics: A Series of Textbooks and Monographs. Taylor and Francis, New York.
- Bazán, J.L., Branco, M.D., and Bolfarine, H., 2006. A skew item response model. *Bayesian Analysis*, 1, 861–892.
- Béguin, A.A. and Glas, C.A.W., 2001. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541–562.
- Casella, G. and George, E.I., 1992. Explaining the Gibbs sampler. *The American Statistician*, 46, 167–174.
- Chen, M.H., Dey, D.K., and Shao, Q.M., 1999a. A new skewed link model for dichotomous quantal response data. *Journal of the American Statistical Association*, 94, 1172–1186.
- Chen, M.H., Ibrahim, J.G., Shao, Q.M., and Weiss, R.E., 2003. Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference*, 111, 57–76.
- Chen, M.H., Ibrahim, J.G., and Yiannoutsos, C., 1999b. Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society B*, 61, 223–243.
- Chib, S. and Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–335.
- Choi, Y.J. and Asilkalkan, A., 2019. R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspective*, 17, 168–175.
- Durante, D., 2019. Conjugate bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106, 765–779.
- Fox, J.P., 2010. *Bayesian Item Response Modeling. Theory and Applications*. Springer.
- Fox, J.P. and Glas, C.A., 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271–288.
- Fu, Z.H., Tao, J., and Shi, N., 2009. Bayesian estimation in the multidimensional three-parameter logistic model. *Journal of Statistical Computation and Simulation*, 79, 819–835.
- Gelfand, A.E. and Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association*, 85, 398–409.
- Ghosh, M., Ghosh, A., Chen, M.H., and Agresti, A., 2000. Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, 88, 99–115.
- Jiang, X., Dey, D.K., Prunier, R., Wilson, A.M., and Holsinger, K.E., 2013. A new class of the flexible link function with application to species co-occurrence in cape floristic region. *The Annals of Applied Statistics*, 7, 2180–2204.

- Karabatsos, G., 2016. Bayesian nonparametric response models. In *Handbook of Item Response Theory. Volume I: Models.* (van der Linden, W.J., editor). CRC Press, Boca Raton, FL, US, pp. 323–336.
- Lord, F.M., 1952. A theory of test scores. *Psychometric Monograph*, 7, 84.
- Lord, F.M., 1980. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum Associates, NJ, US.
- Matteucci, M., Mignani, S., and Veldkamp, B.P., 2012. Prior distributions for item parameters in IRT models. *Communications in Statistics: Theory and Methods*, 41, 2944–2958.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1091.
- Naranjo, L., Pérez, C., and Martín, J., 2015. Bayesian analysis of some models that use the asymmetric exponential power distribution. *Statistics and Computing*, 25, 497–514.
- Patz, R.J. and Junker, B.W., 1999. A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146–178.
- Plummer, M., 2017. *JAGS Version 4.3.0 user Manual.*
- R Core Team, 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.
- Rasch, G., 1961. *On General Laws and the Meaning of Measurement in Psychology.* University of California Press, Berkeley, California, US.
- Reckase, M.D., 2009. *Multidimensional Item Response Theory.* Springer, New York.
- Rupp, A.A., Dey, D.K., and Zumbo, B.D., 2004. To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 11, 424–451.
- Samejima, F., 2000. Logistic positive exponent family of models: Virtue of asymmetric item characteristic curve. *Psychometrika*, 65, 319–335.
- San Martín, E., Jara, A., Rolin, J.M., and Mouchart, M., 2011. On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, 76, 385–409.
- Sinharay, S., Johnson, M.S., and Stern, H.S., 2006. Posterior predictive assessment of item theory models. *Applied Psychological Measurement*, 30, 298–321.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, 64, 583–639.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A., 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society B*, 76, 485–493.
- Spiegelhalter, D.J., Thomas, A., Best, N.G., and Lunn, D., 2003. *WinBUGS Version 1.4 User Manual.*
- Stan Development Team, 2014. *Stan Modeling Language: User’s Guide and Reference Manual.* Version 2.2.0.
- Su, Y.S. and Yajima, M., 2020. *R2jags: Using R to Run ‘JAGS’.* R package version 0.6-1.
- Svetina, D., 2013. Assessing dimensionality of noncompensatory multidimensional item response theory with complex structures. *Educational and Psychological Measurement*, 73, 312–338.
- van der Linden, W.J. and Hambleton, R.K., 1997. *Handbook of Modern Item Response Theory.* Springer, New York.
- Zhu, D. and Zinde-Walsh, V., 2009. Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, 148, 86–99.

INFORMATION FOR AUTHORS

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to the Editors-in-Chief of the ChJS (E-mail: chilean.journal.of.statistics@gmail.com). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a "cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

PREPARATION OF ACCEPTED MANUSCRIPTS

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at <http://soche.cl/chjs/files/ChJS.zip>. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at <http://www.ams.org/mathscinet/msc/>. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

COPYRIGHT

Authors who publish their articles in the ChJS automatically transfer their copyright to the Chilean Statistical Society. This enables full copyright protection and wide dissemination of the articles and the journal in any format. The ChJS grants permission to use figures, tables and brief extracts from its collection of articles in scientific and educational works, in which case the source that provides these issues (Chilean Journal of Statistics) must be clearly acknowledged.