Calidad de Revistas Científicas Españolas
FECYT | Fundación Española para la Ciencia y la Tecnología | 2016

ISSN: 1696-2281
eISSN: 2013–8830
SORT 40 (1) January-June 2016

SORT

Statistics and Operations Research Transactions

Coediting institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Universitat Pompeu Fabra
Universitat de Lleida
Institut d'Estadística de Catalunya

Supporting institutions

Spanish Region of the International Biometric Society
Societat Catalana d'Estadística

Generalitat de Catalunya
Institut d'Estadística de Catalunya

SORT

Volume 40
Number 1
January-June 2016
ISSN: 1696-2281
eISSN: 2013-8830

Articles

SORT 40 (1) January-June 2016, 3-30

The relevance of multi-country input-output tables in measuring emissions trade balance of countries: the case of Spain

Teresa Sanz[1,*], Rocío Yñiguez[1] and José Manuel Rueda-Cantuche[2]

Abstract

As part of national accounts, input-output tables are becoming crucial statistical tools to study the economic, social and environmental impacts of globalization and international trade. In particular, global input-output tables extend the national dimension to the international dimension by relating individual countries' input-output tables among each other, thus providing an opportunity to balance the global economy as a whole. Concerning emissions of greenhouse gases, the relative position that countries hold among their main trade partners at the global level is a key issue in terms of international climate negotiations. With this purpose, we show that (official) Multi-country input-output tables are crucial to analyse the greenhouse gas emission trade balance of individual countries. Spain has a negative trade emissions balance for all three gases analysed, being the most negative balances those associated to the bilateral trade with China, Russia, United States and the rest of the European Union as a whole.

MSC: 91F.

Keywords: WIOD, Emissions Trade Balance, Spain, GHG footprint, GHG.

1. Background and statistical context

The latest meeting of the Group of Experts on National Accounts of the United Nations Economic Commission for Europe (UNECE, 7-9 July 2015), was devoted to data collection and compilation methods in respect to global production activities. It was jointly

* Corresponding author.

[1] University of Seville, Dpt. Economic Analysis and Political Economy, Avda. Ramón y Cajal, 1. 41018 Sevilla. Phone: +954557524/954 554481. Fax: 954557629. mtsanz@us.es/ovando@us.es

[2] European Commission, Joint Research Centre, Institute for Prospective and Technological Studies, Inca Garcilaso, 3, 41092-Edificio EXPO. The views expressed in this paper belong to the authors and should not be attributed to the European Commission or its services.

Received: December 2014
Accepted: November 2015

organized with Eurostat and the Organization for Economic Co-operation and Development (OECD). The meeting was attended by representatives from more than thirty countries worldwide and representatives from the European Commission (EC), International Monetary Fund (IMF), OECD, the United Nations Conference on Trade and Development (UNCTAD), United Nations Statistics Division (UNSD) and World Trade Organization (WTO), among others.

According to the experts at this UNECE meeting, in order to measure global production and global value chains it is no longer sufficient to look only at what a firm does, but to also to consider how the firm does its activities and with whom. For instance, linking business statistics and trade statistics on a micro level should provide new dimensions to the data as long as new balancing challenges at the macro level data (e.g. national accounts). Indeed, statisticians have not always been able to keep up to date with business practices and must find ways to be forward looking and provide the information that meets future policy needs. Traditional measures of trade in goods and services have to be progressively supplemented with information on income and financial flows. Foreign direct investment statistics (FDI) should be further developed and complemented with foreign affiliate statistics (FATS) in order to improve their clarity, usefulness and coverage, and to provide better insights into global value chains.

In this respect, the UNECE Report emanating from this meeting supported new global initiatives, such as the extensions to Trade in Value Added and Global Input-Output Tables (OECD), the construction of the European Multi-Country Input-Output Framework (EC and Eurostat) as well as the elaboration of a new Handbook on a System of Extended International and Global Accounts (UNSD).

Hence, there is no doubt that globalization is currently affecting the way statisticians are measuring national production of countries and international statistical organizations are indeed very busy working on it in order to meet the policy needs at the worldwide level. As national accounts and input-output tables became an integral part of the production activities of national statistical institutes in the past, very soon multi-country and international input-output tables will become a crucial statistical tool to measure global production, trade in value added, environmental footprints and/or employment effects of export activities with official statistics (e.g. carbon footprint estimated by Eurostat).

Bearing all this in mind, we would like to illustrate in this paper the usefulness of global/world input-output tables in measuring the greenhouse gas footprints of individual countries and its external emission trade balance with respect to others. Hopefully, these types of indicators will soon become regularly produced in the future by statisticians using official global input-output tables instead of using other databases produced as one-off projects (e.g. World Input-Output Database, WIOD – www.wiod.org).

This paper is structured in five sections. Following this background, there is an introductory section on the related literature on greenhouse gases emissions footprints. Next, the third section introduces the methodology and the database. The fourth section presents the results obtained and discusses them. The fifth section concludes.

2. Introduction to GHG footprints

Greenhouse gas emissions (GHG) are considered to be one of the main causes of climate change. This is the reason why governments are increasingly making efforts to implement policies aiming to reduce GHG emissions. National climate policies are mainly driven by international negotiations and these are strongly linked to the amount of emissions produced within a country or the so called producer's responsibility principle. Within this context, exporting (producing) countries are responsible for their GHG emissions, irrespective of where the demand for such products comes from.

On the other hand, the interest in the so called consumer's responsibility principle has been growing since Leontief (1970) described the environmental impacts of the final consumer as a negative externality of the production process. This concept has been endorsed by the OECD's Green Growth Strategy (2011). According to this principle, the GHG emissions are allocated according to countries' domestic demand of goods and services, irrespective of where they were produced. Different approaches have been used to analyse this new concept of responsibility, such as general balance models, dynamic models and the analysis of structural decomposition, i.e. Peters and Hertwich (2006), Peters (2008) Peters et al. 2011), Druckman and Jackson (2009), Davis and Caldeira (2010), Zhou and Imura (2011) and Edens et al. (2011), Kanemoto et al. (2012), among others.

Among others, Rueda-Cantuche and Amores (2010) noted that developed countries may reduce their emissions produced but at the same time, they may increase their consumption-based emissions. This is due to the different technologies used in the production processes of developing countries, generally less clean than those of the developed countries. In the end, some environmental policies might result in a global increase in GHG emissions. At the national level, the difference between the production-based emissions and the consumption-based emissions lead to the so called emission trade balance (ETB) of a country or of a certain industry. This analysis will determine the surplus/deficit that a country/industry has. It is expected that developing countries have surpluses and developed countries, deficits.

Within this context, the aim of this paper is to calculate the Emission Trade Balance (ETB) of Spain in 2008 at a worldwide level and bilaterally with respect to 39 countries, 35 industries and one additional region as the "rest of the world" for the three main GHGs (CO_2, N_2O and CH_4). In order to do so, we have used multi-regional input-output analysis (MRIO) and the World Input-Output Database (WIOD) (Dietzenbacher et al., 2013).

Input-output analysis (IOA) has been generally used to study environmental problems (Miller and Blair, 2009). Particularly, there are numerous related studies devoted to the analysis of polluting GHG emissions, i.e. Minx et al., (2009), Su et al. (2010), Chen et al., (2010), Liang et al. (2010), Chang et al. (2010), Zhu et al., (2012) and Mattila et al. (2013), among others.

Likewise, there are also many studies about GHG emissions associated with the international trade of specific countries, such as China, (Liang et al., 2007, Liu et al., 2009, Zhao et al., 2009, Xu et al., 2011, Hongtau et al., 2010, Chen et al., 2010 a, b, Chen and Zhang 2010); Finland (Maenpaa and Siikavirta 2007); Ireland (Llop and Tol, 2012); Italy (Cellura et al., 2013, Mongelli et al.,2006); Japan (Nansai et al., 2009); the United Kingdom (Wiedman et al., 2010, Druckman and Jackson 2009)) and Turkey (Tunç et al., 2007).

The work of Musksgaard and Pedersen (2001) for Denmark was the first one that linked the input-output methodology to the consumer's responsibility principle related to GHG emissions. It was followed by Ahmad and Wyckoff (2003) for OECD countries and Peters and Hertwich (2006) for the Norwegian economy and for three different gases (CO_2, NO_2 and SO_2).

IOA has also been applied to study GHG emissions associated to consumption in the case of Spain. Tarancón and del Rio (2007) used a combination of IOA with sensitivity analyses; Cadarso et al. (2010) study the effect of international trade of the Spanish emissions balace under DTA assumption; Sánchez-Choliz, and Duarte (2004), Serrano and Roca (2008a, 2008b), Serrano and Dietzenbacher (2010) used IOA assuming domestic technology in monetary terms while Arto (2009) and Arto et al., (2012) do the same but in physical terms; Lopez et al. (2013) analyse the existence of pollution haven hypothesis in a bi-regional input-output model and Cadarso et al. (2012) defined a shared responsibility criterion to analyse the impact of international trade in CO_2 emissions on an industrial basis, such as the food industry in Lopez et al. (2015).

But none of them has used a homogeneous multi-country IO database such as WIOD (Dietzenbacher et al., 2013), nor has the analysis been carried out with high industry resolution and bilateral trade flows as in the present study. This work covers 35 industries and 41 different geographical areas for each of the three GHGs considered. Therefore, the originality and interest of this work lies in the details and the extension of the results in terms of higher industry breakdown, homogeneity of the multi-country database, country coverage and pollutants covered (CO_2, CH_4 and N_2O) rather than the topic itself, which has already been addressed in the literature.

3. Methodology and database

3.1. Input-output analysis

Input-output analysis revolves around the so called input-output tables, which reflect the supply and demand of the economy in terms of products, industries and final users. By using the so called Leontief quantity model (Rueda-Cantuche, 2011), the total output of an economy can be broken down into final and intermediate demand, as indicated in (1):

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y} \tag{1}$$

where $\boldsymbol{x}$ is the total industry output vector for n industries $(n \times 1)$; $\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{x}$ is a matrix describing the intermediate uses of industries; $\boldsymbol{A}$ is a matrix $(n \times n)$ of input-output coefficients showing the inputs needed per unit of output by each industry; and $\boldsymbol{y}$ stands for a final demand vector $(n \times 1)$ showing the sum of consumption, investment and exports of all goods and services. Within this framework, we use industry by industry IO tables from the WIOD database (Dietzenbacher et al., 2013) with the same number of industries and commodities (n).

Reordering (1), it yields

$$\boldsymbol{x} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{y} \tag{2}$$

where $\boldsymbol{I}$ is the identity matrix and $(\boldsymbol{I} - \boldsymbol{A})^{-1}$, the so called Leontief inverse matrix that shows the total requirements of the economy for the production of goods and services to satisfy a certain level of final demand. Moreover, with appropriate emission levels ($\boldsymbol{s}$) per unit of total industry outputs ($\boldsymbol{x}$), $\boldsymbol{c} = \boldsymbol{s}\hat{\boldsymbol{x}}^{-1}$ (where $\hat{}$ denotes diagonalization of the vector $\boldsymbol{x}$), the Leontief model can serve to estimate the absolute levels of emissions for the production of a certain level of total output needed to satisfy changes in final demand, e.g. emissions of the car industry to produce vehicles due to changes in households demand. It is important to note that this paper is focused on the production phase of emissions alone and it does not include those emissions derived from the use phase of a product (e.g. households driving cars). That is:

$$\boldsymbol{s} - \hat{\boldsymbol{c}}(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{y} \tag{3}$$

3.2. Multi-regional input-output analysis

Multi-regional input-output analysis is based on a set of interconnected input-output tables of various countries (Miller and Blair, 2009). While equation (3) refers to one single country with n industries, we will express hereafter the same equation for a three-region model with n industries in each region, namely: Spain (u), rest of the EU (r) and rest of the world (w). The result is a fully fledged input-output table with three times n industries and its main components are described below.

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{uu} & \boldsymbol{A}_{ur} & \boldsymbol{A}_{uw} \\ \boldsymbol{A}_{ru} & \boldsymbol{A}_{rr} & \boldsymbol{A}_{rw} \\ \boldsymbol{A}_{wu} & \boldsymbol{A}_{wr} & \boldsymbol{A}_{ww} \end{pmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_{uu} & \boldsymbol{y}_{ur} & \boldsymbol{y}_{uw} \\ \boldsymbol{y}_{ru} & \boldsymbol{y}_{rr} & \boldsymbol{y}_{rw} \\ \boldsymbol{y}_{wu} & \boldsymbol{y}_{wr} & \boldsymbol{y}_{ww} \end{pmatrix}$$

$$\boldsymbol{L} = (\boldsymbol{I} - \boldsymbol{A})^{-1} = \begin{pmatrix} \boldsymbol{L}_{uu} & \boldsymbol{L}_{ur} & \boldsymbol{L}_{uw} \\ \boldsymbol{L}_{ru} & \boldsymbol{L}_{rr} & \boldsymbol{L}_{rw} \\ \boldsymbol{L}_{wu} & \boldsymbol{L}_{wr} & \boldsymbol{L}_{ww} \end{pmatrix} \qquad \hat{\boldsymbol{C}} = \begin{pmatrix} \hat{\boldsymbol{c}}_u & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{c}}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{c}}_w \end{pmatrix}$$

Matrix $\boldsymbol{A}$ and vector $\boldsymbol{y}$ stand for input-output coefficients and final uses, respectively. The subscript on the left corresponds to the exporting region and the subscript on the right refers to the importing region. Doing so, these two elements include bilateral exports and bilateral imports of intermediate and final uses, too. Besides, each of the submatrices of the $\boldsymbol{A}$ matrix has n rows and n columns, so the fully-fledged matrix $\boldsymbol{A}$ is of order $(3n \times 3n)$. For one single final demand category, the matrix $\boldsymbol{Y}$ is therefore of order $(3n \times 3)$.

Moreover, it is straightforward that the Leontief inverse is a square matrix of the same dimension as $\boldsymbol{A}$, being eventually matrix $\hat{\boldsymbol{C}}$ a diagonal matrix with three diagonalized vectors of n-dimension each. The latter corresponds to different emission coefficients by country of origin (or region), which is quite relevant for our analysis. These emission coefficients have been calculated as the total emissions of each country and industry over their corresponding total output, both provided by the WIOD database (Dietzenbacher et al., 2013).

With these new matrices, we re-define equation (3) but also allowing for a fully-fledged decomposition of the final demand by region. Subsequently, equation (4) is split up into as many components as number of regions the model has (i.e. three). As a matter of fact, the sum of all the elements of each component is nothing else but the footprint of each of the regions (e.g. carbon footprint). As in Lopez et al., (2013), Cadarso et al., (2012) or Skelton (2013), we have estimated matrices of emissions (see equation 5), where the sum by rows allocate the responsibility to industries that supply intermediate and final goods and the sum by columns allocate the responsibility to agents/industries that consume them. More precisely, the focus of our analysis is based on the sum of the elements of each row in each of the three fully-fledged matrices of equation (5), which yields three vectors of emissions.

$$\hat{\boldsymbol{C}}(\boldsymbol{I}-\boldsymbol{A})^{-1}\begin{pmatrix}\boldsymbol{y}_u & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{y}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{y}_w\end{pmatrix} =$$

$$\begin{aligned}
&= \begin{pmatrix}\hat{\boldsymbol{c}}_u & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{c}}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{c}}_w\end{pmatrix}\begin{pmatrix}\boldsymbol{L}_{uu} & \boldsymbol{L}_{ur} & \boldsymbol{L}_{uw}\\ \boldsymbol{L}_{ru} & \boldsymbol{L}_{rr} & \boldsymbol{L}_{rw}\\ \boldsymbol{L}_{wu} & \boldsymbol{L}_{wr} & \boldsymbol{L}_{ww}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_{uu} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{y}_{ru} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{y}_{wu}\end{pmatrix}\\
&+ \begin{pmatrix}\hat{\boldsymbol{c}}_u & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{c}}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{c}}_w\end{pmatrix}\begin{pmatrix}\boldsymbol{L}_{uu} & \boldsymbol{L}_{ur} & \boldsymbol{L}_{uw}\\ \boldsymbol{L}_{ru} & \boldsymbol{L}_{rr} & \boldsymbol{L}_{rw}\\ \boldsymbol{L}_{wu} & \boldsymbol{L}_{wr} & \boldsymbol{L}_{ww}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_{ur} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{y}_{rr} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{y}_{wr}\end{pmatrix}\\
&+ \begin{pmatrix}\hat{\boldsymbol{c}}_u & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{c}}_r & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \hat{\boldsymbol{c}}_w\end{pmatrix}\begin{pmatrix}\boldsymbol{L}_{uu} & \boldsymbol{L}_{ur} & \boldsymbol{L}_{uw}\\ \boldsymbol{L}_{ru} & \boldsymbol{L}_{rr} & \boldsymbol{L}_{rw}\\ \boldsymbol{L}_{wu} & \boldsymbol{L}_{wr} & \boldsymbol{L}_{ww}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}_{uw} & \boldsymbol{0} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{y}_{rw} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{y}_{ww}\end{pmatrix}
\end{aligned} \tag{4}$$

Being:

$$\begin{pmatrix} \boldsymbol{y}_u & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{y}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{y}_w \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_{uu}+\boldsymbol{y}_{ur}+\boldsymbol{y}_{uw} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{y}_{ru}+\boldsymbol{y}_{rr}+\boldsymbol{y}_{rw} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{y}_{wu}+\boldsymbol{y}_{wr}+\boldsymbol{y}_{ww} \end{pmatrix}$$

Properly extended, equation (4) becomes into:

$$\begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{wu} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{wu} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{wu} \end{pmatrix} + \begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{wr} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{wr} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{wr} \end{pmatrix} + \begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{ww} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{ww} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{ww} \end{pmatrix}$$

and summing row-wise:

$$\begin{aligned} \begin{pmatrix} \boldsymbol{g}_{uu}^{dom} \\ \boldsymbol{g}_{ru}^{imp} \\ \boldsymbol{g}_{wu}^{imp} \end{pmatrix} + \begin{pmatrix} \boldsymbol{g}_{ur}^{exp} \\ \boldsymbol{g}_{rr}^{dom} \\ \boldsymbol{g}_{wr}^{exp} \end{pmatrix} + \begin{pmatrix} \boldsymbol{g}_{uw}^{exp} \\ \boldsymbol{g}_{rw}^{exp} \\ \boldsymbol{g}_{ww}^{dom} \end{pmatrix} = \\ = \begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{wu} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{wu} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{uu} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{ru} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{wu} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \\ + \begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{wr} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{wr} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{ur} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{rr} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{wr} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \\ + \begin{pmatrix} \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_u\boldsymbol{L}_{uw}\boldsymbol{y}_{ww} \\ \hat{\boldsymbol{c}}_r\boldsymbol{L}_{ru}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rr}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_r\boldsymbol{L}_{rw}\boldsymbol{y}_{ww} \\ \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wu}\boldsymbol{y}_{uw} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{wr}\boldsymbol{y}_{rw} & \hat{\boldsymbol{c}}_w\boldsymbol{L}_{ww}\boldsymbol{y}_{ww} \end{pmatrix} \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{pmatrix} \end{aligned} \tag{5}$$

with the following definitions (only some of them are presented as illustrative purposes):

(a) $\hat{\boldsymbol{c}}_u\boldsymbol{L}_{uu}\boldsymbol{y}_{uu}$ stands for the emissions produced in Spain derived from the Spanish final demand of domestically produced commodities (e.g. purchase of a Spanish car by a Spanish resident);

(b) $\hat{\boldsymbol{c}}_u\boldsymbol{L}_{ur}\boldsymbol{y}_{ru}$ represents the emissions produced in Spain for the production of an exported commodity that will be used by the rest of the EU (r) to produce something else that Spain will import (e.g. exports of Spanish electronic components for the production of Czech cars that will be imported by Spain).

(c) $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uw} \boldsymbol{y}_{wu}$ shows the emissions produced in Spain for the production of an exported commodity that will be used by the rest of the world (w) to produce something else that Spain will import (e.g. exports of Spanish electronic components for the production of American cars that will be imported by Spain).

(d) $\boldsymbol{g}_{uu}^{dom}$ is the sum of (a), (b) and (c); the sum of emissions emitted in Spain coming from the final demand of Spanish residents.

(e) $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{ru} \boldsymbol{y}_{uu}$ stands for the emissions produced in EU countries (r) derived from the imported intermediate inputs needed to satisfy the Spanish final demand of domestically produced commodities (e.g. purchase of a Spanish car by a Spanish resident that involves imports of electronic components from the Czech Republic);

(f) $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rr} \boldsymbol{y}_{ru}$ shows the emissions produced in EU countries (r) to satisfy the Spanish final demand of commodities produced in the EU (e.g. imports of German cars by Spanish residents);

(g) $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rw} \boldsymbol{y}_{wu}$ shows the emissions produced in EU countries (r) to produce an intermediate export to a non-EU country that will serve as input to produce something to be exported to Spain (e.g. purchase of a Japanese car by a Spanish resident that involves imports of electronic components from the Czech Republic);

(h) $\boldsymbol{g}_{ru}^{imp}$ is the sum of (e), (f) and (g); the sum of emissions emitted in the rest of Europe coming from the final demand of Spanish residents.

(i) $\boldsymbol{g}_{wu}^{imp}$ is, analogously, the sum of emissions emitted in the rest of the world coming from the final demand of Spanish residents.

(j) $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uu} \boldsymbol{y}_{ur}$ shows the emissions produced in Spain to satisfy the EU final demand of Spanish commodities (e.g. imports of a Spanish car by a German resident);

(k) $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{ur} \boldsymbol{y}_{rr}$ shows the emissions produced in Spain derived from the imported inputs of the rest of the EU needed to satisfy their own final demand of domestically produced commodities (e.g. purchase of a German car by a German resident that involves imports of electronic components from Spain);

(l) $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uw} \boldsymbol{y}_{wr}$ shows the emissions produced in Spain derived from the imported intermediate inputs of the rest of the world needed to satisfy the final demand of EU residents (e.g. purchase of a Japanese car by a German resident that involves imports of electronic components from Spain);

(m) $\boldsymbol{g}_{ur}^{exp}$ is the sum of (j), (k) and (l); the sum of emissions emitted in Spain coming from the final demand of EU residents.

(n) $\boldsymbol{g}_{uw}^{exp}$ is, similarly, the sum of emissions emitted in Spain coming from the final demand of the rest of the world.

Therefore, the total emissions produced in the region u, is:

$$\boldsymbol{g}_{uu}^{dom} + \boldsymbol{g}_{ur}^{exp} + \boldsymbol{g}_{uw}^{exp} \tag{6}$$

and the total of emissions caused by the final demand of region u (carbon footprint), is:

$$\boldsymbol{g}_{uu}^{dom} + \boldsymbol{g}_{ru}^{imp} + \boldsymbol{g}_{wu}^{imp} \tag{7}$$

The difference between the two is the so called Emission Trade Balance (ETB), which can be calculated here by the difference between the emissions actually produced in Spain (6) and the Spanish footprint (7).

In a bilateral model (i.e. dropping region w in equations 6 and 7), the ETB yields:

$$\boldsymbol{g}_{ur}^{exp} - \boldsymbol{g}_{ru}^{imp}$$

which is equal to (from equation 5):

$$\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uu} \boldsymbol{y}_{ur} + \hat{\boldsymbol{c}}_u \boldsymbol{L}_{ur} \boldsymbol{y}_{rr} - \hat{\boldsymbol{c}}_r \boldsymbol{L}_{rr} \boldsymbol{y}_{ru} - \hat{\boldsymbol{c}}_r \boldsymbol{L}_{ru} \boldsymbol{y}_{uu}$$

And therefore,

$$\hat{\boldsymbol{c}}_u (\boldsymbol{L}_{uu} \boldsymbol{y}_{ur} + \boldsymbol{L}_{ur} \boldsymbol{y}_{rr}) - \hat{\boldsymbol{c}}_r (\boldsymbol{L}_{rr} \boldsymbol{y}_{ru} - \boldsymbol{L}_{ru} \boldsymbol{y}_{uu})$$

where the expressions in parentheses are indeed the sum of intermediate and final exports and imports, respectively. Thus, the ETB (positive or negative) highly depends on both the trade balance and the different pollution (emission) intensity of goods traded in both regions (Rueda-Cantuche, 2011; López et al., 2013).

Furthermore, multi-country input-output tables also allow a detailed separate analysis about trade on intermediate and final goods and services and thus, global value chains in the emissions balance. For instance, the total emissions generated in the country of reference due to Spanish imports of final goods and services ($\boldsymbol{g}_{ru}^{imp}$) can be decomposed into:

(a) Emissions generated in the country of reference for the production of the final goods and services exported to Spain (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rr} \boldsymbol{y}_{ru}$;

(b) Emissions generated in the country of reference for the production of the intermediate inputs that will be exported to Spain for the domestic production of a final good or service demanded by Spanish residents (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{ru} \boldsymbol{y}_{uu}$;

(c) Emissions generated in the country of reference for the production of the intermediate inputs that will be exported to a third country for the domestic production of a final good or service to be exported to Spain (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rw} \boldsymbol{y}_{wu}$;

And similarly, the total emissions produced in Spain due to imports of the country of reference ($\boldsymbol{g}_{ur}^{exp}$) can be split up into:

(a) Emissions produced in Spain for exports of final goods and services - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uu} \boldsymbol{y}_{ur}$;

(b) Emissions produced in Spain for exports of intermediate goods and services to the country of reference for the production of final goods in the same country - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{ur} \boldsymbol{y}_{rr}$;

(c) Emissions produced in Spain for exports of intermediate goods and services to a third country that will use them for the production of goods and services to be exported to the country of reference - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uw} \boldsymbol{y}_{wr}$;

Tables A.2, A.3 and A.4 in the Annex report all these results of the analysis for the three gases, which are described and commented in Section 4.

3.3. Database

The data used in this paper come from the World Input-Output Database (WIOD), as described in Dietzenbacher et al. (2013). This is a free database financed by the European Union and developed with the aim to analyse the effects of globalization on trade patterns, environmental pressures and the socioeconomic development of a large group of countries. The data include world input-output tables for the 27 European Union countries and 13 other non-EU economies and also the corresponding national IO tables. The WIOD database currently covers the period 1995-2011 and includes 35 industries and 59 commodities (see Table A.1 of the Annex I). However, data on energy and emissions have not been updated up to 2011 yet so we had to carry out our analysis with environmental data up to 2009. The selection of the year 2008 was eventually done in order to avoid the use of a year where the economic crisis was hitting hard the European economy.

4. Results and discussion

The description of the results is divided into three blocks. The first block reflects the position of the Spanish emission trade balance (ETB) with the rest of the world for all the three GHG considered. In a second step, the results are broken down into types of gases, countries and polluting industries, describing the situation of Spain with respect to the countries with the largest positive or negative ETB.

4.1. Emission Trade Balance of GHG in Spain: general overview

Spain produced 316.6 million tons of CO_2 equivalents in 2008 (7 tons per capita) and its final demand led to 494 million tons of CO_2 equivalents elsewhere in the same year (10.8 tons per capita). The emission trade balance of Spain of GHG resulted therefore in -177.7 million tons of CO_2 equivalents (3.9 tons per capita, a bit over the EU27 average,

Table 1: *Emission Trade Balance of GHG of Spain (thousand tonnes CO_2-equivalent).*

	GHG produced from Spanish exports of final goods and services	GHG footprint from Spanish final demand of goods and services	Emission Trade Balance of GHG
FRA	10 943.0	8 558.1	2 384.9
PRT	6 244.9	4 417.3	1 827.6
GRC	1 123.0	283.2	839.8
GBR	7 513.9	6 692.0	822.0
SWE	1 141.8	863.7	278.1
CYP	131.7	29.5	102.2
SVN	246.5	150.2	96.3
LUX	123.0	69.9	53.1
MLT	60.9	28.1	32.8
LVA	76.5	53.9	22.5
ESP	225 484.1	225 484.1	0.0
EST	65.8	153.3	−87.5
AUT	771.1	860.7	−89.5
LTU	137.6	334.7	−197.1
MEX	1 471.8	1 699.5	−227.7
HUN	363.4	740.0	−376.6
IRL	575.0	977.3	−402.3
BGR	212.9	632.1	−419.2
SVK	182.9	604.5	−421.6
DNK	602.7	1 051.8	−449.0
FIN	421.2	894.0	−472.8
ROM	527.8	1 071.1	−543.3
CZE	590.4	1 272.6	−682.2
TUR	996.9	1 805.8	−808.9
AUS	613.3	1 471.7	−858.3
BEL	2 059.2	3 005.9	−946.7
ITA	5 963.5	7 289.4	−1 325.9
POL	1 446.0	2 990.5	−1 544.5
JPN	1 207.2	2 930.1	−1 723.0
CAN	1 122.8	2 870.1	−1 747.3
TWN	187.4	1 938.2	−1 750.7
IDN	200.4	2 078.1	−1 877.7
KOR	574.7	2 659.8	−2 085.1
NLD	1 733.8	4 044.6	−2 310.8
DEU	8 209.0	12 685.0	−4 476.0
BRA	892.1	5 810.3	−4 918.2
IND	582.6	6 947.2	−6 364.6
USA	6 766.3	13 686.8	−6 920.5
RUS	1 616.6	20 659.2	−19 042.6
RoW	21 055.7	79 113.2	−58 057.5
CHN	2 385.8	65 456.3	−63 070.4
Total EU27	276 951.8	285 237.5	−8 285.8
Total	316 625.3	494 363.6	−177 738.3

Source: Own elaboration based on data from WIOD (Dietzenbacher et al., 2013).

i.e. 3.2 tons per capita). Spain is the fifth EU country with the largest negative emission trade balance, behind Germany, France, United Kingdom and Italy.

Moreover, Spanish exports of final goods and services to France lead to around 11 million tons of CO_2 equivalent of GHG while Spanish exports to Germany and UK induce 8.2 million and 7.5 million tons of CO_2 equivalents of GHG, respectively. On the other hand, the final demand of Spanish residents (GHG footprint) leads to 65.5 million tons of CO_2 equivalent of GHG in China; followed by Russia and US with 20.7 and 13.7 million tons of CO_2 equivalents (see Table 1).

As a result, the largest positive balances are found in France (24 millions of tons of CO_2 equivalents) and Portugal (18.3 millions of tons of CO_2 equivalents). With respect to the largest negative emission trade balances of Spain, China presents the biggest negative balance (63 million tons of CO_2 equivalents) followed by Russia and US (19 million and 6.9 million of tons of CO_2 equivalents, respectively).For further analysis hereafter, we will limit the analysis to the countries with the largest negative/positive emission trade balance of Spain.

This implies that the GHG emissions originated from the consumption of Spanish residents is bigger than those generated in Spain as a consequence of the foreign demand. As shown in Table 1 and in the Annex II (Figure A.1), China is the country with the biggest negative emission trade balance with respect to Spain, even well above the sum of the EU-27.

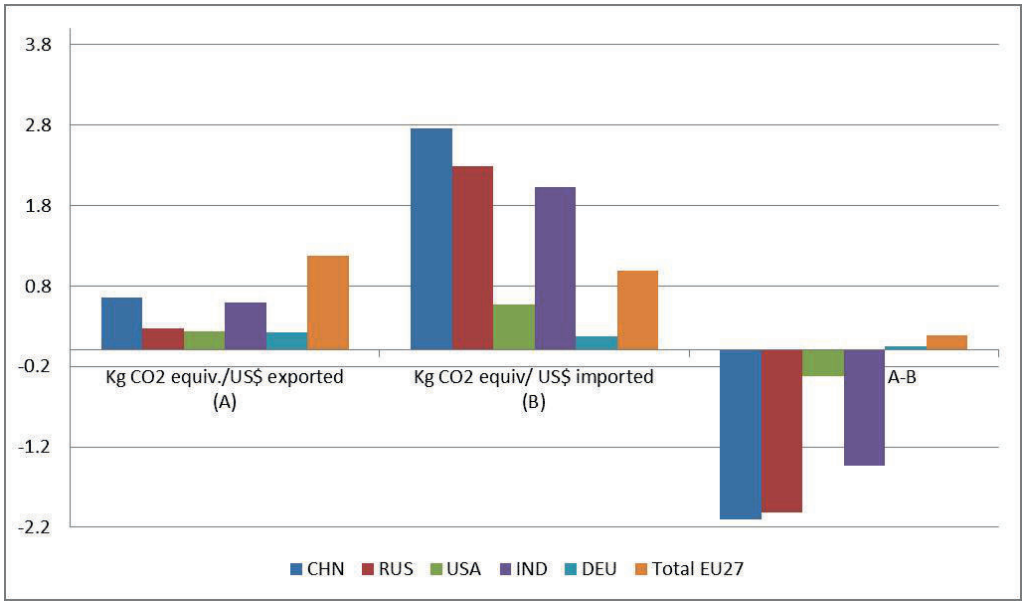

Figure 1: *Comparison of GHG emissions per US dollar in Spain. (Kg CO_2-equivalents/US dollar).*

Source: Based on data from WIOD (Dietzenbacher et al., 2013).

Figure 1 shows GHG emissions per dollar exported (A) and imported (B) by Spain, and the difference between both values (A-B) across some relevant countries and the EU-27 average. Generally speaking, Spanish exports generate less GHG emissions per dollar than Spanish imports, except in the case of the EU-27 average (e.g. Germany). Note that the value of GHG emissions per dollar caused by the production of Chinese and Russian products exported to Spain (i.e. Spanish imports) are remarkably higher than those originated in Spain due to the demand of Spanish products by China and Russia.

4.2. Emission trade balance of GHG in Spain by country of destination

Table A.2 of the Annex I lists, on the one hand, the five countries that contribute most to the negative Spanish ETB in CO_2 emissions, i.e. China, Russia, Germany, the United States and Indonesia. They amount to 47% of the total emissions originated outside Spain due to the imports of Spanish residents. As in Lopez et al. (2013), China is also the country that contributes most to the negative bilateral ETB of Spain. Spanish imports from China account for 25% of the total CH_4 and CO_2 emissions associated with Spanish imports and 14% of N_2O. On the other hand, we show the two countries – France and Portugal – with the largest positive ETB. The emissions associated with the Spanish exports to France and Portugal amounts to 18% of the total emissions produced in Spain to satisfy the total final demand.

Figures A.2, A.3 and A.4 of the Annex II present the results of the bilateral trade emissions of Spain with respect to the rest of the world for the three gases considered: CO_2, CH_4 and N_20, separately. The ETB for CO_2 is positive for 11 countries, which are all EU members. The most prominent positive balances are those of France and Portugal. For CH_4 the situation is similar. The balance is positive for 16 EU countries (e.g. Germany, Italy and Great Britain) and Japan. Finally, in the case of N_2O, the balance is positive for 8 EU countries, Japan and Turkey. As a last remark, 7 EU countries have positive ETB for the three gases, being Great Britain and Portugal the ones that contribute most to the Spanish positive trade balance on GHG emissions (see Figures A.2, A.3 and A.4 of the Annex II).

4.3. Emission trade balance of GHG in Spain by polluting industry

Hereafter, we identify the industries that contribute most to the GHG emissions produced in other countries different from Spain, particularly in those countries where the Spanish carbon footprint is the largest. Analogously, we identify the industries (and countries) that contribute most to the GHG emissions produced in Spain as a result of its imports from other countries. Those GHG emissions are concentrated in seven industries, as it is shown in Tables A.2, A.3 and A.4 of the Annex I.

In Spain, it is interesting to highlight that Electricity is barely traded but nonetheless it is one of the most important sectors in terms of virtual carbon in trade. The reason

is that electricity is generally used to produce goods and services that are eventually traded. In particular, emissions from the Electricity, Gas and Water Supply activities amount to more than half (53.5%) of the CO_2 footprint of Spain in China (column B1 in Table A.2), being 86.6% caused by Spanish imports of Chinese final goods (38.4%, column C1 in Table A.2) and Chinese intermediate goods (48.2%, column D1 in Table A.2). All other emissions (13.4%, column E1 in Table A.2) were due to emissions generated in China for the production of intermediate goods that are exported to third countries, which in turn produce final goods that are consumed by Spanish residents. These results agree with those of Cadarso et al. (2008, 2012). The distribution of CO_2 footprints between final and intermediate goods is similar to other polluting industries (e.g. chemicals, non-metallic mineral and basic metals). However, they do not weight the same as the electricity industry. Cadarso et al.'s results (2008, 2012) suggested that this might be due to the reallocation of production between countries.

The same industry-wide distribution pattern is associated to the emissions of CH_4 and N_2O gases derived from the Spanish demand for final goods produced in China. Particularly, Mining and Quarrying is responsible for almost half (48.1%) of the CH_4 emissions and also the Electricity (38.4%) and Chemicals (36.9%) industries for N_2O emissions.

It is also remarkable that the Agriculture, Hunting, Forestry and Fishing industry is responsible for 26.2% (column B1 of Table A.3) of the CH_4 emissions and 75.6% of the N_2O emissions (column B1 of Table A.4). More than half of these emissions are in both cases caused by the production of Chinese final goods demanded by residents in Spain, being only one third intermediate imported inputs for the domestic production of goods and services demanded by Spanish residents as final goods (columns C1 and D1 in Tables A.3 and A.4 of Annex I).

The second country with the largest negative bilateral ETB (with respect to Spain) is Russia, both for CO_2 and CH_4, although their weight in the total emissions associated with the Spanish imports is much lower than in the case of China: 7.5% for CO_2 and 12.8% for CH_4. In both cases more than 90% of the emissions are explained by a few industries. The most polluting industry in each case is the same as in China: the Electricity industry for CO_2 emissions and Mining and Quarrying activities for CH_4 emissions. Incidentally, Mining and Quarrying is also the second most polluting industry in terms of CO_2 emissions. Although the pattern of types of goods associated with these emissions is somewhat different to China, 70.5% of the emissions associated with the Spanish imports from Russia are caused by the demand for intermediate goods. Besides, Inland Transport industry is responsible for 14.7% and 28.5% of CO_2 and CH_4, respectively, due to pipeline transport services. Differently from China, the relevance of the CO_2 and CH_4 emissions generated in Russia for the production of intermediate goods that will be used by a third country to produce other final goods that Spanish residents will consume, is much higher (over 20%, column E1 in Tables A.2 and A.3).

The third country with the largest negative ETB for CO_2 emissions is Germany, which, however, has a very small but negative N_2O ETB, and a positive CH_4 ETB. The

most polluting industries in terms of CO_2 emissions are the same as those for China plus air transportation services. The relative importance of the contribution of industries to the overall total of emissions is however more spread. The distribution between intermediate and final goods is also similar to that of China.

The list of industries contributing to the United States' (US) emissions associated with Spanish imports is much longer than for the other countries mentioned so far (China, Russia and Germany). Only six industries weight more than 5% in carbon dioxide emissions and they do not sum up even 30% of the overall total, being the most polluting industry the Gas, Water and Electricity supply activities. The distribution pattern between intermediate and final goods is similar to other countries except for Russia, reaching for instance, 78% (sum of columns D1 and E1 in Table A.2) in intermediate goods for Basic metals and fabricated metals. This value is much higher for Russia, i.e. 97%. For N_2O and CH_4 emissions the main source is the Agriculture industry. This industry generates 81.9% and 44.8% of the total emissions of N_2O and CH_4, respectively. Moreover, imports of US final goods are bigger than those of intermediate goods in this industry. As in China, Mining and Quarrying is another relevant emitter of CH_4 gases in the US exports to Spain.

In addition, Brazil is the most polluting country in terms of N_2O and CH_4 emissions coming mainly from the imports of intermediate goods made by the Spanish agricultural industry. France's position is peculiar, since it has a positive ETB in CO_2 and CH_4 and it has, on the other hand, the third largest negative ETB in N_2O emissions; mainly due to the imports of agricultural products (85%) and the imports of chemicals (12%).

Countries with the largest positive emission trade balance in their bilateral trade with Spain are Portugal and France for CO_2, Germany, Italy and United Kingdom (UK) for CH_4 and UK and Portugal for N_2O. In terms of N_2O and CH_4 emissions, Spanish has a surplus in the trade balance of mining and quarrying and agriculture industries. This is mainly due to the fact that the Spanish economy is specialized in exporting agricultural products, while at the same time it does not import large amounts of related natural resources. Exported chemicals products play also a relevant role in terms of N_2O emissions. The same applies to Other Social Services for CH_4 emissions.

CO_2 emissions of Spanish exports (with positive emission trade balance) are spread among several industries but mainly coming from the import demand of France and Portugal (neighboring countries). This demand is concentrated on electricity and demand for intermediate goods of basic and non-metallic minerals.

5. Conclusions

Many studies have addressed the calculation of the GHG footprint of Spain but to our knowledge, none or very few of them has used a homogeneous multi-country IO database, nor has the analysis been carried out with high industry resolution and bilateral country flows as it is done in this paper. Therefore, the originality and interest of

this work lies on the details and the extension of the results in terms of higher industry breakdown, homogeneity of the multi-country database, country coverage and pollutants covered (CO_2, CH_4 and N_2O).

Spain produced 316.6 million tons of CO_2 equivalents in 2008 and its final demand led to 494 million tons of CO_2 equivalents elsewhere in the same year. The emission trade balance of Spain of GHG resulted therefore in -177.7 million tons of CO_2 equivalents. Spain is the fifth EU country with the largest negative emission trade balance, behind Germany, France, United Kingdom and Italy.

Moreover, Spanish exports of final goods and services to France, Germany and UK are those that contribute most to the GHG emissions produced by Spain. On the other hand, the final demand of Spanish residents (GHG footprint) leads to 65.5 million tons of CO_2 equivalent of GHG in China; followed by Russia and US with 20.7 and 13.7 million tons of CO_2 equivalents.

As a result, the largest positive balances are found in France (24 millions of tons of CO_2 equivalents) and Portugal (18.3 millions of tons of CO_2 equivalents), while the largest negative emission trade balances of Spain are found for China, Russia and US. The analysis also gives some details by polluting industry.

Finally, special attention should be devoted to the emissions trade balance between Spain and China. China is the country that produces more CO_2, CH_4 and N_2O emissions due to Spanish imports. In particular, Chinese GHG emissions due to intermediate imported inputs by Spain are much more than those produced for exporting final goods and services to Spain (as in López et al., 2013). This result could be explained by the re-allocation of (less clean) production activities and international supply chains across the world (Cadarso et al., 2012). Interestingly, future work might be focused on whether this trend of re-allocation of production activities to less developed countries will continue in time. Policy options like stimuli of technology transfers and the spread use of cleaner technologies through standard regulations would also be worthwhile to investigate.

Reducing emissions of greenhouse gases (GHG) has become one of the main objectives of the current climate policies of countries. The relative position that countries hold among their main trade partners is also a key issue in terms of international climate negotiations and this paper hopefully contributes to raise the awareness of national statistical institutes and statistical international organizations about the necessary construction of official global multi-country input-output tables that would pave the way for further detailed studies on the economic, social and environmental impacts of globalization and international trade.

Acknowledgements

The first and second authors acknowledge the funding received from the SEJ 132 project of the Andalusian Regional Government, ECO2014-56399-R Project of Spanish Ministry of Economy and Competitiveness and the “Cátedra de Economía de la Energía y

del Medio Ambiente" (Department for Energy Economics and the Environment) at the University of Seville and the "Fundación Roger Torné" (Foundation).

Annex I. Tables

Table A.1: *WIOD Industries and Commodities.*[1]

	WIOD Sectors
1	Agriculture, Hunting, Forestry and Fishing
2	Mining and Quarrying
3	Food, Beverages and Tobacco
4	Textiles and Textile Products
5	Leather, Leather and Footwear
6	Wood and Products of Wood and Cork
7	Pulp, Paper, Paper , Printing and Publishing
8	Coke, Refined Petroleum and Nuclear Fuel
9	Chemicals and Chemical Products
10	Rubber and Plastics
11	Other Non-Metallic Mineral
12	Basic Metals and Fabricated Metal
13	Machinery, Nec
14	Electrical and Optical Equipment
15	Transport Equipment
16	Manufacturing, Nec; Recycling
17	Electricity, Gas and Water Supply
18	Construction
19	Sale, Maintenance and Repair of Motor Vehicles and Motorcycles; Retail Sale of Fuel
20	Wholesale Trade and Commission Trade, Except of Motor Vehicles and Motorcycles
21	Retail Trade, Except of Motor Vehicles and Motorcycles; Repair of Household Goods
22	Hotels and Restaurants
23	Inland Transport
24	Water Transport
25	Air Transport
26	Other Supporting and Auxiliary Transport Activities; Activities of Travel Agencies
27	Post and Telecommunications
28	Financial Intermediation
29	Real Estate Activities
30	Renting of M&Eq and Other Business Activities
31	Public Admin and Defence; Compulsory Social Security
32	Education
33	Health and Social Work
34	Other Community, Social and Personal Services
35	Private Households with Employed Persons

1. Commodities and industries are the same provided that the World IOTs used are square.

Legends to read Tables A.2, A.3 and A.4

A1: Total emissions generated in the country of reference due to Spanish imports of final goods and services (GHG footprints) - $\boldsymbol{g}_{ru}^{imp}$

B1: Cumulated share of A1 over the total amount of emissions (%)

C1: Share of emissions generated in the country of reference for the production of the final goods and services exported to Spain (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rr} \boldsymbol{y}_{ru}$

D1: Share of emissions generated in the country of reference for the production of the intermediate inputs that will be exported to Spain for the domestic production of a final good or service demanded by Spanish residents (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{ru} \boldsymbol{y}_{uu}$

E1: Share of emissions generated in the country of reference for the production of the intermediate inputs that will be exported to a third country for the domestic production of a final good or service to be exported to Spain (%) - $\hat{\boldsymbol{c}}_r \boldsymbol{L}_{rw} \boldsymbol{y}_{wu}$

A2: Total emissions produced in Spain due to imports of the country of reference - $\boldsymbol{g}_{ur}^{exp}$

B2: Cumulated share of A2 over the total amount of emissions (%)

C2: Share of emissions produced in Spain for exports of final goods and services - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uu} \boldsymbol{y}_{ur}$

D2: Share of emissions produced in Spain for exports of intermediate goods and services to the country of reference for the production of final goods in the same country - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{ur} \boldsymbol{y}_{rr}$

E2: Share of emissions produced in Spain for exports of intermediate goods and services to a third country that will use them for the production of goods and services to be exported to the country of reference - $\hat{\boldsymbol{c}}_u \boldsymbol{L}_{uw} \boldsymbol{y}_{wr}$

Table A.2: *Industries with larger CO_2 footprints and commodities. Thousands of tons CO_2, 2008.*

COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE CO_2 EMISSIONS	A1 (Th. tons) g_{ru}^{imp}	B1 (%)	C1 (%) Final	D1 (%) Interm.	E1 (%) Interm.
CHN (-)	TOTAL	50 135	100.0	37.2	49.6	13.2
	TOTAL INDUSTRIES WITH MORE EMISSIONS	39 150	78.1	36.9	49.5	13.6
	Electricity, Gas and Water Supply	26 815	53.5	38.4	48.2	13.4
	Basic Metals and Fabricated Metal	5 826	11.6	32.9	51.7	15.5
	Other Non-Metallic Mineral	3 301	6.6	31.5	58.3	10.2
	Chemicals and Chemical Products	3 208	6.4	36.9	48.0	15.1
RUS (-)	TOTAL	14 450	100.0	4.4	69.7	25.9
	TOTAL INDUSTRIES WITH MORE EMISSIONS	13 482	93.3	4.8	70.5	24.7
	Electricity, Gas and Water Supply	5 850	40.5	5.5	69.2	25.2
	Mining and Quarrying	2 798	19.4	1.6	75.2	23.2
	Inland Transport	2 129	14.7	1.4	72.1	26.5
	Basic Metals and Fabricated Metal	1 868	12.9	3.0	69.4	27.6
	Coke, Refined Petroleum and Nuclear Fuel	836	5.8	22.9	59.7	17.5
DEU (-)	TOTAL	11 170	100.0	35.9	51.6	12.4
	TOTAL INDUSTRIES WITH MORE EMISSIONS	8 779	78.6	33.5	53.6	12.9
	Electricity, Gas and Water Supply	4 174	37.4	39.7	47.6	12.7
	Basic Metals and Fabricated Metal	2 087	18.7	23.3	60.9	15.8
	Chemicals and Chemical Products	1 160	10.4	33.3	53.3	13.4
	Other Non-Metallic Mineral	710	6.4	23.2	67.5	9.3
	Air Transport	647	5.8	38.3	54.1	7.6
USA (-)	TOTAL	10 084	100.0	29.7	50.8	19.5
	TOTAL INDUSTRIES WITH MORE EMISSIONS	7 332	72.7	32.2	49.1	18.7
	Electricity, Gas and Water Supply	2 981	29.6	31.5	47.8	20.8
	Chemicals and Chemical Products	1 256	12.5	44.7	38.0	17.3
	Air Transport	1 043	10.3	28.8	59.2	12.0
	Coke, Refined Petroleum and Nuclear Fuel	843	8.4	35.2	51.1	13.7
	Inland Transport	625	6.2	21.8	56.8	21.5
	Basic Metals and Fabricated Metal	585	5.8	22.0	49.9	28.1
IND (-)	TOTAL	5 178	100.0	35.4	45.2	19.4
	TOTAL INDUSTRIES WITH MORE EMISSIONS	4 095	79.1	68.5	47.7	−16.2
	Electricity, Gas and Water Supply	2 550	49.2	40.7	41.3	17.9
	Basic Metals and Fabricated Metal	687	13.3	56.5	22.8	20.8
	Mining and Quarrying	573	11.1	64.5	10.8	24.7
	Chemicals and Chemical Products	286	5.5	55.4	23.6	21.0

COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE CO_2 EMISSIONS	A2 (Th. tons) g_{ur}^{exp}	B2 (%)	C2 (%) Final	D2 (%) Interm.	E2 (%) Interm.
FRA (+)	TOTAL	8 735	100.0	47.0	46.4	6.6
	TOTAL INDUSTRIES WITH MORE EMISSIONS	7 162	82.0	45.0	48.3	6.8
	Electricity, Gas and Water Supply	2 120	24.3	51.2	41.9	6.9
	Other Non-Metallic Mineral	1 373	15.7	20.3	76.0	3.7
	Coke, Refined Petroleum and Nuclear Fuel	906	10.4	43.7	49.3	7.0
	Basic Metals and Fabricated Metal	905	10.4	37.8	51.9	10.3
	Inland Transport	724	8.3	44.3	47.1	8.6
	Agriculture, Hunting, Forestry and Fishing	608	7.0	86.8	10.0	3.3
	Chemicals and Chemical Products	526	6.0	51.0	40.0	9.0
PRT (+)	TOTAL	4 970	100.0	49.8	48.9	1.3
	TOTAL INDUSTRIES WITH MORE EMISSIONS	4 150	83.5	45.5	53.1	1.4
	Electricity, Gas and Water Supply	1 185	23.8	55.4	43.1	1.5
	Other Non-Metallic Mineral	720	14.5	22.8	76.3	0.9
	Coke, Refined Petroleum and Nuclear Fuel	553	11.1	30.9	68.0	1.1
	Basic Metals and Fabricated Metal	473	9.5	38.0	59.5	2.4
	Agriculture, Hunting, Forestry and Fishing	340	6.8	72.7	26.8	0.5
	Inland Transport	334	6.7	58.1	39.8	2.1
	Chemicals and Chemical Products	288	5.8	42.1	56.1	1.8
	Air Transport	256	5.1	60.2	38.8	1.0

Table A.3: *Industries with larger CH_4 footprints and types of commodities. Tons CH_4, 2008.*

COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE CH_4 EMISSIONS	A1 (tons) g_{ru}^{imp}	B1 (%)	C1 (%) Final	D1 (%) Interm.	E1 (%) Interm.
CHN (-)	TOTAL	542 790	100.0	40.0	48.3	11.7
	TOTAL INDUSTRIES WITH MORE EMISSIONS	534 042	98.4	40.1	48.3	11.6
	Mining and Quarrying	261 244	48.1	34.4	50.6	14.9
	Agriculture, Hunting, Forestry and Fishing	142 116	26.2	56.5	33.0	10.5
	Other Community, Social and Personal Services	130 682	24.1	33.4	60.3	6.2
RUS (-)	TOTAL	287 495	100.0	2.4	73.0	24.7
	TOTAL INDUSTRIES WITH MORE EMISSIONS	274 574	95.5	2.2	73.3	24.5
	Mining and Quarrying	146 779	51.1	1.6	75.2	23.2
	Inland Transport	81 814	28.5	1.4	72.1	26.5
	Electricity, Gas and Water Supply	45 980	16.0	5.5	69.2	25.2
BRA (-)	TOTAL	127 954	100.0	10.2	71.8	18.0
	TOTAL INDUSTRIES WITH MORE EMISSIONS	126 645	99.0	10.2	71.9	17.9
	Agriculture, Hunting, Forestry and Fishing	112 374	87.8	10.7	72.7	16.7
	Mining and Quarrying	8 006	6.3	2.6	70.3	27.0
	Other Community, Social and Personal Services	6 265	4.9	12.0	60.1	27.9
USA (-)	TOTAL	104 801	100.0	35.9	47.8	16.3
	TOTAL INDUSTRIES WITH MORE EMISSIONS	101 962	97.3	36.0	47.8	16.2
	Agriculture, Hunting, Forestry and Fishing	46 960	44.8	47.5	38.8	13.7
	Mining and Quarrying	37 936	36.2	27.8	53.0	19.3
	Other Community, Social and Personal Services	10 578	10.1	23.1	63.9	13.1
	Inland Transport	6 489	6.2	21.8	56.8	21.5
IND (-)	TOTAL	59 613	100.0	33.5	42.0	24.5
	TOTAL INDUSTRIES WITH MORE EMISSIONS	58 575	98.3	33.4	42.0	24.6
	Agriculture, Hunting, Forestry and Fishing	28 941	48.5	44.1	31.1	24.8
	Mining and Quarrying	16 850	28.3	10.8	64.5	24.7
	Other Community, Social and Personal Services	12 784	21.4	39.1	37.0	23.9
COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE CH_4 EMISSIONS	A2 (tons) g_{ur}^{exp}	B2 (%)	C2 (%) Final	D2 (%) Interm.	E2 (%) Interm.
DEU (+)	TOTAL	56 511	100.0	83.3	9.6	7.1
	TOTAL INDUSTRIES WITH MORE EMISSIONS	53 102	94.0	85.7	7.8	6.4
	Agriculture, Hunting, Forestry and Fishing	48 682	86.1	90.2	4.5	5.4
	Other Community, Social and Personal Services	4 419	7.8	37.0	44.9	18.2
ITA (+)	TOTAL	34 812	100.0	70.0	25.0	4.9
	TOTAL INDUSTRIES WITH MORE EMISSIONS	31 434	90.3	73.6	21.7	4.6
	Agriculture, Hunting, Forestry and Fishing	26 808	77.0	80.6	15.2	4.2
	Other Community, Social and Personal Services	4 626	13.3	33.1	59.5	7.4
GBR (+)	TOTAL	35 302	100.0	77.3	14.5	8.3
	TOTAL INDUSTRIES WITH MORE EMISSIONS	32 730	92.7	80.3	11.8	7.9
	Agriculture, Hunting, Forestry and Fishing	28 958	82.0	86.5	6.3	7.2
	Other Community, Social and Personal Services	3 772	10.7	32.5	54.3	13.3

Table A.4: *Industries with larger N_2O footprints and types of commodities. Tons N_2O, 2008.*

COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE N_2O EMISSIONS	A1 (tons) g_{ru}^{imp}	B1 (%)	C1 (%) Final	D1 (%) Interm.	E1 (%) Interm.
CHN (-)	TOTAL	12 652	100.0	51.5	37.7	10.9
	TOTAL INDUSTRIES WITH MORE EMISSIONS	12 268	97.0	52.0	37.2	10.8
	Agriculture, Hunting, Forestry and Fishing	9 561	75.6	56.5	33.0	10.5
	Chemicals and Chemical Products	1 183	9.3	36.9	48.0	15.1
	Other Community, Social and Personal Services	857	6.8	33.4	60.3	6.2
	Electricity, Gas and Water Supply	668	5.3	38.4	48.2	13.4
BRA (-)	TOTAL	6 326	100.0	10.7	72.5	16.8
	TOTAL INDUSTRIES WITH MORE EMISSIONS	6 216	98.3	10.7	72.7	16.7
	Agriculture, Hunting, Forestry and Fishing	6 216	98.3	10.7	72.7	16.7
FRA (-)	TOTAL	5 742	100.0	49.1	44.7	6.2
	TOTAL INDUSTRIES WITH MORE EMISSIONS	5 598	97.5	49.3	44.5	6.2
	Agriculture, Hunting, Forestry and Fishing	4 888	85.1	49.9	44.2	5.9
	Chemicals and Chemical Products	710	12.4	45.1	46.7	8.2
USA (-)	TOTAL	4 523	100.0	45.8	39.7	14.4
	TOTAL INDUSTRIES WITH MORE EMISSIONS	4 259	94.2	47.2	38.7	14.1
	Agriculture, Hunting, Forestry and Fishing	3 704	81.9	47.5	38.8	13.7
	Chemicals and Chemical Products	556	12.3	44.7	38.0	17.3
COUNTRY OF REFERENCE	TOP INDUSTRIES WITH MORE N_2O EMISSIONS	A2 (tons) g_{ur}^{exp}	B2 (%)	C2 (%) Final	D2 (%) Interm.	E2 (%) Interm.
GBR (+)	TOTAL	1 828	100.0	79.9	12.2	7.9
	TOTAL INDUSTRIES WITH MORE EMISSIONS	1 682	92.0	82.7	9.6	7.7
	Agriculture, Hunting, Forestry and Fishing	1 564	85.5	86.5	6.3	7.2
	Chemicals and Chemical Products	118	6.5	32.5	52.8	14.7
PRT (+)	TOTAL	1 713	100.0	70.5	28.9	0.6
	TOTAL INDUSTRIES WITH MORE EMISSIONS	1 589	92.8	69.0	30.4	0.6
	Agriculture, Hunting, Forestry and Fishing	1 475	86.1	72.7	26.8	0.5
	Chemicals and Chemical Products	115	6.7	42.1	56.1	1.8

Annex II. Figures

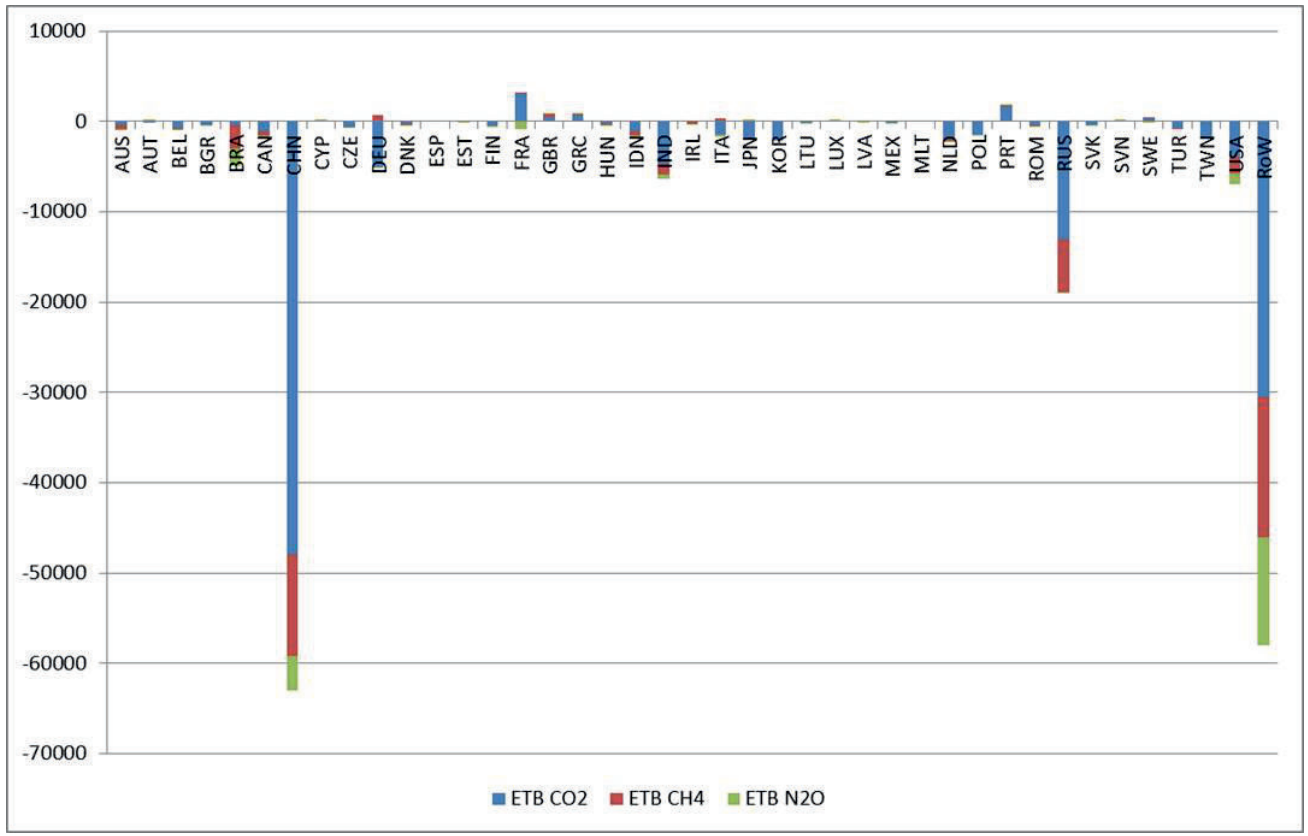

Figure A.1: *GHG Global emission trade balance of Spain. 2008. Thousands of tons of CO_2-equivalents.*

Source: Based on data from WIOD (Dietzenbacher et al., 2013).

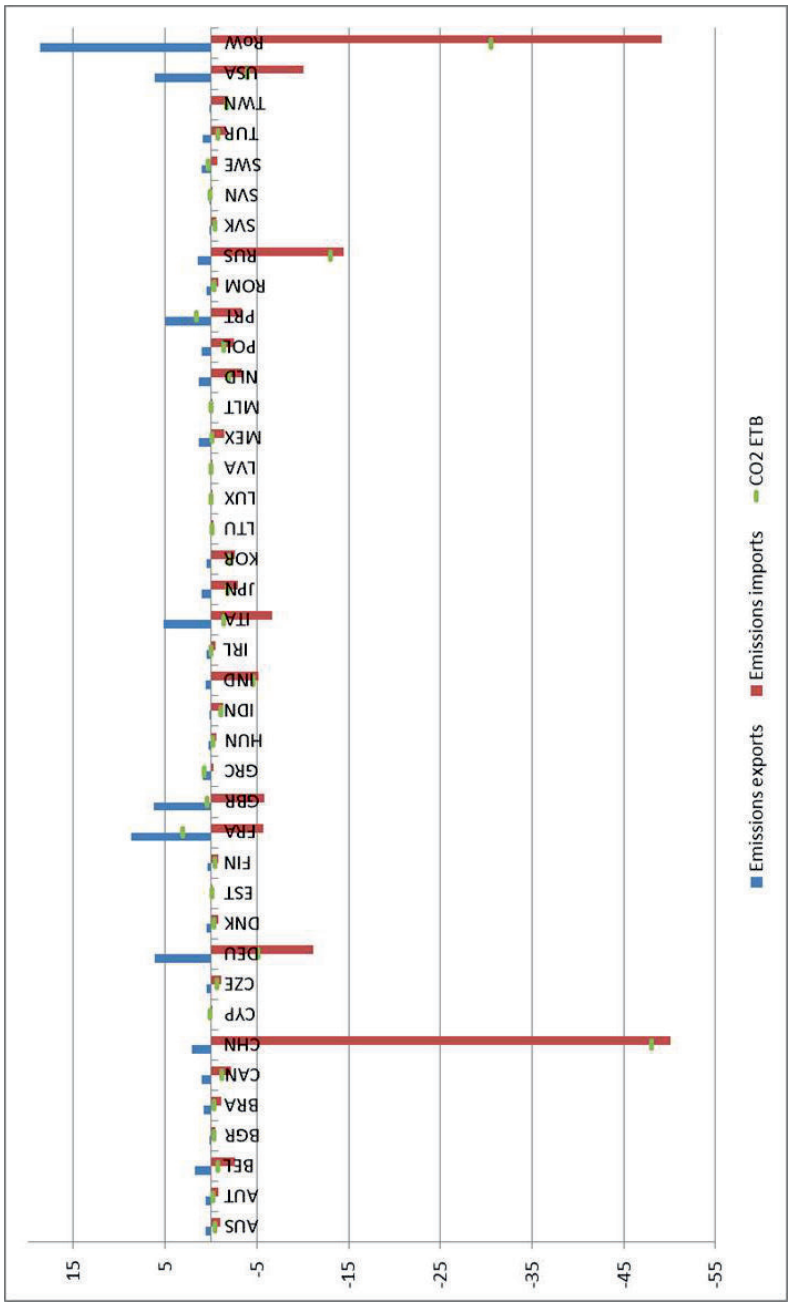

Figure A.2: *Bilateral ETB of CO_2 (Millions of tons) in Spain, 2008, decomposed into emissions associated with imports and exports.*

Source: Based on data from WIOD (Dietzenbacher et al., 2013).

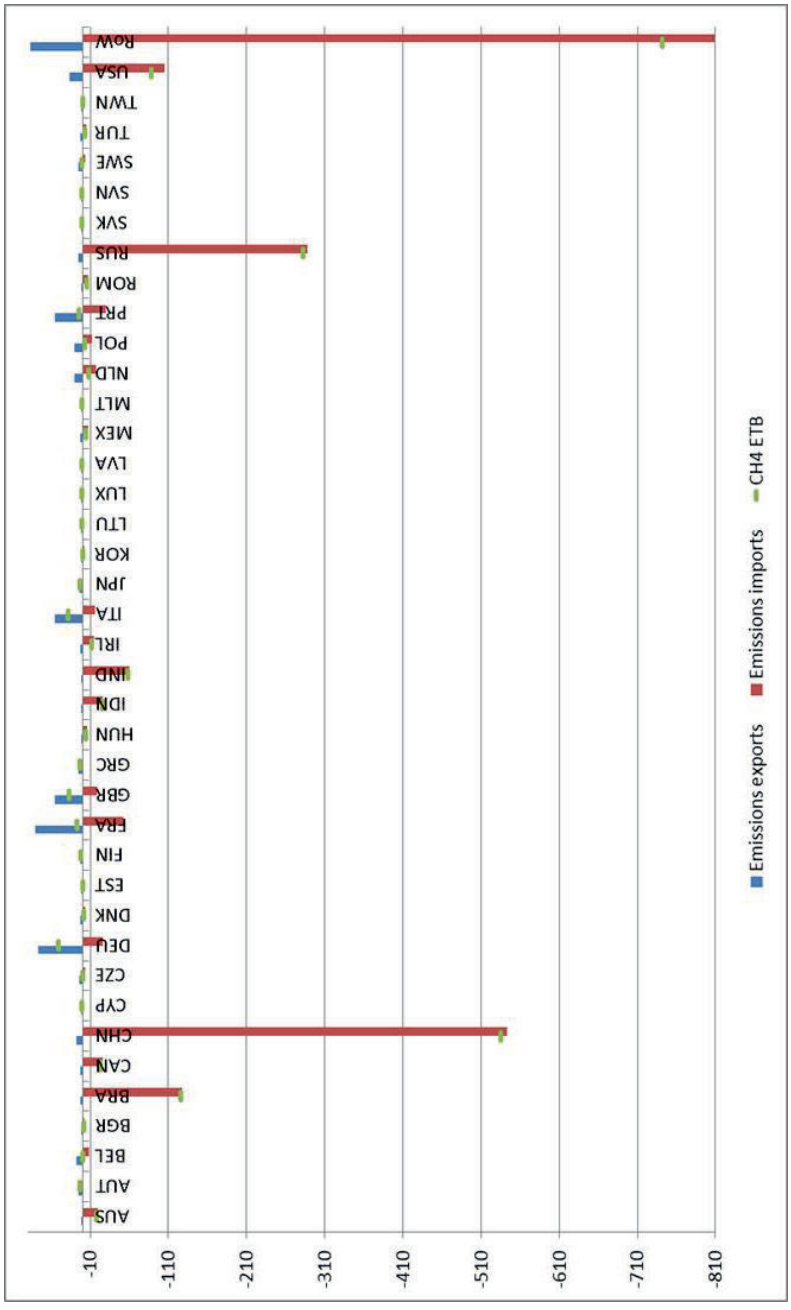

Figure A.3: *Bilateral ETB of CH_4 (Thousands of tons) in Spain, 2008, decomposed into emissions associated with imports and exports.*

Source: Based on data from WIOD (Dietzenbacher et al., 2013).

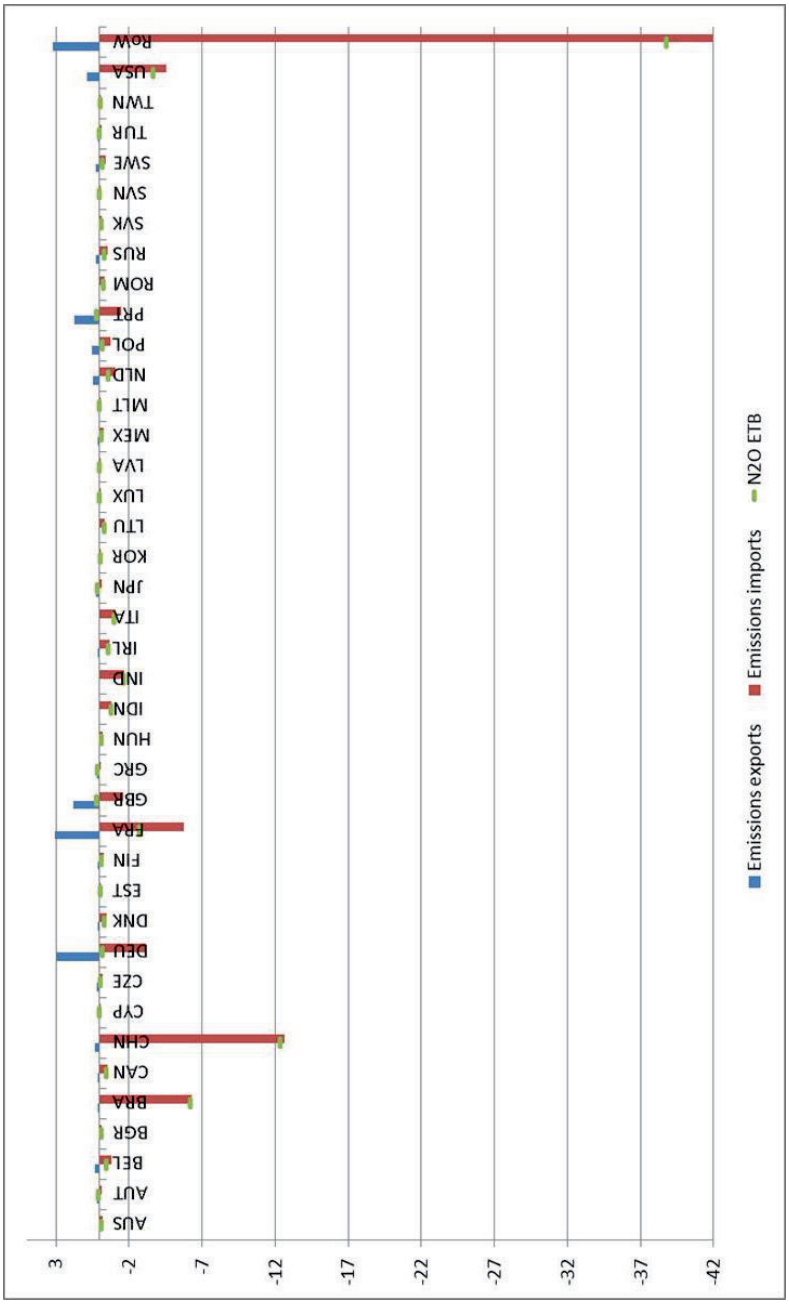

Figure A.4: *Bilateral ETB of N_2O (Thousands of tons) in Spain, 2008, decomposed into emissions associated with imports and exports.*

Source: Based on data from WIOD (Dietzenbacher et al., 2013).

References

Ahmad, N. and Wyckoff, A. (2003). Carbon dioxide emissions embodied in internationaltrade of goods. *OECD Science, Technology and Industry Working Paper*, 15.

Arto, I. (2009). Emisiones de gases efecto invernadero, comercio internacional y hábitos de consumo en España. *Papeles de Economía Española*, 121, 100–111.

Arto, I., Roca, J. and Serrano, M. (2012). Emisiones territoriales y fuga de emisiones. Análisis del caso español. *Revista de la Red Iberoamericana de Economía Ecológica*, 18, 73–87.

Cadarso, M.A., Gómez, N., López, L.A. and Tobarra, M.A. (2008). The EU enlargement and the impact of outsourcing on industrial employment in Spain, 1993-2003. *Structural change and Economics Dynamics*, 19, 95–108.

Cadarso, M.A., López, L.A., Gómez, N., & Tobarra, M.A. (2010). CO_2 emissions of international freight transport and offshoring: Measurement and allocation. *Ecological Economics*, 69, 1682–1694.

Cadarso, MA. and López, L.A. (2012). International trade and shared environmental responsibility by industry. An application to the Spanish Economy. *Ecological Economics*, 83, 221–335.

Cellura, M., Di Gangi, A., Longo, S. and Orioli, A. (2013). An Italian input-output model for the assessment of energy and environmental benefits arising from retrofit actions of buildings. *Energy and Buildings*, 62, 97–106.

Chang, Y., Ries, R. and Wang, Y. (2010). The embodied energy and environmental emissions of construction projects in China: An economic input-output LCA model. *Energy Policy*, 38, 6597–6603.

Chen, G.C. and Zhang, B. (2010a). Greenhouse gas emissions in China 2007: Inventory and input-output analysis. *Energy Policy*, 38, 6180–6193.

Chen, Z.M., Chen, G.C., Zhou, J.B., Jiang, M.M. and Chen, B. (2010b). Ecological input-output modeling for embodied resources and emissions in Chinese economy 2005. *Communications in Nonlinear Science and Numerical Simulation*, 15, 1942–1965.

Dietzenbacher, E., Bart, L. Stehrer, R., Timmer, M. and Vries, G. (2013). The construction of World Input-Output Tables in the WIOD Project. *Economic Systems Research*, 25, 71–98.

Davis, S. and Caldeira, K. (2010). Consumption-based accounting of CO_2 emissions. *PNAS*, 107–12, 5687–5692.

Druckman, A. and Jackson, T. (2009). The carbon footprint of UK households 1990-2004: A socio-economically disaggregated, quasi-multi-regional input-output model. *Ecological Economics*, 66–7, 2066–2077.

Edens, B., Delahaye, R., van Rossum, M. and Schenau, S. (2011). Analysis of changes in Dutch emission trade balance between 1996 and 2007. *Ecological Economics*, 70, 2334–2340.

Kanemoto, K., Lenzen, M., Peters, G. P., Moran, D. D. and Geschke, A. (2012). Frameworks for comparing emissions associated with production, consumption and international trade. *Environmental Science and Technology*, 46, 172–179.

ICEX. http://www.icex.es/icex/cda/controller/pageICEX/0,6558,5518394_5519 205_5548914_0_0_-1,00.html (Last consulted 5.11.13)

Leontief, W. (1970). Environmental repercussions and the economic structure: an input-output approach. *Review of Economics and Statistics*, 52, 262–271.

Liang, Q.M., Fan, Y. and Wei, Y.M. (2007) Multi-regional input-output model for regional energy requirements and CO_2 emissions in China. *Energy Policy*, 35, 1685–1700.

Liang, S., Wang, C. and Zhang, T. (2010). An Improved input-output model for energy analysis. A case study of Suzhou. *Ecological Economics*, 69, 1805–1813.

Liu, H., Dong Qian, J. and Xi, Y. (2009). Comprehensive evaluation of household indirect energy consumption and impacts of alternative energy policies in China by input-output analysis. *Energy Policy*, 37, 3194–3204.

Llop, M. and Tol, R. (2012). Decomposition of industrial greenhouse gas emissions. A subsystem input-output model for the Republic of Ireland. *Journal of Environmental Planning and Management*, 56, 1316–1331.

López, L.A., Arce, G. and Zafrilla, J.E. (2013). Parcelling virtual carbon in the pollution haven hypothesis. *Energy Economics*, 39, 177–186.

López, L.A., Cadarso, M.A., Gómez, N. and Tobarra, M.A. (2015) Food miles, carbon footprint and global value chains for Spanish agriculture: assessing the impact of a carbon border tax. *Journal of Cleaner Production*, 103, 423–436.

Maenpaa, I. and Siikavirta, H. (2007). Greenhouse gases embodied in the international trade and final consumption of Finland: An input-output analysis. *Energy Policy*, 35, 128–143.

Mattila, T., Koskela, S., Seppalla, J. and Maenpaa, I. (2013). Sensitivity analysis of environmentally extended input-output models as a tool for building scenarios of sustainable development. *Ecological Economics*, 86, 148–155.

Miller, R.E. and Blair, P.D. (2009). *Input-Output Analysis, Foundations and Extensions*, 2nd edition. Cambridge University Press, Cambridge.

Minx, J.C., Wiedmann, T., Wood, R., Peters, G.P., Lenzen, M., Owen, A., Scott, K., Barrett, J., Hubacek, K., Baiocchi, G., Paul, A., Dawkins, E., Briggs, J., Guan, D., Suh, S. and Ackerman, F. (2009). Input output analysis and footprinting: an overview of applications. *Economic Systems Research*, 21, 187–216.

Mongelli, I., Tassielli, G. and Notamicola, B. (2006). Global warming agreements, international trade and energy/carbon embodiments: an input-output approach to the Italian case. *Energy Policy*, 34–1, 88–100

Munksgaard, J. and Pedersen, K. (2001). CO_2 accounts for open economies: producer or consumer responsibility? *Energy Policy*, 29, 327–334.

Nansai, K., Kagawa, S., Suh, S., Inaba, R. and Nakajima, K. (2009). Improving the completeness of production carbon footprints using a global link input-output model: the case of Japan. *Economic Systems Research*, 21–3, 267–290.

Peters, G. and Hertwich, E. (2006). Pollution embodied in trade: the Norwegian case. *Global Environmental Change*, 16, 379–389

Peters, G. (2008). From production-based to consumption-based national emission inventories. *Ecological Economics*, 65, 13–23.

Peters, G. Minx, J. Weber, C. and Edenhofer, O. (2011). Growth in emission transfers via international trade from 1990 to 2008. *Proceedings of the National Academy of Sciences*, 108, 8903–8908.

Rueda-Cantuche, J.M. (2011). Comparison of the European Carbon Footprint (2000-2006) from three different perspectives within a Multi-Regional framework: new empirical evidences. In: Costantini, V., Mazzanti, M. and Montini, A. (eds), *Hybrid Economic-Environmental Accounts. Routledge Studies in Ecological Economics*. Oxford, 125–139.

Rueda-Cantuche, J.M. and Amores, A.F. (2010). Consistent and unbiased carbon dioxide emission multipliers: Performance of Danish emission reductions via external trade. *Ecological Economics*, 69, 988–998.

Sánchez-Choliz, J. and Duarte, R. (2004). CO_2 emissions embodied in international trade: evidence for Spain. *Energy Policy*, 32, 1999–2005.

Serrano, M. and Dietzenbacher, E. (2010). Responsibility and trade emission balance: An evaluation of approaches. *Ecological Economics*, 69, 2224–2232.

Serrano, M. and Roca, J. (2008a). Comercio internacional y responsabilidades en las emisiones de gases de efecto invernadero. El caso español 1995-2000. *Economiaz*, 67, 284–301.

Serrano, M. and Roca, J. (2008b). Comercio exterior y contaminación atmosférica en España: un análisis input-output. *Cuadernos aragoneses de economía*, 2nd period, 18, 9–34.

Skelton, A. (2013). EU corporate action as a driver for global emissions abatement: a structural analysis of EU international supply chain carbon dioxide emissions. *Global Environmental Change*, 23, 1795–1806.

Su, B., Huang, H.C., Ang, B.W. and Zhou, P.(2010) Input-output analysis of CO_2 emissions embodied in trade: the effects of industry aggregation. *Energy Economics*, 32, 166–175.

Tarancón, M.A. and Del Riuo, P. (2007). CO_2 emissions and Interindustrial linkages. The case of Spain. *Energy Policy*, 35, 1100–1116.

Tunc, G.P., Asik, S. and Akbostanci, E. (2007). CO_2 responsibility. An input-output approach for the Turkish economy. *Energy Policy*, 35, 855–868.

Wiedmann, T., Wood, R., Minx, J.C., Lenzen, M., Guan, D. and Harris, R. (2010). A carbon footprint time series of the UK-Results from a multi-region input-output model. *Economic Systems Research*, 22–1, 19–42.

Xu, M., Li, R., Critteden, J.C. and Chen, Y. (2011). CO_2 emissions embodied in China's exports from 2002 to 2008: a structural decomposition analysis. *Energy Policy*, 39, 7381–7388.

Zhao, X., Chen, B. and Yang, Z.F. (2009). National water footprint in an input-output framework-a case study of China 2002. *Ecological Modeling*, 220, 245–253.

Zhou, X. and Imura, H. (2011). How does consumer behavior influence regional ecological footprints? An empirical analysis for Chinese regions based on the multi-region input-output model. *Ecological Economics*, 71, 171–179.

Zhu, Q., Peng, X. and Wu, K.(2012). Calculation and decomposition of indirect carbon emissions from residential consumption in China Based on the input-output model. *Energy Policy*, 48, 618–626.

SORT 40 (1) January-June 2016, 31-54

Two alternative estimation procedures for the negative binomial cure rate model with a latent activation scheme

Diego I. Gallardo[1] and Heleno Bolfarine[2]

Abstract

In this paper two alternative estimation procedures based on the EM algorithm are proposed for the flexible negative binomial cure rate model with a latent activation scheme. The Weibull model as well as the log-normal and gamma distributions are also considered for the time-to-event data for the non-destroyed cells. Simulation studies show the satisfactory performance of the proposed methodology. The impact of misspecifying the survival function on both components of the model (cured and susceptible) is also evaluated. The use of the new methodology is illustrated with a real data set related to a clinical trial on Phase III cutaneous melanoma patients.

MSC: 62N01, 62N02, 62P10.

Keywords: Competing risks, EM algorithm, latent activation scheme.

1. Introduction

An implicit assumption with the ordinary survival model is that all individuals under study are susceptible to the event of interest, which is not always true given the improvements in disease treatments experienced in the last decades. For some types of cancer, for example, new treatments have significantly increased the probability that an individual is considered with the disease under control (typically called cured). The proportion of cured individuals after a treatment is usually known as the cure fraction.

[1] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile. diego.gallardo@uantof.cl

[2] Departamento de Estatística, IME, Universidade de São Paulo, São Paulo, Brasil. hbolfar@ime.usp.br

Received: May 2015

Accepted: February 2016

Berkson and Gage (1952) developed a model that became known in the literature as the *mixture model*, which assumes that there is a proportion $1-q_0$ of susceptible individuals and, hence, a proportion q_0 of cured individuals. An alternative route was pursued by Yakovlev and Tsodikov (1996) and Chen et al. (1999). Their approach is based on the assumption that each individual has an unobserved (latent) number M of cells, each capable of triggering the event of interest. This model is known in the literature as the *promotion time cure rate model* and has been the subject of intense research activity. Rodrigues et al. (2009) unify the two approaches considering the negative binomial distribution for the variable M, known in the literature as the *negative binomial cure rate model*. Those models have a common element: both assume that the initial cells will produce the event of interest. In order to relax this assumption, Rodrigues et al. (2012) proposed the so-called destructive weighted Poisson cure rate model in which it is assumed that each one of the initial cells has a probability p of being able to produce the patient's death, so that only $D \leq M$ cells (usually called activated or non-destroyed cells) would remain in effect. Clearly, the case $p = 1$ (i.e., $M = D$) leads to the standard models above. Both destructive and non-destructive models mentioned above assume that one cell is sufficient to produce the event of interest, i.e., the time until the event occurs is considered as the minimum of the times related to each activated cell. This scheme is known as the first activation (FA) scheme.

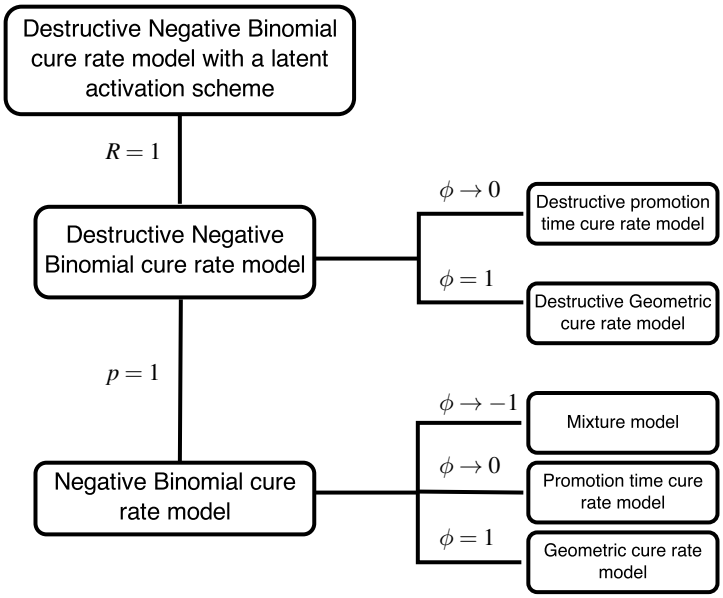

Figure 1: *Summary of some particular cases of the DNB model with a latent activation scheme.*

Cooner et al. (2007) proposed a more general activation scheme in a non-destructive context. This idea was used by Cancho et al. (2013) in the Destructive Negative Bino-

mial (DNB) cure rate model, where the negative binomial distribution with mean θ and variance $\theta(1+\phi\theta)$ is used for the initial number of cells. Accordingly, $\phi > 0 (\phi < 0)$ provides over-dispersion (sub-dispersion), including the Poisson model as particular case for $\phi = 0$. The idea is that the event of interest may be considered as the maximum of the times related to each one of the concurrent cells, i.e., all cells must be activated to produce the event of interest. This scheme is called last activation (LA) scheme. A third activation scheme is proposed assuming that a random number of factors (R) is needed to produce the event of interest, i.e., the time to the event of interest is defined as the R-th order statistics from the times related to the activated cells. A simple specification is to assume the discrete uniform distribution for R on the set $\{1,\dots,D\}$. This scheme is known as the random activation (RA) scheme. Figure 1 depicts a summary of the DNB in Cancho et al. (2013) and some particular cases of the model.

The main focus of this work is to develop two different ways of applying the EM algorithm for maximum likelihood estimation (MLE) for the DNB with different activation schemes. The first way is to compute directly the expected value of M and D, the number of initial and activated cells, respectively, and the second way is to write the model as a *mixture model* and to use the EM algorithm for this alternative version Lu (2010).

The paper is organized as follows. In Section 2 we describe the cutaneous melanoma data set. In Section 3, the DNB model with different activation schemes and some propositions about this model are stated. In Section 4, two estimation procedures based on the EM algorithm are proposed for the model in Section 3. Section 5 reveals results of two simulation studies aiming at investigating parameters recovery and assessing the time-to-event for the non-destroyed cells. Section 6 presents an application to a real data set referring to a clinical trial for patients with melanoma. Finally, in Section 7, the main conclusions and results obtained in this work are presented.

2. Cutaneous melanoma data set

The data set is related to a clinical trial on a Phase III cutaneous melanoma patients available at `http://merlot.stat.uconn.edu/~mhchen/survbook/`, labeled as E1690 data. The clinical trial was conducted by the Eastern Cooperative Oncology Group (see Ibrahim et al. (2001) for details). The incidence of melanoma is one of the highest among most types of cancer, with a high mortality rate even with early detection. The objective of this study was to evaluate a postoperative treatment performance with a high dose of the drug Interferon alpha-2b, in order to prevent recurrence. The study included patients between 1991 to 1995 and follow-up was conducted until 1998.

A characteristic of the disease (as in many other types of cancers) is the presence of a proportion of patients that can lead a normal life, comparable to patients without the disease. In other words, a proportion commonly known as "cured". After deleting patients with incomplete data and missing observation times, the data set is composed of

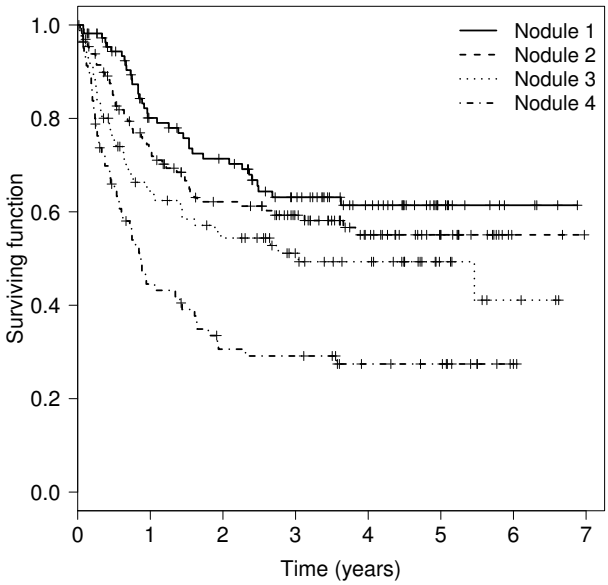

Figure 2: *Kaplan-Meier curves stratified by covariate Nodule.*

$n = 408$ individuals. The collected variables were: Observed time (in years, average = 2.31, median = 1.64, standard deviation = 1.93), treatment (0: control and 1: interferon alfa-2b with 198 and 210 patients respectively), age (in years, average = 48.1, median = 47.2 and standard deviation = 13.1), nodal category (categorical variable with levels 1-4 with 110, 131, 86 and 81 patients in each group, respectively, where 1 indicates the lower risk patients and 4 the higher risk patients) and tumour thickness (in mm, average = 3.98, median = 3.18 and standard deviation = 3.22).

Figure 2 depicts the Kaplan-Meier curves by nodule category, confirming a well pronounced plateau in all nodule categories. In the next Section, we present the model addressed for this particular problem.

3. Model specification

Following Cancho et al. (2013), let M be an unobservable random variable denoting the initial number of competing causes related to the occurrence of the event of interest. For the cutaneous melanoma data set, M represents the number of carcinogenic cells. Assume that M has negative binomial distribution with probability mass function (p.m.f.) given by

$$P(M = m;\theta,\phi) = \frac{\Gamma(\phi^{-1}+m)}{\Gamma(\phi^{-1})m!}\left(\frac{\phi\theta}{1+\phi\theta}\right)^{m}(1+\phi\theta)^{-1/\phi}, \quad m = 0,1,2,\ldots, \qquad (1)$$

where $\theta > 0$, $\phi \geq -1$ and $1 + \phi\theta > 0$. The distribution in (1) is denoted as $M \sim NB\left(\phi, \frac{\phi\theta}{1+\phi\theta}\right)$. Under this parametrization, $\mathbb{E}(M) = \theta$ and $\text{Var}(M) = \theta(1+\phi\theta)$. For this reason, $\phi > 0$ ($\phi < 0$) corresponds to over (under)-dispersion in relation to the Poisson distribution. For $\phi \to 0$, the p.m.f. in (1) is reduced to the p.m.f. of the Poisson distribution and $\phi = 1$ corresponds to the geometric distribution with parameter $1/(1+\theta)$.

Let ζ_j, $j = 1, \ldots, M$ be (conditionally) independent random variables given $M = m$, with Bernoulli distribution and success probability p indicating whether the j-th concurrent cause can produce or not the event. Contextualizing to the medical problem under study, $\zeta_j = 1$ ($\zeta_j = 0$) indicates that the j-th carcinogenic cell was (was not) activated or non-destroyed (destroyed), and each activated carcinogenic cell can produce the metastasis process. The (unobservable) total damaged D is defined as

$$D = \begin{cases} \zeta_1 + \ldots + \zeta_M & \text{, if } M > 0, \\ 0 & \text{, if } M = 0. \end{cases}$$

Note that D represents the total number of activated carcinogenic cells (among the M initials) which are activated. It is immediate that $D \mid M = m \sim Bin(m, p)$ for $m > 0$ and $P(D = 0 \mid M = 0) = 1$. Moreover, it is possible to show that $D \sim NB\left(\phi, \frac{\phi\theta p}{1+\phi\theta p}\right)$ Rodrigues et al. (2011). Define W_j, $j = 1, \ldots, D$ as the time to event for the j-th activated cell produces the metastasis process. Assume that W_j, $j = 1, \ldots, D$, are conditionally independent and identically distributed (*i.i.d.*) given D with common cumulative distribution function $F(\cdot;\lambda)$ and survival function $S(\cdot;\lambda) = 1 - F(\cdot;\lambda)$. Further, assume that $W_1, W_2, \ldots$, are independent of D and M. As discussed in Cooner et al. (2007), cure rate models with latent activation schemes assume that the failure time T^* is generated by the activation times of D latent factors. Thus, $D = 0$ implies $T^* = \infty$ and then the individual is considered cured. If $D > 0$ and it is assumed that R among the D cells are required to produce the event of interest, so the failure time to event is defined by $T^* = W_{(R)}$, where R depends (or not) on D and $W_{(R)}$ denotes the R-th order statistics corresponding to $W_1, \ldots, W_D$.

Assume that the data can be censored to the right. Thus, the observed data can be represented by $T = \min(T^*, C)$ and $\delta = I(T^* \leq C)$, with T^* and C denoting failure and censoring times, respectively, and $I(\cdot)$ the indicator function. Under this scheme and following similar arguments in Cooner et al. (2007), we can write the joint distribution of (T, δ, R, M, D) as

$$\begin{aligned} f(t,\delta,r,m,d;\theta,\phi,p,\lambda) =& f(t,\delta \mid D = d, R = r, \lambda) P(R = r \mid D = d) \times \\ & \times P(D = d \mid M = m; p) P(M = m, \phi, \theta), \end{aligned} \tag{2}$$

where $D \mid M = m; p \sim Bin(m, p)$, $P(M = m; \theta, \phi)$ is given in (1) and

$$f(t,\delta \mid D=d,R=r,\lambda) = \Big\{I(d=0)+I(m\geq d\geq r\geq 1)IB(S(t;\lambda),d-r+1,r)\Big\}^{1-\delta} \times \left\{d\binom{d-1}{r-1}f(t;\lambda)S(t;\lambda)^{d-r}F(t;\lambda)^{r-1}\right\}^{\delta} \quad (3)$$

with $IB(z,a,b)$ denoting the incomplete beta function defined as $IB(z,a,b)=\int_0^z u^{a-1}(1-u)^{b-1}du$. The population survival and density functions can be computed as

$$S_{pop}(t;\theta,\phi,p,\lambda) = P(D=0;\theta,\phi,p)+\sum_{m=1}^{\infty}\sum_{d=1}^{m}\sum_{r=1}^{d} f(t,\delta=0,r,m,d;\theta,\phi,p,\lambda)$$

$$f_{pop}(t;\theta,\phi,p,\lambda) = \sum_{m=1}^{\infty}\sum_{d=1}^{m}\sum_{r=1}^{d} f(t,\delta=1,r,m,d;\theta,\phi,p,\lambda)$$

It is immediate that $q_0 = S_{pop}(\infty;\theta,\phi,p,\lambda) = (1+\phi\theta p)^{-1/\phi}$, so that the cure rate does not depend on the choice of the (conditional) distribution of $R \mid D=d$.

Moreover, to contour the identifiability problems in the sense of Li et al. (2001) and Hanin and Huang (2014) and discussed in Rodrigues et al. (2011) in the context of the destructive weighted Poisson cure rate models, it is necessary to introduce a set of covariates z_{1i} (of dimension r_1) associated with the initial number of cells and z_{2i} (of dimension r_2) related to the activation probabilities for non-destroyed cells by

$$\log\theta_i = z_{1i}^{\top}\beta_1 \quad \text{and} \quad \log\left(\frac{p_i}{1-p_i}\right) = z_{2i}^{\top}\beta_2, \quad i=1,\ldots,n. \quad (4)$$

In addition, z_1 and z_2 shall not simultaneously include intercepts nor share common elements. Henceforth, in order to simplify the notation, define $\psi=(\beta_1,\beta_2,\phi,\lambda)$ as the vector of parameters to be estimated. Three typically used activation schemes are the random activation scheme (RA), first activation scheme (FA) and last activation scheme (LA), for which the p.m.f. for the conditional distribution $P(R=r \mid D=d)$ and the population survival function for DNB are given in Table 1. Those models are denoted by DNB-FA, DNB-LA and DNB-RA, respectively.

Table 1: *Conditional distribution of R given D = d for three activation schemes with DNB.*

Activation scheme	$P(R=r \mid D=d)$	$S_{pop}(t;\psi)$
RA	$\frac{1}{d}I(1\leq r\leq d)$	$q_0+\{1-q_0\}S(t;\lambda).$
FA	$I(r=1)$	$\{1+\phi\theta pF(t;\lambda)\}^{-1/\phi}$
LA	$I(r=d)$	$1+q_0-\{1+\phi\theta pS(t;\lambda)\}^{-1/\phi}$

Under the usual assumptions in survival analysis and right censoring (see Williams and Lagarkos, 1977), the contribution to the (observed) log-likelihood by the i-th individual is given by

$$f(t_i,\delta_i;\psi) = f_{pop}(t_i;\psi)^{\delta_i} S_{pop}(t_i;\psi)^{1-\delta_i}. \tag{5}$$

Based on (2) and (5), the following propositions are now stated.

Proposition 1 *For combinations DNB-FA and DNB-LA it follows that, given D_{obs}, the conditional distribution of R_i degenerates in the distribution of $R_i = 1$ and $R_i = D_i$ respectively. For the combination DNB-RA, that distribution is*

$$P(R_i = r_i \mid D_{obs};\psi) = \begin{cases} \dfrac{\sum_{k=0}^{r_i-1}\binom{r_i-1}{k}(-1)^k \mathbb{E}\left[\frac{S(t_i;\lambda)^{D_i-r_i+k+1}}{D_i(D_i-r_i+k+1)} I(D_i\geq r_i)\right]}{q_{0i}+(1-q_{0i})S(t_i;\lambda)} & , \text{if } \delta_i = 0 \\ \dfrac{F(t_i;\lambda)^{r_i-1}\mathbb{E}\left[\binom{D_i-1}{r_i-1} S(t_i;\lambda)^{D_i-r_i} I(D_i\geq r_i)\right]}{1-q_{0i}} & , \text{if } \delta_i = 1, \end{cases}$$

where $D_i \sim NB\left(\phi, \frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)$, and $r_i = 1,2,\ldots$.

Proof of proposition 1 is presented in the Appendix A.

Proposition 2 *For DNB in (2) and FA and LA schemes in Table 1, $P(D_i = d_i \mid D_{obs};\psi)$, $i = 1,\ldots,n$, have a closed form. Moreover, for the model DNB-FA,*

$$D_i - \delta_i \mid D_{obs};\psi \sim NB\left((\phi^{-1}+\delta_i)^{-1}, \frac{\phi\theta_i p_i S(t_i;\lambda)}{1+\phi\theta_i p_i}\right),$$

and for the DNB-LA

$$D_i - \delta_i \mid D_{obs};\psi \sim \begin{cases} NB\left((\phi^{-1}+1)^{-1}, \frac{\phi\theta_i p_i F(t_i;\lambda)}{1+\phi\theta_i p_i}\right) & , \text{if } \delta_i = 1, \\ a_i NB\left(\phi, \frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right) + (1-a_i) NB\left(\phi, \frac{\phi\theta_i p_i F(t_i;\lambda)}{1+\phi\theta_i p_i}\right) & , \text{if } \delta_i = 0. \end{cases}$$

where $a_i = [1+q_{0i}-(1+\phi\theta_i p_i S(t_i;\lambda))^{-\phi^{-1}}]^{-1}$. For the DNB-RA combination, the conditional distribution is

$$P(D_i = d_i \mid D_{obs};\psi) = \begin{cases} \frac{\sum_{r_i=1}^{d_i}\sum_{k=0}^{r_i}(-1)^k\binom{r_i-1}{k}\frac{S(t_i;\lambda)^{d_i-r_i+k+1}}{d_i(d_i-r_i+k+1)}\frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)^{d_i}}{1+[(1-q_{0i})/q_{0i}]S(t_i;\lambda)} & , \text{if } \delta_i = 0 \\ \frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)^{d_i} I(d_i \geq 1) & , \text{if } \delta_i = 1, \end{cases}$$

Proposition 2 is proved in Appendix B.

Proposition 3 *For DNB in (2) and FA and LA schemes in Table 1, $P(M_i = m_i \mid D_{obs};\psi)$, $i = 1,\ldots,n$, have a closed form. Moreover, for the DNB-FA combination, we have*

$$M_i - \delta_i; D_{obs}, \psi \sim NB\left((\phi^{-1}+\delta_i)^{-1}, \frac{\phi\theta_i(1-p_iF(t_i;\lambda))}{1+\phi\theta_i}\right),$$

and for the DNB-LA,

$$M_i - \delta_i \mid D_{obs};\psi \sim \begin{cases} NB\left((\phi^{-1}+1)^{-1}, \frac{\phi\theta_i(1-p_iS(t_i;\lambda))}{1+\phi\theta_i}\right) & , \text{if } \delta_i = 1, \\ a_iNB\left(\phi, \frac{\phi\theta_i}{1+\phi\theta_i}\right) + (1-a_i)NB\left(\phi, \frac{\phi\theta_i(1-p_iS(t_i;\lambda))}{1+\phi\theta_i}\right) & , \text{if } \delta_i = 0. \end{cases}$$

where $a_i = [1+q_{0i}-(1+\phi\theta_i p_i S(t_i;\lambda))^{-\phi^{-1}}]^{-1}$. For the DNB-RA and $\delta_i = 0$ this conditional distribution is

$$P(M_i = m_i \mid D_{obs},\psi) = \frac{\sum_{d_i=0}^{m_i}\sum_{r_i=1}^{d_i}\sum_{k=0}^{r_i} v_i\left(\frac{p_i}{1-p_i}\right)^{d_i}\left(\frac{\phi\theta_i(1-p_i)}{1+\phi\theta_i}\right)^{m_i}}{1+[(1-q_{0i})/q_{0i}]S(t_i;\lambda)},$$

where $v_i = (-1)^k\binom{r_i-1}{k}\frac{S(t_i;\lambda)^{d_i-r_i+k+1}}{d_i(d_i-r_i+k+1)}\frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})d_i!(m_i-d_i)!}$. On the other hand, for $\delta_i = 1$ we have that

$$P(M_i = m_i \mid D_{obs},\psi) = \frac{[1-(1-p_i)^{m_i}]\frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!}\left(\frac{\phi\theta_i}{1+\phi\theta_i}\right)^{m_i}(1+\phi\theta_i)^{-1/\phi}I(m_i \geq 1)}{1-q_{0i}},$$

Proof of proposition 3 is presented in Appendix C.

Propositions 1-3 are very useful because they allow predicting the initial number of cells, the number of non-destroyed cells and the number of cells necessary to produce the event of interest in each individual. Moreover, they are useful in implementing the EM algorithm, to be discussed now.

Note that the complete log-likelihood function is given by

$$\ell(\psi \mid D_{comp}) = \sum_{i=1}^{n} f(t_i, \delta_i, R_i, M_i, D_i; \psi), \tag{6}$$

with $f(t_i, \delta_i, r_i, m_i, d_i; \psi)$ defined in (2). Specifically, for the DNB-FA the expression in (6), unless to a constant, assumes the form

$$\ell(\psi \mid D_{comp}) = \sum_{i=1}^{n} \Bigg[(D_i - \delta_i) \log S(t_i; \lambda) + \delta_i \log f(t_i; \lambda) + D_i \log(p_i) + M_i \log \theta_i + (M_i - D_i) \log(1 - p_i) + (M_i - \phi^{-1}) \log(1 + \phi\theta_i) \Bigg]. \tag{7}$$

From (7), it is simple to deduce that it is only necessary the expectations of M_i and D_i (given D_{obs}) to implement the E-step of the EM algorithm. Using Propositions 2 and 3, these expectations are

$$\mathbb{E}(M_i \mid D_{obs}; \beta_1, \beta_2, \phi, \lambda) = \delta_i + \frac{(1 + \phi\delta_i)\theta_i(1 - p_i F(t_i; \lambda))}{1 + \phi\theta_i p_i F(t_i; \lambda)} \quad \text{and} \tag{8}$$

$$\mathbb{E}(D_i \mid D_{obs}; \beta_1, \beta_2, \phi, \lambda) = \delta_i + \frac{(1 + \phi\delta_i)\theta_i p_i S(t_i; \lambda)}{1 + \phi\theta_i p_i F(t_i; \lambda)}. \tag{9}$$

On the other hand, the expression (6) for the DNB-LA assumes the form

$$\ell(\psi \mid D_{comp}) = \sum_{i=1}^{n} \Bigg[(1 - \delta_i) \log\left(1 - I(D_i \geq 1) F(t_i; \lambda)^{D_i}\right) + \delta_i \Big(\log D_i + \log f(t_i; \lambda) + (D_i - 1) \log F(t_i; \lambda) \Big) + D_i \log p_i + (M_i - D_i) \log(1 - p_i) + M_i \log \theta_i + (M_i - \phi^{-1}) \log(1 + \phi\theta_i) \Bigg]. \tag{10}$$

However, the expectation of $\log\left(1 - I(D_i \geq 1) F(t_i; \lambda)^{D_i}\right)$ does not have a closed form, hindering the application of the EM algorithm in this way. Finally, using a RA scheme the log-likelihood function of the model is even more complex, making it difficult the implementation of the EM algorithm in this form. For this reason, a second way is proposed to perform the estimation procedure in those models.

Following Tsodikov et al. (2003) and Rodrigues et al. (2009), all cure rate models can be expressed as a mixture model, i.e.,

$$S_{pop}(t;\psi) = q_0 + (1-q_0)S^*(t;\psi), \tag{11}$$

where $S^*(t;\psi)$ represents the survival function for susceptible individuals and q_0 is the cure rate. Table 2 presents this function for the three activation schemes considered in this work.

Table 2: *Survival and hazard functions for susceptible individuals for the DNB mixture model with three activation schemes.*

Act. Scheme	RA	FA	LA
$S^*(t;\psi)$	$S(t;\lambda)$	$\frac{(1+\phi\theta pF(t;\lambda))^{-1/\phi} - q_0}{1-q_0}$	$\frac{1-(1+\phi\theta pS(t;\lambda))^{-1/\phi}}{1-q_0}$
$h^*(t;\psi)$	$h(t;\lambda)$	$\frac{\theta pf(t;\lambda)(1+\phi\theta pF(t;\lambda))^{-1/\phi-1}}{(1+\phi\theta pF(t;\lambda))^{-1/\phi} - q_0}$	$\frac{\theta pf(t;\lambda)(1+\phi\theta pS(t;\lambda))^{-1/\phi-1}}{1-(1+\phi\theta pS(t;\lambda))^{-1/\phi}}$

Let Y_i the binary variable that indicates whether the individual is susceptible or cured ($Y_i = 1$ and $Y_i = 0$, respectively). Following Lu (2010), the complete log-likelihood function for this model is

$$\ell_c(\psi) = \sum_{i=1}^{n} \left[Y_i \log(1-q_{0i}) + (1-Y_i)\log q_{0i} + Y_i \log S^*(t_i;\psi) + \delta_i Y_i \log h^*(t_i;\psi) \right], \tag{12}$$

and the expected value for Y_i given D_{obs} is

$$\mathbb{E}(Y_i \mid D_{obs};\psi) = \delta_i + (1-\delta_i)\frac{(1-q_{0i})S^*(t_i;\psi)}{q_{0i} + (1-q_{0i})S^*(t_i;\psi)}. \tag{13}$$

Equations (12) and (13) provides a second way to implement the EM algorithm in any cure rate model, in particular, for the DNB with different activation schemes.

4. Estimation

In this Section it is discussed some inferential procedures for the parameters of the DNB with the three activation schemes discussed in Section 3. Parameter estimation is approached using the maximum likelihood method.

In Cancho et al. (2013), the estimation procedure was based on the direct maximization of the observed likelihood function given by

$$\ell(\psi \mid D_{obs}) = \sum_{i=1}^{n} \Big[\log S_{pop}(t_i;\psi) + \delta_i \log h_{pop}(t_i;\psi) \Big], \tag{14}$$

where $S_{pop}(\cdot)$ and $h_{pop}(\cdot)$ depend on the activation scheme used in Table 1. However, maximization of (14) is not simple because it is a function that involves all parameters.

The EM algorithm Dempster et al. (1977) is a very popular maximization alternative used to obtain the maximum likelihood estimators when the model has missing data. A further discussion about the EM algorithm in comparison with the direct maximization of the log-likelihood function is performed in MacDonald (2014). In the cure rate context, we found many recent works using this algorithm. For instance, Balakrishnan and Pal (2012, 2013, 2015) and Gallardo et al. (2016). Two different ways of applying this algorithm in the model considered will be presented in next subsection.

4.1. EM algorithm: implementation 1

Consider initially only the combination DNB-FA, i.e., $R = 1$. Moreover, it is assumed that ϕ is fixed. The first way to apply the EM algorithm in this model is to compute the expected values for M_i and D_i, $i = 1,\dots,n$ given D_{obs} and the parameters values in last iteration, namely $\psi^{(k-1)}$. Those values are denoted by $\widetilde{D}_i^{(k)}$ and $\widetilde{M}_i^{(k)}$, respectively, and they can be computed using equations (8) and (9). Then, it is necessary to replace those values in the complete log-likelihood function given in (7) and maximize it in relation to ψ. The algorithm is summarized as follows.

- **E-step**: For $i = 1,\dots,n$, compute

$$\widetilde{D}_i^{(k)} = \delta_i + \frac{(1+\phi\delta_i)\theta_i^{(k-1)} p_i^{(k-1)} S(t_i;\lambda^{(k-1)})}{1+\phi\theta_i^{(k-1)} p_i^{(k-1)} F(t_i;\lambda^{(k-1)})} \quad \text{and}$$

$$\widetilde{M}_i^{(k)} = \delta_i + \frac{(1+\phi\delta_i)\theta_i^{(k-1)} (1-p_i^{(k-1)} F(t_i;\lambda^{(k-1)}))}{1+\phi\theta_i^{(k-1)} p_i^{(k-1)} F(t_i;\lambda^{(k-1)})}.$$

- **M-step**: Given $\widetilde{D}^{(k)} = (\widetilde{D}_1^{(k)},\dots,\widetilde{D}_n^{(k)})$ and $\widetilde{M}^{(k)} = (\widetilde{M}_1^{(k)},\dots,\widetilde{M}_n^{(k)})$, find $\beta_1^{(k)}$, $\beta_2^{(k)}$ and $\lambda^{(k)}$ that maximize $Q_1(\beta_1 \mid \psi^k)$, $Q_2(\beta_2 \mid \psi^k)$ and $Q_3(\lambda \mid \psi^k)$ in relation to β_1, β_2 and λ, respectively, where

$$Q_1(\beta_1 \mid \psi^{(k)}) = \sum_{i=1}^{n} \left\{ \widetilde{M}_i^{(k)} \log\theta_i - \theta_i + \left(\widetilde{M}_i^{(k)} - \phi^{-1}\right) \log(1+\phi\theta_i) \right\}, \tag{15}$$

$$Q_2(\beta_2 \mid \psi^{(k)}) = \sum_{i=1}^{n} \left\{ \widetilde{D}_i^{(k)} \log(p_i) + \left(\widetilde{M}_i^{(k)} - \widetilde{D}_i^{(k)}\right) \log(1-p_i) \right\}, \tag{16}$$

$$Q_3(\lambda \mid \psi^{(k)}) = \sum_{i=1}^{n} \left\{ \left(\widetilde{D}_i^{(k)} - \delta_i\right) \log S(t_i;\lambda) + \delta_i \log f(t_i;\lambda) \right\}. \tag{17}$$

Then, define $\psi^{(k)} = \left(\beta_1^{(k)}, \beta_2^{(k)}, \lambda^{(k)}\right)$. The advantage of this approach is that functions in (15), (16) and (17) can be maximized separately with respect to β_1, β_2 and λ, respectively, instead of the joint maximization as occurs with the observed log-likelihood. Steps M and E are repeated until a suitable convergence rule is satisfied. For instance, $||\psi^{(k)} - \psi^{(k-1)}|| < \epsilon$, where $||\psi^{(k)} - \psi^{(k-1)}||$ represents the euclidian distance between $\psi^{(k)}$ and $\psi^{(k-1)}$ and ϵ is a prefixed value. For instance, we use $\epsilon = 0.0001$.

4.2. EM algorithm: implementation 2

For this approach, three activation schemes are considered in Table 1. As discussed in Section 3, models DBN-FA, DBN-LA and DBN-RA can be expressed as the mixture model with survival function for susceptible individuals given by $S^*(\cdot \mid \psi)$, according to Table 2, and cure rate given by $q_{0i} = (1 + \phi\theta_i p_i)^{-1/\phi}$ that is common for the three models.

Proceeding similarly as in the last procedure, the algorithm is summarized next.

- **E-step**: For $i = 1, \ldots, n$, compute

$$\widetilde{Y}_i^{(k)} = \delta_i + (1 - \delta_i)\frac{(1 - q_{0i}^{(k-1)})S^*\left(t_i; \psi^{(k-1)}\right)}{q_{0i}^{(k-1)} + (1 - q_{0i}^{(k-1)})S^*\left(t_i; \psi^{(k-1)}\right)}.$$

- **M-step**: Given $\widetilde{Y}^{(k)} = (\widetilde{Y}_1^{(k)}, \ldots, \widetilde{Y}_n^{(k)})$, find $\psi^{(k)}$ that maximizes

$$Q(\psi) = \sum_{i=1}^{n} \left[\widetilde{Y}_i^{(k)} \log(1 - q_{0i}^{(k)}) + (1 - \widetilde{Y}_i^{(k)}) \log q_{0i}^{(k)} + \widetilde{Y}_i^{(k)} \log S^*(t_i; \psi^{(k)}) + \delta_i \widetilde{Y}_i^{(k)} \log h^*(t_i; \psi^{(k)}) \right].$$

Then, steps M and E are repeated until a suitable convergence rule is satisfied. The advantage of this approach in relation to directly maximizing the observed log-likelihood in (14) is that the latent variables Y_i are completely observed for individuals with failure times because $\delta_i = 1$ implies $Y_i = 1$ (i.e., a failure time guarantees that the individual is susceptible). This information is lost when an approach based on the observed log-likelihood is used because the vector $Y = (Y_1, \ldots, Y_n)$ is removed when summing over $\{0,1\}^n$. Consequently, implementing the M-step, for fixed β_1 and β_2, which consists in maximizing the function $Q(\cdot)$ with respect to λ is easier than maximizing the observed log-likelihood function in (14). Thus, it seems more advantageous to use the EM algorithm over than a direct maximization of the observed log-likelihood function. Note

Table 3: *Distributions used for modelling the survival function of the non-destroyed cells.*

Distribution	$S(w;\lambda)$	$f(w;\lambda)$
Weibull	$\exp(-e^{\alpha}w^{\nu})$	$\nu w^{\nu-1}\exp(\alpha-e^{\alpha}w^{\nu})$
LN	$1-\Phi\left(\frac{\log(w)-\alpha}{\nu}\right)$	$\frac{1}{\nu w}\phi\left(\frac{\log(w)-\alpha}{\nu}\right)$
Gamma	$1-\frac{\gamma(\alpha,\nu w)}{\Gamma(\alpha)}$	$\frac{\nu^{\alpha}}{\Gamma(\alpha)}w^{\alpha-1}e^{-\nu w}$

NOTE: $\phi(\cdot)$ and $\Phi(\cdot)$ represent the density and the cumulative function of standard normal distribution. $\gamma(\cdot,\cdot)$ represents the lower incomplete gamma function.

that the EM procedures does not depend on a specific survival function considered for non-destroyed cells. In this work, it is used the Weibull, log-normal (LN) and gamma distributions with parametrizations in Table 3, where $\lambda=(\alpha,\nu)$.

Henceforth, the distribution of $S(\cdot \mid \lambda)$ will be specified jointly with the activation mechanism. For instance, DNB-FA/Weibull, DNB-LA/LN, DNB-RA/gamma, etc. Note that the asymptotic variances for the MLEs could be estimated using the inverse of the Hessian matrix (matrix of second derivatives of the log-likelihood function). The observed information matrix is then obtained from the Hessian matrix evaluated in the MLEs. The elements of the Hessian matrix are presented in the Appendix of Cancho et al. (2013) with the Weibull model considered for the times of the non-destroyed cells. Expressions relatives for the LN and gamma models will not be presented because they are slight modifications for the Weibull model.

Remark 1

1. *In the first version of the EM algorithm, it is assumed that ϕ is fixed. However, it is possible to relax this assumption by constructing a profile log-likelihood for ϕ and picking the value that maximize that function. On the other hand, the standard error for the estimator of ϕ can be estimated via Jackknife (Miller, 1974).*
2. *To avoid maximization problems with the constraint $1+\phi\theta>0$ (presented after eq. (1)), we use the same approach used by Cancho et al. (2013) considering $\phi\geq 0$, i.e., the over-dispersed case.*
3. *The maximization involved in the M-steps can be performed using software R (*R Development Core Team, 2015*), among others. The computational programs used in this work are available from the authors upon request.*
4. *Differently from the direct maximization of the log-likelihood function, the EM algorithm allows to obtain predictions for the number of initial cells and activated cells for each individual (M_i and D_i, $i=1,\ldots,n$, respectively) in the version 1 and to the chance for cure for each individual (Y_i, $i=1,\ldots,n$), in the version 2.*

5. Simulation studies

In this Section, two simulation studies are presented. The first study assess the performance of the two procedures through different elements as bias and coverage probabilities. The second study is designed to evaluate whether the AIC and BIC (Akaike's and Bayesian information) criteria are able to correctly pick the distribution for the non-destroyed cells, given the correct activation scheme.

5.1. Parameters recovery

For simulation purposes, the covariates z_1 and z_2 were drawn from the Bernoulli distribution with success probability 0.5. As discussed in Section 3, both vectors should not

Table 4: *Average of parameter estimates, standard errors (se), root of mean squared errors ($\sqrt{MSE}$) and coverage probability of 95% (CP) using the way 2 of defining the EM algorithm for DNB-FA, DNB-LA and DNB-RA models considering Weibull distribution for time-to-event of the non-destroyed cells. (CA denotes censoring average with their respective standard errors).*

DNB-FA									
		$n = 200$				$n = 400$			
Parameter	True	average	se	$\sqrt{MSE}$	CP	average	se	$\sqrt{MSE}$	CP
β_1	1.0	1.021	0.287	0.240	0.943	1.008	0.202	0.165	0.943
β_{20}	−0.5	−0.473	0.453	0.381	0.954	−0.484	0.304	0.259	0.947
β_{21}	0.5	0.577	0.519	0.489	0.970	0.524	0.386	0.309	0.961
ϕ	1.0	1.078	0.254	0.227	0.935	1.042	0.152	0.134	0.941
α	−1.3	−1.333	0.177	0.167	0.905	−1.317	0.124	0.114	0.914
ν	1.5	1.530	0.191	0.125	0.986	1.517	0.133	0.086	0.986
CA		0.636	0.039			0.612	0.024		
DNB-LA									
β_1	1.0	1.040	0.288	0.307	0.914	1.022	0.222	0.207	0.940
β_{20}	−0.5	−0.460	0.514	0.446	0.931	−0.496	0.307	0.295	0.947
β_{21}	0.5	0.646	0.801	0.694	0.923	0.542	0.420	0.402	0.950
ϕ	1.0	1.081	0.267	0.239	0.937	1.032	0.131	0.129	0.943
α	−1.3	−1.308	0.226	0.189	0.938	−1.305	0.158	0.132	0.942
ν	1.5	1.523	0.222	0.152	0.975	1.513	0.155	0.105	0.971
CA		0.660	0.034			0.659	0.024		
DNB-RA									
β_1	1.0	1.025	0.302	0.274	0.916	1.010	0.212	0.191	0.920
β_{20}	−0.5	−0.469	0.491	0.418	0.946	−0.485	0.315	0.276	0.937
β_{21}	0.5	0.615	0.654	0.590	0.960	0.530	0.429	0.360	0.950
ϕ	1.0	1.064	0.297	0.276	0.939	1.037	0.142	0.131	0.945
α	−1.3	−1.320	0.187	0.158	0.940	−1.313	0.132	0.111	0.939
ν	1.5	1.526	0.202	0.134	0.984	1.513	0.142	0.092	0.985
CA		0.632	0.034			0.632	0.024		

incorporate intercept at the same time. Thus, only z_2 has an intercept term. It is chosen $\beta_1 = 1, \beta_{20} = -0.5$ and $\beta_{21} = 0.5$, implying cure rates 0.73, 0.67, 0.49 and 0.42 for profiles $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$ respectively. Parameters related to the time-to-event for non-destroyed cells where chosen as $\alpha = -1.3, \nu = 1.5$ for the Weibull model, $\alpha = 0.8, \nu = 0.4$ for the Log-Normal model and $\alpha = 3.5, \nu = 1.5$ for the gamma model. Those parameters were used with FA, LA and RA schemes. We assume $\phi = 1$ in all cases.

For scheme FA, the two methods exposed in Section 4 were used with sample sizes $n = 200$ and $n = 400$. For schemes LA and RA, the second method exposed in Section 4 was used with sample sizes $n = 200$ and $n = 400$. In each case, 10,000 replicates were considered. Tables 4 shows part of the results for the simulations. We report the average of the estimates obtained (average), the mean of the asymptotic standard errors (se), the root of the mean squared error ($\sqrt{MSE}$) and the asymptotic coverage probability with 95% (CP). Main conclusions are that the two ways of implementing the EM algorithm provide close results relation to average, se, $\sqrt{MSE}$ and CP for the three activation schemes. Results also reveals that the estimates are closer to the true values and $\sqrt{MSE}$ is decreased as n increases, suggesting that estimators are consistent. On the other hand, the se is greater than $\sqrt{MSE}$, suggesting that the standard errors are overestimated. Despite this, the CP are closer to the nominal value.

5.2. Misspecification of the distribution for the non-destroyed concurrent cells

In the survival analysis literature, it is common to consider the Weibull distribution as the survival model for the time-to-event for the non-destroyed cells because of its appropriateness in many medical and biological contexts. However, to the best of our knowledge, we were unable to trace studies on the effects on both susceptible and cured parts of the model, of an incorrect specification of the survival function for the time-to-event for the non-destroyed cells.

Bearing this in mind, a simulation study is conducted using the same specification for parameters used in the last subsection. The three activation schemes mentioned in Section 3 and the Weibull, LN and gamma distributions for the time-to-event for non-destroyed cells were used. For each activation scheme/distribution combination, 10,000 samples were simulated and, for each sample, parameter estimates were computed (including $S(\cdot \mid \lambda)$). Then, the mean and MSE of the estimates were computed for each parameter and for the cure rate. Additionally, the mean and MSE for the expected times for the non-destroyed cells were also computed. Furthermore, the AIC and BIC criteria were computed for the three distribution and which was the model choice based on those criteria. Since they provide similar results, data on AIC was presented. Results for FA scheme are shown in Table 5. It is expected that a wrong choice for $S(\cdot \mid \lambda)$ increases the bias and the MSE for the expected activation time for non-destroyed cells. However,

Table 5: *Estimated bias and MSE for cure rate and expected values for the non-destroyed cells in DNB-FA with different activation schemes.*

			Cure rate		$\mathbb{E}(W)$		
	True		First Activation Scheme				
n	Distribution	Distribution	bias	MSE	bias	MSE	% AIC
200	Weibull	Weibull	0.001	0.004	−0.049	0.129	0.912
		Log-Normal	0.087	0.014	−7.909	195.7	0.080
		Gamma	0.008	0.902	0.689	0.902	0.008
	Log-Normal	Weibull	0.005	0.004	0.337	0.138	0.025
		Log-Normal	0.000	0.004	0.294	0.118	0.932
		Gamma	0.294	0.004	−24.1	689.0	0.043
	Gamma	Weibull	0.000	0.004	2.974	8.886	0.083
		Log-Normal	0.019	0.005	2.321	5.857	0.086
		Gamma	0.001	0.004	−0.722	6.089	0.831
400	Weibull	Weibull	0.002	0.002	−0.030	0.084	0.920
		Log-Normal	0.094	0.012	−6.908	88.6	0.038
		Gamma	0.008	0.002	0.782	0.787	0.042
	Log-Normal	Weibull	0.005	0.002	0.336	0.125	0.001
		Log-Normal	0.000	0.002	0.301	0.106	0.923
		Gamma	0.000	0.002	−22.5	549.7	0.075
	Gamma	Weibull	0.000	0.002	2.974	8.865	0.074
		Log-Normal	0.019	0.002	2.363	5.754	0.074
		Gamma	0.000	0.002	−0.313	2.371	0.852

the wrong choice also impacts on the cure rate estimates. Except for the gamma model, the AIC and BIC criteria chose the correct model for more than 90% of generate samples, suggesting that those criteria are appropriate to this purpose. For other activation schemes, similar results are obtained.

6. Application

In this section we analyze the cutaneous melanoma data set described in Section 2. Models DNB-FA, DNB-LA and DNB-RA were fitted to the data, with the survival functions from the Weibull, LN and gamma distributions used as survival functions for the time-to-event for the non-destroyed cells. To avoid identifiability problems, the covariates treatment, age, nodule and thickness were incorporated into the model through the θ and p parameters. All possible combinations of covariates preserving identifiability were considered and the combination that provided the least AIC and BIC criteria was

Table 6: *AIC/BIC criteria for E1690 data set using the DNB with different activation schemes.*

	Activation Scheme		
$S(\cdot \mid \lambda)$	FA	LA	RA
Weibull	**827.6/863.7**	854.3/890.4	842.7/878.8
LN	834.6/870.7	851.8/888.0	842.0/878.1
gamma	828.0/864.1	854.5/890.6	841.8/877.9

selected, leading to the one assigning nodule and tumour thickness to θ and treatment to p (see equation (4)). Given that all considered patients have cutaneous melanoma, it is reasonable to assume that the nodule category is related to the number of initial cells (most advance stage, more initial cells) and the same with tumour thickness (greater tumour, more initial cells). On the other hand, treatment can be interpreted as an element that determines the chance of such cells be activated (patients receiving the treatment have reduced their probability of initial activation of the initial cells). Table 6 shows the AIC and BIC vales for those combinations of covariates. Based on those criteria, the DNB-FA/Weibull model was chosen as the one presenting the best fit. On the other hand, it makes sense to use this activation scheme in a biological context, because just one cell can trigger the metastasis process. The estimates for this model are presented in Table 7.

Table 7: *Parameter estimates for the DNB-FA/Weibull model.*

Parameter	est	se	|est|/se
$\beta_{1,nodule1}$	0.4690	0.4565	1.03
$\beta_{1,nodule2}$	1.5143	0.3661	4.14
$\beta_{1,nodule3}$	2.1539	0.4044	5.32
$\beta_{1,nodule4}$	3.0702	0.4210	7.29
$\beta_{1,thickness}$	0.0858	0.0473	1.81
$\beta_{2,treatment}$	-0.7965	0.4064	1.96
ϕ	3.1807	0.0785	
α	-1.3142	0.1977	
ν	1.5372	0.0273	

The estimated means of the initial number of cells are $1.60 \times 1.09^{thickness}$ (nodule 1), $4.55 \times 1.09^{thickness}$ (nodule 2), $8.62 \times 1.09^{thickness}$ (nodule 3) and $21.55 \times 1.09^{thickness}$ (nodule 4) and the probability of activation of those cells is 0.5 for patients in control group and 0.31 for patients in the treatment group.

Finally Figure 3 shows the estimated mean of non-destroyed cells (D) for each patients stratified by control and treatment group. Note that the estimated means of D vary on both group, agreeing with the fact that the treatment is effective. On the other hand, it is possible to conclude that patients with nodule 4 have more estimated non-destroyed

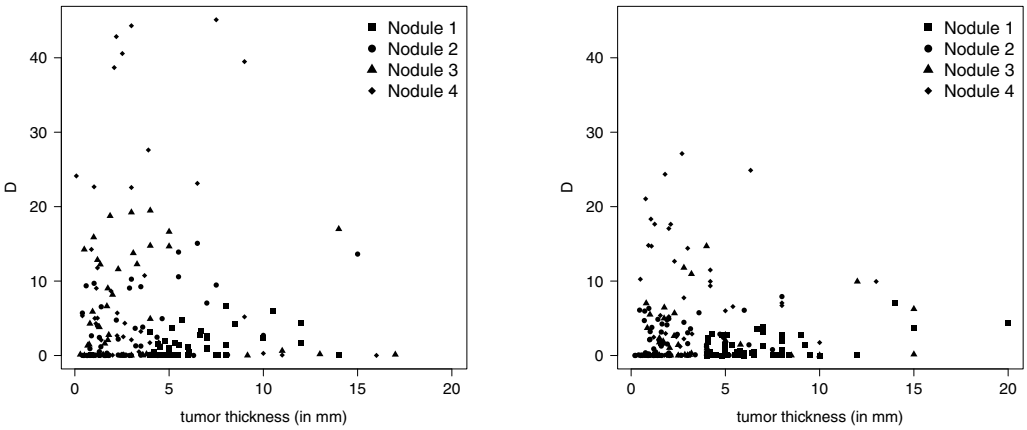

Figure 3: *Predicted means of the conditional distributions for all patients under the DNB-FA/Weibull model for the number of activated cells (D), stratified by nodule categories and belonging to the control group (left panel) and to the treatment group (right panel), respectively.*

cells. This is expected because patients in this stage of disease are more susceptible to die faster than patients in others stages of disease.

7. Final discussion

In this paper, an alternative estimation procedure based on the EM algorithm is proposed for the destructive Negative Binomial cure rate model introduced in Cancho et al. (2013). Two different ways of implementing the algorithm are investigated. Simulation studies indicate that those procedures work satisfactorily. It also investigated other alternatives (besides the Weibull distribution) for the survival function for the time for non-destroyed cells $S(\cdot \mid \lambda)$, and through the use of simulation studies evaluating the performances of the AIC/BIC criteria to correctly choose the model that provides the best fit to the data. Using simulation studies we assess the performances of the AIC/BIC criteria to correctly choose the model that provides the best fit to the data. However, a wrong choice for $S(\cdot \mid \lambda)$ can lead to incorrect estimates in both, the parameters related to the cure rate and the ones related to the survival function of the time-to-event for non-destroyed cells. Thus, precision loss is incurred if the wrong model is selected, that is, one has to be careful when selecting the working model. For this reason, it will be proposed non-parametric frameworks to estimate $S(\cdot \mid \lambda)$. Finally, the proposed approach was illustrated using real data related to a clinical trial on Phase III cutaneous melanoma patients.

Acknowledgments

We thank two referees and Editor for comments and suggestions that substantially improved the presentation of the work. The research of Diego I. Gallardo was supported by FAPESP grant number 2013/23684-2.

8. Appendix: Proofs of propositions

8.1. Appendix A: Proposition 1

For DNB-FA and DNB-LA the result is trivial. On the other hand, note it is possible to show that the marginal distribution of $D_i \mid \theta_i, p_i, \phi$ is $NB(\theta_i p_i, \phi)$. Thus, for the DNB-RA we have that for $r_1 \in \{1,2,\dots,\}$

$$P(R_i = r_i \mid D_{obs}, \psi) = \frac{\sum_{d_i=r_i}^{\infty} f(t_i, \delta_i \mid D_i = d_i, R_i = r_i)}{[(1-q_{0i})f(t_i;\lambda)]^{\delta_i}[q_{0i} + (1-q_{0i})S(t_i;\lambda)]^{1-\delta_i}},$$

where $f(t_i, \delta_i \mid D_i = d_i, R_i = r_i)$ is defined in (3). For $\delta_i = 1$, the expression takes the form

$$P(R_i = r_i \mid D_{obs}, \psi) = \frac{1}{(1-q_{0i})} \sum_{d_i=r_i}^{\infty} d_i \binom{d_i - 1}{r_i - 1} S(t_i;\lambda)^{d_i - r_i} F(t_i;\lambda)^{r_i - 1} P(D_i = d_i; \theta_i, p_i, \phi)$$

$$= \frac{F(t_i;\lambda)^{r_i-1}}{(1-q_{0i})} \mathbb{E}\left[D_i \binom{D_i - 1}{r_i - 1} S(t_i;\lambda)^{D_i - r_i} I(D_i \geq r_i) \right].$$

For $\delta_i = 0$,

$$P(R_i = r_i \mid D_{obs}, \psi) = \frac{\sum_{d_i=r_i}^{\infty} IB(S(t_i;\lambda), d_i - r_i + 1, r_i) P(D_i = d_i; \theta_i, p_i, \phi)}{q_{0i} + (1-q_{0i})S(t_i;\lambda)}.$$

On the other hand, by using the binomial theorem, it can be shown that $IB(S(t_i;\lambda), d_i - r_i + 1, r_i) = \sum_{k=0}^{r_i-1} \binom{r_i-1}{k} (-1)^k \frac{S(t_i;\lambda)^{d_i - r_i + k + 1}}{d_i - r_i + k + 1}$. In other words,

$$P(R_i = r_i \mid D_{obs}, \psi) = \frac{\sum_{k=0}^{r_i-1} \binom{r_i-1}{k} (-1)^k \mathbb{E}\left[\frac{S(t_i;\lambda)^{D_i - r_i + k + 1}}{D_i - r_i + k + 1} I(D_i \geq r_i) \right]}{q_{0i} + (1-q_{0i})S(t_i;\lambda)}.$$

8.2. Appendix B: proposition 2

Consider now the DNB-FA model ($R_i = 1$, $i = 1,\ldots,n$). Thus, by (2) and (5) the expression $P(D_i = d_i \mid D_{obs}, \psi)$ assumes the following form

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{S(t_i;\lambda)^{d_i-\delta_i}[d_i f(t_i;\lambda)]^{\delta_i}\frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)^{d_i}(1+\phi\theta_i p_i)^{-\phi^{-1}}}{(\theta_i p_i f(t_i;\lambda))^{\delta_i}\left(1+\phi\theta_i p_i F(t_i;\lambda)\right)^{-(\phi^{-1}+\delta_i)}}$$

$$= \frac{\Gamma((\phi^{-1}+\delta_i)+d_i-\delta_i)}{\Gamma(\phi^{-1}+\delta_i)(d_i-\delta_i)!}\theta_{1i}^{d_i-\delta_i}\left(1-\theta_{1i}\right)^{(\phi^{-1}+\delta_i)},$$

i.e., $D_i - \delta_i \mid D_{obs} \mid \psi \sim NB\left((\phi^{-1}+\delta_i)^{-1}, \theta_{1i}\right)$, where $\theta_{1i} = \frac{\phi\theta_i p_i S(t_i;\lambda)}{1+\phi\theta_i p_i}$. For the DNB-LA, $R_i = D_i$, $i = 1,\ldots,n$ and then

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{\left\{d_i F(t_i;\lambda)^{d_i-1} f(t_i;\lambda)\right\}^{\delta_i}\left\{1-F(t_i;\lambda)^{d_i}\right\}^{1-\delta_i}}{\left\{\theta_i p_i f(t_i;\lambda)\left(1+\phi\theta_i p_i S(t_i;\lambda)\right)^{-(\phi^{-1}+1)}\right\}^{\delta_i}}\times$$

$$\times\frac{\frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)^{d_i}(1+\phi\theta_i p_i)^{-\phi^{-1}}}{\left\{1+q_{0i}-\left(1+\phi\theta_i p_i S(t_i;\lambda)\right)^{-\phi^{-1}}\right\}^{1-\delta_i}}.$$

For $\delta_i = 1$, this expression takes the form

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{\Gamma((\phi^{-1}+1)+(d_i-1))}{\Gamma(\phi^{-1}+1)(d_i-1)!}\theta_{2i}^{d_i-1}\left(1-\theta_{2i}\right)^{-(\phi^{-1}+1)},$$

i.e., $(D_i - 1) \mid D_{obs}, \psi \sim NB\left((\phi^{-1}+1)^{-1}, \theta_{2i}\right)$, where $\theta_{2i} = \frac{\phi\theta_i p_i F(t_i;\lambda)}{1+\phi\theta_i p_i}$. For $\delta_i = 0$, this expression is reduced to

$$P(D_i = d_i \mid D_{obs}, \psi) = a_i\frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right)^{d_i}(1+\phi\theta_i p_i)^{-\phi^{-1}}$$

$$+(1-a_i)\frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1})d_i!}\theta_{2i}^{d_i}\left(1-\theta_{2i}\right)^{\phi^{-1}},$$

where $a_i = (1+q_{0i}-(1+\phi\theta_i p_i S(t_i;\lambda))^{-\phi^{-1}})^{-1}$, i.e., $D_i \mid D_{obs}, \psi \sim a_i NB\left(\phi, \frac{\phi\theta_i p_i}{1+\phi\theta_i p_i}\right) + (1-a_i)NB(\phi, \theta_{2i})$. Finally, for DNB-RA we have that

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{\sum_{r_i=1}^{d_i} \left\{ d_i \binom{d_i-1}{r_i-1} f(t_i;\lambda) S(t_i;\lambda)^{d_i-r_i} F(t_i;\lambda)^{r_i-1} \right\}^{\delta_i} \frac{1}{d_i}}{[q_{0i} f(t_i;\lambda)]^{\delta_i} [q_{0i} + (1 - q_{0i} S(t_i;\lambda))]^{1-\delta_i}}$$
$$\times \{IB(S(t_i;\lambda), d_i - r_i + 1, r_i)\}^{1-\delta_i}$$
$$\frac{\Gamma(\phi^{-1} + d_i)}{\Gamma(\phi^{-1}) d_i!} \left(\frac{\phi\theta_i p_i}{1 + \phi\theta_i p_i} \right)^{d_i} (1 + \phi\theta_i p_i)^{-\phi^{-1}}.$$

For $\delta_i = 1$, it is immediate that

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{\Gamma(\phi^{-1} + d_i)}{\Gamma(\phi^{-1}) d_i!} \left(\frac{\phi\theta_i p_i}{1 + \phi\theta_i p_i} \right)^{d_i} I(d_i \geq 1),$$

i.e., $(D_i - 1) \mid D_{obs}, \psi \sim NB(\phi, \frac{\phi\theta_i p_i}{1+\phi\theta_i p_i})$. Finally, for $\delta_i = 0$, using the binomial theorem, the expression is reduced to

$$P(D_i = d_i \mid D_{obs}, \psi) = \frac{\sum_{r_i=1}^{d_i} \sum_{k=0}^{r_i} (-1)^k \binom{r_i-1}{k} \frac{S(t_i;\lambda)^{d_i-r_i+k+1}}{d_i(d_i-r_i+k+1)} \frac{\Gamma(\phi^{-1}+d_i)}{\Gamma(\phi^{-1}) d_i!} \left(\frac{\phi\theta_i p_i}{1+\phi\theta_i p_i} \right)^{d_i}}{1 + [(1 - q_{0i})/q_{0i}] S(t_i;\lambda)}.$$

8.3. Appendix C: proposition 3

Considering the DNB-FA model ($R_i = 1,\ i = 1, \ldots, n$), and by (2) and (5) the expression $P(M_i = m_i \mid D_{obs}, \psi)$ assume the following form

$$P(M_i = m_i \mid D_{obs}, \psi) = \frac{\sum_{d_i=\delta_i}^{m_i} S(t_i;\lambda)^{d_i-\delta_i} [d_i f(t_i;\lambda)]^{\delta_i} \binom{m_i}{d_i} p_i^{d_i} (1 - p_i)^{m_i - d_i}}{(\theta_i p_i f(t_i;\lambda))^{\delta_i} (1 + \phi\theta_i p_i F(t_i;\lambda))^{-(\phi^{-1}+\delta_i)}} \times$$
$$\times \frac{\Gamma(\phi^{-1} + m_i)}{\Gamma(\phi^{-1}) m_i!} \left(\frac{\phi\theta_i}{1 + \phi\theta_i} \right)^{m_i} (1 + \phi\theta_i)^{-\phi^{-1}}$$
$$= \frac{\Gamma((\phi^{-1} + \delta_i) + m_i - \delta_i)}{\Gamma(\phi^{-1} + \delta_i)(m_i - \delta_i)!} \theta_{3i}^{m_i - \delta_i} (1 - \theta_{3i})^{(\phi^{-1}+\delta_i)},$$

i.e., $M_i - \delta_i \mid D_{obs} \mid \psi \sim NB\left((\phi^{-1} + \delta_i)^{-1}, \theta_{3i}\right)$, where $\theta_{3i} = \frac{\phi\theta_i(1 - p_i F(t_i;\lambda))}{1+\phi\theta_i}$. For the DNB-LA, $R_i = D_i$, $i = 1, \ldots, n$ and then

$$P(M_i = m_i \mid D_{obs}, \psi) = \sum_{d_i=\delta_i}^{m_i} \left[\frac{\left\{ d_i F(t_i;\lambda)^{d_i-1} f(t_i;\lambda) \right\}^{\delta_i} \left\{ 1 - F(t_i;\lambda)^{d_i} \right\}^{1-\delta_i}}{\left\{ \theta_i p_i f(t_i;\lambda) \left(1 + \phi\theta_i p_i S(t_i;\lambda)\right)^{-(\phi^{-1}+1)} \right\}^{\delta_i}} \times \right.$$

$$\left. \times \frac{\binom{m_i}{d_i} p_i^{d_i} (1-p_i)^{m_i-d_i} \frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!} \left(\frac{\phi\theta_i}{1+\phi\theta_i}\right)^{m_i} (1+\phi\theta_i)^{-\phi^{-1}}}{\left\{ 1 + q_{0i} - \left(1 + \phi\theta_i p_i S(t_i;\lambda)\right)^{-\phi^{-1}} \right\}^{1-\delta_i}} \right]$$

For $\delta_i = 1$, this expression is reduced to

$$P(M_i = m_i \mid D_{obs}, \psi) = \frac{\Gamma((\phi^{-1}+1)+m_i-1)}{\Gamma(\phi^{-1}+1)(m_i-1)!} \theta_{4i}^{m_i-1} \left(1-\theta_{4i}\right)^{(\phi^{-1}+1)},$$

i.e., $(M_i - 1) \mid D_{obs}, \psi \sim NB\left((\phi^{-1}+1)^{-1}, \theta_{4i}\right)$, where $\theta_{4i} = \frac{\phi\theta_i(1-p_i S(t_i;\lambda))}{1+\phi\theta_i}$. For $\delta_i = 0$, this expression takes the form

$$P(M_i = m_i \mid D_{obs}, \psi) = a_i \frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!} \left(\frac{\phi\theta_i}{1+\phi\theta_i}\right)^{m_i} (1+\phi\theta_i)^{-\phi^{-1}}$$

$$+ (1-a_i) \frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!} \theta_{4i}^{m_i} \left(1-\theta_{4i}\right)^{\phi^{-1}},$$

where $a_i = (1 + q_{0i} - (1+\phi\theta_i p_i S(t_i;\lambda))^{-\phi^{-1}})^{-1}$, i.e., $M_i \mid D_{obs}, \psi \sim a_i NB\left(\phi, \frac{\phi\theta_i}{1+\phi\theta_i}\right) + (1-a_i) NB(\phi, \theta_{4i})$. Finally, for DNB-RA we have that

$$P(M_i = m_i \mid D_{obs}, \psi) = \frac{\sum_{d_i=\delta_i}^{m_i} \sum_{r_i=1}^{d_i} \left\{ d_i \binom{d_i-1}{r_i-1} f(t_i;\lambda) S(t_i;\lambda)^{d_i-r_i} F(t_i;\lambda)^{r_i-1} \right\}^{\delta_i} \times \frac{1}{d_i}}{[q_{0i} f(t_i;\lambda)]^{\delta_i} [q_{0i} + (1 - q_{0i} S(t_i;\lambda))]^{1-\delta_i}}$$

$$\times \left\{ IB(S(t_i;\lambda), d_i - r_i + 1, r_i) \right\}^{1-\delta_i} \times$$

$$\times \binom{m_i}{d_i} p_i^{d_i} (1-p_i)^{m_i-d_i} \frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!} \left(\frac{\phi\theta_i}{1+\phi\theta_i}\right)^{m_i} (1+\phi\theta_i)^{-\phi^{-1}}.$$

For $\delta_i = 1$, the expression is reduced to

$$P(M_i = m_i \mid D_{obs}, \psi) = \left[1 - (1-p_i)^{m_i}\right] \frac{\Gamma(\phi^{-1}+m_i)}{\Gamma(\phi^{-1})m_i!} \left(\frac{\phi\theta_i}{1+\phi\theta_i}\right)^{m_i} I(m_i \geq 1).$$

Finally, for $\delta_i = 0$,

$$P(M_i = m_i \mid D_{obs}, \psi) = \frac{\sum_{d_i=0}^{m_i} \sum_{r_i=1}^{d_i} \sum_{k=0}^{r_i} v_i \left(\frac{p_i}{1-p_i}\right)^{d_i} \left(\frac{\phi\theta_i(1-p_i)}{1+\phi\theta_i}\right)^{m_i}}{1 + [(1-q_{0i})/q_{0i}]S(t_i;\lambda)}.$$

References

Balakrishnan, N. and Pal, S. (2009). EM algorithm-based likelihood estimation for some cure rate models. *Journal of Statistical Theory and Practice*, 6, 698–724.

— (2013). Expectation maximization-based likelihood inference for flexible cure rate models with Weibull lifetimes. *Statistical Methods in Medical Research.* DOI:10.1177/0962280213491641.

— (2015). An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood-and information-based methods. *Computational Statistics*, 30, 151–189.

Berkson, J. and Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47, 501–515.

Cancho, V., Bandyopadhyay, D., Louzada, F. and Yiqi, B. (2013). The destructive negative binomial cure rate model with a latent activation scheme. *Statistical Methodology*, 13, 48–68.

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.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B*, 39, 1–38.

Gallardo, D.I., Bolfarine, H. and Pedroso-de-Lima, A.C. (2016). An EM algorithm for estimating the destructive weighted Poisson cure rate model. *Journal of Statistical Computation and Simulation*, 86, 1497–1515.

Hanin, L. and Huang, L.S. (2014). Identifiability of cure models revisited. *Multivariate Data Analysis*, 130, 261–274.

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001). *Bayesian Survival Analysis*, Springer, New York.

Li, C.S., Taylor, J. and Sy, J. (2001). Identifiability of cure models. *Statistics and Probability Letters*, 54, 389–395.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure estimation. *Statistica Sinica*, 20, 661–674.

MacDonald, I.L. (2014). Numerical Maximisation of Likelihood: A Neglected Alternative to EM? *International Statistical Review*, 82, 296–308.

Miller, R.G. (1974). The jackknife: a review. *Biometrika*, 61, 1–15.

R Development Core Team. (2015). *R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing*. ISBN 3-900051-07-0, version 3.0.3.

Rodrigues, J., Cancho, V.G., Castro, M.A. and Louzada-Neto, F. (2009). On the unification of the long-term survival models. *Statistics and Probability Letters*, 79, 753–759.

Rodrigues, J., Castro, M., Balakrishnan, N. and Cancho, V.G. (2011). Destructive weighted Poisson cure rate models. *Lifetime Data Analysis*, 17, 333–346.

Tsodikov, A.D., Ibrahim, J.G. and Yakovlev, A.Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 2003, 1063–1078.

Williams, J.S. and Lagakos, S.W. (1977). Models for censored survival analysis: constant-sum and variable-sum models. *Biometrika*, 64, 215–224.

Yakovlev, A.Y. and Tsodikov, A.D. (1996). *Stochastic Models of Tumour Latency and their Biostatistical Applications.* World Scientific, New Jersey.

SORT 40 (1) January-June 2016, 55-88

A test for normality based on the empirical distribution function

Hamzeh Torabi[1], Narges H. Montazeri[1] and Aurea Grané[2]

Abstract

In this paper, a goodness-of-fit test for normality based on the comparison of the theoretical and empirical distributions is proposed. Critical values are obtained via Monte Carlo for several sample sizes and different significance levels. We study and compare the power of forty selected normality tests for a wide collection of alternative distributions. The new proposal is compared to some traditional test statistics, such as Kolmogorov-Smirnov, Kuiper, Cramér-von Mises, Anderson-Darling, Pearson Chi-square, Shapiro-Wilk, Shapiro-Francia, Jarque-Bera, SJ, Robust Jarque-Bera, and also to entropy-based test statistics. From the simulation study results it is concluded that the best performance against asymmetric alternatives with support on the whole real line and alternative distributions with support on the positive real line is achieved by the new test. Other findings derived from the simulation study are that SJ and Robust Jarque-Bera tests are the most powerful ones for symmetric alternatives with support on the whole real line, whereas entropy-based tests are preferable for alternatives with support on the unit interval.

MSC: 62F03, 62F10.

Keywords: Empirical distribution function, entropy estimator, goodness-of-fit tests, Monte Carlo simulation, Robust Jarque-Bera test, Shapiro-Francia test, SJ test; test for normality.

1. Introduction

Let $X_1,\ldots,X_n$ be a n independent an identically distributed (iid) random variables with continuous cumulative distribution function (cdf) $F(.)$ and probability density function (pdf) $f(.)$. All along the paper, we will denote the order statistic by $(X_{(1)},\ldots,X_{(n)})$. Based on the observed sample $x_1,\ldots,x_n$, we are interested in the following goodness-of-fit test for a location-scale family:

[1] Statistics Department, Yazd University, 89175-741, Yazd, Iran, htorabi@yazd.ac.ir, nmontazeri@stu.yazd.ac.ir

[2] Statistics Department, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe, Spain, aurea.grane@uc3m.es

Received: April 2015
Accepted: February 2016

$$\begin{cases} H_0: F \in \mathscr{F} \\ H_1: F \notin \mathscr{F} \end{cases} \tag{1}$$

where $\mathscr{F} = \left\{F_0(.;\boldsymbol{\theta}) = F_0\left(\frac{x-\mu}{\sigma}\right) \mid \boldsymbol{\theta} = (\mu,\sigma) \in \Theta\right\}$, $\Theta = \mathbb{R} \times (0,\infty)$ and μ and σ are unspecified. The family $\mathscr{F}$ is called location-scale family, where $F_0(.)$ is the standard case for $F_0(.;\boldsymbol{\theta})$ for $\boldsymbol{\theta} = (0,1)$. Suppose that $f_0(x;\boldsymbol{\theta}) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$ is the corresponding pdf of $F_0(x;\boldsymbol{\theta})$.

The goodness-of-fit test problem for location-scale family described in (1) has been discussed by many authors. For instance, Zhao and Xu (2014) considered a random distance between the sample order statistic and the quasi sample order statistic derived from the null distribution as a measure of discrepancy. On the other hand, Alizadeh and Arghami (2012) used a test based on the minimum Kullback-Leibler distance. The Kullback-Leibler divergence measure is a special case of a ϕ-divergence measure (2) for $\phi(x) = x\log(x) - x + 1$ (see p. 5 of Pardo, 2006 for details). Also ϕ-divergence is a special case of the ϕ-disparity measure. The ϕ-disparity measure between two pdf's f_0 and f is defined by

$$D_\phi(f_0, f) = \int \phi\left(\frac{f_0(x;\boldsymbol{\theta})}{f(x)}\right) f(x)\, dx, \tag{2}$$

where $\phi : (0,\infty) \to [0,\infty)$ is assumed to be continuous, decreasing on $(0,1)$ and increasing on $(1,\infty)$, with $\phi(1) = 0$ (see p. 29 of Pardo, 2006 for details). In ϕ-divergence, ϕ is a convex function.

Inspired by this idea, in this paper we propose a goodness-of-fit statistic to test (1) by considering a new proximity measure between two continuous cdf's. The organization of the paper is as follows. In Section 2 we define the new measure H_n and study its properties as a goodness-of-fit statistic. In Section 3 we propose a normality test based on H_n and find its critical values for several sample sizes and different significance levels. In Section 4 we review forty normality tests, including the most traditional ones such as Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling, Shapiro-Wilk, Shapiro-Francia, Pearson Chi-square, among others, and in Section 5 we compare their performances to that of our proposal through a wide set of alternative distributions. We also provide an application example where the Kolmogorov-Smirnov test fails to detect the non normality of the sample.

2. A new discrepancy measure

In this section we define a discrepancy measure between two continuous cdf's and study its properties as a goodness-of-fit statistic.

Definition 2.1 *Let X and Y be two absolutely continuous random variables with cdf's F_0 and F, respectively. We define*

$$D(F_0,F)=\int_{-\infty}^{\infty} h\left(\frac{1+F_0(x;\boldsymbol{\theta})}{1+F(x)}\right)dF(x)=E_F\left[h\left(\frac{1+F_0(X;\boldsymbol{\theta})}{1+F(X)}\right)\right], \tag{3}$$

where $E_F[.]$ is the expectation under F and $h:(0,\infty)\to\mathbb{R}^+$ is assumed to be continuous, decreasing on $(0,1)$ and increasing on $(1,\infty)$ with an absolute minimum at $x=1$ such that $h(1)=0$.

Lemma 2.2 *$D(F_0,F)\geq 0$ and equality holds if and only if $F_0=F$, almost everywhere.*

Proof. Using the non-negativity of function h, we have $D(F_0,F)\geq 0$. It is clear that $F_0=F$ implies $D(F_0,F)=0$. Conversely, if $D(F_0,F)=0$, since h has an absolute minimum at $x=1$, then $F_0=F$. ∎

Let us return to the goodness-of-fit test problem for a location-scale family described in (1). Firstly, we estimate μ and σ by their maximum likelihood estimators (MLEs), i.e., $\hat{\mu}$ and $\hat{\sigma}$, respectively, and we take $z_i=(x_i-\hat{\mu})/\hat{\sigma}$, $i=1,\ldots,n$. Note that in this family, $F_0(x_i;\hat{\mu},\hat{\sigma})=F_0(z_i)$. Secondly, consider the empirical distribution function (EDF) based on data x_i, that is

$$F_n(t)=\frac{1}{n}\sum_{j=1}^{n}\mathrm{I}_{[x_j\leq t]},$$

where I_A denotes the indicator of an event A. Then, our proposal is based on the ratio of the standard cdf under H_0 and the EDF based on the x_i's. Using (3) with $F=F_n$, $D(F_0,F_n)$ can be written as

$$\begin{aligned}\mathrm{H}_n:=D(F_0,F_n)&=\int_{-\infty}^{\infty} h\left(\frac{1+F_0(x;\hat{\mu},\hat{\sigma})}{1+F_n(x)}\right)dF_n(x)\\&=\frac{1}{n}\sum_{i=1}^{n}h\left(\frac{1+F_0(x_{(i)};\hat{\mu},\hat{\sigma})}{1+F_n(x_{(i)})}\right)\\&=\frac{1}{n}\sum_{i=1}^{n}h\left(\frac{1+F_0(z_{(i)})}{1+i/n}\right)\end{aligned}$$

Under H_0, we expect that $F_0(t;\hat{\mu},\hat{\sigma})\approx F_n(t)$, for every $t\in\mathbb{R}$ and $1+F_0(t;\hat{\mu},\hat{\sigma})\approx 1+F_n(t)$. Note that, since $h(1)=0$, we expect that $h\big((1+F_0(t))/(1+F_n(t))\big)\approx 0$ and

thus H_n will take values close to zero when H_0 is true. Therefore, it seems justifiable that H_0 must be rejected for large values of H_n. Some standard choices for h are: $h(x) = (x-1)^2/(x+1)^2, x\log(x)-x+1, (x-1)\log(x), |x-1|$ or $(x-1)^2$ (for more examples, see p. 6 of Pardo, 2006 for details).

Proposition 2.3 *The support of* H_n *is* $[0, \max(h(1/2), h(2))]$.

Proof. Since $F_0(.)$ and F_n are cdf's and take values in $[0,1]$, we have that

$$1/2 \leq \frac{1+F_0(y)}{1+F_n(y)} \leq 2, \quad y \in \mathbb{R}.$$

Thus

$$0 \leq h\left(\frac{1+F_0(y)}{1+F_n(y)}\right) \leq \max(h(1/2), h(2))$$

Finally, since $\mathrm{H_n}$ is the mean of $h(.)$ over the transformed data, the result is obtained. ■

Proposition 2.4 *The test statistic based on* H_n *is invariant under location-scale transformations.*

Proof. The location-scale family is invariant under the location-scale transformations of the form $g_{c,r}(X_1,\ldots,X_n) = (rX_1+c,\ldots,rX_n+c)$, $c \in \mathbb{R},\ r > 0$, which induces similar transformations on $\Theta : g_{c,r}(\boldsymbol{\theta}) = (r\mu+c, r\sigma)$ (See Shao, 2003). The estimator $T_0(X_1,\ldots,X_n)$ for μ is location-scale invariant if

$$T_0(rX_1+c,\ldots,rX_n+c) = rT_0(X_1,\ldots,X_n)+c, \quad \forall r>0, c \in \mathbb{R},$$

and the estimator $T_1(X_1,\ldots,X_n)$ for σ is location-scale invariant if

$$T_1(rX_1+c,\ldots,rX_n+c) = rT_1(X_1,\ldots,X_n), \quad \forall r>0, c \in \mathbb{R}.$$

We know that MLE of μ and σ are location-scale invariant for μ and σ, respectively. Therefore under H_0, the distribution of $Z_i = (X_i - \hat{\mu})/\hat{\sigma}$ does not depend on μ and σ.

If G_n is the EDF based on data z_i, then

$$G_n(z_i) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}_{[z_j \leq z_i]} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{I}_{[x_j \leq x_i]} = F_n(x_i),$$

therefore

$$\mathrm{H}_n = \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_{(i)};\hat{\mu},\hat{\sigma})}{1+F_n(x_{(i)})}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(z_{(i)})}{1+G_n(z_{(i)})}\right).$$

Since the statistic H_n is a function of z_i, $i = 1,\ldots,n$, is location-scale invariant. As a consequence, the null distribution of H_n does not depend on the parameters μ and σ. ■

Proposition 2.5 *Let F_1 be an arbitrary continuous cdf in H_1. Then under the assumption that the observed sample have cdf F_1, the test based on H_n is consistent.*

Proof. Based on Glivenko-Cantelli theorem, for n large enough, we have that $F_n(x) \simeq F_1(x)$, for all $x \in \mathbb{R}$. Also $\hat{\mu}$ and $\hat{\sigma}$ are MLEs of μ and σ, respectively, and hence are consistent. Therefore

$$\begin{aligned}
\mathrm{H}_n &= \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_{(i)};\hat{\mu},\hat{\sigma})}{1+F_n(x_{(i)})}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_i;\hat{\mu},\hat{\sigma})}{1+F_n(x_i)}\right) \\
&\simeq \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_i;\hat{\mu},\hat{\sigma})}{1+F_1(x_i)}\right) \simeq \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_i,\mu,\sigma)}{1+F_1(x_i)}\right) \\
&\to E_{F_1}\left[h\left(\frac{1+F_0(X,\mu,\sigma)}{1+F_1(X)}\right)\right] =: D(F_0,F_1), \text{ as } n\to\infty,
\end{aligned}$$

where $E_{F_1}[.]$ is the expectation under F_1, and μ and σ^2 are, respectively, the expectation and variance of F_1. Note that the convergence holds by the law of large numbers and $D(F_0,F_1)$ is a divergence between F_0 and F_1. So the test based on H_n is consistent. ■

3. A normality test based on H_n

Many statistical procedures are based on the assumption that the observed data are normally distributed. Consequently, a variety of tests have been developed to check the validity of this assumption. In this section, we propose a new normality test based on H_n.

Consider again the goodness-of-fit testing problem described in (1), where now $f_0(x;\mu,\sigma) = 1/\sqrt{2\pi\sigma^2}e^{-(x-\mu)^2/2\sigma^2}$, $x \in \mathbb{R}$, in which $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown, and $F_0(.;\mu,\sigma)$ is the corresponding cdf, where $F_0(.)$ is the standard case for $F_0(.;0,1)$.

First we estimate μ and σ by their maximum likelihood estimators (MLEs), i.e., $\hat{\mu} = \bar{x} = 1/n\sum_{i=1}^{n} x_i$ and $\hat{\sigma}^2 = s^2 = 1/(n-1)\sum_{i=1}^{n}(x_i-\bar{x})^2$, respectively. Let $z_i = (x_i-\bar{x})/s$, $i = 1,\ldots,n$. Then, the test statistic for normality is:

$$\mathrm{H}_n = \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(x_{(i)},\bar{x},s)}{1+F_n(x_{(i)})}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(\frac{1+F_0(z_{(i)})}{1+i/n}\right), \tag{4}$$

where

$$h(x) = \left(\frac{x-1}{x+1}\right)^2. \tag{5}$$

Note that $h:(0,\infty)\to\mathbb{R}^+$ is decreasing on $(0,1)$ and increasing on $(1,\infty)$ with an absolute minimum at $x=1$ such that $h(1)=0$ (see Figure 1). We selected this function h, because based on simulation study, it is more powerful than other functions h. For example, we considered $h_2(x) := x\log(x)-x+1$ for comparison with $h_1(x) := \left(\frac{x-1}{x+1}\right)^2$ (see Tables 6 and 7).

Corollary 3.1 *The support of* H_n *is* $[0, 0.11]$.

Proof. From Proposition 2.3 and Figure 1, $\max(h(1/2), h(2)) = 0.11$. ∎

Table 1 contains the upper critical values of H_n, which have obtained by Monte Carlo from 100000 simulated samples for different sample sizes n and significance levels $\alpha = 0.01, 0.05, 0.1$.

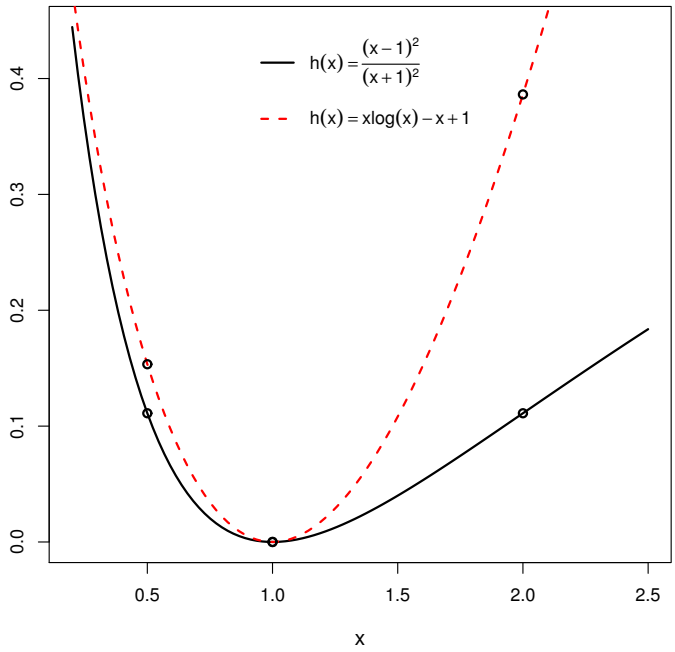

Figure 1: *Plot of function h.*

Table 1: *Critical values of* H_n *for* $\alpha = 0.01, 0.05, 0.1$.

n / α	5	6	7	8	9	10	15	20	25	30	40	50
0.01	.0039	.0035	.0030	.0026	.0023	.0021	.0014	.0011	.0008	.0007	.0005	.0004
0.05	.0030	.0026	.0022	.0019	.0017	.0016	.0010	.0007	.0006	.0005	.0004	.0003
0.10	.0026	.0022	.0019	.0016	.0015	.0013	.0009	.0006	.0005	.0004	.0003	.0002

Remember that, H_n is expected to take values close to zero when H_0 is true. Hence, H_0 will be rejected for large values of H_n. Also H_n is invariant under location-scale transformations and consistent under the assumption H_1, respectively, from Propositions 2.4 and 2.5.

4. Normality tests under evaluation

Comparison of the normality tests has received attention in the literature The goodness-of-fit tests have been discussed by many authors including Shapiro et al. (1968), Poitras (2006), Yazici and Yolacan (2007), Krauczi (2009), Romao et al. (2010), Yap and Sim (2010) and Alizadeh and Arghami (2011).

In this section we consider a large number (forty) of recent and classical statistics that have been used to test normality and in Section 5 we compare their performances with that of H_n. In the following we prefer to keep the original notation for each statistic. Concerning the notation, let $x_1, x_2, \ldots, x_n$ be a random sample of size n and $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ the corresponding order statistic. Also consider the sample mean, variance, skewness and kurtosis, defined by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad \sqrt{b_1} = \frac{m_3}{(m_2)^{3/2}}, \qquad b_2 = \frac{m_4}{(m_2)^2},$$

respectively, where the j-th central moment m_j is given by $m_j = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^j$ and finally consider $z_{(i)} = (x_{(i)} - \bar{x})/s$, for $i = 1, \ldots, n$.

1. Vasicek's entropy estimator (Vasicek, 1976):

$$\mathrm{KL}_{mn} = \frac{\exp\{\mathrm{HV}_{mn}\}}{s}$$

where

$$\mathrm{HV}_{mn} = \frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\}, \tag{6}$$

$m < n/2$ is a positive integer and $X_{(i)} = X_{(1)}$ if $i < 1$ and $X_{(i)} = X_{(n)}$ if $i > n$. H_0 is rejected for small values of KL. Vasicek (1976) showed that the maximum power for KL was typically attained by choosing $m = 2$ for $n = 10$, $m = 3$ for $n = 20$ and $m = 4$ for $n = 50$. The lower-tail 5%-significance values of KL for $n = 10, 20$ and 50 are 2.15, 2.77 and 3.34, respectively.

2. Ebrahimi's entropy estimator (Ebrahimi, Pflughoeft and Soofi, 1994):

$$\mathrm{TE}_{mn} = \frac{\exp\{\mathrm{HE}_{mn}\}}{s},$$

where

$$\mathrm{HE}_{mn} = \frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\}, \tag{7}$$

and $c_i = (1 + \frac{i-1}{m})\mathrm{I}_{[1,m]}(i) + 2\mathrm{I}_{[m+1,n-m]}(i) + (1 + \frac{n-i}{m})\mathrm{I}_{[n-m+1,n]}(i)$. Ebrahimi et al. (1994) proved the linear relationship between their estimator and (6). Thus for fixed values of n and m, the tests based on (6) and (7) have the same power.

3. Nonparametric distribution function of Vasicek's estimator:

$$\mathrm{TV}_{mn} = \log\sqrt{2\pi\hat{\sigma}_v^2} + 0.5 - \mathrm{HV}_{mn},$$

where HV_{mn} was defined in (6), $\hat{\sigma}_v^2 = \mathrm{Var}_{g_v}(X)$, and

$$g_v(x) = \begin{cases} 0 & x < \xi_1 \ \text{ or } \ x > \xi_{n+1}, \\ \frac{2m}{n(x_{(i+m)} - x_{(i-m)})} & \xi_i < x \le \xi_{i+1} \ \ i = 1, \ldots, n, \end{cases}$$

where $\xi_i = \left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right)/2m$. H_0 is rejected for large values of TV_{mn}. (See Park, 2003).

4. Nonparametric distribution function of Ebrahimi estimator:

$$\mathrm{TE}_{mn} = \log\sqrt{2\pi\hat{\sigma}_e^2} + 0.5 - \mathrm{HE}_{mn},$$

where HE_{mn} was defined in (7), $\hat{\sigma}_e^2 = \mathrm{Var}_{g_e}(X)$ and

$$g_e(x) = \begin{cases} 0 & x < \eta_1 \ \text{ or } \ x > \eta_{n+1} \\ \frac{1}{n(\eta_{i+1} - \eta_i)} & \eta_i < x \le \eta_{i+1} \ \ i = 1, \ldots, n, \end{cases}$$

with

$$\eta_i = \begin{cases} \xi_{m+1} - \frac{1}{m+k-1}\sum_{k=i}^{m}(x_{(m+k)} - x_{(1)}) & 1 \le i \le m, \\ \frac{1}{2m}\left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right) & m+1 \le i \le n-m+1, \\ \xi_{n-m+1} + \frac{1}{n+m-k+1}\sum_{k=n-m+2}^{i}(x_{(n)} - x_{(k-m-1)}) & n-m+2 \le i \le n+1, \end{cases}$$

and $\xi_i = \left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right)/2m$. H_0 is rejected for large values of TE_{mn}. (See Park, 2003).

5. Nonparametric distribution function of Alizadeh and Arghami estimator (Alizadeh Noughabi and Arghami, 2010, 2013):

$$\mathrm{TA}_{mn} = \log\sqrt{2\pi\hat{\sigma}_a^2} + 0.5 - \mathrm{HA}_{mn},$$

where

$$\mathrm{HA}_{mn} = \frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{n}{a_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right\},$$

with $a_i = \mathrm{I}_{[1,m]}(i) + 2\mathrm{I}_{[m+1,n-m]}(i) + \mathrm{I}_{[n-m+1,n]}(i)$, $\hat{\sigma}_a^2 = \mathrm{Var}_{g_a}(X)$ and

$$g_a(x) = \begin{cases} 0 & x < \eta_1 \ \text{ or } \ x > \eta_{n+1}, \\ \frac{1}{n(\eta_{i+1} - \eta_i)} & \eta_i < x \le \eta_{i+1} \ \ i = 1, \ldots, n, \end{cases}$$

with

$$\eta_i = \begin{cases} \xi_{m+1} - \frac{1}{m}\sum_{k=i}^{m}(x_{(m+k)} - x_{(1)}) & 1 \le i \le m, \\ \frac{1}{2m}\left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right) & m+1 \le i \le n-m+1, \\ \xi_{n-m+1} + \frac{1}{m}\sum_{k=n-m+2}^{i}(x_{(n)} - x_{(k-m-1)}) & n-m+2 \le i \le n+1, \end{cases}$$

and $\xi_i = \left(x_{(i-m)} + \cdots + x_{(i+m-1)}\right)/2m$. Also $m = [\sqrt{n}+1]$. H_0 is rejected for large values of TA_{mn}. The upper-tail 5%-significance values of TA for $n = 10, 20$ and 50 are 0.4422, 0.2805 and 0.1805, respectively.

6. Dimitriev and Tarasenko's entropy estimator (Dimitriev and Tarasenko, 1973):

$$\mathrm{TD}_{mn} = \frac{\exp\{\mathrm{HD}_{mn}\}}{s}$$

where

$$\mathrm{HD}_{mn} = -\int_{-\infty}^{\infty} \ln(\hat{f}(x))\hat{f}(x)\,dx,$$

where $\hat{f}(x)$ is the kernel density estimation of $f(x)$ given by

$$\hat{f}(X_i) = \frac{1}{nh}\sum_{j=1}^{n} k\left(\frac{X_i - X_j}{h}\right), \tag{8}$$

where k is a kernel function satisfying $\int_{-\infty}^{\infty} k(x)\,dx = 1$ and h is a bandwidth. The kernel function k being the standard normal density function and the bandwidth $h = 1.06\hat{\sigma}n^{-1/5}$. H_0 is rejected for small values of TD_{mn}.

7. Corea's entropy estimator (Corea, 1995):

$$\mathrm{TC}_{mn} = \frac{\exp\{\mathrm{HC}_{mn}\}}{s},$$

where

$$\mathrm{HC}_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \ln\left\{\frac{\sum_{j=i-m}^{i+m}\left(X_{(j)} - \tilde{X}_{(i)}\right)(j-i)}{n\sum_{j=i-m}^{i+m}\left(X_{(j)} - \tilde{X}_{(i)}\right)^2}\right\}$$

and $\tilde{X}_{(i)} = \sum_{j=i-m}^{i+m} X_{(j)}/(2m+1)$. H_0 is rejected for small values of TC_{mn}.

8. Van Es's entropy estimator (Van Es, 1992):

$$\mathrm{TEs}_{mn} = \frac{\exp\{\mathrm{HEs}_{mn}\}}{s},$$

where

$$\mathrm{HEs}_{mn} = \frac{1}{n-m}\sum_{i=1}^{n-m}\left\{\ln\left(\frac{n+1}{m}(X_{(i+m)} - X_{(i)})\right)\right\} + \sum_{k=m}^{n}\frac{1}{k} + \ln(m) - \ln(n+1).$$

H_0 is rejected for small values of TEs_{mn}.

9. Zamanzade and Arghami's entropy estimator (Zamanzade and Arghami, 2012):

$$\mathrm{TZ1}_{mn} = \frac{\exp\{\mathrm{HZ1}_{mn}\}}{s},$$

where $\mathrm{HZ1}_{mn} = \frac{1}{n}\sum_{i=1}^{n}\ln(b_i)$, with

$$b_i = \frac{X_{(i+m)} - X_{(i-m)}}{\sum_{j=k_1(i)}^{k_2(i)-1}(\hat{f}(X_{(j+1)}) + \hat{f}(X_{(j)}))(X_{(j+1)} - X_{(j)})/2} \tag{9}$$

where $\hat{f}$ is defined as in (8) with the kernel function k being the standard normal density function and the bandwidth $h = 1.06\hat{\sigma}n^{-1/5}$. H_0 is rejected for small values of TZ1. For $n = 10, 20$ and 50, the lower-tail 5%-significance critical values are 3.403, 3.648 and 3.867.

10. Zamanzade and Arghami's entropy estimator (Zamanzade and Arghami, 2012):

$$\mathrm{TZ2}_{mn} = \frac{\exp\{\mathrm{HZ2}_{mn}\}}{s},$$

where $\mathrm{HZ2}_{mn} = \sum_{i=1}^{n} w_i \ln(b_i)$, being coefficients b_i's were defined in (9) and

$$w_i = \begin{cases} (m+i-1)/\sum_{i=1}^{n} w_i & 1 \le i \le m, \\ 2m/\sum_{i=1}^{n} w_i & m+1 \le i \le n-m, \\ (n-i+m)/\sum_{i=1}^{n} w_i & n-m+1 \le i \le n, \end{cases} \qquad i = 1, \ldots, n,$$

are weights proportional to the number of points used in computation of b_i's. H_0 is rejected for small values of TZ2. For $n = 10, 20$ and 50, the lower-tail 5%-significance critical values are 3.321, 3.520 and 3.721.

11. Zhang and Wu's statistics (Zhang and Wu, 2005):

$$\mathrm{Z_K} = \max_{1 \le i \le n}\left[(i-0.5)\ln\frac{i-0.5}{nF_0(Z_{(i)})} + (n-i+0.5)\ln\frac{n-i+0.5}{n(1-F_0(Z_{(i)}))}\right],$$

$$\mathrm{Z_C} = \sum_{i=1}^{n}\left(\log\frac{(1/F_0(Z_{(i)})-1)}{(n-0.5)/(i-0.75)-1}\right)^2,$$

and

$$\mathrm{Z_A} = -\sum_{i=1}^{n}\left(\frac{\log F_0(Z_{(i)})}{n-i+0.5} + \frac{\log(1-F_0(Z_{(i)})}{i-0.5}\right),$$

The null hypothesis H_0 is rejected for large values of the three test statistics.

12. Classical test statistics for normality based skewness and kurtosis from D'Agostino and Pearson (D'Agostino and Pearson, 1973):

$$\sqrt{b_1} = \frac{m_3}{(m_2)^{3/2}}, \qquad b_2 = \frac{m_4}{(m_2)^2},$$

The null hypothesis H_0 is rejected for both small and large values of the two test statistics.

13. Transformed skewness and kurtosis statistic from D'Agostino et al. (1990):

$$\mathrm{K}^2 = \left[Z(\sqrt{b_1})\right]^2 + [Z(b_2)]^2,$$

where

$$Z(\sqrt{b_1}) = \frac{\log(Y/c + \sqrt{(Y/c)^2+1})}{\sqrt{\log(w)}},$$
$$Z(b_2) = \left[\left(1 - \frac{2}{9A}\right) - \sqrt[3]{\frac{1-2/A}{1+y\sqrt{2/(A-4)}}}\right]\sqrt{\frac{9A}{2}},$$

where

$$c_1 = 6 + 8/c_2(2/c_2 + \sqrt{1+4/c_2^2}),$$

$$c_2 = (6(n^2-5n+2)/(n+7)(n+9))\sqrt{6(n+3)(n+5)/n(n-2)(n-3)},$$

$$c_3 = (b_2 - 3(n-1)/(n+1))/\sqrt{24n(n-2)(n-3)/(n+1)^2(n+3)(n+5)}.$$

and

$$Y = \sqrt{b_1}\sqrt{\frac{(n+1)(n+3)}{6(n-2)}}, \qquad w^2 = \sqrt{2\beta_2 - 1} - 1,$$
$$\beta_2 = \frac{3(n^2+27n-70)(n+1)(n+3)}{(n-2)(n+5)(n+7)(n+9)}; \quad c = \sqrt{\frac{2}{(w^2-1)}}.$$

Transformed skewness $Z(\sqrt{b_1})$ and transformed kurtosis $Z(b_2)$ is obtained by D'Agostino (1970) and Anscombe and Glynn (1983), respectively. The null hypothesis H_0 is rejected for large values of K^2.

14. Transformed skewness and kurtosis statistic by Doornik and Hansen (1994):

$$\mathrm{DH} = \left[Z(\sqrt{b_1})\right]^2 + z_2^2,$$

where

$$z_2 = \left[\left(\frac{\xi}{2a}\right)^{1/3} - 1 + \frac{1}{9a}\right]\sqrt{9a},$$

and

$$\xi = (b_2 - 1 - b_1)2k,$$

$$k = \frac{(n+5)(n+7)(n^3+37n^2+11n-313)}{12(n-3)(n+1)(n^2+15n-4)},$$

$$a = \frac{(n+5)(n+7)\left((n-2)(n^2+27n-70)+b_1(n-7)(n^2+2n-5)\right)}{6(n-3)(n+1)(n^2+15n-4)},$$

Transformed kurtosis z_2 is obtained by Shenton and Bowman (1977). The null hypothesis H_0 is rejected for large values of DH.

15. Bonett and Seier's statistic (Bonett and Seier, 2002):

$$\mathrm{Z_w} = \frac{\sqrt{n+2}(\hat{w}-3)}{3.54},$$

where $\hat{w} = 13.29\left(\ln\sqrt{m_2} - \log\left(n^{-1}\sum_{i=1}^{n}|x_i - \bar{x}|\right)\right)$. H_0 is rejected for both small and large values of $\mathrm{Z_w}$.

16. D'Agostino's statistic (D'Agostino, 1971):

$$\mathrm{D} = \frac{\sum_{i=1}^{n}(i-(n+1)/2)X_{(i)}}{n^2\sqrt{\sum_{i=1}^{n}\left(x_{(i)}-\bar{X}\right)^2}},$$

H_0 is rejected for both small and large values of D.

17. Chen and Shapiro's statistic (Chen and Shapiro, 1995):

$$\mathrm{QH} = \frac{1}{(n-1)s}\sum_{i=1}^{n-1}\frac{X_{(i+1)}-X_{(i)}}{M_{(i+1)}-M_{(i)}},$$

where $M_i = \Phi^{-1}((i-0.375)/(n+0.25))$, where Φ is the cdf of a standard normal random variable. H_0 is rejected for small values of QH.

18. Filliben's statistic (Filliben, 1975):

$$\mathrm{r} = \frac{\sum_{i=1}^{n} x_{(i)} M_{(i)}}{\sqrt{\sum_{i=1}^{n} M_{(i)}^2}\sqrt{(n-1)s^2}},$$

where $M_{(i)} = \Phi^{-1}(m_{(i)})$ and $m_{(1)} = 1-0.5^{1/n}$, $m_{(n)} = 0.5^{1/n}$ and $m_{(i)} = (i-0.3175)/(n+0.365)$ for $i = 2,\ldots,n-1$. H_0 is rejected for small values of r.

19. del Barrio et al.'s statistic (del Barrio et al., 1999):

$$\mathrm{R}_n = 1 - \frac{\left(\sum_{k=1}^{n} X_{(k)} \int_{(k-1)/n}^{k/n} F_0^{-1}(t)\, dt\right)^2}{m_2},$$

where m_2 is the sample standardized second moment. H_0 is rejected for large values of R_n.

20. Epps and Pulley statistic (Epps and Pulley, 1983):

$$\mathrm{T}_{\mathrm{EP}} = \frac{1}{\sqrt{3}} + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\exp\left(\frac{-(X_j - X_k)^2}{2m_2}\right) - \frac{\sqrt{2}}{n}\sum_{j=1}^{n}\exp\left(\frac{-(X_j-\bar{X})^2}{4m_2}\right),$$

where m_2 is the sample standardized second moment. H_0 is rejected for large values of T_{EP}.

21. Martinez and Iglewicz's statistic (Martinez and Iglewicz, 1981):

$$\mathrm{I}_n = \frac{\sum_{i=1}^{n}(X_i - M)^2}{(n-1)S_b^2},$$

where M is is the sample median and

$$S_b^2 = \frac{n\sum_{|\tilde{Z}_i|<1}(X_i - M)^2(1-\tilde{Z}_i^2)^4}{\left(\sum_{|\tilde{Z}_i|<1}(1-\tilde{Z}_i^2)(1-5\tilde{Z}_i^2)\right)^2},$$

with $\tilde{Z}_i = (X_i - M)/(9A)$ for $|\tilde{Z}_i| < 1$ and $\tilde{Z}_i = 0$ otherwise, and A is the median of $|X_i - M|$. H_0 is rejected for large values of $\mathrm{I_n}$.

22. deWet and Venter statistic (de Wet and Venter, 1972):

$$\mathrm{E}_n = \sum_{i=1}^{n} \left(X_{(i)} - \bar{X} - s\Phi^{-1}\left(\frac{i}{n+1}\right)\right)^2 \Big/ s^2.$$

H_0 is rejected for large values of $\mathrm{E_n}$.

23. Optimal test (Csörgo and Révész, 1971):

$$\mathrm{M}_n = \sum_{i=1}^{n} \left(X_{(i)} - \bar{X} - s\Phi^{-1}\left(\frac{i}{n+1}\right)\right)^2 \phi\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)\left[\Phi^{-1}\left(\frac{i}{n+1}\right)\right]^{\lambda-1}.$$

H_0 is rejected for large values of M_n.

24. Pettitt statistic (Pettitt, 1977):

$$\mathrm{Q}_n = \sum_{i=1}^{n} \left(\Phi\left(\frac{X_{(i)} - \bar{X}}{s}\right) - \frac{i}{n+1}\right)^2 \left[\phi\left(\Phi^{-1}\left(\frac{i}{n+1}\right)\right)\right]^{-2}.$$

H_0 is rejected for large values of $\mathrm{Q_n}$.

25. Three test statistics from LaRiccia (1986):

$$\mathrm{T}_{1n} = C_{1n}^2/(s^2 B_{1n}), \qquad \mathrm{T}_{2n} = C_{2n}^2/(s^2 B_{2n}), \qquad \mathrm{T}_{3n} = \mathrm{T}_{1n} + \mathrm{T}_{2n},$$

where

$$C_{1n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[W_1\left(\frac{i}{n+1}\right) - A_{1n}\right] X_{(i)},$$
$$C_{2n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[W_2\left(\frac{i}{n+1}\right) - A_{2n}\Phi^{-1}\left(\frac{i}{n+1}\right)\right] X_{(i)},$$

Also $W_1(u) = [\Phi^{-1}(u)]^2 - 1$ and $W_2(u) = [\Phi^{-1}(u)]^3 - 3\Phi^{-1}(u)$. The constants A_{1n}, A_{2n}, B_{1n} and B_{2n} are given in Table 1 from LaRiccia (1986). For all three statistics H_0 is rejected for large value.

26. Kolmogorov-Smirnov's (Lilliefors) statistic (Kolmogorov, 1933):

$$\mathrm{KS} = \max\left\{ \max_{1\le j\le n}\left[\frac{j}{n} - F_0(Z_{(j)})\right], \max_{1\le j\le n}\left[F_0(Z_{(j)}) - \frac{j-1}{n}\right]\right\}.$$

Lilliefors (1967) computed estimated critical points for the Kolmogorov-Smirnov's test statistic for testing normality when mean and variance estimated.

27. Kuiper's statistic (Kuiper, 1962):

$$\mathrm{V} = \max_{1\leq j\leq n}\left[\frac{j}{n} - F_0(Z_{(j)})\right] + \max_{1\leq j\leq n}\left[F_0(Z_{(j)}) - \frac{j-1}{n}\right].$$

Louter and Kort (1970) computed estimated critical points for the Kuiper test statistic for testing normality when mean and variance estimated.

28. Cramér-von Mises' statistic (Cramér, 1928 and von Mises, 1931):

$$\mathrm{W}^2 = \frac{1}{12n} + \sum_{j=1}^{n}\left(F_0(Z_{(j)}) - \frac{2j-1}{2n}\right)^2.$$

29. Watson's statistic (Watson, 1961):

$$\mathrm{U}^2 = \mathrm{W}^2 - n\left(\frac{1}{n}\sum_{j=1}^{n}F_0(Z_{(j)}) - \frac{1}{2}\right)^2.$$

30. Anderson-Darling's statistic (Anderson, 1954):

$$\mathrm{A}^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log(F_0(Z_{(i)})) + \log\left(1 - F_0(Z_{(n-i+1)})\right)\right).$$

These classical tests are based on the empirical distribution function and H_0 is rejected for large values of KS, V, W^2, U^2 and A^2.

31. Pearson's chi-square statistic (D'Agostino and Stephens, 1986):

$$\mathrm{P} = \sum_i (C_i - E_i)^2/E_i,$$

where C_i is the number of counted and E_i is the number of expected observations (under H_0) in class i. The classes are build is such a way that they are equiprobable under the null hypothesis of normality. The number of classes used for the test is $\lceil 2n^{2/5}\rceil$ where $\lceil . \rceil$ is ceiling function.

32. Shapiro-Wilk's statistic (Shapiro and Wilk, 1965):

$$\mathrm{SW} = \frac{\left(\sum_{i=1}^{[n/2]} a_{(n-i+1)} \left(X_{(n-i+1)} - X_{(i)}\right)\right)^2}{\sum_{i=1}^{n} \left(X_{(i)} - \bar{X}\right)^2},$$

where coefficients a_i's are given by

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}, \tag{10}$$

and $m^T = (m_1, \ldots, m_n)$ and V are, respectively, the vector of expected values and the covariance matrix of the order statistic of n iid random variables sampled from the standard normal distribution. H_0 is rejected for small values of SW.

33. Shapiro-Francia's statistic (Shapiro and Francia, 1972) is a modification of SW. It is defined as

$$\mathrm{SF} = \frac{\left(\sum_{i=1}^{n} b_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_{(i)} - \bar{X})^2},$$

where

$$(b_1, \ldots, b_n) = \frac{m^T}{(m^T m)^{1/2}}$$

and m is defined as in (10). H_0 is rejected for small values of SF.

34. SJ statistic discussed in Gel, Miao and Gastwirth (2007). It is based on the ratio of the classical standard deviation $\hat{\sigma}$ and the robust standard deviation J_n (average absolute deviation from the median (MAAD)) of the sample data

$$\mathrm{SJ} = \frac{s}{J_n}, \tag{11}$$

where $J_n = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |X_i - M|$ and M is the sample median. H_0 is rejected for large values of SJ.

35. Jarque-Bera's statistic (Jarque and Bera, 1980, 1987):

$$\mathrm{JB} = \frac{n}{6} b_1 + \frac{n}{24} (b_2 - 3)^2,$$

where $\sqrt{b_1}$ and b_2 are the sample skewness and sample kurtosis, respectively. H_0 is rejected for large values of JB.

36. Robust Jarque-Bera's statistic (Gel and Gastwirth, 2008):

$$\text{RJB} = \frac{n}{C_1}\left(\frac{m_3}{J_n^3}\right)^2 + \frac{n}{C_2}\left(\frac{m_4}{J_n^4} - 3\right)^2,$$

where J_n is defined as in (11), C_1 and C_2 are positive constants. For a 5%-significance level, $C_1 = 6$ and $C_2 = 64$ according to Monte Carlo simulations. H_0 is rejected for large values of RJB.

5. Simulation study

In this section we study the power of the normality test based on H_n and compare it with a large number of recent and classical normality tests. To facilitate comparisons of the power of the present test with the powers of the mentioned tests, we select two sets of alternative distributions:

Set 1. Alternatives listed in in Esteban et al. (2001).

Set 2. Alternatives listed in Gan and Koehler (1990) and Krauczi (2009).

Set 1 of alternative distributions

Following Esteban et al. (2001) we consider the following alternative distributions, that can be classified in four groups:

Group I: Symmetric distributions with support on $(-\infty, \infty)$:

- Standard Normal (N);
- Student's t (t) with 1 and 3 degrees of freedoms;
- Double Exponential (DE) with parameters $\mu = 0$ (location) and $\sigma = 1$ (scale);
- Logistic (L) with parameters $\mu = 0$ (location) and $\sigma = 1$ (scale);

Group II: Asymmetric distributions with support on $(-\infty, \infty)$:

- Gumbel (Gu) with parameters $\alpha = 0$ (location) and $\beta = 1$ (scale);
- Skew Normal (SN) with with parameters $\mu = 0$ (location), $\sigma = 1$ (scale) and $\alpha = 2$ (shape);

Group III: Distributions with support on $(0,\infty)$:

- Exponential (Exp) with mean 1;
- Gamma (G) with parameters $\beta = 1$ (scale) and $\alpha = .5, 2$ (shape);
- Lognormal (LN) with parameters $\mu = 0$ and $\sigma = .5, 1, 2$;
- Weibull (W) with parameters $\beta = 1$ (scale) and $\alpha = .5, 2$ (shape);

Group IV: Distributions with support on $(0,1)$:

- Uniform (Unif);
- Beta (B) with parameters (2,2), (.5,.5), (3,1.5) and (2,1).

Set 2 of alternative distributions

Gan and Koehler (1990) and Krauczi (2009) considered a battery of "difficult alternatives" for comparing normality tests. We also consider them in order to evaluate the sensitivity of the proposed test. Let U and Z denote a $[0,1]$-Uniform and a Standard Normal random variable, respectively.

- Contaminated Normal distribution (CN) with parameters $(\lambda, \mu_1, \mu_2, \sigma)$ given by the cdf $F(x) = (1-\lambda)F_0(x, \mu_1, 1) + \lambda F_0(x, \mu_2, \sigma)$;
- Half Normal (HN) distribution, that is, the distribution of $|Z|$.
- Bounded Johnson's distribution (SB) with parameters (γ, δ) of the random variable $e^{(Z-\gamma)/\delta}/(1+e^{(Z-\gamma)/\delta})$;
- Unbounded Johnson's distribution (UB) with parameters (γ, δ) of the random variable $\sinh((Z-\gamma)/\delta)$;
- Triangle type I (Tri) with density function $f(x) = 1 - |t|$, $-1 < t < 1$;
- Truncated Standard Normal distribution at a and b (TN);
- Tukey's distribution (Tu) with parameter λ of the random variable $U^\lambda - (1-U)^\lambda$.
- Cauchy distribution with parameters $\mu = 0$ (location), $\sigma = 1$ (scale).
- Chi-squared distribution χ^2 with k degrees of freedom.

Tables 2-3 contain the skewness $(\sqrt{\beta_1})$ and kurtosis (β_2) of the previous sets of alternative distributions. Alternatives in *Set 2* are roughly ordered and grouped in five groups according to their skewness and kurtosis values in Table 3. These groups correspond to: symmetric short tailed, symmetric closed to normal, asymmetric short tailed, asymmetric long tailed. Figure 2 illustrates some of the possible shapes of the pdf's of the alternatives in *Set 1* and *Set 2*.

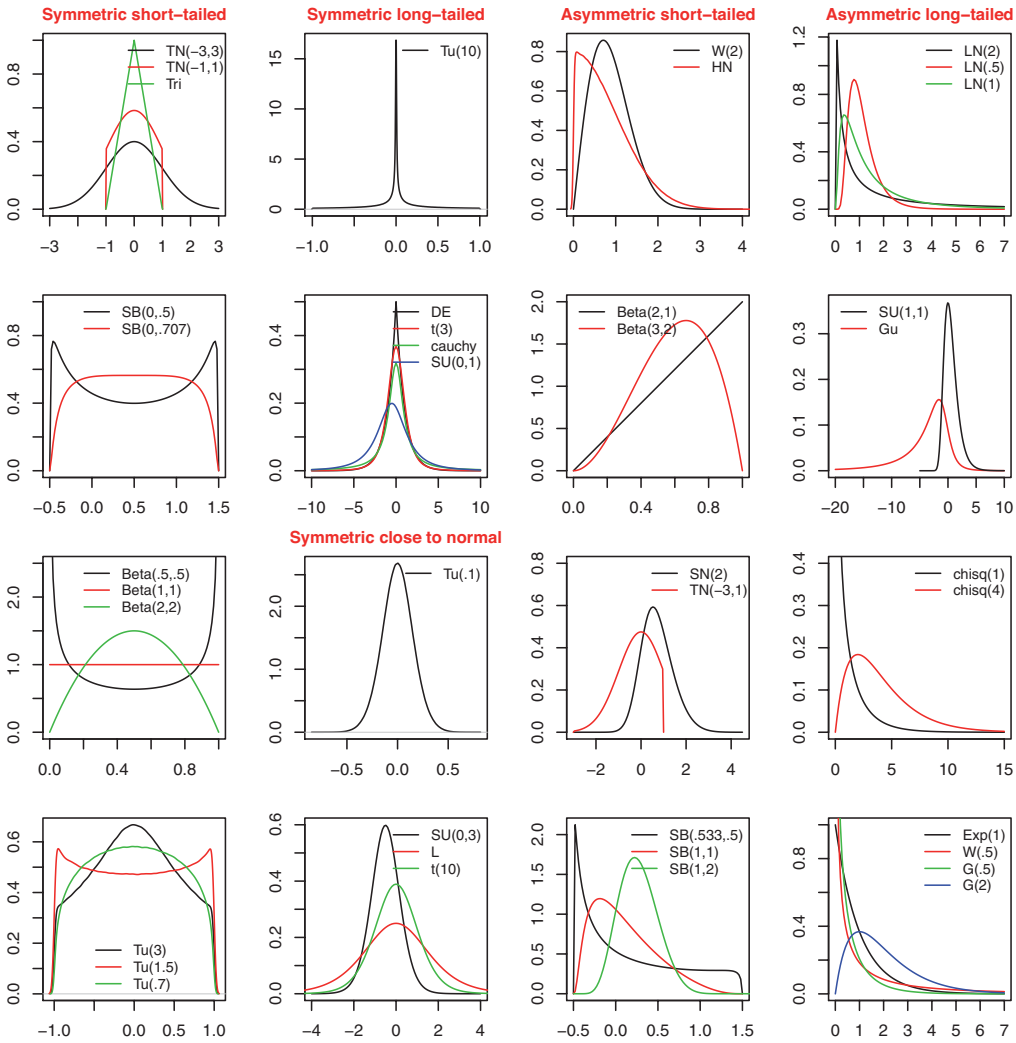

Figure 2: *Plots of alternative distributions in* Set 1 *and* Set 2.

Tables 4-5 contain the estimated value of $\mathrm{H_n}$ (for $h(x) = (x-1)^2/(x+1)^2$ and $h(x) = x\log(x) - x + 1$, respectively), for each alternative distribution, computed as the average value from 10000 simulated samples of sizes $n = 10, 20, 50, 100, 1000$. In the last row of these tables $(n = \infty)$), we show the value of $D(F_0, F_1)$ computed with the the command `integrate` in R Software, with (μ) and (σ^2) being the expectation and variance of F_1, respectively. *These tables show consistency of the test statistic* H_n.

Tables 6-7 report the power of the 5% significance level of forty normality tests based on the statistics considered in Section 4 for the *Set 1* of alternatives.

Tables 8-9 contain the power of the 5% significance level test of normality based on the most powerful statistics and the alternatives listed in *Set 2*.

Table 2: *Skewness and kurtosis of alternative distributions in* Set 1.

	Group I				Group II		Group III								Group IV				
	t(1)	t(3)	L	DE	Gu	SN(2)	Exp	G(2)	G(.5)	LN(1)	LN(2)	LN(.5)	W(.5)	W(2)	Unif	B(2,2)	B(.5,.5)	B(3,.5)	B(2,1)
$\sqrt{\beta_1}$	0	0	0	0	1.30	.45	2	1.41	2.83	6.18	414.36	1.75	6.62	.63	0	0	0	−1.575	−.57
β_2	—	—	4.2	6	5.4	.31	9	6	15	113.94	9220560	8.90	87.72	3.25	1.8	2.14	1.5	5.22	2.4

Table 3: *Skewness and kurtosis of alternative distributions in* Set 2.

	Symmetric													Asymmetric							
	Short tailed							Close to Normal			Long tailed			Short tailed					Long tailed		
	Tu	Tu	Tu	SB	Tri	TN	TN	Tu	SU	t	Tu	SU	caushy	TN	SB	SB	SB	HN	SU	χ^2	χ^2
	(.7)	(1.5)	(3)	(0,.5)		(−1,1)	(−3,3)	(.1)	(0,3)	(10)	(10)	(0,1)		(−3,1)	(1,1)	(1,2)	(.533,.5)		(1,1)	(1)	(4)
$\sqrt{\beta_1}$	0	0	0	0	0	0	0	0	0	0	0	0	0	−.55	.73	.28	.65	.97	−5.37	2.83	1.41
β_2	1.92	1.75	2.06	1.63	2.4	1.94	2.84	3.21	3.53	4	5.38	36.2	∞	2.78	2.91	2.77	2.13	3.78	93.4	15	6

Table 4: *Estimated value of* H_n *with* $h_1(x) = (x-1)^2/(x+1)^2$ *under* H_1*, based on 10000 simulations for several values of n.*

	Group I				Group II		Group III								Group IV				
	t(1)	t(3)	L	DE	Gu	SN(2)	Exp	G(2)	G(.5)	LN(1)	LN(2)	LN(.5)	W(.5)	W(2)	Unif	B(2,2)	B(.5,.5)	B(3,.5)	B(2,1)
n																			
10	—	.0011	.00086	.0010	.0011	.00092	.0017	.0013	.0025	.00226	.0040	.0013	.0035	.00097	.0009	.00082	.0012	.0013	.0008
20	—	.0007	.00043	.0006	.0007	.00047	.0014	.0009	.0023	.00213	.0045	.0009	.0036	.00054	.0005	.00041	.0008	.0011	.0005
50	—	.0005	.00018	.0004	.0004	.00022	.0012	.0006	.0022	.00211	.0052	.0007	.0037	.00028	.0003	.00019	.0006	.0011	.0003
100	—	.0004	.00011	.0003	.0003	.00013	.0011	.0006	.0022	.00215	.0056	.0006	.0039	.00019	.0003	.00012	.0006	.0011	.0003
1000	—	.0004	.00004	.0002	.0002	.00006	.0010	.0005	.0021	.00226	.0066	.0005	.0040	.00012	.0002	.00006	.0005	.0011	.0002
∞	—	.0004	.00003	.0002	.0002	.00005	.0010	.0005	.0021	.00228	.0074	.0005	.0040	.00011	.0002	.00006	.0005	.0011	.0002

Table 5: *Estimated value of* H_n *with* $h_2(x) = x\log(x) - x + 1$ *under* H_1*, based on 10000 simulations for several values of n.*

	Group I				Group II		Group III								Group IV				
	t(1)	t(3)	L	DE	Gu	SN(2)	Exp	G(2)	G(.5)	LN(1)	LN(2)	LN(.5)	W(.5)	W(2)	Unif	B(2,2)	B(.5,.5)	B(3,.5)	B(2,1)
n																			
10	—	.0021	.00167	.0019	.0022	.0018	.0034	.0027	.0048	.0044	.0077	.0026	.0065	.0020	.0019	.0017	.0025	.0027	.0017
20	—	.0014	.00086	.0012	.0013	.0009	.0028	.0017	.0045	.0042	.0088	.0018	.0070	.0010	.0011	.0009	.0017	.0024	.0010
50	-	.0010	.00037	.0007	.0008	.0004	.0023	.0013	.0044	.0042	.0106	.0013	.0075	.0005	.0006	.0004	.0013	.0023	.0006
100	-	.0009	.00021	.0006	.0006	.0003	.0022	.001	.0043	.0043	.0113	.0012	.0079	.0004	.0005	.0003	.0012	.0023	.0005
1000	—	.0009	.00007	.0004	.0005	.0001	.0021	.0009	.0043	.0046	.0139	.0010	.0084	.0002	.0004	.0001	.0011	.0023	.0004
∞	—	.0009	.00006	.0004	.0005	.0001	.0021	.0009	.0043	.0047	.0163	.0010	.0084	.0002	.0004	.0001	.0010	.0023	.0004

Table 6: *Power comparisons for the normality test for* Set 1 *of alternative distributions,* $\alpha = 0.05$, $n = 10$.

	Group		I				II		III								IV				
	altern.	N	t(1)	t(3)	L	DE	Gu	SN	Exp	G(2)	G(.5)	LN(1)	LN(2)	LN(.5)	W(.5)	W(2)	Unif	B(2,2)	B(.5,.5)	B(3,.5)	B(2,1)
1	KL	.048	.442	.091	.051	.091	.101	.058	.416	.179	.782	.552	.938	.181	.931	.075	.167	.082	.512	.108	.173
2	TV	.048	.375	.082	.048	.053	.092	.055	.397	.151	.762	.519	.933	.144	.923	.073	**.181**	.084	**.514**	.656	.170
3	TE	.052	.460	.112	.058	.077	.111	.059	.454	.185	.794	.581	.945	.181	.935	.074	.158	.071	.481	.686	.164
4	TA	.053	.507	.134	.065	.094	.124	.062	.477	.213	**.810**	.616	**.951**	.208	**.940**	.080	.129	.064	.451	**.704**	.162
5	TD	.051	.583	.201	.087	.163	.154	.071	.394	.222	.631	.565	.869	.249	.813	.076	.028	.025	.080	.065	.093
6	TC	.054	.409	.083	.047	.057	.097	.053	.404	.173	.786	.542	.936	.171	.926	.071	.170	**.086**	.489	.110	**.182**
7	TEs	.049	.591	.167	.074	.140	.113	.062	.330	.158	.679	.485	.892	.176	.876	.064	.061	.037	.238	.064	.092
8	TZ1	.053	.632	.212	.089	.177	.145	.068	.359	.209	.581	.524	.846	.229	.784	.074	.030	.025	.078	.061	.081
9	TZ2	.051	.638	.216	.091	.181	.144	.066	.353	.205	.572	.516	.840	.228	.776	.073	.026	.023	.060	.058	.076
10	Z_K	.055	.587	.174	.075	.154	.126	.071	.352	.180	.636	.509	.885	.192	.842	.079	.078	.053	.221	.510	.109
11	Z_C	.053	.580	.183	.079	.154	.157	.074	.450	.245	.740	.606	.926	.248	.898	.089	.094	.044	.336	.621	.130
12	Z_A	.053	.608	.199	.083	.167	.162	.071	.457	.246	.744	.612	.928	.255	.901	.086	.050	.032	.204	.621	.115
13	$\sqrt{b_1}$	.057	.587	.219	.096	.184	.165	.073	.372	.226	.557	.532	.928	.247	.751	.088	.019	.024	.035	.437	.083
14	b_2	.053	.536	.170	.073	.136	.113	.060	.227	.148	.340	.353	.907	.159	.508	.072	.115	.057	.270	.235	.092
15	K^2	.058	.592	.220	.096	.190	.154	.073	.314	.197	.467	.464	.754	.221	.662	.082	.020	.021	.065	.336	.067
16	DH	.055	.625	.207	.084	.183	.130	.067	.344	.183	.590	.507	.860	.195	.797	.069	.071	.037	.238	.467	.093
17	Z_W	.055	.501	.150	.068	.130	.075	.088	.125	.091	.181	.210	.416	.097	.311	.055	.100	.056	.215	.123	.073
18	D	.051	.584	.175	.071	.142	.111	.060	.270	.146	.478	.434	.799	.168	.717	.064	.042	.044	.039	.335	.061
19	QH	.053	.598	.189	.081	.159	.160	.075	.455	.245	.742	.609	.928	.250	.901	.090	.094	.046	.321	.625	.135
20	r	.054	.635	.214	.088	.187	.160	.074	.421	.231	.692	.578	.907	.245	.868	.089	.042	.031	.164	.561	.099
21	R_n	.054	.609	.196	.083	.167	.162	.075	.448	.244	.733	.604	.924	.251	.894	.090	.077	.042	.276	.613	.125
22	T_{EP}	.053	.602	.200	.088	.170	.167	.077	.427	.244	.663	.587	.891	.256	.842	.070	.054	.040	.152	.538	.115
23	I_n	.055	.157	.151	.084	.151	.120	.066	.209	.149	.199	.215	.100	.151	.134	.070	.024	.025	.043	.207	.065
24	E_n	.055	.638	.218	.089	.193	.158	.073	.407	.226	.670	.567	.898	.240	.852	.082	.035	.028	.126	.536	.091
25	M_n	.054	.631	.226	.095	.198	.147	.071	.326	.189	.524	.484	.808	.214	.733	.073	.014	.020	.029	.385	.061
26	Q_n	.053	.604	.175	.074	.152	.141	.071	.426	.220	.728	.585	.923	.222	.894	.081	.094	.051	.285	.610	.130
27	T_{1n}	.054	.516	.179	.083	.145	.173	.072	.475	.264	.726	.626	.918	.274	.884	.095	.036	.030	.093	.605	.114
28	T_{2n}	.053	.555	.168	.072	.155	.075	.055	.106	.075	.167	.206	.453	.088	.326	.049	.090	.046	.284	.105	.060
29	T_{3n}	.057	.647	.225	.093	.204	.146	.070	.360	.199	.625	.518	.882	.216	.831	.074	.039	.026	.203	.487	.076
30	KS	.053	.581	.164	.073	.148	.124	.072	.312	.170	.545	.469	.828	.182	.761	.078	.066	.051	.163	.424	.103
31	V	.050	.593	.163	.071	.143	.119	.065	.365	.180	.662	.530	.894	.188	.856	.074	.087	.054	.240	.540	.108
32	W^2	.052	.624	.186	.080	.164	.143	.073	.396	.210	.674	.562	.898	.220	.860	.082	.083	.050	.236	.552	.116
33	U^2	.052	.618	.178	.076	.159	.135	.071	.382	.200	.661	.547	.893	.211	.853	.081	.091	.056	.260	.540	.120
34	A^2	.051	.619	.190	.083	.165	.147	.073	.417	.225	.670	.578	.911	.233	.877	.085	.086	.048	.268	.580	.126
35	P	.042	.531	.148	.083	.136	.127	.080	.397	.200	.704	.545	.903	.199	.878	.087	.086	.061	.229	.594	.136
36	SW	.052	.597	.187	.082	.159	.159	.075	.451	.245	.740	.608	.927	.248	.899	.088	.090	.045	.312	.622	.133
37	SF	.054	.631	.214	.088	.185	.161	.074	.426	.234	.701	.584	.912	.248	.872	.085	.047	.033	.183	.571	.104
38	SJ	.055	**.655**	.217	.096	**.211**	.121	.068	.253	.147	.429	.416	.756	.176	.660	.060	.012	.021	.022	.285	.046
39	JB	.059	.600	.223	.096	.192	.149	.075	.352	.219	.532	.511	.804	.242	.731	.087	.016	.021	.029	.396	.073
40	RJB	.056	.644	**.228**	**.097**	.205	.165	.072	.485	.189	.504	.470	.784	.214	.700	.076	.015	.021	.025	.345	.061
h_2	H_n	.051	.596	.173	.074	.150	.190	.091	.504	.285	.780	.659	.940	.290	.918	.114	.074	.046	.218	.331	.054
h_1	H_n	.051	.587	.169	.073	.144	**.199**	**.095**	**.516**	**.296**	.784	**.665**	.942	**.301**	.920	**.119**	.079	.049	.220	.300	.047

Table 7: *Power comparisons for the normality test for* Set 1 *of alternative distributions,* $\alpha = 0.05$, $n = 20$.

	Group		I				II		III								IV				
	altern.	N	t(1)	t(3)	L	DE	Gu	SN	Exp	G(2)	G(.5)	LN(1)	LN(2)	LN(.5)	W(.5)	W(2)	Unif	B(2,2)	B(.5,.5)	B(3,.5)	B(2,1)
1	KL	.045	.737	.165	.051	.091	.198	.073	.846	.457	.992	.927	.999	.404	**1.00**	.132	.442	.131	.914	.224	**.438**
2	TV	.047	.684	.121	.046	.062	.176	.067	.830	.429	.992	.910	**1.00**	.364	**1.00**	.126	**.443**	**.136**	**.910**	.980	.428
3	TE	.047	.786	.205	.064	.129	.237	.079	.865	.508	**.993**	.934	**1.00**	.445	**1.00**	.143	.391	.112	.891	**.984**	.423
4	TA	.048	.858	.301	.095	.229	.279	.101	**.870**	.533	**.993**	**.937**	**1.00**	.485	**1.00**	.145	.258	.064	.824	.983	.358
5	TD	.049	.872	.371	.134	.304	.310	.102	.790	.507	.959	.909	.997	.517	.995	.148	.084	.028	.408	.129	.221
6	TC	.047	.687	.138	.043	.070	.185	.076	.836	.443	.991	.919	.999	.386	.999	.133	.438	.135	.902	.225	.432
7	TEs	.054	.871	.330	.114	.271	.195	.073	.646	.322	.955	.825	.997	.360	.997	.089	.076	.027	.460	.069	.131
8	TZ1	.056	.885	.377	.133	.309	.294	.099	.745	.459	.947	.895	.996	.470	.994	.123	.099	.028	.442	.114	.200
9	TZ2	.062	.900	.402	.147	.344	.282	.096	.688	.416	.915	.865	.994	.445	.987	.110	.028	.013	.145	.079	.130
10	Z_K	.055	.861	.308	.109	.252	.251	.088	.797	.438	.983	.906	.992	.423	.999	.118	.132	.054	.512	.952	.253
11	Z_C	.050	.844	.333	.121	.249	.313	.104	.838	.529	.983	.931	.999	.520	.999	.159	.231	.052	.782	.953	.307
12	Z_A	.052	.864	.347	.124	.268	.323	.108	.866	**.559**	.989	.943	.999	.541	.999	.166	.142	.032	.674	.967	.318
13	$\sqrt{b_1}$	.052	.775	.345	.135	.286	.324	.114	.708	.471	.891	.869	.990	.508	.979	.151	.006	.008	.013	.762	.125
14	b_2	.049	.832	.333	.111	.239	.181	.076	.365	.230	.544	.600	.877	.279	.787	.093	.324	.109	.683	.316	.122
15	K^2	.048	.849	.370	.139	.282	.267	.100	.570	.371	.777	.781	.967	.418	.936	.119	.133	.030	.491	.587	.093
16	DH	.050	.871	.382	.141	.316	.258	.089	.730	.429	.941	.888	.997	.444	.994	.110	.101	.024	.494	.855	.186
17	Z_w	.049	.853	.326	.108	.280	.120	.062	.203	.135	.340	.427	.756	.173	.602	.059	.225	.089	.539	.160	.111
18	D	.051	.882	.347	.119	.276	.202	.075	.517	.280	.805	.758	.984	.330	.963	.086	.094	.075	.031	.607	.067
19	QH	.053	.862	.327	.115	.251	.313	.103	.841	.533	.983	.933	.999	.520	.999	.157	.229	.059	.761	.957	.326
20	r	.053	.895	.389	.145	.325	.311	.108	.794	.492	.970	.911	.998	.504	.999	.145	.073	.019	.460	.916	.207
21	R_n	.054	.875	.353	.128	.281	.320	.108	.833	.528	.981	.931	.999	.524	.999	.158	.176	.045	.683	.946	.292
22	T_{EP}	.054	.868	.332	.115	.257	.309	.104	.778	.502	.954	.912	.998	.507	.995	.147	.130	.043	.478	.888	.266
23	I_n	.053	.144	.268	.145	.286	.216	.091	.387	.289	.286	.310	.038	.313	.084	.095	.004	.006	.013	.382	.070
24	E_n	.053	.901	.398	.150	.337	.302	.105	.763	.467	.959	.899	.998	.488	.997	.135	.038	.013	.326	..891	.169
25	M_n	.050	.894	.409	.153	.339	.274	.099	.661	.395	.897	.841	.992	.431	.984	.112	.005	.004	.025	.771	.087
26	Q_n	.053	.874	.311	.105	.257	.277	.092	.847	.508	.988	.932	.999	.486	.999	.142	.176	.051	.663	.963	.333
27	T_{1n}	.050	.656	.255	.106	.179	**.345**	.111	.838	.569	.972	.938	.999	**.565**	.999	.178	.029	.018	.082	.924	.246
27	T_{2n}	.049	.866	.343	.116	.296	.100	.059	.150	.101	.269	.362	.734	.141	.554	.049	.311	.079	.773	.109	.119
27	T_{3n}	.050	.897	.387	.143	.330	.278	.096	.779	.453	.973	.905	.999	.466	.999	.121	.174	.032	.732	.926	.225
30	KS	.056	.847	.268	.089	.227	.214	.084	.595	.338	.884	.799	.992	.349	.985	.109	.102	.056	.377	.761	.192
31	V	.052	.863	.273	.090	.236	.199	.073	.697	.352	.955	.859	.998	.348	.997	.093	.148	.063	.495	.885	.205
32	W^2	.056	.880	.308	.105	.265	.254	.091	.732	.420	.954	.883	.998	.429	.996	.123	.149	.056	.517	.882	.237
33	U^2	.055	.878	.297	.099	.261	.225	.083	.694	.381	.942	.862	.997	.391	.995	.113	.167	.064	.554	.863	.230
34	A^2	.054	.880	.324	.110	.268	.279	.094	.780	.463	.968	.906	.999	.467	.998	.133	.179	.056	.624	.917	.269
35	P	.049	.777	.182	.067	.144	.141	.063	.656	.282	.956	.827	.998	.267	.994	.074	.082	.053	.272	.880	.162
36	SW	.054	.867	.337	.119	.266	.317	.105	.840	.534	.982	.933	.999	.526	.999	.160	.208	.053	.738	.954	.314
37	SF	.053	.893	.383	.143	.318	.313	.107	.802	.498	.973	.915	.998	.507	.999	.148	.086	.022	.499	.922	.220
38	SJ	.053	**.915**	.404	.147	**.377**	.188	.078	.416	.234	.676	.695	.959	.301	.911	.064	.003	.006	.002	.435	.033
39	JB	.050	.864	.384	.146	.300	.285	.104	.630	.404	.840	.825	.984	.448	.964	.125	.003	.004	.006	.677	.080
40	RJB	.054	.906	**.410**	**.159**	.354	.266	.099	.563	.357	.790	.787	.977	.410	.949	.106	.002	.004	.004	.594	.061
h_2	H_n	.055	.874	.302	.100	.254	.322	.116	.832	.540	.982	.933	.999	.525	.999	.176	.154	.061	.524	.734	.140
h_1	H_n	.053	.869	.293	.097	.244	.330	**.120**	.835	.546	.983	.934	.999	.532	.999	**.181**	.156	.062	.525	.709	.126

Table 8: *Power comparisons for the normality test for* Set 2 *of alternative distributions,* $\alpha = 0.05$, $n = 10$.

	Symmetric													Asymmetric							
	Short tailed							Close to Normal			Long tailed			Short tailed					Long tailed		
	Tu (.7)	Tu (1.5)	Tu (3)	SB (0,.5)	Tri	TN (-1,1)	TN (-3,3)	Tu (.1)	SU (0,3)	t (10)	Tu (10)	SU (0,1)	caushy	TN (-3,1)	SB (1,1)	SB (1,2)	SB (.533,.5)	HN	SU (1,1)	χ^2 (1)	χ^2 (4)
TV	**.122**	**.205**	**.095**	**.313**	**.057**	**.125**	**.053**	.044	.045	.057	.281	.086	.369	.027	.130	.055	**.442**	.206	.260	.766	.171
TA	.086	.155	.068	.244	.043	.089	.047	.050	.051	.048	.460	.162	.496	**.095**	.143	.051	.431	.177	.357	**.814**	.225
Z_A	.040	.060	.033	.101	.031	.037	.043	.062	.067	.080	.523	.258	.608	.071	.133	.050	.296	.196	.438	.755	.254
$\sqrt{b_1}$	.018	.019	.017	.024	.025	.017	.040	.061	.069	.084	.372	.255	.574	.061	.101	.042	.130	.156	.412	.566	.221
r	.035	.050	.021	.086	.031	.034	.044	.064	.068	.087	.595	.276	.634	.070	.123	.050	.253	.177	.436	.704	.239
R_n	.057	.094	.042	.149	.035	.054	.046	.061	.066	.075	.552	.251	.609	.073	.142	.054	.323	.194	.436	.743	.251
T_{EP}	.046	.067	.037	.101	.035	.047	.050	.063	.068	.085	.480	.180	.602	.072	.143	.057	.271	.193	**.448**	.684	.256
E_n	.029	.038	.026	.067	.030	.028	.044	.063	.068	.090	.602	.278	.639	.067	.114	.048	.224	.168	.433	.682	.232
M_n	.015	.017	.017	.020	.043	.016	.045	.064	**.075**	.100	.521	.287	.634	.061	.087	.044	.115	.137	.397	.550	.200
T_{1n}	.032	.041	.026	.060	.031	.030	.044	.061	.063	.069	.338	.220	.517	.077	.140	.051	.253	.204	.441	.739	.266
T_{3n}	.029	.047	.026	.088	.030	.030	.045	.064	.072	.038	.579	.288	.645	.083	.093	.046	.200	.139	.405	.644	.198
A^2	.063	.102	.049	.154	.037	.063	.050	.060	.064	.068	.630	.245	.620	.075	.137	.054	.319	.182	.422	.715	.234
SW	.064	.109	.049	.170	.036	.064	.046	.060	.064	.074	.532	.242	.598	.078	.144	.054	.345	.199	.433	.751	.253
SF	.037	.055	.031	.093	.030	.036	.044	.063	.067	.082	.588	.270	.630	.070	.124	.050	.261	.179	.435	.709	.238
SJ	.014	.016	.019	.015	.032	.016	.047	**.067**	.072	**.093**	**.678**	.290	**.660**	.049	.070	.046	.086	.101	.360	.442	.157
RJB	.015	.016	.016	.018	.026	.016	.045	.064	.073	**.093**	.569	**.290**	.645	.057	.088	.044	.105	.132	.394	.504	.195
H_n	.066	.094	.052	.141	.047	.061	**.053**	.060	.064	.060	.625	.222	.592	.026	**.208**	**.073**	.416	**.262**	.264	.807	**.315**

Table 9: *Power comparisons for the normality test for* Set 2 *of alternative distributions,* $\alpha = 0.05$, $n = 20$.

	Symmetric													Asymmetric							
	Short tailed							Close to Normal			Long tailed			Short tailed					Long tailed		
	Tu (.7)	Tu (1.5)	Tu (3)	SB (0,.5)	Tri	TN (-1,1)	TN (-3,3)	Tu (.1)	SU (0,3)	t (10)	Tu (10)	SU (0,1)	caushy	TN (-3,1)	SB (1,1)	SB (1,2)	SB (.533,.5)	HN	SU (1,1)	χ^2 (1)	χ^2 (4)
TV	**.291**	**.515**	**.188**	**.729**	**.075**	**.268**	**.051**	.042	.047	.048	.724	.159	.683	**.180**	.314	.070	**.877**	.458	.547	**.993**	.433
TA	.131	.310	.083	.531	.036	.122	.040	.051	.065	.091	.909	.376	.853	.171	.307	.057	.807	**.477**	.678	.992	.515
Z_A	.064	.168	.040	.343	.020	.057	.037	.060	.077	.103	.721	.421	.859	.154	.305	.058	.709	.462	.714	.989	.541
$\sqrt{b_1}$	.005	.007	.008	.009	.011	.006	.035	.065	.084	.113	.354	.401	.771	.111	.190	.050	.174	.307	.708	.882	.446
r	.034	.084	.021	.193	.017	.030	.037	.066	.085	.109	.851	.480	.890	.105	.230	.052	.534	.360	.720	.966	.472
R_n	.085	.198	.050	.385	.028	.078	.038	.059	.077	.102	.817	.440	.872	.135	.282	.058	.681	.414	.721	.980	.509
T_{EP}	.073	.149	.045	.267	.034	.065	.042	.058	.071	.090	.807	.417	.866	.129	.284	.062	.580	.368	.722	.952	.488
E_n	.019	.047	.014	.114	.015	.019	.036	.067	.087	.112	.859	.494	.897	.094	.205	.049	.444	.329	.712	.957	.450
M_n	.003	.005	.006	.010	.011	.004	.034	.067	.090	.116	.774	.501	.894	.076	.140	.041	.191	.253	.675	.895	.381
T_{1n}	.018	.029	.017	.046	.020	.019	.040	.057	.070	.089	.261	.292	.645	.152	.301	.058	.448	.436	**.723**	.971	.547
T_{3n}	.074	.212	.043	.409	.037	.070	.039	.061	.083	.110	.775	.482	.896	.098	.205	.045	.646	.333	.695	.971	.433
A^2	.105	.206	.060	.374	.040	.092	.048	.057	.070	.084	.906	.423	.878	.117	.266	.064	.651	.359	.704	.970	.459
SW	.108	.250	.067	.452	.034	.100	.040	.058	.077	.097	.805	.424	.866	.143	.305	.063	.723	.435	.719	.982	.522
SF	.041	.102	.025	.222	.018	.036	.038	.065	.084	.106	.848	.477	.888	.109	.242	.054	.561	.372	.722	.970	.482
SJ	.002	.001	.005	.004	.018	.003	.037	.066	.086	.115	**.930**	**.509**	**.917**	.039	.065	.044	.054	.109	.594	.669	.227
RJB	.002	.002	.004	.003	.011	.003	.036	**.068**	**.092**	**.121**	.819	.507	.902	.065	.119	.041	.091	.206	.666	.784	.348
H_n	.095	.176	.058	.308	.041	.082	.049	.056	.065	.078	.914	.384	.867	.056	**.345**	**.083**	.719	.441	.574	.981	**. 527**

Table 10: *Ranking from first to the fifth of average powers computed from values in Tables 6-7 for* Set 1 *of alternative distributions.*

	Group I		Group II		Group III		Group IV	
	Symmetric $(-\infty,\infty)$		Asymmetric $(-\infty,\infty)$		Asymmetric$(0,\infty)$		(0,1)	
Rank	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$
1	SJ	SJ	H_n	T_{1n}	H_n	Z_A	TV	TV
2	RJB	RJB	T_{1n}	H_n	TV	T_{1n}	TE	TE
3	T_{3n}	M_n	T_{EP}	Z_A	A	H_n	TV	TA
4	M_n	TZ2	$\sqrt{b_1}$	R_n	T_{1n}	SW	Z_C	QH
5	E_n	E_n	R_n	SW	Z_A	QH	QH	Z_C

Table 11: *Ranking from first to the fifth of average powers computed from values in Tables 8-9 for* Set 2 *of alternative distributions.*

	Symmetric						Asymmetric			
Rank	Short tailed		Close to Normal		Long tailed		Short tailed		Long tailed	
	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$
1	TV	TV	M_n	RJB	SJ	SJ	H_n	H_n	T_{1n}	T_{1n}
2	TA	TA	SJ	M_n	RJB	RJB	TA	TV	SW	SW
3	SW	R_n	RJB	SJ	A^2	SF	TV	TA	R_n	R_n
4	H_n	SW	SF	SF	SF	A^2	SW	SW	H_n	TA
5	A^2	A^2	SW	T_{3n}	T_{3n}	M_n	R_n	R_n	TA	H_n

Tables 10-11 contain the ranking from first to the fifth of the average powers computed from the values in Tables 6-7 and 8-9, respectively. By average powers we can select the tests that are, on average, most powerful against the alternatives from the given groups.

Power against an alternative distribution has been estimated by the relative frequency of values of the corresponding statistic in the critical region for 10000 simulated samples of size n = 10, 20. The maximum reached power is indicated in bold. For computing the estimated powers of the new test, R software is used. We also use R software for computing Pearson chi-square and Shapiro-Francia tests by the package (nortest), command `pearson.test` and `sf.test`, respectively, and also the package (lawstat), command `sj.test` and `rjb.test` for SJ and Robast Jarque-Bera tests, respectively. For the entropy-based test statistics, powers are taken from Zamanzadeh and Arghami (2012) and Alizadeh and Arghami (2011, 2013). In the case of the test based on H_n, we also consider $h_2(x) := x\log(x) - x + 1$ for comparison with $h_1(x) := \left(\frac{x-1}{x+1}\right)^2$.

Results and recommendations

Based on these comparisons, the following recommendations can be formulated for the application of the evaluated statistics for testing normality in practice.

***Set 1* of alternative distributions** (Tables 6-7 and 10): In Group I, for $n = 10$ and 20, it is seen that the tests based on SJ, RJB, T_{3n}, TZ2, M_n and E_n are the most powerful whereas the tests based on I_n, TV, TC and KL are the least powerful. The difference of powers between KL and the others is substantial. In Group II, for $n = 10$ and 20, it is seen that the tests based on H_n, T_{1n}, T_{EP}, R_n, Z_A and $\sqrt{b_1}$ are the most powerful whereas those based on T_{2n}, TV, TC, Kl and Z_w are the least powerful. In Group III, the most powerful tests for $n = 10$ are those based on H_n, TV, TA and T_{1n}, and for $n = 20$, those based on Z_A, T_{1n}, H_n and SW are the most powerful. On the other hand, the least powerful tests are those based on I_n and Z_w are the least powerful. Finally, in group IV, the results are not in favour of the proposed tests. In this group, for $n = 10$ and 20, the most powerful tests are those based on TV, TE, TA, Z_C, Z_A and r, whereas the tests based on TZ_2, SJ and RJB are the least powerful. The SJ and RJB show very poor sensitivity against symmetric distributions in $[0,1]$ such as Unif, $B(2,2)$ or $B(.5,.5)$. For example, for $n = 20$, in the case of the $[0,1]$-Unif alternative, the SJ test has a power of .002 while even the H_n test has a power of .156. From Tables 6-7 one can see that the proportion of times that the SJ and RJB statistics lie below the 5% point of the null distribution are greater than those of the H_n statistic.

Note that for the proposed test, the maximum power in Group II and III was typically attained by choosing h_1.

From the simulation study implemented for *Set 1* of alternative distributions we can lead to different conclusions from that existing in the literature. New and existing results are reported in Table 12.

Table 12: *Comparison of most powerful tests in Groups I–IV, according to Alizadeh and Arghami (2011, 2013) and Zamanzade and Arghami (2012) with new simulation results.*

Alizadeh and Arghami (2011)	JB	SW	KL[a] or SW	KL
Alizadeh and Arghami (2013)	A^2	SW	TA	TV[b]
Zamanzadeh and Arghami (2012)	TZ2	TZ2 or TD	TZ1, KL or TD	KL or TC
New simulation study	SJ or RJB	H_n or T_{1n}	H_n or Z_A	TV or TE

[a] Statistic based on Vasicek's estimator
[b] Statistic using nonparametric distribution of Vasicek's estimato

***Set 2* of alternative distributions** (Tables 8-9 and 11): For symmetric short-tailed distributions, it is seen that the tests based on TV, TA and SW are the most powerful. For symmetric close to normal and symmetric long tailed distributions, RJB, JB and M_n are the most powerful. For asymmetric short tailed distributions, H_n, TV and TA are the

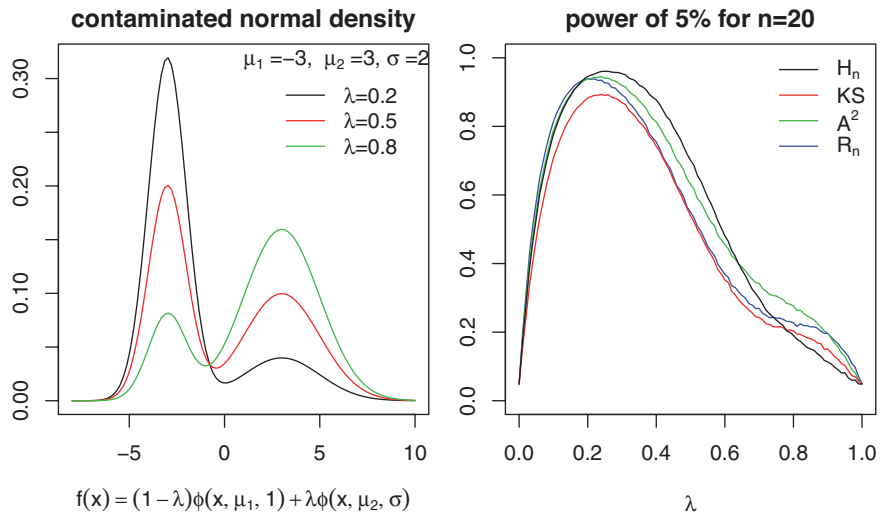

Figure 3: *Left panel: Probability density functions of Contaminated Normal distribution for several values of the parameter* λ*. Right panel: Power of the tests based on* H_n*,* KS*,* A^2 *and* R_n *as a function of* λ *against alternative* $CN(\lambda, \mu_1 = -3, \mu_2 = 3, \sigma = 2)$.

most powerful. Finally, for asymmetric long tailed distributions, T_{1n}, SW and R_n are the most powerful. It is also worth mentioning that the differences between the power of tests based on TV and H_n in $TN(-3,3)$ alternative are not considerable.

In Figure 3 we compare the power of the tests based on H_n, KS, A^2 and R_n against a family of Contaminated Normal alternatives $CN(\lambda, \mu_1 = -3, \mu_2 = 3, \sigma = 2)$. The left panel of Figure 3 contains the probability density functions of Contaminated Normal alternatives $CN(\lambda, \mu_1 = -3, \mu_2 = 3, \sigma = 1)$, for $\lambda = .2, .5, .8$, whereas the right panel contains the power comparisons for $n = 20$ and $\alpha = 0.05$. We can see the good power results of H_n for $0.2 < \lambda < 0.6$.

In general, we can conclude that the proposed test H_n has good performance and therefore can be used in practice.

Numerical example

Finally, we illustrate the performance of the new proposal through the analysis of a real data set. One of the most famous tests of normality among practitioners is the Kolmogorov-Smirnov test, mostly because it is available in any statistical software. However, one of its drawbacks is the low power against several alternatives (see also Grané and Fortiana, 2003; Grané, 2012; Grané and Tchirina, 2013).We would like to emphasize this fact through a numerical example.

Armitage and Berry (1987) provided the weights in ounces of 32 newborn babies(see also data set 3 of Henry, 2002, p. 342). The approximate ML estimators of $\hat{\mu} = 111.75$ and $\hat{\sigma} = \sqrt{331.03} = 18.19$. Also sample skewness and kurtosis are $\sqrt{b_1} = -.64$ and

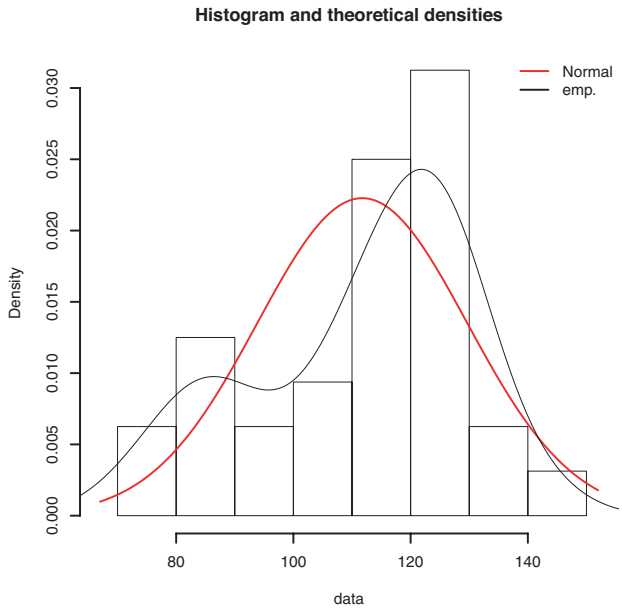

Figure 4: *Histogram and theoretical (normal) distribution for ounces of 32 newborn babies data.*

$b_2 = 2.33$, respectively. From the histogram of these data it can be observed that the birth weights are skewed to the left and may be bimodal (see Figure 4).

When fitting the normal distribution to these data, we find that the KS (Kolmogorov-Smirnov) test does not reject the null hypothesis providing a p-value of 0.093. However with the H_n statistic we are able to reject the null hypothesis of normality at a 5% significance level, since we obtain $H_n = .0006$ and the corresponding critical value for $n = 32$ is .00047. Also associated p-values of the H_n, SW (Shapiro-Wilk) and SF (Shapiro-Francia) tests are .015, .024 and .036, respectively. Thus, the non-normality is more pronounced by the new test at 5% level. In Appendix, we provide an R software program, to calculate the H_n statistics, the critical points and corresponding p-value.

6. Conclusions

In this paper we propose a statistic to test normality and compare its performance with 40 recent and classical tests for normality and a wide collection of alternative distributions. As expected (Janssen, 2000), the simulation study shows that none of the statistics under evaluation can be considered to be the best one for all the alternative distributions studied. However, the tests based on RJB or SJ have the best performance for symmetric distributions with the support on $(-\infty, \infty)$ and the same happens to TV or TA for distributions with the support on $(0, 1)$. Regarding our proposal, H_n and also T_{1n} are the most powerful for asymmetric distributions with the support on $(-\infty, \infty)$ and distributions with the support on $(0, \infty)$, mainly for small sample sizes.

Acknowledgements

This work has been partially supported by research grant project MTM2014-56535-R (Spanish Ministry of Economy and Competitiveness). The authors are thankful to two Referees and the Editor, whose helpful comments and suggestions contributed to improve the quality of the paper.

Appendix

```
h=function(x) (x-1)^2/(x+1)^2
Hn=function(x) {x=sort(x);n=length(x);
F=pnorm(x, mean(x), sd(x)*sqrt(n/(n-1)))+1;
Fn=1:n/n+1; mean(h(F/Fn))}

##weights in ounces of 32 newborn babies,
data=c(72,80,81,84,86,87,92,94,103,106,107,111,112,115,116,118,
119,122,123,123,114,125,126,126,126,127,118,128,128,132,133,142)
Hn(data)  ## statistics
n=length(data); B=10000; x=matrix(rnorm(n*B, 0, 1), nrow=B, ncol=n)
H0=apply(x, 1, Hn); Q=quantile(H0, .95); Q  ## critical point
length(H0[H0>Hn(data)])/B  ##p-value
```

References

Alizadeh Noughabi, H. and Arghami, N.R. (2012). General treatment of goodness-of-fit tests based on Kullback-Leibler information. *Journal of Statistical Computation and Simulation*, 83, 1–14.

Alizadeh Noughabi, H. and Arghami, N.R. (2010). A new estimator of entropy. *Journal of the Iranian Statistical Society*, 9, 53–64.

Alizadeh Noughabi, H. and Arghami, N.R. (2013). Goodness-of-fit tests based on correcting moments of entropy estimators. *Communications in Statistics-Simulation and Computation*, 42, 499–513.

Alizadeh Noughabi, H. and Arghami, N.R. (2011). Monte carlo comparison of seven normality tests. *Journal of Statistical Computation and Simulation*, 81, 965–972.

Anderson, T.W. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, 765–769.

Anscombe, F.J. and Glynn, W.J. (1983). Distribution of kurtosis statistic b_2 for normal statistics. *Biometrika*, 70, 227–234.

Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research*. Blackwell Scientific Publications, Oxford.

Bonett, D.G. and Seier, E. (2002). A test of normality with high uniform power. *Computational Statistics and Data Analysis*, 40, 435–445.

Chen, L. and Shapiro, S.S. (1995). An alternative test for normality based on normalized spacings. *Journal of Statistical Computation and Simulation*, 53, 269–287.

Corea, J.C. (1995). A new estimator of entropy. *Communications in Statistics-Theory and Methods*, 24, 2439–2449.

Cramer, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 11, 141–180.

Csörgo, M. and Révész, P. (1971). *Strong Approximations in Probability and Statistics*. Academic Press, New York.

D'Agostino, R.B. and Stephens, M. (1986). *Goodness-of-fit Techniques*. New York: Marcel Dekker, Inc.

D'Agostino, R.B. and Pearson, E.S. (1973). Tests for departure from normality. Empirical results for the distributions of b_2 and $sqrtb_1$. *Biometrika*, 60, 613–622.

D'Agostino, R.B., Belanger, A. and D'Agostino R.B.Jr. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44, 316–321.

D'Agostino, R.B. (1970). Transformation to Normality of the Null Distribution of g_1. *Biometrika*, 57, 679–681.

D'Agostino, R.B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, 58, 341–348.

del Barrio, E., Cuesta-Albertos, J.A., Matrán, C. and Rodríguez-Rodríguez, J.M. (1999). Tests of goodness of fit based on the L2-wasserstein distance. *The Annals of Statistics*, 27, 1230–1239.

de Wet, T. and Venter, J.H. (1972). Asymptotic distributions of certain tests criteria of normality. *South African Statistical Journal*, 6, 135–149.

Dimitriev, Y.G. and Tarasenko, F.P. (1973). On the estimation functions of the probability density and its derivatives. *Theory of Probability & Its Applications*, 18, 628–633.

Doornik, J.A. and Hansen, H. (1994). An omnibus test for univariate and multivariate normality. *Economics Working Papers*, Nuffield College.

Ebrahimi, N., Pflughoeft, K. and Soofi, S.E. (1994). Two measures of sample entropy. *Statistics and Probability Letters*, 20, 225–234.

Epps, T.W. and Pulley, L.B. (1983). A test for normality based on the empirical characteristic function. *Biometrika*, 70, 723–726.

Esteban, M.D., Castellanos, M.E., Morales, D. and Vajda, I. (2001). Monte carlo comparison of four normality tests using different entropy estimates. *Communications in Statistics-Simulation and Computation*, 30, 761–285.

Filliben, J.J. (1975). The probability plot correlation coefficient test for normality. *Technometrics*, 17, 111–117.

Gan, F.F. and Koehler, K.J. (1990). Goodness of fit tests based on P-P probability plots. *Technometrics*, 32, 289–303.

Gel, Y.R., Miao, W. and Gastwirth, J.L. (2007). Robust directed tests of normality against heavy-tailed alternatives. *Computational Statistics and Data Analysis*, 51, 2734–2746.

Gel, Y.R. and Gastwirth, J.L. (2008). A robust modification of the Jarque-Bera test of normality. *Economics Letters*, 99, 30–32.

Grané, A. and Fortiana, J. (2003). Goodness of fit tests based on maximum correlations and their orthogonal decompositions. *Journal of the Royal Statistical Society. Series B: Methodological*, 65, 115–126.

Grané, A. (2012). Exact goodness of fit tests for censored data. *Annals of the Institute of Statistical Mathematics*, 64, 1187–1203.

Grané, A. and Tchirina, A. (2013). Asymptotic properties of a goodness-of-fit test based on maximum correlations. *Statistics*, 47, 202–215.

Henry, C. T. (2002). *Testing for Normality*. Marcel Dekker, New York.

Janssen, A. (2000). Global power functions of godness-of-fit tests. *Annals of Statististics*, 28, 239–253.

Jarque, C.M. and Bera, A.K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6, 255–259.

Jarque, C.M. and Bera, A.K. (1987). A Test for Normality of Observations and Regression Residuals. *International Statistical Review/Revue Internationale de Statistique*, 55, 163–172.

Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di dislibuzione. *Giornale dell'Istitu Italiano degli Attuari*, 4, 83–91.

Krauczi, E. (2009). A study of the quantile correlation test for normality. *TEST*,18, 156–165.

Kuiper, N.H. (1962). Test concerning random points on a circle. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen Series A*, 63, 38–47.

LaRiccia, V. (1986). Optimal goodness-of-fit tests for normality against skewness and kurtosis alternatives. *Journal of Statistical Planning and Inference*, 13, 67–79.

Lilliefors, H.W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62, 399–402.

Louter, A.S. and Koerts, J. (1970). On the Kuiper test for normality with mean and variance unknown. *Statistica Neerlandica*, 24, 83–87.

Martinez, J.and Iglewicz, B. (1981). A test for departure from normality based on a biweight estimator of scale. *Biometrika*, 68, 331–333.

Mises, R. von (1931). *Wahrscheinlichkeitsrechnung und ihre Anwendung in der Statistik und theoretischen Physik*. Deuticke, Leipzig and Vienna.

Pardo, L. (2006). *Statistical Inference based on Divergence Measures*. New York: Taylor Francis Group.

Park, S. and Park, D. (2003). Correcting moments for goodness-of-fit tests based on two entropy estimates. *Journal of Statistical Computation and Simulation*, 73, 685–694.

Pettitt, A.N. (1977). A Cramer-von Mises type goodness of fit statistic related to $\sqrt{b_1}$ and b_2. *Journal of the Royal Statistical Society. Series B: Methodological*, 39, 364–370.

Poitras, G. (2006). More on the correct use of omnibus tests for normality. *Economics Letters*, 90, 304–309.

Romao, X., Delgado, R. and Costa, A. (2010). An empirical power comparison of univariate goodness-of-fit tests for normality. *Journal of Statistical Computation and Simulation*, 80, 545–591.

Shao, J. (2003). *Mathematical Statistics*. New York: Springer, Verlag.

Shapiro, S.S., Wilk, M.B. and Chen, M.H.J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, 63, 1343–1372.

Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.

Shapiro, S.S. and Francia, R.S. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association*, 67, 215–216.

Shenton, L. and Bowman, K. (1977). A bivariate model for the distribution of b_1 and b_2. *Journal of the American Statistical Association*, 72, 206–211.

Van Es, B. (1992). Estimating functionals related to a density by class of statistics based on spacing. *Scandinavian Journal of Statistics*, 19, 61–72.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B: Methodological*, 38, 54–59.

Watson, G.S. (1961). Goodness-of-fit tests on a circle. *Technometrics*, 48, 109–114.

Yap, B.W. and Sim, C.H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81, 2141–2155.

Yazici, B. and Yolacan, S. (2007). A comparision of various tets of normality. *Journal of Statistical Computation and Simulation*, 77, 175–183.

Zamanzade, E. and Arghami, N.R. (2012). Testing normality based on new entropy estimators. *Journal of Statistical Computation and Simulation*, 82, 1701–1713.

Zhang, J. and Wu, Y. (2005). Likelihood-ratio tests for normality. *Computational Statistics and Data Analysis*, 49, 709–721.

Zhao, J. and Xu, X. (2014). Goodness-of-fit tests for location scale families based on random distance. *Journal of Statistical Computation and Simulation*, 84, 739–752.

SORT 40 (1) January-June 2016, 89-112

Point and interval estimation for the logistic distribution based on record data

A. Asgharzadeh[1,*], R. Valiollahi[2] and M. Abdi[3]

Abstract

In this paper, based on record data from the two-parameter logistic distribution, the maximum likelihood and Bayes estimators for the two unknown parameters are derived. The maximum likelihood estimators and Bayes estimators can not be obtained in explicit forms. We present a simple method of deriving explicit maximum likelihood estimators by approximating the likelihood function. Also, an approximation based on the Gibbs sampling procedure is used to obtain the Bayes estimators. Asymptotic confidence intervals, bootstrap confidence intervals and credible intervals are also proposed. Monte Carlo simulations are performed to compare the performances of the different proposed methods. Finally, one real data set has been analysed for illustrative purposes.

MSC: 62G30, 62F10, 62F15, 62E15.

Keywords: Logistic distribution, record data, maximum likelihood estimator, Bayes estimator, Gibbs sampling.

1. Background and statistical context

Let $\{Y_i, i \geq 1\}$ be a sequence of independent and identically distributed (iid) random variables with cumulative distribution function (cdf) $G(y;\theta)$ and probability density function (pdf) $g(y;\theta)$, where θ is a vector of parameters. An observation Y_j is called an upper record value if $Y_j > Y_i$ for all $i = 1,2,\ldots,j-1$. An analogous definition can be given for lower record values. Generally, if $\{U(n),\ n \geq 1\}$ is defined by

$$U(1) = 1, \qquad U(n) = \min\{j : j > U(n-1),\ Y_j > Y_{U(n-1)}\},$$

* Corresponding author: a.asgharzadeh@umz.ac.ir

[1] Department of Statistics, Faculty of Mathematical Sciences, University of Mazandaran, Babolsar, Iran.

[2] Department of Mathematics, Statistics and Computer Science, Semnan University, Semnan, Iran.

[3] Department of Mathematics and Soft Computing, Higher Education Complex of Bam, Bam, Iran.

Received: February 2014
Accepted: April 2016

for $n \geq 2$, then the sequence $\{Y_{U(n)},\ n \geq 1\}$ provides a sequence of upper record statistics. The sequence $\{U(n),\ n \geq 1\}$ represents the record times.

Suppose we observe the first m upper record values $Y_{U(1)} = y_1, Y_{U(2)} = y_2, \cdots, Y_{U(m)} = y_m$ from the cdf $G(y;\theta)$ and pdf $g(y;\theta)$. Then, the joint pdf of the first m upper record values is given (see Ahsanullah, 1995) by

$$h(\mathbf{y};\theta) = g(y_m;\theta) \prod_{i=1}^{m-1} \frac{g(y_i;\theta)}{1-G(y_i;\theta)}, \quad -\infty < y_1 < y_2 < \ldots < y_m < \infty, \qquad (1.1)$$

where $\mathbf{y} = (y_1, \ldots, y_m)$. The marginal pdf of the n*th* record $Y_{U(n)}$ is

$$h_n(y;\theta) = \frac{[-\ln(1-G(y;\theta))]^{n-1}}{(n-1)!} g(y;\theta).$$

The definition of record statistics was formulated by Chandler (1952). These statistics are of interest and important in many real life problems involving weather, economics, sports data and life testing studies. In reliability and life testing experiments, many products fail under stress. For example, an electronic component ceases to function in an environment of too high temperature, a wooden beam breaks when sufficient perpendicular force is applied to it, and a battery dies under the stress of time. Hence, in such experiments, measurements may be made sequentially and only the record values (lower or upper) are observed. For more details and applications of record values, one may refer to Arnold et al. (1998) and Nevzorov (2001).

The logistic distribution has been used for growth models in the biological sciences, and is used in a certain type of regression known as the logistic regression. It has many applications in technological problems including reliability, studies on income, graduation of mortality statistics, modeling agriculture production data, and analysis of categorical data. The shape of the logistic distribution is very similar to that of the normal distribution, but it is more peaked in the center and has heavier tails than the normal distribution. Because of the similarity of the two distributions, the logistic model has often been selected as a substitute for the normal model. For more details and other applications, see Balakrishnan (1992) and Johnson et al. (1995).

Although extensive work has been done on inferential procedures for logistic distribution based on complete and censored data, but not much attention has been paid on inference based on record data. In this article, we consider the point and interval estimation of the unknown parameters of the logistic distribution based on record data. We first consider the maximum likelihood estimators (MLEs) of the unknown parameters. It is observed the MLEs can not be obtained in explicit forms. We present a simple method of deriving explicit MLEs by approximating the likelihood function. We further consider the Bayes estimators of the unknown parameters and it is observed the Bayes estimators and the corresponding credible intervals can not be obtained in explicit forms. We use an

approximation based on the Gibbs sampling procedure to compute the Bayes estimators and the corresponding credible intervals.

The rest of the paper is organized as follows. In Section 2, we discuss the MLEs of the unknown parameters of the logistic distribution. In Section 3, we provide the approximate maximum likelihood estimators (AMLEs). Bayes estimators and the corresponding credible intervals are provided in Section 4. The Fisher information and different confidence confidence intervals are presented in Section 5. Finally, in Section 4, one numerical example and a Monte Carlo simulation study are given to illustrate the results.

2. Maximum likelihood estimation

Let the failure time distribution be a logistic distribution with probability density function (pdf)

$$g(y;\mu,\sigma) = \frac{e^{-(y-\mu)/\sigma}}{\sigma(1+e^{-(y-\mu)/\sigma})^2}\ , \quad -\infty < y < \infty, \quad \mu \in R, \quad \sigma > 0, \tag{2.1}$$

and cumulative distribution function (cdf)

$$G(y;\mu,\sigma) = \frac{1}{1+e^{-(y-\mu)/\sigma}}\ , \quad -\infty < y < \infty, \quad \mu \in R, \quad \sigma > 0. \tag{2.2}$$

Consider the random variable $X = (Y-\mu)/\sigma$. Then, X has the standard logistic distribution with pdf and cdf as

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2}\ , \quad -\infty < x < \infty, \tag{2.3}$$

and

$$F(x) = \frac{1}{1+e^{-x}}\ , \quad -\infty < x < \infty, \tag{2.4}$$

respectively. Note that $g(y;\mu,\sigma) = \frac{1}{\sigma} f((y-\mu)/\sigma)$ and $G(y;\mu,\sigma) = F((y-\mu)/\sigma)$. It should also be noted that $f(x)$ and $F(x)$ satisfy the following relationships:

$$f(x) = F(x)[1-F(x)], \qquad f'(x) = f(x)[1-2F(x)]. \tag{2.5}$$

Suppose we observe the first m upper record values $Y_{U(1)} = y_1, Y_{U(2)} = y_2, \cdots, Y_{U(m)} = y_m$ from the logistic distribution with pdf (2.1) and cdf (2.2). The likelihood function is

given by

$$L(\mu,\sigma) = g(y_m,\mu,\sigma)\prod_{i=1}^{m-1}\frac{g(y_i;\mu,\sigma)}{1-G(y_i;\mu,\sigma)}. \tag{2.6}$$

By using Eqs. (2.3), (2.4) and (2.5), the likelihood function may be rewritten as

$$L(\mu,\sigma) = \sigma^{-m} f(x_m)\prod_{i=1}^{m-1} F(x_i)\ , \tag{2.7}$$

where $x_i = (y_i-\mu)/\sigma$. Subsequently, the log-likelihood function is

$$\ln L(\mu,\sigma) = -m\ln\sigma + \ln f(x_m) + \sum_{i=1}^{m-1}\ln F(x_i). \tag{2.8}$$

Again, by using Eq. (2.5), we derive the likelihood equations for μ and σ from (2.8), as

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\mu} = -\frac{1}{\sigma}\left[m - F(x_m) - \sum_{i=1}^{m} F(x_i)\right] = 0, \tag{2.9}$$

$$\frac{\partial \ln L(\mu,\sigma)}{\partial\sigma} = -\frac{1}{\sigma}\left[m + \sum_{i=1}^{m} x_i - x_m F(x_m) - \sum_{i=1}^{m} x_i F(x_i)\right] = 0. \tag{2.10}$$

The MLES $\widehat{\mu}$ and $\widehat{\sigma}$, respectively of μ and σ, are solution of the system of Eqs. (2.9) and (2.10). They can not be obtained in closed forms and so some iterative methods such as Newton's method are required to compute these estimators.

3. Approximate maximum likelihood estimation

It is observed that the likelihood equations (2.9) and (2.10) do not yield explicit estimators for the MLEs, because of the presence of the term $F(x_i)$, $i = 1,\ldots,m$, and they have to be solved by some iterative methods. However, as mentioned by Tiku and Akkaya (2004), solving the likelihood equations by iterative methods can be problematic for reasons of (i) multiple roots, (ii) nonconvergence of iterations, or (iii) convergence to wrong values. Moreover, these methods are usually very sensitive to their initiate values. Here, we present a simple method to derive approximate MLEs for μ and σ by linearizing the term $F(x_i)$ using Taylor series expansion. Approximate solutions for MLEs have been discussed in the book by Tiku and Akkaya (2004) for several specific distributions.

Balakrishnan and Aggarwala (2000), Balakrishnan and Kannan (2000), Balakrishnan and Asgharzadeh (2005), Agharzadeh (2006), Raqab et al. (2010) and Asgharzadeh et al. (2011) used approximate solutions for the MLEs, when the data are progressively censored.

We approximate the term $F(x_i)$ by expanding it in a Taylor series around $E(X_i) = \delta_i$. From Arnold et al. (1998), it is known that

$$F(X_i) \stackrel{d}{=} U_i,$$

where U_i is the i-th record statistic from the uniform $U(0,1)$ distribution. We then have

$$X_i \stackrel{d}{=} F^{-1}(U_i),$$

and hence

$$\delta_i = E(X_i) \approx F^{-1}(E(U_i)).$$

From Arnold et al. (1998), it is known that

$$E(U_i) = 1 - \left(\frac{1}{2}\right)^{i+1}, \quad i = 1, \ldots, m.$$

Since, for the standard logistic distribution, we have

$$F^{-1}(u) = \ln\left(\frac{u}{1-u}\right),$$

we can approximate δ_i by $F^{-1}[1 - (\frac{1}{2})^{i+1}] = \ln(2^{i+1} - 1)$.

Now, by expanding the function $F(x_i)$ around the point δ_i and keeping only the first two terms, we have the following approximation

$$\begin{aligned} F(x_i) &\simeq F(\delta_i) + (x_i - \delta_i) f(\delta_i) \\ &= \alpha_i + \beta_i x_i, \end{aligned} \tag{3.1}$$

where

$$\alpha_i = F(\delta_i) - \delta_i f(\delta_i),$$

and

$$\beta_i = f(\delta_i) \geq 0,$$

for $i = 1, \cdots, m$.

Using the expression in (3.1), we approximate the likelihood equations in (2.9) and (2.10) by

$$\frac{\partial \ln L^*(\mu,\sigma)}{\partial \mu} = -\frac{1}{\sigma}\left[m-(\alpha_m+\beta_m x_m)-\sum_{i=1}^{m}(\alpha_i+\beta_i x_i)\right]=0, \tag{3.2}$$

$$\frac{\partial \ln L^*(\mu,\sigma)}{\partial \sigma} = -\frac{1}{\sigma}\left[m+\sum_{i=1}^{m}x_i - x_m(\alpha_m+\beta_m x_m)-\sum_{i=1}^{m}x_i(\alpha_i+\beta_i x_i)\right]=0, \tag{3.3}$$

which can be rewritten as

$$\left[m-\alpha_m-\sum_{i=1}^{m}\alpha_i\right]-\frac{1}{\sigma}\left[\beta_m y_m+\sum_{i=1}^{m}\beta_i y_i\right]+\frac{1}{\sigma}\left[\beta_m+\sum_{i=1}^{m}\beta_i\right]\mu=0, \tag{3.4}$$

$$m+\frac{1}{\sigma}\left[(\sum_{i=1}^{m}y_i-\alpha_m y_m-\sum_{i=1}^{m}\alpha_i y_i)+\frac{(\beta_m y_m+\sum_{i=1}^{m}\beta_i y_i)\,(\alpha_m+\sum_{i=1}^{m}\alpha_i-m)}{\beta_m+\sum_{i=1}^{m}\beta_i}\right]$$

$$+\frac{1}{\sigma^2}\left[-(\beta_m y_m^2+\sum_{i=1}^{m}\beta_i y_i^2)+\frac{(\beta_m y_m+\sum_{i=1}^{m}\beta_i y_i)^2}{\beta_m+\sum_{i=1}^{m}\beta_i}\right]=0, \tag{3.5}$$

respectively. By solving the quadratic equation in (3.5) for σ, we obtain the approximate MLE of σ as

$$\widetilde{\sigma}=\frac{-A+\sqrt{A^2-4mB}}{2m}, \tag{3.6}$$

where

$$A=(\sum_{i=1}^{m}y_i-\alpha_m y_m-\sum_{i=1}^{m}\alpha_i y_i)+\frac{(\beta_m y_m+\sum_{i=1}^{m}\beta_i y_i)\,(\alpha_m+\sum_{i=1}^{m}\alpha_i-m)}{\beta_m+\sum_{i=1}^{m}\beta_i}, \tag{3.7}$$

$$B=-(\beta_m y_m^2+\sum_{i=1}^{m}\beta_i y_i^2)+\frac{(\beta_m y_m+\sum_{i=1}^{m}\beta_i y_i)^2}{\beta_m+\sum_{i=1}^{m}\beta_i}. \tag{3.8}$$

Now, by using (3.4), we obtain the approximate MLE of μ as

$$\tilde{\mu}=C+D\tilde{\sigma}, \tag{3.9}$$

where

$$C = \frac{\beta_m y_m + \sum_{i=1}^{m} \beta_i y_i}{\beta_m + \sum_{i=1}^{m} \beta_i}, \qquad D = \frac{\alpha_m + \sum_{i=1}^{m} \alpha_i - m}{\beta_m + \sum_{i=1}^{m} \beta_i}. \tag{3.10}$$

Note that Eq. (3.5) has two roots but since $B \leq 0$, only one root in (3.6) is admissible. The proof of $B \leq 0$ is given in Appendix A.

Note that, the AMLE method has an advantage over the MLE method as the former provides explicit estimators. The AMLEs in (3.6) and (3.9) can be used as good starting values for the iterative solution of the likelihood equations (2.9) and (2.10) to obtain the MLEs. As mentioned in Tiku and Akkaya (2004), the AMLEs of the location an scale parameters μ and σ are asymptotically equivalent to the corresponding MLEs for any location-scale distribution. This is due to the asymptotic equivalence of the approximate likelihood and the likelihood equations. The approximate MLEs have all desirable asymptotic properties of MLEs. They are asymptotically unbiased and efficient. They have also robustness properties for all the three types distributions: skew, short-tailed symmetric and long-tailed symmetric distributions. For more details, see Tiku and Akkaya (2004).

4. Bayesian estimation and credible intervals

In this section, the Bayes estimators of the unknown parameters μ and σ are derived under the squared error loss function. Further, the corresponding credible intervals of μ and σ are also obtained. It is assumed that joint prior distribution for μ and σ is in the form

$$\pi(\mu,\sigma) = \pi_1(\mu|\sigma)\pi_2(\sigma),$$

where σ has an inverse gamma prior $IG(a,b)$, with the pdf

$$\pi_2(\sigma) \propto e^{-\frac{b}{\sigma}} \sigma^{-(a+1)}, \quad \sigma > 0, \;\; a,b > 0,$$

and μ given σ has the logistic prior with parameters μ_0 and σ

$$\pi_1(\mu|\sigma) = \frac{e^{-\frac{\mu-\mu_0}{\sigma}}}{\sigma\left[1+e^{-\frac{\mu-\mu_0}{\sigma}}\right]^2},$$

This joint prior is suitable for deriving the posterior distribution in a location and scale parameter estimation.

From (2.6), for the logistic distribution, the likelihood function of μ and σ for the given record sample $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ is given by

$$L(\mu,\sigma|\mathbf{y}) = e^{-\frac{y_m-\mu}{\sigma}}\sigma^{-m}\frac{\prod_{i=1}^{m}(1+e^{-\frac{y_i-\mu}{\sigma}})^{-1}}{1+e^{-\frac{y_m-\mu}{\sigma}}}. \tag{4.1}$$

By combining the likelihood function in (4.1) and the joint prior distribution, we obtain the joint posterior distribution of μ and σ as

$$\pi(\mu,\sigma|\mathbf{y}) \propto e^{-\frac{b+y_m-\mu_0}{\sigma}}\sigma^{-(m+a+2)}\ \frac{\prod_{i=1}^{m}(1+e^{-\frac{y_i-\mu}{\sigma}})^{-1}}{\left[1+e^{-\frac{y_m-\mu}{\sigma}}\right]\left[1+e^{-\frac{\mu-\mu_0}{\sigma}}\right]^2}. \tag{4.2}$$

Therefore, the Bayes estimators of μ and σ are respectively obtained as

$$\widehat{\mu}_{BS} = E(\mu|\mathbf{y}) = k\int_{-\infty}^{\infty}\int_{0}^{\infty}\mu\, e^{-\frac{b+y_m-\mu_0}{\sigma}}\sigma^{-(m+a+2)}\ \frac{\prod_{i=1}^{m}(1+e^{-\frac{y_i-\mu}{\sigma}})^{-1}}{\left[1+e^{-\frac{y_m-\mu}{\sigma}}\right]\left[1+e^{-\frac{\mu-\mu_0}{\sigma}}\right]^2}d\sigma d\mu,$$

and

$$\widehat{\sigma}_{BS} = E(\sigma|\mathbf{y}) = k\int_{-\infty}^{\infty}\int_{0}^{\infty}e^{-\frac{b+y_m-\mu_0}{\sigma}}\sigma^{-(m+a+1)}\ \frac{\prod_{i=1}^{m}(1+e^{-\frac{y_i-\mu}{\sigma}})^{-1}}{\left[1+e^{-\frac{y_m-\mu}{\sigma}}\right]\left[1+e^{-\frac{\mu-\mu_0}{\sigma}}\right]^2}d\sigma d\mu,$$

where k is the normalizing constant.

It is seen that the Bayes estimators can not be obtained in closed forms. In what follows, similarly as in Kundu (2007, 2008), we provide the approximate Bayes estimators using a rejection-sampling within the Gibbs sampling procedure. Note that the joint posterior distribution of μ and σ given $\mathbf{y}$ in (4.2), can be written as

$$\pi(\mu,\sigma|\mathbf{y}) \propto g_1(\sigma|\mathbf{y})g_2(\mu|\sigma,\mathbf{y}). \tag{4.3}$$

Here $g_1(\sigma|\mathbf{y})$ is an inverse gamma density function with the shape and scale parameters as $m+a+1$ and $b+y_m-\mu_0$, respectively, and $g_2(\mu|\sigma,\mathbf{y})$ is a proper density function given by

$$g_2(\mu|\sigma,\mathbf{y}) \propto \frac{\prod_{i=1}^{m}(1+e^{-\frac{y_i-\mu}{\sigma}})^{-1}}{\left[1+e^{-\frac{y_m-\mu}{\sigma}}\right]\left[1+e^{-\frac{\mu-\mu_0}{\sigma}}\right]^2}. \tag{4.4}$$

To obtain the Bayes estimates using the Gibbs sampling procedure, we need the following result.

Theorem 1. *The conditional distribution of μ given σ and $\mathbf{y}$, $g_2(\mu|\sigma,\mathbf{y})$, is log-concave.*

Proof: See the Appendix B.

Thus, the samples of μ can be generated from (4.4) using the method proposed by Devroye (1984). Now, using Theorem 1, and adopting the method of Devroye (1984), we can generate the samples (μ,σ) from the posterior density function (4.3), using the Gibbs sampling procedure as follows:

1. Generate σ_1 from $g_1(.|\mathbf{y})$.
2. Generate μ_1 from $g_2(.|\sigma_1,\mathbf{y})$ using the method developed by Devroye (1984).
3. Repeat steps 1 and 2 N times and obtain $(\mu_1,\sigma_1),\cdots,(\mu_N,\sigma_N)$.

Note that in step 2, we use the Devroye algorithm as follows:

i) Compute $c = g_2(m|\sigma,\mathbf{y})$. (m is the mode of $g_2(.|\sigma,\mathbf{y})$).

ii) Generate U uniform on [0,2], and V uniform on [0,1].

iii) If $U \leq 1$ then $\mu = U$ and $T = V$, else $\mu = 1 - \ln(U-1)$ and $T = V(U-1)$.

iv) Let $\mu = m + \frac{\mu}{c}$.

v) If $T \leq \frac{g_2(\mu|\sigma,\mathbf{y})}{c}$, then μ is a sample from $g_2(.|\sigma,\mathbf{y})$, else go to Step (ii).

Now, the Bayesian estimators of μ and σ under the squared error loss function are obtained as

$$\widehat{\mu}_{BS} = \frac{\sum_{j=1}^{N}\mu_j}{N}, \qquad \widehat{\sigma}_{BS} = \frac{\sum_{j=1}^{N}\sigma_j}{N}. \tag{4.5}$$

Now we obtain the credible intervals of μ and σ using the idea of Chen and Shao (1999). To compute the credible intervals of μ and σ, we generate $\mu_1,\ldots,\mu_N$ and $\sigma_1,\ldots,\sigma_N$ as described above. We then order $\mu_1,\ldots,\mu_N$ and $\sigma_1,\ldots,\sigma_N$ as $\mu_{(1)},\ldots,\mu_{(N)}$ and $\sigma_{(1)},\ldots,\sigma_{(N)}$. Then, the $100(1-\gamma)\%$ credible intervals μ and σ can be constructed as

$$\left(\mu_{(\frac{\gamma}{2}N)},\ \mu_{((1-\frac{\gamma}{2})N)}\right), \qquad \left(\sigma_{(\frac{\gamma}{2}N)},\ \sigma_{((1-\frac{\gamma}{2})N)}\right). \tag{4.6}$$

5. Fisher information and different confidence intervals

In this section, we derive the Fisher information matrix based on the likelihood as well as the approximate likelihood functions. Using the Fisher information matrix and based on the asymptotic distribution of MLEs, we can obtain the asymptotic confidence intervals of μ and σ. We further, propose two confidence intervals based on the bootstrap method.

5.1. Fisher information

From (2.9) and (2.10), the expected Fisher information matrix of $\boldsymbol{\theta} = (\mu, \sigma)$ is

$$I(\boldsymbol{\theta}) = -\begin{pmatrix} E(\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu^2}) & E(\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu \partial \sigma}) \\ E(\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \sigma \partial \mu}) & E(\frac{\partial^2 \ln L(\mu,\sigma)}{\partial \sigma^2}) \end{pmatrix} = \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}, \tag{5.1}$$

where

$$\begin{aligned} I_{11} &= \frac{1}{\sigma^2}\left[E[f(X_m)] + \sum_{i=1}^{m} E[f(X_i)]\right], \\ I_{12} &= \frac{1}{\sigma^2}\left[m - E[F(X_m)] - \sum_{i=1}^{m} E[F(X_i)] - E[X_m f(X_m)] - \sum_{i=1}^{m} E[X_i f(X_i)]\right], \\ I_{22} &= -\frac{1}{\sigma^2}\left[m + 2\sum_{i=1}^{m} E[X_i(1-F(X_i))] - 2E[X_m F(X_m)]\right. \\ &\qquad \left. - E[X_m^2 f(X_m)] - \sum_{i=1}^{m} E[X_i^2 f(X_i)]\right]. \end{aligned}$$

Similarly, the expected approximate Fisher information matrix of $\boldsymbol{\theta} = (\mu, \sigma)$ is obtained to be

$$I^*(\boldsymbol{\theta}) = -\begin{pmatrix} E(\frac{\partial^2 \ln L^*(\mu,\sigma)}{\partial \mu^2}) & E(\frac{\partial^2 \ln L^*(\mu,\sigma)}{\partial \mu \partial \sigma}) \\ E(\frac{\partial^2 \ln L^*(\mu,\sigma)}{\partial \sigma \partial \mu}) & E(\frac{\partial^2 \ln L^*(\mu,\sigma)}{\partial \sigma^2}) \end{pmatrix} = \begin{pmatrix} I_{11}^* & I_{12}^* \\ I_{12}^* & I_{22}^* \end{pmatrix}, \tag{5.2}$$

where

$$\begin{aligned} I_{11}^* &= \frac{1}{\sigma^2}\left[\beta_m + \sum_{i=1}^{m} \beta_i\right], \\ I_{12}^* &= -\frac{1}{\sigma^2}\left[m - \alpha_m - \sum_{i=1}^{m} \alpha_i - 2\beta_m E[X_m] - 2\sum_{i=1}^{m} \beta_i E[X_i]\right], \end{aligned}$$

$$I_{22}^{*} = -\frac{1}{\sigma^2}\left[m + 2\sum_{i=1}^{m}(1-\alpha_i)E[X_i] - 2\alpha_m E[X_m] - 3\beta_m E[X_m^2] - 3\sum_{i=1}^{m}\beta_i E[X_i^2]\right].$$

From Ahsanullah (1995), since

$$E[X_1] = 0, \quad E[X_i] = \sum_{l=2}^{i}\zeta(l), \quad i \geq 2,$$

and

$$E[X_i^2] = 2i\sum_{l=2}^{i+1}\zeta(l) - i(i+1) + \sum_{l=2}^{\infty}\frac{B_l}{(l+1)^i},$$

where $\zeta(.)$ is Riemann zeta function $\zeta(n) = \sum_{k=1}^{\infty} k^{-n}$ and for $n \geq 2$

$$B_n = \frac{1}{n}(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}),$$

we can derive the elements of Fisher information matrix in (5.2). Now, to derive the elements of Fisher information matrix in (5.1), we need to calculate the expectations $E[f(X_i)]$, $E[F(X_i)]$, $E[X_i(1-F(X_i))]$, $E[X_i f(X_i)]$, $E[X_i F(X_i)]$ and $E[X_i^2 f(X_i)]$. We use the following lemma to compute these expectations.

Lemma 1. *Let $X_1 < X_2 < \cdots < X_m$ is the first m upper record values from the standard logistics distribution with pdf (2.3). Then we have*

$$E[f(X_i)] = \frac{1}{2^i} - \frac{1}{3^i}, \tag{5.3}$$

$$E[F(X_i)] = 1 - \frac{1}{2^i}, \tag{5.4}$$

$$E[X_i f(X_i)] = \sum_{l=1}^{\infty}\left[\frac{1}{l(l+3)^i} - \frac{1}{l(l+2)^i}\right] + i\left[\frac{1}{2^i} - \frac{1}{3^i}\right], \tag{5.5}$$

$$E\left[X_i(1-F(X_i))\right] = \frac{i}{2^{i+1}} - \sum_{l=1}^{\infty}\frac{1}{l(l+2)^i}, \tag{5.6}$$

and

$$E\left[X_i^2 f(X_i)\right] = \sum_{l=1}^{\infty}\left[\frac{1}{l^2(2l+2)^i} - \frac{1}{l^2(2l+3)^i}\right] + 2\sum\sum_{1\le l<k<\infty}\left[\frac{1}{lk(l+k+2)^i} - \frac{1}{lk(l+k+3)^i}\right] + 2i\sum_{l=1}^{\infty}\left[\frac{1}{l(l+3)^{i+1}} - \frac{1}{l(l+2)^{i+1}}\right] + i(i+1)\left[\frac{1}{2^{i+2}} - \frac{1}{3^{i+2}}\right]. \tag{5.7}$$

Proof. See the Appendix C.

Moreover, $E[X_iF(X_i)]$ can be obtained from the expression

$$E[X_iF(X_i)] = E[X_i] - E[X_i(1-F(X_i))].$$

It should be mentioned here that the loss of information due to using record data instead of the complete logistic data can be discussed by comparing the Fisher information contained in record data with that of the Fisher information contained in the complete data. Since $\boldsymbol{\theta} = (\mu,\sigma)$ is a vector parameter, the comparison is not a trivial issue. One method is that to compare the Fisher information matrices for the two data using their traces. Based on a given data, the trace of Fisher information matrix of $\boldsymbol{\theta} = (\mu,\sigma)$ is the sum of the Fisher information measures of μ, when σ is known, and σ, when μ is known. For the logistic distribution, the Fisher information matrix of $\boldsymbol{\theta} = (\mu,\sigma)$ based on the first m record observations can be obtained from (5.1). On the other hand, the Fisher information matrix based on the m complete logistic observations is (see Nadarajah (2004))

$$J(\boldsymbol{\theta}) = \begin{pmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{pmatrix},$$

where

$$J_{11} = \frac{m}{3\sigma^2}\left(\frac{\pi^2}{3}+1\right),$$
$$J_{12} = J_{12} = -\frac{m}{\sigma^2},$$
$$J_{22} = \frac{m}{3\sigma^2}.$$

Table 1: *The trace of the Fisher information matrix based on complete and record observations for different values of m.*

	Complete observations	Record observations
$m = 2$	3.526	3.149
$m = 3$	5.289	4.502
$m = 5$	8.816	6.916
$m = 10$	17.633	12.131
$m = 15$	26.450	19.175
$m = 20$	35.265	27.917

We have computed the traces of the corresponding Fisher information matrices for both data and the results are reported in Table 1. From Table 1, as expected, we see that the Fisher information contained in the m complete observations is greater than that the Fisher information contained in the m record observations.

5.2. Different confidence intervals

Now, the variances of the MLEs $\widehat{\mu}$ and $\widehat{\sigma}$, can be approximated by inverting the Fisher information matrix in (5.1), *i.e*,

$$\begin{pmatrix} Var(\widehat{\mu}) & Cov(\widehat{\mu},\widehat{\sigma}) \\ Cov(\widehat{\mu},\widehat{\sigma}) & Var(\widehat{\sigma}) \end{pmatrix} \approx \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}^{-1}. \tag{5.8}$$

The approximate asymptotic variance covariance matrices are valid only if asymptotic normality holds. For the asymptotic normality, the certain regularity conditions must be satisfied (see, for example, the conditions in Theorem 4.17 of Shao (2003)). These conditions mainly relate to differentiability of the density and the ability to interchange differentiation and integration. In most reasonable problems, the regularity conditions are often satisfied. Since the logistic distribution satisfies all the the regularity conditions (see Shao (2005), Pages 198-200), we can obtain the approximate $100(1-\gamma)\%$ confidence intervals of μ and σ using the asymptotic normality of MLEs as

$$\left(\widehat{\mu} - z_{1-\gamma/2}\sqrt{\widehat{Var(\widehat{\mu})}} \ , \ \widehat{\mu} + z_{1-\gamma/2}\sqrt{\widehat{Var(\widehat{\mu})}}\right), \tag{5.9}$$

and

$$\left(\widehat{\sigma} - z_{1-\gamma/2}\sqrt{\widehat{Var(\widehat{\sigma})}} \ , \ \widehat{\sigma} + z_{1-\gamma/2}\sqrt{\widehat{Var(\widehat{\sigma})}}\right). \tag{5.10}$$

Similarly, the approximate confidence intervals can be obtained based on the AMLEs also, by inverting the approximate Fisher information in (5.2).

Now, we present two confidence intervals based on the parametric bootstrap methods: (i) percentile bootstrap method (we call it Boot-p) based on the idea of Efron (1982), (ii) bootstrap-t method (we refer to it as Boot-t) based on the idea of Hall (1988). The algorithms for these two bootstrap procedures are briefly described as follows.

(i) Boot-p method:

1. Estimate μ and σ, say $\widehat{\mu}$ and $\widehat{\sigma}$, from sample based on the MLE procedure.
2. Generate a bootstrap sample $\{X_1^*,\cdots,X_m^*\}$, using $\widehat{\mu}$ and $\widehat{\sigma}$. Obtain the bootstrap estimates of μ and σ, say $\widehat{\mu}^*$ and $\widehat{\sigma}^*$ using the bootstrap sample.
3. Repeat Step 2 NBOOT times.
4. Order $\widehat{\mu}_1^*,\cdots,\widehat{\mu}_{NBOOT}^*$ as $\widehat{\mu}_{(1)}^*,\cdots,\widehat{\mu}_{(NBOOT)}^*$ and $\widehat{\sigma}_1^*,\cdots,\widehat{\sigma}_{NBOOT}^*$ as $\widehat{\sigma}_{(1)}^*,\cdots,\widehat{\sigma}_{(NBOOT)}^*$. Then, the approximate $100(1-\gamma)\%$ confidence intervals for μ and σ become, respectively, as

$$\left(\widehat{\mu}_{Boot-p}^*(\frac{\gamma}{2}) \ , \ \widehat{\mu}_{Boot-p}^*(1-\frac{\gamma}{2})\right), \qquad \left(\widehat{\sigma}_{Boot-p}^*(\frac{\gamma}{2}) \ , \ \widehat{\sigma}_{Boot-p}^*(1-\frac{\gamma}{2})\right). \quad (5.11)$$

(ii) Boot-t method:

1. Estimate μ and σ, say $\widehat{\mu}$ and $\widehat{\sigma}$, from sample based on the MLE method.
2. Generate a bootstrap sample $\{X_1^*,\cdots,X_m^*\}$, using $\widehat{\mu}$ and $\widehat{\sigma}$ and obtain the bootstrap estimates of μ and σ, say $\widehat{\mu}^*$ and $\widehat{\sigma}^*$ using the bootstrap sample.
3. Determine

$$T_\mu^* = \frac{(\widehat{\mu}^*-\widehat{\mu})}{\sqrt{\widehat{Var}(\widehat{\mu}^*)}}, \qquad T_\sigma^* = \frac{(\widehat{\sigma}^*-\widehat{\sigma})}{\sqrt{\widehat{Var}(\widehat{\sigma}^*)}},$$

   where $\widehat{Var}(\widehat{\mu}^*)$ and $\widehat{Var}(\widehat{\sigma}^*)$ are obtained using (5.8)
4. Repeat Steps 2 and 3 NBOOT times.
5. Define $\widehat{\mu}_{Boot-t}^* = \widehat{\mu}+\sqrt{\widehat{Var}(\widehat{\mu}^*)}T_\mu^*$ and $\widehat{\sigma}_{Boot-t}^* = \widehat{\sigma}+\sqrt{\widehat{Var}(\widehat{\sigma}^*)}T_\sigma^*$. Order $\widehat{\mu}_1^*,\cdots,\widehat{\mu}_{NBOOT}^*$ as $\widehat{\mu}_{(1)}^*,\cdots,\widehat{\mu}_{(NBOOT)}^*$ and $\widehat{\sigma}_1^*,\cdots,\widehat{\sigma}_{NBOOT}^*$ as $\widehat{\sigma}_{(1)}^*,\cdots,\widehat{\sigma}_{(NBOOT)}^*$. Then, the approximate $100(1-\gamma)\%$ confidence intervals for μ and σ become respectively as

$$\left(\widehat{\mu}_{Boot-t}^*(\frac{\gamma}{2}) \ , \ \widehat{\mu}_{Boot-t}^*(1-\frac{\gamma}{2})\right), \qquad \left(\widehat{\sigma}_{Boot-t}^*(\frac{\gamma}{2}) \ , \ \widehat{\sigma}_{Boot-t}^*(1-\frac{\gamma}{2})\right). \quad (5.12)$$

6. Data analysis and simulation

In this section, we analyze a real data set to illustrate the estimation methods presented in the preceding sections. Further, a Monte Carlo simulation study is conducted to compare the performance of proposed estimators.

6.1. Data analysis

The following data are the total annual rainfall (in inches) during March recorded at Los Angeles Civic Center from 1973 to 2006 (see the website of Los Angeles Almanac: www.laalman-ac.com/weather/we08aa.htm).

```
2.70  3.78  4.83  1.81  1.89  8.02  5.85  4.79  4.10  3.54
8.37  0.28  1.29  5.27  0.95  0.26  0.81  0.17  5.92  7.12
2.74  1.86  6.98  2.16  0.00  4.06  1.24  2.82  1.17  0.32
4.31  1.17  2.14  2.87
```

The Los Angeles rainfall data have been used earlier by some authors. See for example, Raqab (2006), Madi and Raqab (2007) and Raqab et al. (2010).

We analyzed the above rainfall data by using the logistic distribution with $\mu = 2.905$ and $\sigma = 1.367$. It is observed that the Kolmogorov-Smirnov (KS) distance and the corresponding p-value are respectively

$$\text{KS} = 0.1066, \qquad \text{and} \qquad \text{p-value} = 0.8120.$$

Hence the logistic model (2.1) fits quite well to the above data.

For the above data, we observe the following five upper record values

```
2.70   3.78   4.83   8.02   8.37
```

We shall use the above rainfall records to obtain the different estimators discussed in this paper. Here, we have $m = 5, A = -3.644, B = -1.436, C = 4.089$ and $D = -0.742$. From (3.6), we obtain the AMLE of σ as

$$\tilde{\sigma} = \frac{-A + \sqrt{A^2 - 4mB}}{2m} = 1.012.$$

Now, by using (3.9), the AMLE of μ becomes

$$\tilde{\mu} = C + D\tilde{\sigma} = 3.338.$$

The MLEs of μ and σ are then respectively as $\widehat{\mu} = 2.929$ and $\widehat{\sigma} = 0.998$. Note that the MLEs were obtained by solving the nonlinear equations (2.9) and (2.10) using the Maple package, in which the AMLEs were used as starting values for the iterations. To ensure that the solution $(\widehat{\mu} = 2.929, \widehat{\sigma} = 0.998)$ of the likelihood equations (2.9) and (2.10) is indeed a maximum, it must be shown that the matrix of second-order partial derivatives (Hessian matrix)

$$H = \begin{pmatrix} \frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu^2} & \frac{\partial^2 \ln L(\mu,\sigma)}{\partial \mu \partial \sigma} \\ \frac{\partial^2 \ln L(\mu,\sigma)}{\partial \sigma \partial \mu} & \frac{\partial^2 \ln L(\mu,\sigma)}{\partial \sigma^2} \end{pmatrix},$$

is a negative definite when $\mu = \widehat{\mu}$ and $\sigma = \widehat{\sigma}$. Based on the above rainfall records and for $\widehat{\mu} = 2.929$ and $\widehat{\sigma} = 0.998$, the Hessian matrix is

$$H = \begin{pmatrix} -0.5857 & 0.4156 \\ 0.4156 & -5.0194 \end{pmatrix},$$

which can be shown that is negative definite. Therefore, we have indeed found a maximum. On the other hand, we have also plotted the likelihood function of μ and σ for the given record data in Figure 1. From Figure 1, one can observe that the likelihood surface has curvature in both μ and σ directions. This leads to the interpretation that MLEs of μ and σ are exist and unique.

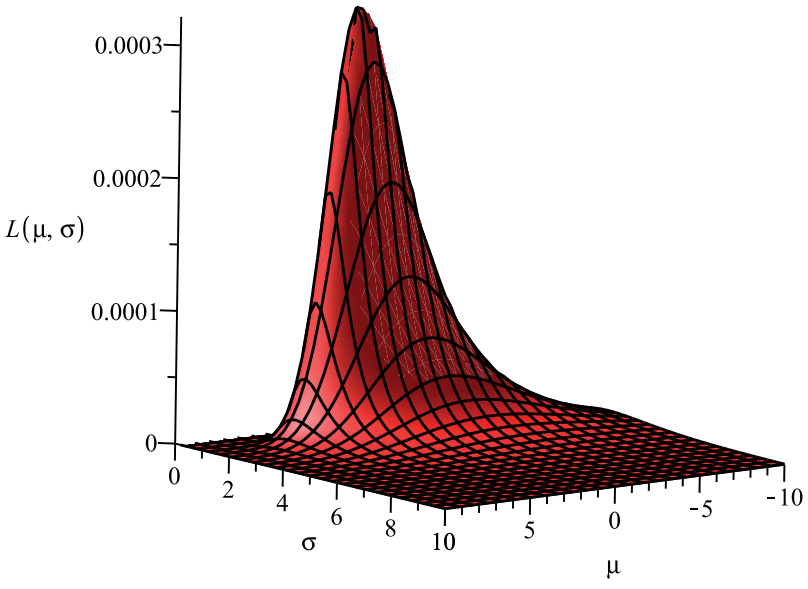

Figure 1: *Likelihood function of μ and σ.*

Table 2: *Point and interval estimators of μ and σ.*

	Point estimators			95% Confidence intervals				
	MLE	AMLE	Bayes	MLE	AMLE	p-boot	t-boot	Bayes
μ	2.929	3.338	3.649	(2.059,3.798)	(1.785,4.889)	(0.643,5.554)	(1.501,3.641)	(2.831,4.239)
σ	0.998	1.012	1.370	(0.625,1.371)	(0.689,1.335)	(0.251,0.936)	(0.719,0.992)	(0.592,1.604)

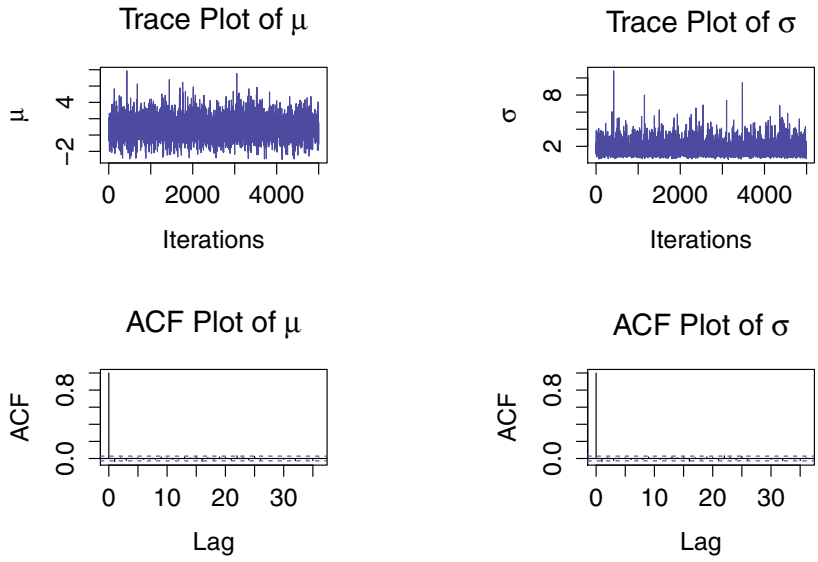

Figure 2: *Trace and autocorrelation plots of μ and σ.*

We also computed the Bayes estimators of μ and σ using Gibbs sampling procedure. To compute the Bayes estimators, since we do not have any prior information, we have used very small (close to zero) values of the hyper-parameters on σ, i.e. $a = b = 0.00001$. In this case, the prior on σ is a proper prior but it is almost improper. Since μ_0 is a location parameter for the logistic prior of μ given σ, without loss of generality, we assumed that $\mu_0 = 0$. For Gibbs sampling procedure we use $N = 5000$ and we have checked the convergence of generated samples of μ and σ. We have used the graphical diagnostics tools like trace plots and autocorrelation function (ACF) plots for this purpose. Figure 2 shows the trace plots and ACF plots for the parameters. The trace plots look like a random scatter and show the fine mixing of the chains for both parameters μ and σ. ACF plots show that chains have very low autocorrelations. Based on these plots, we can fairly conclude that convergence has been attained.

We also computed different confidence intervals namely the approximate confidence intervals based MLEs and AMLEs, p-boot and t-boot confidence intervals and credible intervals. All results are reported in Table 2.

6.2. Simulation study

In this section, a Monte Carlo simulation is conducted to compare the performance of the different estimators. In this simulation, we have randomly generated 1000 upper record sample $X_1, X_2, \ldots, X_m$ from the standard logistic distribution (i.e., $\mu = 0$ and $\sigma = 1$) and then computed the MLEs, AMLEs and Bayes estimators of μ and σ. We then compared the performances of these estimators in terms of biases, and mean square errors (MSEs). For computing Bayes estimators, we take $\mu_0 = 0$. We use both non-informative and informative priors for the scale parameter σ. In case of non-informative prior, we take $a = b = 0$. We call it as Prior 1. For the informative prior, we chose $a = 3$ and $b = 1$. We call it as Prior 2. Clearly Prior 2 is more informative than the non-informative Prior 1.

In Table 3, for different values of m, we reported the average biases, and MSEs of the MLEs, AMLEs and Bayes estimators over 1000 replications. All the computations are performed using Visual Maple (V16) package.

Table 3: *Biases and MSEs of the MLEs, AMLEs and Bayes estimators for different values of m.*

		Estimation of μ				Estimation of σ			
		MLE	AMLE	Bayes		MLE	AMLE	Bayes	
				Prior 1	Prior 2			Prior 1	Prior 2
$m = 2$	Bias	−0.732	−0.749	−0.635	−0.608	0.362	0.386	0.310	0.286
	MSE	2.619	2.654	2.574	2.543	0.510	0.538	0.497	0.467
$m = 3$	Bias	−0.653	−0.681	−0.568	−0.534	0.284	0.297	0.261	0.242
	MSE	2.468	2.492	2.419	2.397	0.451	0.468	0.416	0.402
$m = 5$	Bias	−0.558	−0.579	−0.488	−0.443	0.142	0.166	0.123	0.107
	MSE	2.129	2.171	1.938	1.903	0.109	0.139	0.087	0.073
$m = 10$	Bias	−0.313	−0.366	−0.265	−0.244	0.067	0.084	0.041	0.016
	MSE	1.567	1.636	1.510	1.482	0.059	0.067	0.049	0.041
$m = 15$	Bias	−0.238	−0.250	−0.197	−0.170	0.059	0.063	0.053	0.048
	MSE	1.148	1.176	1.125	1.107	0.043	0.051	0.034	0.027
$m = 20$	Bias	−0.150	−0.177	−0.121	−0.104	0.033	0.045	0.021	0.018
	MSE	0.999	1.024	0.956	0.937	0.024	0.033	0.018	0.011

From Table 3, we observe that the AMLEs and the MLEs are almost identical in terms of both bias and MSEs. The AMLEs are almost as efficient as the MLEs for all sample sizes. Comparing the two Bayes estimators based on two priors 1 and 2, it is observed that the Bayes estimators based on prior 2 perform better than the Bayes estimators based on non-informative prior 1. In addition, the Bayes estimators perform better than the classical estimators MLEs and AMLEs. It is also noted as m increases, the performances of all estimators better in terms of biases and MSEs.

We also computed the 95% confidence/credible intervals for μ and σ based on the asymptotic distributions of the MLEs and AMLEs. We further computed Boot-p, and

Table 4: *Average confidence/credible lengths and coverage probabilities for different values of m.*

			MLE	AMLE	p-boot	t-boot	Bayes	
							Prior 1	Prior 2
Estimation of μ	$m=2$	Length	1.964	1.972	1.937	1.925	1.916	1.892
		Cov. Prob.	0.937	0.936	0.938	0.939	0.939	0.940
	$m=3$	Length	1.729	1.741	1.709	1.686	1.681	1.669
		Cov. Prob.	0.938	0.937	0.940	0.941	0.940	0.941
	$m=5$	Length	1.411	1.424	1.384	1.377	1.358	1.345
		Cov. Prob.	0.939	0.937	0.941	0.942	0.941	0.943
	$m=10$	Length	1.097	1.110	1.068	1.046	1.028	1.009
		Cov. Prob.	0.941	0.940	0.943	0.943	0.943	0.944
	$m=15$	Length	0.804	0.811	0.794	0.783	0.752	0.739
		Cov. Prob.	0.943	0.942	0.944	0.945	0.945	0.946
	$m=20$	Length	0.653	0.673	0.634	0.625	0.605	0.590
		Cov. Prob.	0.945	0.943	0.945	0.947	0.947	0.948
Estimation of σ	$m=2$	Length	1.310	1.328	1.286	1.279	1.271	1.260
		Cov. Prob.	0.939	0.936	0.939	0.940	0.939	0.941
	$m=3$	Length	1.186	1.197	1.172	1.164	1.152	1.140
		Cov. Prob.	0.941	0.939	0.941	0.941	0.942	0.943
	$m=5$	Length	0.924	0.931	0.907	0.902	0.894	0.887
		Cov. Prob.	0.942	0.940	0.943	0.944	0.944	0.945
	$m=10$	Length	0.716	0.724	0.701	0.694	0.680	0.671
		Cov. Prob.	0.943	0.941	0.943	0.944	0.945	0.946
	$m=15$	Length	0.543	0.560	0.530	0.522	0.516	0.505
		Cov. Prob.	0.944	0.942	0.944	0.945	0.946	0.948
	$m=20$	Length	0.375	0.383	0.366	0.359	0.352	0.345
		Cov. Prob.	0.946	0.945	0.945	0.947	0.948	0.949

Boot-t confidence intervals, and the credible intervals. Table 4 presents the average confidence/credible lengths and the corresponding coverage probability over 1000 replications. The nominal level for the confidence intervals is 0.95 in each case.

From Table 4, the length of the 95% confidence interval based on the asymptotic distribution of the MLE, is slightly smaller than the corresponding length of the interval based on the asymptotic distribution of the AMLE. We also observe that the Bayesian credible intervals work slightly better than the bootstrap and asymptotic confidence intervals in terms of both confidence length and coverage probability. Also, Boot-t confidence intervals perform very similarly to the Bayesian credible intervals. The bootstrap confidence intervals work better than the asymptotic confidence intervals. The Boot-t confidence intervals perform better than the Boot-p confidence intervals. Also, it is observed that all the simulated coverage probabilities are very close to the nominal level

95%. Also, for all interval estimators, the confidence lengths and the simulated coverage percentages decrease as m increases.

Overall speaking, from Tables 3 and 4, we would recommend the use of Bayesian method for point and interval estimation, especially when reliable prior information about the logistic parameters is available.

Appendix A

To prove $B \leq 0$, we need to show that

$$\frac{(\beta_m y_m + \sum_{i=1}^{m} \beta_i y_i)^2}{\beta_m + \sum_{i=1}^{m} \beta_i} \leq (\beta_m y_m^2 + \sum_{i=1}^{m} \beta_i y_i^2),$$

or equivalently

$$2\beta_m y_m \sum_{i=1}^{m} \beta_i y_i + \left(\sum_{i=1}^{m} \beta_i y_i\right)^2 \leq \beta_m \sum_{i=1}^{m} \beta_i y_i^2 + \beta_m y_m^2 \sum_{i=1}^{m} \beta_i + \left(\sum_{i=1}^{m} \beta_i\right)\left(\sum_{i=1}^{m} \beta_i y_i^2\right). \tag{A.1}$$

We can rewrite (A.1) as

$$\beta_m \left(\sum_{i=1}^{m} \beta_i [2y_m y_i]\right) + \left(\sum_{i=1}^{m} \beta_i y_i\right)^2 \leq \beta_m \left(\sum_{i=1}^{m} \beta_i [y_i^2 + y_m^2]\right) + \left(\sum_{i=1}^{m} \beta_i\right)\left(\sum_{i=1}^{m} \beta_i y_i^2\right). \tag{A.2}$$

Now, since $y_i^2 + y_j^2 \geq 2y_i y_j$, we have

$$\beta_m \left(\sum_{i=1}^{m} \beta_i [2y_m y_i]\right) \leq \beta_m \left(\sum_{i=1}^{m} \beta_i [y_i^2 + y_m^2]\right), \tag{A.3}$$

and

$$\left(\sum_{i=1}^{m} \beta_i y_i\right)^2 \leq \left(\sum_{i=1}^{m} \beta_i\right)\left(\sum_{i=1}^{m} \beta_i y_i^2\right). \tag{A.4}$$

Now by using (A.3) and (A.4), (A.2) is true and the proof is thus obtained.

Appendix B (Proof of Theorem 1)

The log-likelihood function of $g_2(\mu|\sigma)$ is

$$\ln g_2(\mu|\sigma,\mathbf{y}) \propto -\sum_{i=1}^{m}\ln(1+e^{-(y_i-\mu)/\sigma}) - \ln(1+e^{-(y_m-\mu)/\sigma}) - 2\ln(1+e^{-(\mu-\mu_0)/\sigma}).$$

The second derivative of $\ln g_2(\mu|\sigma,\mathbf{y})$ is obtained as

$$-\frac{1}{\sigma^2}\left[\sum_{i=1}^{m}\frac{e^{-(y_i-\mu)/\sigma}}{(1+e^{-(y_i-\mu)/\sigma})^2} + \frac{e^{-(y_m-\mu)/\sigma}}{(1+e^{-(y_m-\mu)/\sigma})^2} + \frac{2\,e^{-(\mu-\mu_0)/\sigma}}{(1+e^{-(\mu-\mu_0)/\sigma})^2}\right],$$

which is negative. So, the result follows.

Appendix C (Proof of Lemma 1)

Using the relation (2.5), we have

$$\begin{aligned}
E\left[f(X_i)\right] &= E\left[F(X_i)(1-F(X_i))\right] \\
&= \int_{-\infty}^{\infty} F(x)[1-F(x)]\frac{[-\ln(1-F(x))]^{i-1}}{(i-1)!}f(x)dx \\
&= \int_{0}^{1} u(1-u)\frac{[-\ln(1-u)]^{i-1}}{(i-1)!}du \\
&= \int_{0}^{\infty}(1-e^{-t})e^{-2t}\frac{t^{i-1}}{(i-1)!}dt = \frac{1}{2^i}-\frac{1}{3^i},
\end{aligned}$$

and

$$\begin{aligned}
E\left[F(X_i)\right] &= \int_{-\infty}^{\infty} F(x)\frac{[-\ln(1-F(x))]^{i-1}}{(i-1)!}f(x)dx \\
&= \int_{0}^{1} u\frac{[-\ln(1-u)]^{i-1}}{(i-1)!}du \\
&= \int_{0}^{\infty}(1-e^{-t})e^{-t}\frac{t^{i-1}}{(i-1)!}dt = 1-\frac{1}{2^i}.
\end{aligned}$$

We also have

$$\begin{aligned} E\left[X_i f(X_i)\right] &= E\left[X_i F(X_i)(1-F(X_i))\right] \\ &= \int_{-\infty}^{\infty} xF(x)[1-F(x)]\frac{[-\ln(1-F(x))]^{i-1}}{(i-1)!}f(x)dx \\ &= \int_{0}^{1} [\ln u - \ln(1-u)]u(1-u)\frac{[-\ln(1-u)]^{i-1}}{(i-1)!}du, \end{aligned}$$

since $F^{-1}(u) = \ln u - \ln(1-u)$. Setting $t = -\ln(1-u)$, we get

$$\begin{aligned} E\left[X_i f(X_i)\right] &= \int_{0}^{\infty} \ln(1-e^{-t})e^{-2t}(1-e^{-t})\frac{t^{i-1}}{(i-1)!}dt + \int_{0}^{\infty} e^{-2t}(1-e^{-t})\frac{t^{i}}{(i-1)!}dt \\ &= \sum_{l=1}^{\infty}\left[\frac{1}{l(l+3)^i} - \frac{1}{l(l+2)^i}\right] + i\left[\frac{1}{2^i} - \frac{1}{3^i}\right]. \end{aligned}$$

The two other expectations $E\left[X_i(1-F(X_i))\right]$ and $E[X_i^2 f(X_i)]$, can be obtained in the same manner using the binomial expansion and writing $\ln(1-e^{-t}) = -\sum_{l=1}^{\infty}\frac{e^{-lt}}{l}$.

Acknowledgements

We would like to thank the editor for his encouragement and the referees for their constructive suggestions and comments that substantially improved the paper.

References

Ahsanullah, M. (1995). *Record Values-Theory and Applications*. New York, University Press of America Inc.

Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (1998). *Records*. New York, John Wiley and Sons.

Asgharzadeh, A. (2006). Point and interval estimation for a generalized logistic distribution under progressive Type II censoring. *Communications in Statistics-Theory and Methods* 35, 1685–1702.

Asgharzadeh, A., Valiollahi, R. and Raqab, M. Z. (2011). Stress-strength reliability of Weibull distribution based on progressively censored samples. *SORT*, 35, 103–124.

Balakrishnan, N. (1992). *Handbook of Logistic Distribution*. New York, Dekker.

Balakrishnan, N. and Aggarwala, R. (2000). *Progressive Censoring: Theory Methods and Applications*. Birkhäuser, Boston.

Balakrishnan, N. and Asgharzadeh, A. (2005). Inference for the scaled half-logistic distribution based on progressively Type II censored samples. *Communications in Statistics-Theory and Methods,* 34, 73–87.

Balakrishnan, N., Chan, P.S. (1998). On the normal record values and associated inference. *Statistics and Probability Letters*, 39, 73–80.

Balakrishnan, N., Kannan, N. (2000). Point and interval estimation for the parameters of the logistic distribution based on progressively Type-2 censored samples. *Handbook of Statistics*, 20, 431–456.

Chandler, K. N. (1952). The distribution and frequency of record values. *Journal of the Royal Statistical Society*, B14, 220–228.

Chen, M. H., Shao, Q. M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals, *Journal of Computational and Graphical Statistics,* 8, 69–92.

Devroye, L. (1984). A simple algorithm for generating random variates with a log-concave density. *Computing*, 33, 247–257.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Re-Sampling Plans. *Philadelphia, PA: SIAM, vol. 38, CBMS-NSF Regional Conference Series in Applied Mathematics.*

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Annals of Statistics,* 16, 927–953.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions, vol. 2*, 2-nd edition. New York, John Wiley and Sons.

Kundu, D. (2007). On hybrid censoring Weibull distribution. *Journal of Statistical Planning and Inference*, 137, 2127–2142.

Kundu, D. (2008). Bayesian inference and life testing plan for the Weibull distribution in presence of progressive censoring. *Technometrics*, 50,144–154.

Madi, M. T. and Raqab, M. Z. (2007). Bayesian prediction of rainfall records using the generalized exponential distribution. *Environmetrics*, 18, 541–549.

Nadarajah, S. (2004). Information matrix for logistic distributions. *Mathematical and Computer Modelling*, 40, 953–958.

Nevzorov, V. (2001). *Records: Mathematical Theory*. In: Translation of Mathematical Monographs, vol. 194. Amer. Math. Soc. Providence, RI, USA.

Raqab, M. Z. (2006). Nonparametric prediction intervals for the future rainfall records. *Environmetrics*, 17, 457–464.

Raqab, M. Z., Asgharzadeh, A. and Valiollahi. R. (2010). Prediction for Pareto distribution based on progressively Type-II censored samples. *Computational Statistics and Data Analysis*, 54, 1732–1743.

Tiku, M. L. and Akkaya A. D. (2004) *Robust Estimation and Hypothesis Testing*. New Delhi, New Age International.

Shao, J. (2003) *Mathematical Statistics*. Second edition, New Work, Springer-Verlag.

Shao, J. (2005) *Mathematical Statistics: Exercises and Solutions*. New York, Springer-Verlag.

SORT 40 (1) January-June 2016, 113-138

A goodness-of-fit test for the multivariate Poisson distribution

F. Novoa-Muñoz[1,*] and M.D. Jiménez-Gamero[2]

Abstract

Bivariate count data arise in several different disciplines and the bivariate Poisson distribution is commonly used to model them. This paper proposes and studies a computationally convenient goodness-of-fit test for this distribution, which is based on an empirical counterpart of a system of equations. The test is consistent against fixed alternatives. The null distribution of the test can be consistently approximated by a parametric bootstrap and by a weighted bootstrap. The goodness of these bootstrap estimators and the power for finite sample sizes are numerically studied. It is shown that the proposed test can be naturally extended to the multivariate Poisson distribution.

MSC: 62F03; 62F05; 62F12; 62F40.

Keywords: Bivariate Poisson distribution, goodness-of-fit, empirical probability generating function, parametric bootstrap, weighted bootstrap, multivariate Poisson distribution.

1. Introduction

Univariate count data appear in many real life situations and the univariate Poisson distribution is frequently used to model this kind of data (see for example Haight, 1967; Johnson and Kotz 1969; Sahai and Khurshid, 1993). Gürtler and Henze (2000) present a wide variety of procedures for testing goodness-of-fit (gof) for the univariate Poisson distribution.

In practice, bivariate count data appear in different areas of knowledge and the bivariate Poisson distribution (BPD), being a generalization of the Poisson distribution, plays a key role in modelling them, provided that such data present a positive correlation.

* Corresponding author: fnovoa@ubiobio.cl

[1] Departamento de Estadística, Universidad del Bío-Bío, Chile.

[2] Departamento de Estadística e I.O., Universidad de Sevilla, Spain.

Received: January 2015
Accepted: April 2016

Different authors have given a definition for the BPD (see for example Kocherlakota and Kocherlakota, 1992). In this article we will work with the one that has received more attention (see for example Holgate, 1964; Johnson, Kotz and Balakrishnan, 1997). Let

$$X_1 = Y_1 + Y_3 \quad \text{and} \quad X_2 = Y_2 + Y_3,$$

where Y_1, Y_2 and Y_3 are mutually independent Poisson random variables with means $\theta_1' = \theta_1 - \theta_3 > 0$, $\theta_2' = \theta_2 - \theta_3 > 0$ and $\theta_3 > 0$, respectively. The joint distribution of the vector (X_1, X_2) is called BPD with parameter $\theta = (\theta_1, \theta_2, \theta_3)$, $(X_1, X_2) \sim BP(\theta)$ for short. In the statistical literature on gof tests for the BPD, which is not so rich as in the univariate case, we found the following: the tests given by Crockett (1979), Loukas and Kemp (1986), Rayner and Best (1995) – these three tests are not consistent against all fixed alternatives – and, more recently, the tests in Novoa-Muñoz and Jiménez-Gamero (2014) (hereafter abbreviated to NJ).

The two tests in NJ are consistent against all fixed alternatives. The results in Janssen (2000) assert that the global power function of any nonparametric test is flat on balls of alternatives except for alternatives coming from a finite dimensional subspace. Because of this reason, it is interesting to propose new gof tests able to detect different sets of alternatives.

This paper presents a consistent gof test for the BPD. It is based on the following: since the probability generating function (pgf) of the BPD is the unique pgf satisfying certain system of partial differential equations, and the empirical probability generating function (epgf) consistently estimates the pgf, the epgf should approximately satisfy such system. The proposed test statistic is a function of the coefficients of the polynomials of an empirical version of that system. The asymptotic behaviour of the proposed test under alternatives is shared with the ones in NJ. An advantage of the test proposed in this paper over those in NJ is that its application does not entail the choice of a weight function, which is rather arbitrary.

The null distribution of the test statistic can be consistently approximated by a parametric bootstrap as well as by means of a weighted bootstrap. The finite sample performance of the proposed test is investigated by means of a simulation study, where the goodness of the proposed approximations is numerically studied and the test is compared, in terms of power, to the tests cited above. The numerical power study reveals that, as expected from the results in Janssen (2000), there is no test yielding the highest power against all considered alternatives. In most cases, the power of the proposed test is quite close to the highest one; in other cases, the proposed test is the most powerful. In addition, from a computational point of view, the test proposed in this paper is more efficient than its competitors.

The work is organized as follows. Section 2 introduces the test statistic and derives its asymptotic null distribution. Since the asymptotic null distribution does not provide a useful means of approximating the null distribution of the test statistic, Section 3 stud-

ies two bootstrap estimators. Specifically, it is shown that the parametric bootstrap and a conveniently defined weighted bootstrap estimators produce consistent null distribution estimators. This Section also studies the power of the resulting tests against fixed alternatives. Section 4 deals with the practical implementation of the bootstrap null distribution estimators as well as other related issues. Section 5 reports a summary of the results of a simulation study carried out to examine the finite sample performance of the tests and to compare them with the existing ones. All stated results are valid for $\theta_3 > 0$. Section 6 deals with the case $\theta_3 = 0$. Section 7 shows how the proposed technique can be applied to the general multivariate case. All proofs are relegated to the last section.

Hereinafter we shall use the following notation: all vectors are row vectors and $v^{\intercal}$ is the transposed of the row vector v; for any vector v, v_k denotes its kth coordinate, and $\|v\|$ its Euclidean norm; $\mathbb{N}_0 = \{0,1,2,3,\ldots\}$; $I\{A\}$ denotes the indicator function of the set A; P_θ denotes the probability law of the BPD with parameter θ; P denotes the probability law of the data; E_θ denotes expectation with respect to the probability function P_θ; E denotes expectation with respect to the true probability function of the data; P_* denote the probability law, given the data; all limits in this work are taken as $n \to \infty$; $\xrightarrow{L}$ denotes convergence in distribution; $\xrightarrow{P}$ denotes convergence in probability; $\xrightarrow{a.s.}$ denotes almost sure (a.s.) convergence; for any function $h : S \subset \mathbb{R}^m \to \mathbb{R}$, for some fixed $m \in \mathbb{N}$, we will denote

$$D^{a_1\cdots a_m} h(u) = \frac{\partial^k}{\partial u_1^{a_1} \cdots \partial u_m^{a_m}} h(u),$$

$\forall a_1, \ldots, a_m \in \mathbb{N}_0$ such that $k = a_1 + \cdots + a_m$.

2. The test statistic and its asymptotic null distribution

Let $\boldsymbol{X}_1 = (X_{11}, X_{12}), \boldsymbol{X}_2 = (X_{21}, X_{22}), \ldots, \boldsymbol{X}_n = (X_{n1}, X_{n2})$ be independent identically distributed (iid) from a random vector $\boldsymbol{X} = (X_1, X_2) \in \mathbb{N}_0^2$. Based on the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$, the objective is to test the hypothesis

$$H_0 : \ (X_1, X_2) \sim BP(\theta_1, \theta_2, \theta_3), \ \text{for some } (\theta_1, \theta_2, \theta_3) \in \Theta,$$

against the alternative

$$H_1 : \ (X_1, X_2) \nsim BP(\theta_1, \theta_2, \theta_3), \ \forall\, (\theta_1, \theta_2, \theta_3) \in \Theta,$$

where $\Theta = \{(\theta_1,\theta_2,\theta_3) \in \mathbb{R}^3 \,:\, \theta_1 > \theta_3,\ \theta_2 > \theta_3,\ \theta_3 > 0\}$. Since the distribution of a random vector $\boldsymbol{X} = (X_1,X_2) \in \mathbb{N}_0^2$ is determined by its pgf $g(u) = E\left(u_1^{X_1}u_2^{X_2}\right)$, $u = (u_1,u_2) \in [0,1]^2$, and the joint pgf of a random vector $\boldsymbol{X} \sim BP(\theta)$ is

$$g(u;\theta) = E_\theta(u_1^{X_1}u_2^{X_2}) = \exp\{\theta_1(u_1-1)+\theta_2(u_2-1)+\theta_3(u_1-1)(u_2-1)\}, \quad (1)$$

testing H_0 vs H_1 is equivalent to testing

$$H_0:\ g(u) = g(u;\theta),\ \forall u \in [0,1]^2,\ \text{for some } (\theta_1,\theta_2,\theta_3) \in \Theta,$$

versus

$$H_1:\ g(u) \neq g(u;\theta),\ \text{for some } u \in [0,1]^2,\ \forall (\theta_1,\theta_2,\theta_3) \in \Theta.$$

Proposition 2 in NJ shows that $g(u_1,u_2;\theta)$ is the only pgf in $G_2 = \{g : [0,1]^2 \to \mathbb{R}$, such that g is a pgf and $\frac{\partial}{\partial u_1}g(u_1,u_2)$ and $\frac{\partial}{\partial u_2}g(u_1,u_2)$ exist $\forall (u_1,u_2) \in [0,1]^2\}$ satisfying the following system,

$$D_i(u;\theta) = 0, \quad i = 1,2, \quad \forall u \in [0,1]^2,$$

where

$$\begin{aligned} D_1(u;\theta) &= \frac{\partial}{\partial u_1}g(u_1,u_2) - \{\theta_1+\theta_3(u_2-1)\}\,g(u_1,u_2), \\ D_2(u;\theta) &= \frac{\partial}{\partial u_2}g(u_1,u_2) - \{\theta_2+\theta_3(u_1-1)\}\,g(u_1,u_2). \end{aligned}$$

Now we consider the following empirical versions of the functions $D_i(u;\theta)$, $i = 1,2$,

$$\begin{aligned} D_{1n}(u;\hat\theta) &= \frac{\partial}{\partial u_1}g_n(u_1,u_2) - \{\hat\theta_1+\hat\theta_3(u_2-1)\}\,g_n(u_1,u_2), \\ D_{2n}(u;\hat\theta) &= \frac{\partial}{\partial u_2}g_n(u_1,u_2) - \{\hat\theta_2+\hat\theta_3(u_1-1)\}\,g_n(u_1,u_2), \end{aligned}$$

where $\hat\theta = (\hat\theta_1,\hat\theta_2,\hat\theta_3)$ is a consistent estimator of θ and $g_n(u_1,u_2)$ is the epgf associated to the data,

$$g_n(u_1,u_2) = \frac{1}{n}\sum_{i=1}^{n} u_1^{X_{i1}}u_2^{X_{i2}}.$$

Proposition 1 in NJ shows that $g(u)$ and its derivatives can be consistently estimated by the epgf and the derivatives of the epgf, respectively. Thus, if H_0 is true then $D_{1n}(u;\hat{\theta})$ and $D_{2n}(u;\hat{\theta})$ should be close to 0, $\forall u \in [0,1]^2$. This proximity to 0 can be interpreted in several ways. For example, NJ interpreted this proximity as

$$S_{n,w}(\hat{\theta}) = n \int \{D_{1n}(u;\hat{\theta})^2 + D_{2n}(u;\hat{\theta})^2\} w(u) du \approx 0, \tag{2}$$

where $w(u)$ is a non-negative function on $[0,1]^2$.

Here we present another interpretation, reasoning as in Nakamura and Pérez-Abreu (1993) for the univariate case. With this aim, observe that

$$D_{in}(u;\hat{\theta}) = \sum_{r\geq 0} \sum_{s\geq 0} d_i(r,s;\hat{\theta}) u_1^r u_2^s, \ \ i = 1,2, \tag{3}$$

where

$$d_1(r,s;\hat{\theta}) = (r+1)p_n(r+1,s) - (\hat{\theta}_1 - \hat{\theta}_3)p_n(r,s) - \hat{\theta}_3 p_n(r,s-1),$$
$$d_2(r,s;\hat{\theta}) = (s+1)p_n(r,s+1) - (\hat{\theta}_2 - \hat{\theta}_3)p_n(r,s) - \hat{\theta}_3 p_n(r-1,s),$$

and

$$p_n(r,s) = \frac{1}{n} \sum_{k=1}^{n} I(X_{k1} = r, X_{k2} = s)$$

is the relative frequency of the pair (r,s). Thus, $D_{in}(u;\hat{\theta}) = 0$, $\forall u \in [0,1]^2$, $i = 1,2$, if and only if the coefficient of $u_1^r u_2^s$ in the right hand side of (3) is null, $\forall r,s \geq 0$, $i = 1,2$. This leads us to consider the following statistic for testing H_0,

$$W_n(\hat{\theta}) = \sum_{r\geq 0} \sum_{s\geq 0} \{d_1(r,s;\hat{\theta})^2 + d_2(r,s;\hat{\theta})^2\} = \sum_{r,s=0}^{M} \{d_1(r,s;\hat{\theta})^2 + d_2(r,s;\hat{\theta})^2\}, \tag{4}$$

where $M = \max\{X_{(n)1}, X_{(n)2}\}$, $X_{(n)k} = \max_{1\leq i\leq n} X_{ik}$, $k = 1,2$.

Taking into account that

$$d_k(r,s;\hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \phi_{krs}(\boldsymbol{X}_i;\hat{\theta}), \quad k = 1,2,$$

with

$$\phi_{1rs}(x;\theta) = (r+1)I(x_1 = r+1, x_2 = s) - (\theta_1 - \theta_3)I(x_1 = r, x_2 = s) - \theta_3 I(x_1 = r, x_2 = s-1),$$
$$\phi_{2rs}(x;\theta) = (s+1)I(x_1 = r, x_2 = s+1) - (\theta_2 - \theta_3)I(x_1 = r, x_2 = s) - \theta_3 I(x_1 = r-1, x_2 = s),$$

where $x = (x_1, x_2)$, the statistic $W_n(\hat{\theta})$ can be expressed as follows,

$$W_n(\hat{\theta}) = \frac{1}{n^2} \sum_{i,j=1}^{n} h(\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\theta}),$$

with

$$\begin{aligned}
h(x,y;\theta) &= h_1(x,y;\theta) + h_2(x,y;\theta),\\
h_1(x,y;\theta) &= \textstyle\sum_{r\geq 0}\sum_{s\geq 0} \phi_{1rs}(x;\theta)\phi_{1rs}(y;\theta)\\
&= \{x_1^2 + (\theta_1-\theta_3)^2 + \theta_3^2\} I(x_1 = y_1, x_2 = y_2) - (\theta_1-\theta_3)x_1 I(x_1 = y_1+1, x_2 = y_2)\\
&\quad -\theta_3 x_1 I(x_1 = y_1+1, x_2 = y_2+1) + (\theta_1-\theta_3)\theta_3 I(x_1 = y_1, x_2 = y_2+1)\\
&\quad -(\theta_1-\theta_3)y_1 I(y_1 = x_1+1, y_2 = x_2) - \theta_3 y_1 I(y_1 = x_1+1, y_2 = x_2+1)\\
&\quad +(\theta_1-\theta_3)\theta_3 I(y_1 = x_1, y_2 = x_2+1),\\
h_2(x,y;\theta) &= \textstyle\sum_{r\geq 0}\sum_{s\geq 0} \phi_{2rs}(x;\theta)\phi_{2rs}(y;\theta)\\
&= \{x_2^2 + (\theta_2-\theta_3)^2 + \theta_3^2\} I(x_1 = y_1, x_2 = y_2) - (\theta_2-\theta_3)x_2 I(x_1 = y_1, x_2 = y_2+1)\\
&\quad -\theta_3 x_2 I(x_1 = y_1+1, x_2 = y_2+1) + (\theta_2-\theta_3)\theta_3 I(x_1 = y_1+1, x_2 = y_2)\\
&\quad -(\theta_2-\theta_3)y_2 I(y_1 = x_1, y_2 = x_2+1) - \theta_3 y_2 I(y_1 = x_1+1, y_2 = x_2+1)\\
&\quad +(\theta_2-\theta_3)\theta_3 I(y_1 = x_1+1, y_2 = x_2),
\end{aligned}$$

where $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

In order to give a sound justification of $W_n(\hat{\theta})$ as a test statistic for testing H_0 we next derive its a.s. limit.

Theorem 1 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid from $\boldsymbol{X} = (X_1, X_2) \in \mathbb{N}_0^2$ with $E(X_k^2) < \infty$, $k = 1, 2$. Let $p(r,s) = P(X_1 = r, X_2 = s)$. If $\hat{\theta} \xrightarrow{a.s.} \theta$, for some $\theta \in \mathbb{R}^3$, then*

$$W_n(\hat{\theta}) \xrightarrow{a.s.} \sum_{r,s\geq 0} \left\{ a_1(r,s;\theta)^2 + a_2(r,s;\theta)^2 \right\} = \eta(P;\theta),$$

where

$$a_1(r,s;\theta) = (r+1)p(r+1,s) - (\theta_1-\theta_3)p(r,s) - \theta_3 p(r,s-1),$$
$$a_2(r,s;\theta) = (s+1)p(r,s+1) - (\theta_2-\theta_3)p(r,s) - \theta_3 p(r-1,s).$$

Note that $\eta(P;\theta) \geq 0$ and, taking into account that

$$D_k(u;\theta) = \sum_{r\geq 0}\sum_{s\geq 0} a_k(r,s;\theta)u_1^r u_2^s, \quad k=1,2,$$

it follows that $\eta(P;\theta)=0$ if and only if H_0 is true. Thus, a reasonable test for testing H_0 should reject the null hypothesis for large values of $W_n(\hat{\theta})$. Now, to determine what are large values we must calculate its null distribution, or at least an approximation to it.

We first try to estimate the null distribution of $W_n(\hat{\theta})$ by means of its asymptotic null distribution. In order to derive it, it will be assumed that the estimator $\hat{\theta}$ is asymptotically linear, as expressed in the next assumption.

Assumption 1 *Under H_0, if $\theta=(\theta_1,\theta_2,\theta_3)\in\Theta$ denotes the true parameter value, then*

$$\sqrt{n}\left(\hat{\theta}-\theta\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell\left(\boldsymbol{X}_i;\theta\right)+\mathbf{o}_P(1),$$

where $\ell:\mathbb{N}_0^2\times\Theta\longrightarrow\mathbb{R}^3$ is such that $E_\theta\left\{\ell\left(\boldsymbol{X}_1;\theta\right)\right\}=\boldsymbol{0}$ and $J(\theta)=E_\theta\left\{\ell\left(\boldsymbol{X}_1;\theta\right)^{\intercal}\ell\left(\boldsymbol{X}_1;\theta\right)\right\}<\infty$.

Assumption 1 is not restrictive at all since it is fulfilled by some commonly used estimators such as the moment estimator, the maximum likelihood estimator, the double zero estimator, the even points estimator and the conditional even points estimator (see Kocherlakota and Kocherlakota, 1992, and Papageorgiou and Loukas, 1988).

The next result gives the asymptotic null distribution of $W_n(\hat{\theta})$.

Theorem 2 *Let $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ be iid from $\boldsymbol{X}=(X_1,X_2)\sim BP(\theta_1,\theta_2,\theta_3)$. Suppose that Assumption 1 holds. Then*

$$nW_n(\hat{\theta}) \xrightarrow{L} \sum_{j\geq 1}\lambda_j\chi_{1j}^2,$$

where $\chi_{11}^2,\chi_{12}^2,\ldots$ are independent χ^2 variates with one degree of freedom and the set $\{\lambda_j\}$ are the non-null eigenvalues of the operator $C(\theta)$ defined on the function space $\{\tau:\mathbb{N}_0^2\to\mathbb{R},$ such that $E_\theta\left[\tau^2(\boldsymbol{X})\right]<\infty,\forall\theta\in\Theta\}$, as follows

$$C(\theta)\tau(x) = E_\theta\{K(x,\boldsymbol{X};\theta)\tau(\boldsymbol{X})\},$$

with $K(x,y;\theta)=h(x,y;\theta)+\ell(x;\theta)\mu(y;\theta)^{\intercal}+\ell(y;\theta)\mu(x;\theta)^{\intercal}+\ell(x;\theta)S(\theta)\ell(y;\theta)^{\intercal}$, $\mu(x;\theta)=(\mu_1(x;\theta),\mu_2(x;\theta),\mu_3(x;\theta))$,

$$\begin{aligned}
\mu_1(x;\theta) &= -x_1 P_\theta(x_1-1,x_2)+\theta_3 P_\theta(x_1,x_2+1)+(\theta_1-\theta_3)P_\theta(x_1,x_2),\\
\mu_2(x;\theta) &= -x_2 P_\theta(x_1,x_2-1)+\theta_3 P_\theta(x_1+1,x_2)+(\theta_2-\theta_3)P_\theta(x_1,x_2),\\
\mu_3(x;\theta) &= -\mu_1(x;\theta)-x_1 P_\theta(x_1-1,x_2-1)+\theta_3 P_\theta(x_1,x_2)+(\theta_1-\theta_3)P_\theta(x_1,x_2-1)\\
&\quad -\mu_2(x;\theta)-x_2 P_\theta(x_1-1,x_2-1)+\theta_3 P_\theta(x_1,x_2)+(\theta_2-\theta_3)P_\theta(x_1-1,x_2),
\end{aligned}$$

$S(\theta)=\sum_{r,s\geq 0} S_{rs}(\theta)$,

$$S_{rs}(\theta)=\begin{pmatrix} a^2 & 0 & a(b-a)\\ 0 & a^2 & a(c-a)\\ a(b-a) & a(c-a) & (b-a)^2+(c-a)^2 \end{pmatrix},$$

$a=P_\theta(r,s)$, $b=P_\theta(r,s-1)$, $c=P_\theta(r-1,s)$,

The asymptotic null distribution of $W_n(\hat{\theta})$ does not provide a useful approximation to its null distribution since it depends on the unknown true value of θ. Even if θ were known or replaced by an appropriate estimator, to determine the eigenvalues of an operator is a rather hard problem.

So, we next study two further ways of approximating it: a parametric bootstrap (PB) estimator and a weighted bootstrap (WB) estimator.

3. Approximating the null distribution

3.1. Parametric bootstrap

Let $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ be iid taking values in $\mathbb{N}_0^2$ such that $\hat{\theta}=\hat{\theta}(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n)\in\Theta$. Let $\boldsymbol{X}_1^*,\boldsymbol{X}_2^*,\ldots,\boldsymbol{X}_n^*$ be iid from a population with distribution $BP\big(\hat{\theta}\big)$, given $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$, and let $W_n^*(\hat{\theta}^*)$ be the bootstrap version of $W_n(\hat{\theta})$ obtained by replacing $\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n$ and $\hat{\theta}=\hat{\theta}\big(\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\big)$ by $\boldsymbol{X}_1^*,\boldsymbol{X}_2^*,\ldots,\boldsymbol{X}_n^*$ and $\hat{\theta}^*=\hat{\theta}\big(\boldsymbol{X}_1^*,\boldsymbol{X}_2^*,\ldots,\boldsymbol{X}_n^*\big)$, respectively, in the expression of $W_n(\hat{\theta})$. To prove that the PB can be used to consistently approximate the null distribution of $W_n(\hat{\theta})$, we will assume the following, which is a bit stronger than Assumption 1.

Assumption 2 *Assumption 1 holds and the functions ℓ and J satisfy*

(1) $\sup_{\vartheta\in\Theta_0} E_\vartheta\left[\|\ell(\boldsymbol{X};\vartheta)\|^2 I\{\|\ell(\boldsymbol{X};\vartheta)\|>\gamma\}\right]\to 0$, *as* $\gamma\to\infty$, *where* $\Theta_0\subseteq\Theta$ *is an open neighborhood of* θ.

(2) $\ell(\boldsymbol{X};\vartheta)$ *and* $J(\vartheta)$ *are continuous as functions of* ϑ *at* $\vartheta=\theta$ *and* $J(\vartheta)$ *is finite* $\forall\vartheta\in\Theta_0$.

Theorem 3 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid from $\boldsymbol{X} = (X_1, X_2) \in \mathbb{N}_0^2$. Suppose that Assumption 2 holds and that $\hat{\theta} \xrightarrow{a.s.} \theta$, for some $\theta \in \Theta$. Then*

$$\sup_{x \in \mathbb{R}} \left| P_* \left\{ nW_n^*(\hat{\theta}^*) \leq x \right\} - P_\theta \left\{ nW_n(\hat{\theta}) \leq x \right\} \right| \xrightarrow{a.s.} 0.$$

Let $w_{n,\alpha}^* = \inf\{x : P_*(W_n^*(\hat{\theta}^*) \geq x) \leq \alpha\}$ be the α upper percentile of the PB distribution of $W_n(\hat{\theta})$ and let W_{obs} be the observed value of the test statistic. From Theorem 3, the test function

$$\Psi_{PB}^* = \begin{cases} 1, & \text{if } W_n(\hat{\theta}) \geq w_{n,\alpha}^*, \\ 0, & \text{otherwise,} \end{cases}$$

or equivalently, the test that rejects H_0 when $p^* = P_* \left(W_n^*(\hat{\theta}^*) \geq W_{obs} \right) \leq \alpha$, is asymptotically correct in the sense that $P_\theta(\Psi_{PB}^* = 1) \to \alpha$.

3.2. Weighted bootstrap

From the proof of Theorem 2, when H_0 is true, we have that $nW_n(\hat{\theta}) = nW_{1n}(\theta) + o_P(1)$, where

$$nW_{1n}(\theta) = \frac{1}{n} \sum_{i,j=1}^{n} K(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta),$$

which converges in law to $W_0 = \sum_{j \geq 1} \lambda_j \chi_{1j}^2$. As observed before, the greatest difficulty with W_0 is to determine the set $\{\lambda_j\}$. Nevertheless, Delhing and Mikosch (1994) have shown that the eigenvalues $\{\lambda_j\}$ can be consistently (a.s.) approximated by the eigenvalues of the matrix

$$H_n = \left(\frac{1}{n} K(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta) \right)_{1 \leq i,j \leq n},$$

say $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$. Therefore, we could approximate the null distribution of $nW_{1n}(\hat{\theta})$ (and thus that of $nW_n(\hat{\theta})$) through the conditional distribution, given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, of

$$nW_{1n}^* = \sum_{j=1}^{n} \hat{\lambda}_j \chi_{1j}^2.$$

This is tantamount to approximate the null distribution of $nW_{1n}(\hat{\theta})$ by means of the conditional distribution, given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, of

$$W_1^* = \frac{1}{n}\sum_{i,j=1}^{n} K(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta)\xi_i\xi_j,$$

where $\xi_1, \ldots, \xi_n$ are iid from a standard normal distribution, $N(0,1)$, independent of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, that is, by means of the WB distribution of $nW_{1n}(\hat{\theta})$, in the sense of Burke (2000). The main problem with this approach is that $K(x,y;\theta)$ is unknown because it depends on θ, which is unknown, and because it also depends on $\ell(x;\theta)$, which is usually unknown. To overcome this problem we replace θ by $\hat{\theta}$ and $\ell(x;\theta)$ by $\hat{\ell}(x;\hat{\theta})$ which is assumed to satisfy

$$\begin{aligned} &\frac{1}{n}\sum_{j=1}^{n} \|\ell_1(\boldsymbol{X}_j;\theta) - \hat{\ell}(\boldsymbol{X}_j;\hat{\theta})\|^2 \xrightarrow{P} 0, \\ &\text{with } E\{\|\ell_1(\boldsymbol{X};\theta)\|^2\} < \infty \text{ and } \ell_1(x;\theta) = \ell(x;\theta) \text{ if } H_0 \text{ is true.} \end{aligned} \tag{5}$$

So, instead of $nW_{1n}^*(\hat{\theta})$ we consider

$$nW_{2n}^*(\hat{\theta}) = \sum_{j=1}^{n} \tilde{\lambda}_j \chi_{1j}^2,$$

where $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$ are the eigenvalues of the matrix

$$\hat{H}_n = \left(\frac{1}{n}\hat{K}(\boldsymbol{X}_i, \boldsymbol{X}_j; \theta)\right)_{1\le i,j\le n},$$

with $\hat{K}(x,y;\theta) = h(x,y;\theta) + \hat{\ell}(x;\theta)\mu(y;\theta)^\intercal + \hat{\ell}(y;\theta)\mu(x;\theta)^\intercal + \hat{\ell}(x;\theta)S(\theta)\hat{\ell}(y;\theta)^\intercal$. The next theorem gives the limit of the conditional distribution of $nW_{2n}^*(\hat{\theta})$, given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

Theorem 4 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid from $\boldsymbol{X} = (X_1, X_2) \in \mathbb{N}_0^2$ with $E(X_k^2) < \infty$, $k = 1, 2$. Suppose that $\hat{\theta} \xrightarrow{P} \theta$, for some $\theta \in \Theta$ and that (5) holds. Then,*

$$\sup_x \left| P_*\left\{nW_{2n}^*(\hat{\theta}) \le x\right\} - P\{W_1 \le x\}\right| \xrightarrow{P} 0, \tag{6}$$

where $W_1 = \sum_{j\ge 1} \lambda_{1j}\chi_{1j}^2$, $\{\lambda_{1j}\}$ are the non-null eigenvalues of the operator $C_1(\theta)$ defined on the function space $\{\tau : \mathbb{N}_0^2 \to \mathbb{R}$, such that $E\left[\tau^2(\boldsymbol{X})\right] < \infty\}$, as follows

$$C_1(\theta)\tau(x) = E\{K_1(x, \boldsymbol{X}; \theta)\tau(\boldsymbol{X})\},$$

with $K_1(x,y;\theta) = h(x,y;\theta) + \ell_1(x;\theta)\mu(y;\theta)^\intercal + \ell_1(y;\theta)\mu(x;\theta)^\intercal + \ell_1(x;\theta)S(\theta)\ell_1(y;\theta)^\intercal$.

Remark 1 If in addition to the assumptions in Theorem 4 we assume that $\hat{\theta} \xrightarrow{a.s.} \theta$ and that the limit in (5) is a.s., then the convergence in (6) is a.s.

Remark 2 The result in Theorem 4 keeps on being true if instead of using the raw multipliers, $\xi_1, \ldots, \xi_n$, we use the centered multipliers, $\xi_1 - \bar{\xi}, \ldots, \xi_n - \bar{\xi}$, as suggested in Burke (2000) and Kojadinovic and Yan (2012).

Let $w^*_{2,n,\alpha} = \inf\{x : P_*(W^*_{2n}(\hat{\theta}) \geq x) \leq \alpha\}$ be the α upper percentile of the WB distribution of $W_n(\hat{\theta})$. From Theorems 2 and 4, the test function

$$\Psi^*_{WB} = \begin{cases} 1, & \text{if } W_n(\hat{\theta}) \geq w^*_{2,n,\alpha}\,, \\ 0, & \text{otherwise,} \end{cases}$$

or equivalently, the test that rejects H_0 when $p^* = P_*\left(W^*_{2n}(\hat{\theta}) \geq W_{obs}\right) \leq \alpha$, is asymptotically correct.

3.3. Behaviour against alternatives

This subsections shows that, in contrast to the tests given by Crockett (1979), Loukas and Kemp (1986) and Rayner and Best (1995), the tests Ψ^*_{PB} and Ψ^*_{WB} are consistent, that is, they are able to detect any fixed alternative.

As an immediate consequence of Theorems 1 and 3 (Theorems 1 and 4) the next result gives the asymptotic power of the test Ψ^*_{PB} (Ψ^*_{WB}) against fixed alternatives.

Corollary 1 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid from $\boldsymbol{X} \in \mathbb{N}_0^2$ with pgf $g(u)$. Suppose that assumptions in Theorems 1 and 3 hold. If $\eta(P;\theta) > 0$, then $P(\Psi^*_{PB} = 1) \to 1$.*

Corollary 2 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be iid from $\boldsymbol{X} \in \mathbb{N}_0^2$ with pgf $g(u)$. Suppose that assumptions in Theorems 1 and 4 hold. If $\eta(P;\theta) > 0$, then $P(\Psi^*_{WB} = 1) \to 1$.*

It can be shown that the proposed tests are also able to detect local alternatives converging to the null at the rate $n^{-1/2}$. The statement and the proof of this result are quite similar to those of Theorem 4 in NJ, for the PB, and of Theorem 4 in Jiménez-Gamero and Kim (2015), for the WB. So, in order to save space, we omit it.

Although the tests Ψ^*_{PB} and Ψ^*_{WB} both asymptotically correct and consistent, their power for finite sample sizes differ. This point will be numerically studied by simulation in Section 5.

4. Some practical considerations

4.1. Bootstrap algorithms

In practice, the exact bootstrap estimator of the null distribution of $W_n(\hat{\theta})$ cannot be calculated. As usual, we approximate it by simulation as follows:

PB algorithm

1. Estimate θ through $\hat{\theta}$ and compute the observed value of the test statistic W_{obs}.
2. For some large integer B, repeat for every $b \in \{1,\ldots,B\}$:
   (a) Generate $\boldsymbol{X}^{*b} = (\boldsymbol{X}_1^{*b}, \boldsymbol{X}_2^{*b}, \ldots, \boldsymbol{X}_n^{*b})$, where $\boldsymbol{X}_1^{*b}, \boldsymbol{X}_2^{*b}, \ldots, \boldsymbol{X}_n^{*b}$ are iid from a $BP(\hat{\theta})$.
   (b) Calculate the test statistic evaluated at $\boldsymbol{X}^{*b}$, obtaining $W_n^{*b}(\hat{\theta}^{*b})$.
3. Approximate the p-value by $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} I\{W_n^{*b}(\hat{\theta}^{*b}) > W_{obs}\}$.

In contrast to the PB distribution, the exact WB estimator of the null distribution of $W_n(\hat{\theta})$ can be calculated by using some numerical approximation method, as for example Imhof's (1961) method. Thus, to calculate the WB distribution of $W_n(\hat{\theta})$ we can proceed as follows:

WB algorithm 1

1. Estimate θ through $\hat{\theta}$ and compute the observed value of the test statistic W_{obs}.
2. Calculate $m_{ij} = \hat{K}(\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\theta})$, $1 \le i \le j \le n$. Note that $m_{ji} = m_{ij}$.
3. Calculate the eigenvalues of $\hat{H}_n$, $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$.
4. Approximate the p-value by $\hat{p} = P_*\left(\sum_{j=1}^{n} \tilde{\lambda}_j \chi_{1j}^2 > W_{obs}\right)$.

The WB estimator can be also approximated by simulation as follows:

WB algorithm 2

1. Estimate θ through $\hat{\theta}$ and compute the observed value of the test statistic W_{obs}.
2. Calculate $m_{ij} = \hat{K}(\boldsymbol{X}_i, \boldsymbol{X}_j; \hat{\theta})$, $1 \le i \le j \le n$. Note that $m_{ji} = m_{ij}$.
3. For some large integer B, repeat for every $b \in \{1,\ldots,B\}$:
   (a) Generate n iid $N(0,1)$ variates $\xi_1, \ldots, \xi_n$.
   (b) Calculate $W_{2n}^{*b}(\hat{\theta}) = \frac{1}{n^2}\sum_{i,j} \xi_i \xi_j m_{ij}$ (or $W_{2n}^{*b}(\hat{\theta}) = \frac{1}{n^2}\sum_{i,j} (\xi_i - \bar{\xi})(\xi_j - \bar{\xi}) m_{ij}$, as observed in Remark 2).
4. Approximate the p-value by $\hat{p} = \frac{1}{B}\sum_{b=1}^{B} I\{W_{2n}^{*b}(\hat{\theta}) > W_{obs}\}$.

4.2. Point estimators

All above theory assumes that the considered estimator $\hat{\theta}$ satisfies Assumption 1. Commonly used estimators such as maximum likelihood estimators (MLE) and method of moment estimators (MME) satisfy it. Lemmas 1 and 3 in Jiménez-Gamero and Kim (2015) show that the functions ℓ associated to MLEs and MMEs can be approximated by $\hat{\ell}$ satisfying (5), and give the expressions of such approximations. Specifically, if θ is estimated by means of its MLE, then a choice for $\hat{\ell} = \hat{\ell}_{ML}$ satisfying (5) is

$$\hat{\ell}_{ML}((x_1,x_2);\theta) = \left(x_1-\theta_1, x_2-\theta_2, \theta_3\left(\frac{P_\theta(x_1-1,x_2)}{P_\theta(x_1,x_2)} + \frac{P_\theta(x_1,x_2-1)}{P_\theta(x_1,x_2)} - 2\right)\right.$$
$$\left. + f(\theta)\left(\frac{P_\theta(x_1-1,x_2-1)}{P_\theta(x_1,x_2)} - \frac{P_\theta(x_1-1,x_2)}{P_\theta(x_1,x_2)} - \frac{P_\theta(x_1,x_2-1)}{P_\theta(x_1,x_2)} + 1\right)\right),$$

where

$$f(\theta) = \frac{\theta_3^2(\theta_1+\theta_2-2\theta_3)(Q-1) - \theta_3^2 + (\theta_1-2\theta_3)(\theta_2-2\theta_3)}{(\theta_1\theta_2-\theta_3^2)(Q-1) - \theta_1 - \theta_2 + 2\theta_3},$$
$$Q = \sum_{i,j\in\mathbb{N}_0} \frac{P_\theta(i-1,j-1)^2}{P_\theta(i,j)}.$$

If θ is estimated by means of its MME, then a choice for $\hat{\ell} = \hat{\ell}_{MM}$ satisfying (5) is

$$\hat{\ell}_{MM}((x_1,x_2);\theta) = (x_1-\theta_1, x_2-\theta_2, -\theta_2(x_1-\theta_1) - \theta_1(x_2-\theta_2) + x_1x_2 - \theta_3 - \theta_1\theta_2).$$

5. Finite sample performance

The properties so far studied are asymptotic. To study the finite sample performance of the proposed tests, we conducted a simulation experiment. In this section we briefly describe it and display a summary of the results obtained. All computations in this paper were performed by using programs written in the R language (R Development Core Team, 2015).

We started by comparing the proposed approximations to the null distribution of the test statistic $W_n(\hat{\theta})$ from the point of view of the required time to get a p-value. Several values of θ_1, θ_2 and θ_3 were considered. We observed that the value of θ_3 has almost no influence in the required computation time. In contrast, the values of θ_1 and θ_2 have a high impact. We also tried two methods to estimate the parameters: maximum likelihood (ML) and the method of moments (MM), and observed that the choice of the method has little mark on the consumed time. The method used to estimate the null distribution has a high repercussion on the consumed time. In order to value some of these facts,

Table 1: *CPU time (in seconds) to get a p-value with* $B = 1000$.

ML		$n = 50$			$n = 100$			$n = 200$		
θ_1	θ_2	PB	WB2	WB1	PB	WB2	WB1	PB	WB2	WB1
1	1	8.37	0.11	0.07	9.31	0.25	0.15	10.25	0.72	0.46
1	3	13.15	0.14	0.09	17.20	0.27	0.18	19.72	0.74	0.49
3	3	23.57	0.14	0.10	30.12	0.27	0.17	41.15	0.74	0.46
3	10	57.27	0.22	0.22	60.90	0.36	0.29	87.06	0.82	0.59
10	10	132.32	0.28	0.22	187.57	0.36	0.30	277.43	0.86	0.60
10	50	188.64	2.08	2.17	317.47	2.34	2.42	449.14	3.29	2.89
50	50	621.02	2.29	2.40	1160.52	2.48	2.43	2340.78	3.45	3.03
MM		$n = 50$			$n = 100$			$n = 200$		
θ_1	θ_2	PB	WB2	WB1	PB	WB2	WB1	PB	WB2	WB1
1	1	7.69	0.11	0.07	9.82	0.25	0.15	10.53	0.72	0.45
1	3	11.62	0.13	0.10	14.89	0.26	0.17	21.18	0.73	0.47
3	3	25.73	0.14	0.10	31.81	0.28	0.17	43.13	0.74	0.46
3	10	69.31	0.22	0.20	60.15	0.36	0.28	79.80	0.81	0.57
10	10	88.90	0.27	0.22	195.39	0.38	0.31	278.02	0.85	0.58
10	50	174.27	2.07	2.17	280.04	2.31	2.43	462.41	3.26	2.91
50	50	717.87	2.28	2.24	1172.18	2.48	2.38	2402.07	3.43	2.89

Table 1 displays the CPU consumed time (in seconds) to get a p-value for several values of θ_1 and θ_2. The value of θ_3 was set so that the correlation coefficient between the variables, $\rho = \theta_3/\sqrt{\theta_1\theta_2}$, is equal to 0.5. To calculate the PB approximation and the approximation in WB algorithm 2 we took $B = 1000$. There is almost no difference in using the raw multipliers and the centered multipliers in WB algorithm 2. To calculate the p-value of the approximation in WB algorithm 1 we used the function `imhof` of the package `CompQuadForm` of the R language (Duchesne and Lafaye De Micheaux, 2010). From the results in this table it becomes evident that the PB is much more time consuming than the WB, specially for large values of θ_1, θ_2 and the sample size. There are small differences between WB algorithm 1 and WB algorithm 2.

We then studied the goodness of the proposed bootstrap approximations to the null distribution of the test statistic for finite sample sizes. With this aim, we generated 1000 samples of size $n = 50, 100, 200, 300$ from a $BP(\theta_1, \theta_2, \theta_3)$, for several values of θ_1 and θ_2, with θ_3 such that ρ equals to 0.25 and 0.75, in order to examine the approximations for low and high correlated data, respectively, when $\theta_1 = \theta_2$, and $\rho = 0.25$ for $\theta_1 \neq \theta_2$ ($\rho = 0.75$ was not considered because it gives values of θ_3 out of the parametric space for the tried values of $\theta_1 \neq \theta_2$). Because of the results in Table 1, for $\theta_1 = \theta_2 = 50$, the PB was only tried for $n = 50, 100$. For $\theta_1 = \theta_2 = 50$ the WB was also tried for greater sample sizes. For each sample, the p-values were calculated with $B = 500$. The p-values obtained with the WB approximation calculated by means of simulation (WB algorithm 2 with raw and centered multipliers) and numerical approximation (WB

Table 2: *Simulation results for the type I error probability for nominal levels $\alpha = 5\%$ and 10%.*

		Ψ^*_{PB}				Ψ^*_{WB}			
		ML		MM		ML		MM	
	n	5%	10%	5%	10%	5%	10%	5%	10%
$\theta_1 = 1$	50	3.1	7.5	3.5	7.6	3.3	7.8	3.0	7.8
$\theta_2 = 1$	100	3.9	8.6	4.0	9.2	4.2	9.1	3.9	8.8
$\rho = 0.25$	200	4.4	8.4	4.5	8.6	4.8	8.8	4.8	8.6
	300	5.0	10.3	5.5	10.7	4.9	10.3	5.3	10.5
$\theta_1 = 1$	50	3.2	7	3.7	8.5	4.3	8.3	4.6	9.6
$\theta_2 = 1$	100	4.5	8.4	4.1	8.9	4.9	9.1	4.5	9.8
$\rho = 0.75$	200	4.3	8.7	4.7	9.1	4.5	9.2	4.4	9.3
	300	4.7	9.4	4.8	9.6	5.1	10.1	4.4	9.4
$\theta_1 = 1$	50	3.4	7.7	3.6	7.5	2.9	6.6	2.9	6.7
$\theta_2 = 3$	100	3.4	8.3	3.8	8.3	2.8	8.3	2.7	8.3
$\rho = 0.25$	200	4.2	9.1	4.5	9.1	4.4	9.0	4.5	9.1
	300	4.5	9.8	4.6	9.3	4.2	9.7	4.0	9.9
$\theta_1 = 3$	50	3.4	7.7	3.4	7.2	1.6	5.6	1.8	5.3
$\theta_2 = 3$	100	3.3	8.0	3.2	7.7	2.3	6.0	2.3	5.9
$\rho = 0.25$	200	4.4	7.7	4.2	8.0	3.5	7.6	3.2	7.3
	300	4.2	9.6	4.5	9.3	3.7	8.4	3.8	8.6
$\theta_1 = 3$	50	3.7	7.4	4.0	8.8	3.1	7.1	2.9	6.7
$\theta_2 = 3$	100	3.8	8.1	4.0	8.9	3.3	7.6	2.8	7.2
$\rho = 0.75$	200	5.5	10.2	5.5	9.9	5.4	10.5	5.2	10.3
	300	5.3	9.8	5.1	9.7	4.8	10.0	4.4	9.6
$\theta_1 = 3$	50	4.7	9.4	4.3	9.1	2.3	5.8	2.3	5.2
$\theta_2 = 10$	100	5.1	9.1	4.8	8.6	3.0	6.3	3.0	6.4
$\rho = 0.25$	200	5.4	9.9	5.9	9.9	4.3	9.0	4.3	8.6
	300	5.5	10.6	5.6	10.7	4.3	9.7	4.2	10.0
$\theta_1 = 10$	50	3.6	7.3	3.3	6.8	1.0	3.1	1.0	3.0
$\theta_2 = 10$	100	3.9	7.7	4.0	7.3	1.1	3.8	1.1	3.9
$\rho = 0.25$	200	4.4	9.6	4.7	9.2	2.3	6.7	2.3	6.5
	300	4.8	9.5	4.7	8.9	3.2	6.7	3.2	6.9

		Ψ^*_{PB}				Ψ^*_{WB}			
		ML		MM		ML		MM	
	n	5%	10%	5%	10%	5%	10%	5%	10%
$\theta_1 = 10$	50	3.8	7.9	3.9	8.2	1.7	4.2	1.5	3.9
$\theta_2 = 10$	100	3.7	7.9	3.7	8.5	2.0	5.5	1.9	5.3
$\rho = 0.75$	200	4.0	8.3	3.8	8.9	2.6	7.1	2.6	7.0
	300	5.5	9.8	5.3	9.5	4.7	8.7	4.6	8.7
$\theta_1 = 10$	50	3.8	8.4	2.9	7.9	0.0	0.7	0.0	0.6
$\theta_2 = 50$	100	3.9	8.5	3.7	8.0	0.4	2.1	0.4	2.0
$\rho = 0.25$	200	4.2	9.4	3.8	9.0	0.8	3.9	0.9	3.9
	300	5.9	10.0	5.8	10.1	2.3	7.0	2.5	7.0
$\theta_1 = 50$	50	1.1	3.9	1.1	3.6	0.0	0.1	0.0	0.1
$\theta_2 = 50$	100	1.2	4.0	1.2	3.5	0.1	0.2	0.1	0.2
$\rho = 0.25$	200					0.2	1.5	0.4	1.4
	300					0.2	2.2	0.2	2.2
	500					0.5	2.2	0.5	2.2
	1000					2.8	6.2	2.8	6.2
	2000					3.7	6.9	3.8	6.9
	3000					3.7	7.6	3.7	7.7
	4000					4.9	10.0	4.9	10.1
$\theta_1 = 50$	50	3.2	7.0	1.2	3.5	0.1	0.2	0.1	0.2
$\theta_2 = 50$	100	3.8	8.5	3.7	8.3	0.1	0.9	0.1	0.8
$\rho = 0.75$	200					0.0	0.9	0.0	1.0
	300					0.3	1.7	0.3	1.7
	500					1.5	5.5	1.5	5.6
	1000					2.3	7.2	2.3	7.3
	2000					4.0	8.5	3.9	8.3
	3000					5.1	9.1	5.1	9.2
	4000					4.7	9.3	4.6	9.2

algorithm 1 with Imhof's method) were, as expected, quite close. As for raw multipliers versus centered multipliers, a bit better results are obtained when using the centered multipliers. Table 2 displays the fraction of estimated p-values less than or equal to 0.05 and 0.10, which are the estimated type I error probabilities for $\alpha = 0.05$ and 0.10, respectively by using PB and WB with centered multipliers. From the results in this table it can be concluded that both approximations give rise to conservative tests for small sample sizes. As the values of θ_1 and θ_2 increase, the tests become more conservative, specially the one based on the WB approximation. For example, when $\theta_1 = \theta_2 = 50$ and $\rho = 0.25$, the sample size required to get empirical levels close to the nominal values is $n = 4000$. For $\theta_1 = \theta_2 = 50$ and $\rho = 0.75$, $n = 3000$ is enough. In general, better results (in the sense of closeness to the nominal values) are obtained for $\rho = 0.75$ than for $\rho = 0.25$. Finally, it is also observed a bit better results when the parameter is estimated by the maximum likelihood estimator.

To study the power we repeated the above experiment for samples with size $n = 50$ from the following alternatives: bivariate binomial distribution $BB(m; p_1, p_2, p_3)$, where $p_1 + p_2 - p_3 \leq 1$, $p_1 \geq p_3, p_2 \geq p_3$ and $p_3 > 0$; bivariate Hermite distribution $BH(\mu, \sigma^2; \lambda_1, \lambda_2, \lambda_3)$, where $\mu > \sigma^2(\lambda_1 + \lambda_2 + \lambda_3)$; bivariate logarithmic series distribution $BLS(\lambda_1, \lambda_2, \lambda_3)$, where $0 < \lambda_1 + \lambda_2 + \lambda_3 < 1$; bivariate Neyman type A distribution $BNTA(\lambda; \lambda_1, \lambda_2, \lambda_3)$, where $0 < \lambda_1 + \lambda_2 + \lambda_3 \leq 1$; bivariate Poisson distribution mixtures of the form $pBP(\theta) + (1-p)BP(\lambda), 0 < p < 1$, denoted by $BPP(p; \theta, \lambda)$; and (X_1, X_2) with $X_1 = \max\{Y_1, Y_3\}$ and $X_2 = |Y_1 - Y_3|$ (type 1), $X_1 = \max\{Y_2, Y_3\}$ and $X_2 = |Y_2 - Y_3|$ (type 2), $X_1 = \max\{Y_1, Y_3\}$ and $X_2 = \min\{Y_2, Y_3\}$ (type 3), $X_1 = \max\{Y_2, Y_3\}$ and $X_2 = \min\{Y_1, Y_3\}$ (type 4), $X_1 = \max\{Y_1, Y_3\}$ and $X_2 = \max\{Y_2, Y_3\}$ (type 5), where Y_1, Y_2, Y_3 are independent variables taking values in $\mathbb{N}_0$ whose distribution are binomial $B(m; p)$, negative binomial $BN(m; p)$, Poisson $P(\lambda)$ and uniform on $1, 2, \ldots, m$, $U(m)$. The values of the parameters were chosen so that the expectations $E(X_1)$ and $E(X_2)$ are small for the PB and the WB not to be excessively conservative. In this part of the simulation experiment we only considered the maximum likelihood estimator of the parameter.

In addition to the tests proposed in this paper, Ψ^*_{PB} and Ψ^*_{WB}, we also considered the tests given in Crockett (1979) (denoted by T), Loukas and Kemp (1986) (denoted by I_B), Rayner and Best (1995) (denoted by NI_B) and NJ (denoted by R_n and S_n, with weight function $w(u) = 1$). Table 3 displays the alternatives considered and the estimated power for nominal significance level $\alpha = 0.05$. Looking at this table we conclude that the tests Ψ^*_{PB}, Ψ^*_{WB}, R_n and S_n are able to detect all considered alternatives while, as expected, the other tests cannot, specially the tests based on I_B and NI_B. For the alternatives in the first half of Table 3 we see that the powers of the new tests, R_n and S_n are quite close; while for the other alternatives the tests proposed in this paper are more powerful than R_n and S_n. We also compared these tests from a computational point of view. From the results in Table 1 we saw that, in this respect, Ψ^*_{WB} is more efficient than Ψ^*_{PB}. Since R_n and S_n are both based on a PB, for the comparisons to be fair, we compared Ψ^*_{PB}, R_n and S_n. Table 4 reports the ratio of the average CPU to get a p-value. Clearly, regarding the required computing time, Ψ^*_{PB} is more efficient than R_n and S_n.

Table 3: *Simulation results for the power.*

Alternative	$E(X_1)$	$\frac{var(X_1)}{E(X_1)}$	$E(X_2)$	$\frac{var(X_2)}{E(X_2)}$	ρ	R_n	S_n	Ψ^*_{PB}	Ψ^*_{WB}	T	I_B	NI_B
$BB(1, 0.45, 0.02, 0.01)$	0.450	0.550	0.020	0.980	0.014	0.953	0.961	0.944	1.000	0.263	0.000	0.000
$BB(1, 0.55, 0.03, 0.02)$	0.550	0.450	0.030	0.970	0.041	0.998	0.999	0.998	1.000	0.768	0.000	0.000
$BB(2, 0.71, 0.04, 0.03)$	1.420	0.290	0.080	0.960	0.018	0.997	0.993	1.000	1.000	0.999	0.000	0.000
$BH(0.99, 1, 0.66, 0.10, 0.10)$	0.752	1.768	0.198	1.202	0.446	0.963	0.985	0.996	0.982	0.747	0.809	0.842
$BH(1.40, 1, 1.00, 0.26, 0.12)$	1.568	1.800	0.532	1.271	0.430	0.967	0.983	0.999	0.999	0.795	0.849	0.899
$BH(1.50, 1, 1.00, 0.38, 0.10)$	1.650	1.733	0.720	1.320	0.411	0.939	0.971	0.989	0.968	0.771	0.849	0.887
$BLS(0.30, 0.01, 0.11)$	1.298	0.409	0.380	0.827	0.303	1.000	1.000	1.000	1.000	0.830	0.012	0.006
$BLS(0.40, 0.01, 0.02)$	1.311	0.426	0.094	0.959	0.039	1.000	1.000	1.000	1.000	0.763	0.008	0.007
$BLS(0.50, 0.01, 0.02)$	1.465	0.641	0.085	0.083	0.093	1.000	1.000	1.000	1.000	0.446	0.027	0.025
$BNTA(0.1, 0.01, 0.01, 0.93)$	0.094	1.940	0.094	1.940	0.995	0.797	0.793	0.860	0.225	0.565	0.007	0.599
$BNTA(0.1, 0.01, 0.01, 0.92)$	0.093	1.930	0.093	1.930	0.994	0.810	0.810	0.869	0.222	0.580	0.005	0.617
$BNTA(0.1, 0.01, 0.01, 0.95)$	0.096	1.960	0.096	1.960	0.995	0.835	0.833	0.884	0.188	0.604	0.003	0.629
$BPP(0.30; (0.2, 0.2, 0.1), (0.9, 0.9, 0.5))$	0.690	1.149	0.690	1.149	0.665	0.846	0.844	0.623	0.729	0.548	0.003	0.001
$BPP(0.31; (0.2, 0.2, 0.1), (1.0, 1.2, 0.9))$	0.752	1.182	0.890	1.240	0.909	0.726	0.652	0.673	0.755	0.527	0.004	0.001
$BPP(0.35; (0.2, 0.2, 0.1), (0.9, 0.9, 0.6))$	0.655	1.170	0.655	1.170	0.778	0.805	0.799	0.618	0.722	0.516	0.001	0.000
$B\_B\_P(2, 0.1; 2, 0.1; 0.9)$, type 1	0.983	0.854	0.870	0.935	0.937	0.229	0.486	0.992	0.938	0.054	0.045	0.058
$B\_P\_P(2, 0.1; 0.9; 0.9)$, type 1	0.981	0.850	0.867	0.930	0.936	0.224	0.481	0.991	0.926	0.081	0.06	0.058
$B\_U\_U(2, 0.1; 2; 2)$, type 1	1.070	0.566	0.941	0.643	0.902	0.066	0.265	1.000	1.000	0.483	0.000	0.000
$B\_B\_P(2, 0.1; 2, 0.1; 2.1)$, type 2	2.126	0.950	1.953	1.063	0.962	0.329	0.455	0.965	0.773	0.060	0.075	0.158
$B\_B\_U(2, 0.1; 2, 0.1; 2)$, type 2	1.069	0.566	0.940	0.641	0.901	0.073	0.289	1.000	1.000	0.491	0.000	0.000
$B\_P\_U(2, 0.1; 0.1; 2)$, type 2	1.032	0.618	0.966	0.659	0.949	0.095	0.448	1.000	1.000	0.350	0.000	0.001
$B\_BN\_P(2, 0.1; 2, 0.1; 2.1)$, type 3	2.129	0.950	2.032	0.997	0.951	0.370	0.516	1.000	0.995	0.028	0.080	0.134
$P\_BN\_B(0.1; 2, 0.1; 2, 0.1)$, type 3	0.281	0.825	0.196	0.906	0.796	0.120	0.350	0.838	0.851	0.007	0.051	0.018
$P\_BN\_P(0.6; 2, 0.1; 0.6)$, type 3	0.976	0.725	0.586	1.003	0.644	0.123	0.205	0.848	0.825	0.177	0.066	0.026
$BN\_B\_B(2, 0.1; 2, 0.1; 2, 0.15)$, type 4	0.446	0.695	0.296	0.848	0.750	0.033	0.146	0.855	0.914	0.063	0.006	0.001
$BN\_B\_P(2, 0.1; 2, 0.1; 0.4)$, type 4	0.537	0.790	0.394	0.993	0.833	0.176	0.355	0.927	0.850	0.085	0.087	0.053
$BN\_B\_P(2, 0.1; 2, 0.1; 0.5)$, type 4	0.623	0.809	0.493	1.002	0.872	0.180	0.407	0.971	0.899	0.114	0.087	0.053
$B\_U\_B(2, 0.1; 1; 2, 0.15)$, type 5	0.445	0.700	0.662	0.406	0.428	0.994	0.996	0.982	0.995	0.940	0.000	0.000
$B\_U\_U(2, 0.1; 1; 1)$, type 5	0.605	0.427	0.752	0.248	0.462	1.000	1.000	1.000	1.000	1.000	0.000	0.000
$B\_U\_U(2, 0.1; 3; 3)$, type 5	1.557	0.741	2.126	0.404	0.597	0.300	0.187	0.953	0.952	0.580	0.000	0.000

Table 4: *Ratio of average CPU time (in seconds).*

	$n=30$	$n=50$	$n=70$	$n=100$	$n=200$	$n=300$
R_n/Ψ^*_{PB}	73.50	75.71	77.44	80.19	79.69	79.20
S_n/Ψ^*_{PB}	5.01	11.07	20.20	43.73	145.28	303.92

Table 5: *Results for the real data sets.*

	Plants	Health
$R_{n,(0,0)}$	0.003	0.000
$R_{n,(1,0)}$	0.005	0.000
$R_{n,(0,1)}$	0.010	0.000
$S_{n,(0,0)}$	0.005	0.002
$S_{n,(1,0)}$	0.009	0.000
$S_{n,(0,1)}$	0.011	0.000
Ψ^*_{PB}	0.049	0.000
$\hat{\theta}_n$	(0.64000, 0.94000, 0.19852)	(0.30173, 1.21830, 0.12518)

To end this section, Ψ^*_{PB} is applied to two real data sets. The first one were first given and analysed by Holgate (1966), and refers to the number of plants of the species *Lacistema aggregatum* and *Protium guianense* in each of 100 contiguous quadrats. Crockett (1979), Loukas and Kemp (1986), Rayner and Best (1995) and NJ tested the data for agreement with the bivariate Poisson model, they all concluded the data were not well modelled by a BPD. The second data set were analysed in Karlis and Tsiamyrtzis (2008), who used two variables, the number of consultations with a doctor or a specialist (X_1) and the total number of prescribed and non-prescribed medications used in past 2 days (X_2), from the Australian Health survey for 1977–1978. The sample size was quite large ($n = 5190$). These authors assumed that (X_1, X_2) has a BPD. NJ tested these data sets for agreement with the bivariate Poisson model, concluding that they were not well modelled by a BPD. The p-values obtained by applying the test proposed in this paper to these two real data sets are 0.049 and 0.000, respectively, in agreement with the previous analyses.

Table 6: *Simulations results for the type I error probabilities when $\theta_3 = 0$.*

	$\theta_1=\theta_2=1$				$\theta_1=\theta_2=3$				$\theta_1=\theta_2=10$			
	ML		MM		ML		MM		ML		MM	
n	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
100	3.4	7.4	3.5	7.9	3.1	7.3	3.0	7.3	1.0	3.4	0.9	3.4
200	4.2	8.0	4.3	9.1	3.4	8.0	3.3	7.9	2.2	6.6	2.3	6.6
300	4.4	8.7	4.6	9.4	3.7	8.5	3.7	8.5	3.7	7.9	3.7	8.0

6. Case $\theta_3 = 0$

The case $\theta_3 = 0$ has been excluded from H_0 because it is a boundary point. It is well-known (see, for example Andrews, 1999, Self and Liang, 1987, and the references therein) that in such a case the MLE is not asymptotically normally distributed and thus Assumption 1 is not satisfied. Moreover, Andrews (2000) have proven that the bootstrap does not provides a consistent estimator of the distribution. Therefore, the theory so far developed is not valid for $\theta_3 = 0$.

Next we give two possible ways of dealing with this case. A first way consist in applying the method in Feng and McCulloch (1992), which proposed to enlarge the parametric space to $\theta \in \mathbb{R}^3$, so that negative values for $\hat{\theta}_3$ are allowed. With this approach all required assumptions in our theory are satisfied. The only problem with this solution is how to apply in practice the PB approximation because it implies the generation of samples from a $BP(\theta_1, \theta_2, \theta_3)$ distribution with $\theta_3 < 0$. Nevertheless, the WB approximation can be applied. Table 6 gives the result of a small simulation that studies the goodness of this solution. Observe that the results are quite close to those obtained for $\theta_3 > 0$.

Another possible way of dealing with this case is to adapt the alternatives to the usual bootstrap proposed in Andrews (2000). Two of them consists in subsampling, while the other two are based on testing if the parameter is in the boundary. For the later methods we could calculate a confidence interval for θ_3 and look if it contains 0 by applying, for example, the method in Feng and McCulloch (1992) but, as recognized by the authors, it requires rather large sample sizes. Note that testing for $\theta_3 = 0$ is tantamount to having two independent Poisson variables. Another way of investigating the independence of the marginal distributions is by applying the classical χ^2-test. Nevertheless, such test requires the data to be grouped in classes, and the decision could depend on the grouping. In our view, there is a need of a test for independence of variables taking values on $\mathbb{N}_0$, which will be the topic of a future research.

If it can be reasonably assumed that the variables are independent, then by using Raikov's theorem (which states that the sum of two independent non-negative random variables has a Poisson distribution if and only if both random variables have the Poisson distribution), testing gof for an independent Poisson model is equivalent to testing gof to the sum of the components to a univariate Poisson model. In the statistical literature there is a variety of test for testing gof to a univariate Poisson model (see, for example, the review in Gürtler and Henze, 2000).

7. The general *m*-variate case

This section shows that the proposed test can be extended to the general m-variate case, for any $m \geq 2$. Let

$$X_1 = Y_1 + Y_{m+1}, \;\; X_2 = Y_2 + Y_{m+1}, \;\; \ldots, \;\; X_m = Y_m + Y_{m+1},$$

where $Y_1, Y_2, \ldots, Y_{m+1}$ are mutually independent Poisson random variables with means $\theta_1' = \theta_1 - \theta_{m+1} > 0, \ldots, \theta_m' = \theta_m - \theta_{m+1} > 0$ and $\theta_{m+1} > 0$, respectively. The joint distribution of the vector $(X_1, X_2, \ldots, X_m)$ is called a m-variate Poisson distribution with parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_{m+1})$ (see Johnson, Kotz and Balakrishnan, 1997). The joint pgf of $(X_1, X_2, \ldots, X_m)$ is

$$g(u;\theta) = \exp\left\{\sum_{i=1}^{m} \theta_i (u_i - 1) + \theta_{m+1}\left(\prod_{i=1}^{m} u_i - \sum_{i=1}^{m} u_i + m - 1\right)\right\}, \;\; \forall u \in \mathbb{R}^m. \qquad (7)$$

Now, the objective is to test the hypothesis

$$H_{0m} : (X_1, X_2, \ldots, X_m) \text{ has a } m\text{-variate Poisson distribution.}$$

In order to extend the proposed test to the general m-variate case we will use the following result in Proposition 3 in NJ which states that $g(u;\theta)$ is the only pgf in $G_m = \{g : [0,1]^m \to \mathbb{R}$, such that g is a pgf and $\frac{\partial}{\partial u_i} g(u_1, u_2, \ldots, u_m)$ exists $\forall u \in [0,1]^m$, $1 \leq i \leq m\}$ satisfying the following system,

$$D_i(u;\theta) = 0, \;\; 1 \leq i \leq m, \qquad (8)$$

$\forall u \in [0,1]^m$, where $D_i(u;\theta) = \dfrac{\partial}{\partial u_i} g(u) - \left\{\theta_i + \theta_{m+1}\left(\prod_{j \neq i} u_j - 1\right)\right\} g(u), \;\; 1 \leq i \leq m.$

Let $(X_1, X_2, \ldots, X_m) \in \mathbb{N}_0^m$ be a random vector and let $g(u_1, u_2, \ldots, u_m) = E\left(u_1^{X_1} u_2^{X_2} \cdots u_m^{X_m}\right)$ its pgf. Then, taking into account that

$$g(u) = \sum_{r_1, r_2, \ldots, r_m \geq 0} u_1^{r_1} u_2^{r_2} \cdots u_m^{r_m} p(r_1, r_2, \ldots, r_m),$$

where $p(r_1, r_2, \ldots, r_m) = P(X_1 = r_1, X_2 = r_2, \ldots, X_m = r_m)$, we can write

$$D_i(u;\theta) = \sum_{r_1, r_2, \ldots, r_m \geq 0} \Big\{(r_i + 1) p(r_1, \ldots, r_{i-1}, r_i + 1, r_{i+1}, \ldots, r_m) - (\theta_i - \theta_{m+1}) p(r_1, r_2, \ldots, r_m)$$
$$- \theta_{m+1} p(r_1 - 1, \ldots, r_{i-1} - 1, r_i, r_{i+1} - 1, \ldots, r_m - 1)\Big\} u_1^{r_1} u_2^{r_2} \cdots u_m^{r_m}, \;\; 1 \leq i \leq m.$$

Let $D_{in}(u;\hat{\theta})$ denote the empirical counterpart of $D_i(u;\theta)$ obtained by replacing the pgf g by the epgf g_n and θ by a consistent estimator $\hat{\theta}$, $1 \leq i \leq m$. If H_{0m} is true then the functions $D_{in}(u;\hat{\theta})$, $1 \leq i \leq m$, should be close to 0, $\forall u \in [0,1]^m$. This proximity to zero can be interpreted as we did in Section 2, for the bivariate case. Observe that

$$D_{in}(u;\hat{\theta}) = \sum_{r_1,r_2,\ldots,r_m \geq 0} d_i(r_1,r_2,\ldots,r_m;\hat{\theta}) u_1^{r_1} u_2^{r_2} \cdots u_m^{r_m}, \;\; 1 \leq i \leq m,$$

where

$$\begin{aligned} d_i(r_1,r_2,\ldots,r_m;\hat{\theta}) &= (r_i+1) p_n(r_1,\ldots,r_{i-1},r_i+1,r_{i+1},\ldots,r_m) \\ &\quad - (\hat{\theta}_i - \hat{\theta}_{m+1}) p_n(r_1,r_2,\ldots,r_m) \\ &\quad - \hat{\theta}_{m+1}\, p_n(r_1-1,\ldots,r_{i-1}-1,r_i,r_{i+1}-1,\ldots,r_m-1), \;\; 1 \leq i \leq m, \end{aligned}$$

and $p_n(r_1,r_2,\ldots,r_m) = \frac{1}{n}\sum_{k=1}^n I(X_{k1}=r_1, X_{k2}=r_2,\ldots,X_{km}=r_m)$ is the relative frequency of $(r_1,r_2,\ldots,r_m)$. Therefore, $D_{in}(u;\hat{\theta}) = 0$, $\forall u \in [0,1]^m$, $1 \leq i \leq m$, if and only if the coefficients of $u_1^{r_1} u_2^{r_2} \cdots u_m^{r_m}$ in the previous expansions are null, $\forall r_1, r_2, \ldots, r_m \geq 0$. This leads us to consider the following statistic for testing H_{0m},

$$W_{m,n}(\hat{\theta}) = \sum_{r_1,r_2,\ldots,r_m \geq 0} \left\{ \sum_{i=1}^m d_i(r_1,r_2,\ldots,r_m;\hat{\theta})^2 \right\} = \sum_{r_1,r_2,\ldots,r_m=0}^{M} \left\{ \sum_{i=1}^m d_i(r_1,r_2,\ldots,r_m;\hat{\theta})^2 \right\},$$

where $M = \max\{X_{(n)1}, X_{(n)2}, \ldots, X_{(n)m}\}$, $X_{(n)k} = \max_{1 \leq i \leq n} X_{ik}$, $1 \leq k \leq m$. Similar results to those stated in Sections 2 and 3 for the bivariate case can be established for $W_{m,n}(\hat{\theta})$.

8. Proofs

Here we give a sketch of the proofs of the results in Sections 2 and 3. A detailed derivation of the results can be obtained from the authors upon request.

Proof of Theorem 1 Observe that

$$d_1(r,s;\hat{\theta}) = d_1(r,s;\theta) - (\hat{\theta}_1 - \theta_1) p_n(r,s) + (\hat{\theta}_3 - \theta_3)\{p_n(r,s) - p_n(r,s-1)\}$$

and

$$\sum_{r,s \geq 0} d_1(r,s;\theta)^2 = \frac{1}{n^2} \sum_{i \neq j} h_1(\boldsymbol{X}_i, \boldsymbol{X}_j;\theta) + \frac{1}{n^2} \sum_{i=1}^n h_1(\boldsymbol{X}_i, \boldsymbol{X}_i;\theta).$$

By the SLLN,

$$\frac{1}{n}\sum_{i=1}^{n} h_1(\boldsymbol{X}_i,\boldsymbol{X}_i;\theta) \xrightarrow{a.s.} E\left\{\sum_{r,s\geq 0}\phi_{1rs}(\boldsymbol{X}_1;\theta)^2\right\} < \infty.$$

By the SLLN for U-statistics (Theorem 5.4 in Serfling, 1980),

$$\frac{1}{n^2}\sum_{i\neq j} h_1(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta) \xrightarrow{a.s.} E\{h_1(\boldsymbol{X}_1,\boldsymbol{X}_2;\theta)\} = \sum_{r,s\geq 0} a_1(r,s;\theta)^2.$$

Therefore,

$$\sum_{r,s\geq 0} d_1(r,s;\theta)^2 \xrightarrow{a.s.} \sum_{r,s\geq 0} a_1(r,s;\theta)^2.$$

Since $p_n(r,s)^2 \leq p_n(r,s)$, $\forall r,s \geq 0$, and $\sum_{r,s\geq 0} p_n(r,s) = 1$, we have

$$(\hat{\theta}_1-\theta_1)^2 \sum_{r,s\geq 0} p_n(r,s)^2 \leq (\hat{\theta}_1-\theta_1)^2 = o(1),$$

and analogously,

$$(\hat{\theta}_3-\theta_3)^2 \sum_{r,s\geq 0} \{p_n(r,s)-p_n(r,s-1)\}^2 = o(1).$$

Thus,

$$\sum_{r,s\geq 0} d_1(r,s;\hat{\theta})^2 \xrightarrow{a.s.} \sum_{r,s\geq 0} a_1(r,s;\theta)^2. \tag{9}$$

Following similar steps we get

$$\sum_{r,s\geq 0} d_2(r,s;\hat{\theta})^2 \xrightarrow{a.s.} \sum_{r,s\geq 0} a_2(r,s;\theta)^2. \tag{10}$$

Finally, the result is obtained from (9) and (10). ■

Proof of Theorem 2 Let us consider the separable Hilbert space of functions $\mathscr{H} = \{g : \mathbb{N}_0 \to \mathbb{R}$, so that $\|g\|^2_{\mathscr{H}} = \sum_{r\geq 0}\sum_{s\geq 0} g(r,s)^2 < \infty\}$. We have that

$$\sqrt{n}d_k(r,s;\hat{\theta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{krs}(\boldsymbol{X}_i;\theta) + \sqrt{n}(\hat{\theta}-\theta)\hat{v}_k(r,s)^{\intercal}, \quad k=1,2,$$

with $\hat{v}_1(r,s) = (-p_n(r,s), 0, p_n(r,s) - p_n(r,s-1))$ and $\hat{v}_2(r,s) = (0, -p_n(r,s), p_n(r,s) - p_n(r-1,s))$. From Assumption 1 and the SLLN, we get that

$$\sqrt{n}d_k(r,s;\hat{\theta}) = \sqrt{n}d_{1k}(r,s;\theta) + R_k(r,s), \quad k=1,2,$$

with

$$d_{1k}(r,s;\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\phi_{krs}(\boldsymbol{X}_i;\theta) + \ell(\boldsymbol{X}_i;\theta)v_k(r,s;\theta)^{\intercal}\right\}, \quad k=1,2,$$

$$v_1(r,s;\theta) = (-P_\theta(r,s), 0, P_\theta(r,s) - P_\theta(r,s-1)),$$

$$v_2(r,s;\theta) = (0, -P_\theta(r,s), P_\theta(r,s) - P_\theta(r-1,s)),$$

and $\|R_k\|_{\mathscr{H}} = o_P(1)$, $k=1,2$. From the CLT in Hilbert spaces (see, for example, van der Vaart and Wellner, 1996, pp. 50–51), it follows that $\|\sqrt{n}d_{1k}\|^2_{\mathscr{H}} = O_P(1)$, $k=1,2$, and therefore

$$nW_n(\hat{\theta}) = \|\sqrt{n}d_{1k}\|^2_{\mathscr{H}} + \|\sqrt{n}d_{12}\|^2_{\mathscr{H}} + o_P(1).$$

Routine calculations show that

$$\|\sqrt{n}d_{1k}\|^2_{\mathscr{H}} + \|\sqrt{n}d_{12}\|^2_{\mathscr{H}} = \frac{1}{n}\sum_{i,j=1}^{n}K(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta).$$

The result is achieved by applying Theorem 6.4.1.B in Serfling (1980) to $\frac{1}{n}\sum_{i,j=1}^{n}K(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta)$. ■

Proof of Theorem 3 Following similar steps to those given in the proof of Theorem 2 but instead of applying the CLT for iid random elements taking values in $\mathscr{H}$, we apply a CLT for triangular arrays, such as Theorem 1.1 in Kundu et al. (2000). ■

Proof of Theorem 4 $nW^*_{2n}(\hat{\theta})$ can be expressed as $nW^*_{2n}(\hat{\theta}) = W^*_1 + W^*_2 + 2W^*_3 + W^*_4$, where

$$W^*_1 = \frac{1}{n}\sum_{i,j=1}^{n}K(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta)\xi_i\xi_j,$$

$$W_2^* = \frac{1}{n}\sum_{i,j=1}^{n}\{h(\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\theta}) - h(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta)\}\xi_i\xi_j,$$

$$W_3^* = \frac{1}{n}\sum_{i,j=1}^{n}\{\hat{\ell}(\boldsymbol{X}_i;\hat{\theta})\mu(\boldsymbol{X}_j;\hat{\theta})^{\intercal} - \ell_1(\boldsymbol{X}_i;\theta)\mu(\boldsymbol{X}_j;\theta)^{\intercal}\}\xi_i\xi_j,$$

$$W_4^* = \frac{1}{n}\sum_{i,j=1}^{n}\{\hat{\ell}(\boldsymbol{X}_i;\hat{\theta})S(\hat{\theta})\hat{\ell}(\boldsymbol{X}_j;\hat{\theta})^{\intercal} - \ell_1(\boldsymbol{X}_i;\theta)S(\theta)\ell_1(\boldsymbol{X}_j;\theta)^{\intercal}\}\xi_i\xi_j.$$

From the results in Delhing and Mikosch (1994),

$$\sup_x |P_*\{W_1^* \leq x\} - P\{W_1 \leq x\}| \xrightarrow{a.s.} 0.$$

Thus, to show the result it suffices to see that $W_k^* = o_{P_*}(1)$ in probability, $k = 2,3,4$. We first deal with W_2^*. Observe that

$$E_*(W_2^{*2}) \leq M\frac{1}{n^2}\sum_{i,j=1}^{n}\{h(\boldsymbol{X}_i,\boldsymbol{X}_j;\hat{\theta}) - h(\boldsymbol{X}_i,\boldsymbol{X}_j;\theta)\}^2,$$

for some positive $M > 0$. From the assumptions made, the right-hand side of the above expression is $o_P(1)$. Therefore, $W_2^* = o_{P_*}(1)$ in probability. As for W_3^*, we have that $W_3^* = W_{31}^* W_{32}^{*\top} + W_{33}^* W_{34}^{*\top}$, with

$$W_{31}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\hat{\ell}(\boldsymbol{X}_i;\hat{\theta}) - \ell_1(\boldsymbol{X}_i;\theta)\}\xi_i,$$

$$W_{32}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mu(\boldsymbol{X}_i;\hat{\theta})\xi_i,$$

$$W_{33}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell_1(\boldsymbol{X}_i;\theta)\xi_i,$$

$$W_{34}^* = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\mu(\boldsymbol{X}_i;\hat{\theta}) - \mu(\boldsymbol{X}_i;\theta)\}\xi_i.$$

From the assumptions made, $E_*(W_{31}^{*2}) = o_P(1)$, $E_*(W_{32}^{*2})$ is bounded in probability and $E_*(W_{33}^{*2})$ is bounded a.s.. Now taking into account that

$$\frac{\partial}{\partial\theta_1}P_\theta(r,s) = P_\theta(r-1,s) - P_\theta(r,s),$$

$$\frac{\partial}{\partial\theta_2}P_\theta(r,s) = P_\theta(r,s-1) - P_\theta(r,s),$$

$$\frac{\partial}{\partial\theta_3}P_\theta(r,s) = P_\theta(r-1,s-1) - P_\theta(r-1,s) - P_\theta(r,s-1) + P_\theta(r,s),$$

it follows that

$$\sup_{r,s\in\mathbb{N}_0} |P_{\hat\theta}(r,s) - P_\theta(r,s)| \le M\|\hat\theta - \theta\|, \tag{11}$$

for some positive $M > 0$. This implies that $E_*(W_{34}^{*2}) = o_P(1)$. Therefore, $W_3^* = o_{P_*}(1)$ in probability. By using (11) and the assumptions made, it readily follows that $W_4^* = o_{P_*}(1)$ in probability. This concludes the proof. ■

Acknowledgements

The authors thank the anonymous referees for their valuable time and careful comments, which improved the quality of this paper. F. Novoa-Muñoz wishes to thank his institution, University of Bío-Bío (Chile) and the scholarship given by the Chilean Ministry of Education through the Superior Education MECESUP Program 2, which make his doctorate studies possible. M.D. Jiménez-Gamero acknowledges financial support from grant MTM2014-55966-P of the Spanish Ministry of Economy and Competitiveness.

References

Andrews, D.W.K. (1999) Estimation when a parameter is on a boundary. *Econometrica*, 67, 1341–1383.

Andrews, D.W.K. (2000) Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68, 399–405.

Burke, M.D. (2000). Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap. *Statistics & Probability Letters*, 46, 13–20.

Crockett, N. G. (1979). A quick test of fit of a bivariate distribution. In *Interactive Statistics*, D. McNeil (ed.), 185–191. Amsterdam: North-Holland.

Delhing, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51, 392–413.

Duchesne, P. and Lafaye De Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54, 858–862.

Feng, Z., McCulloch, C.E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statistics & Probability Letters*, 13, 325–332.

Gürtler, N. and Henze, N. (2000). Recent and classical goodness-of-fit tests for the Poisson distribution. *Journal of Statistical Planning and Inference*, 90, 207–225.

Haight, F. A. (1967). *Handbook of the Poisson Distribution*. New York: John Wiley & Sons.

Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, 51, 241–245.

Holgate, P. (1966). Bivariate generalizations of Neyman's type a distribution. *Biometrika*, 53, 241–245.

Janssen, A. (2000). Global power functions of goodness of fit tests. *The Annals of Statistics*, 28, 239–253.

Imhof, J.P. (1961) Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 419–426.

Jiménez-Gamero, M.D., Kim, H.-M. (2015) Fast goodness-of-fit tests based on the characteristic function. *Computational Statistics and Data Analysis*, 89, 172–191.

Johnson, N. L. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Wiley, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley, New York.

Karlis, D. and Tsiamyrtzis, P. (2008). Exact Bayesian modelling for bivariate Poisson data and extensions. *Statistics and Computing*, 18, 27–40.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Marcel Dekker, Inc, New York.

Kojadinovic, I. and Yan, J. (2012) Goodness-of-fit testing based on a weighted bootstrap: A fast large sample alternative to the parametric bootstrap, *Canadian Journal of Statistics*, 40, 2012, 480–500.

Kundu, S., Majumdar, S. and Mukherjee, K. (2000). Central limits theorems revisited. *Statistics & Probability Letters*, 47, 265–275.

Loukas, S. and Kemp, C.D. (1986). The index of dispersion test for the bivariate Poisson distribution. *Biometrics*, 42, 941–948.

Nakamura, M. and Pérez-Abreu, V. (1993). Use of an empirical probability generating function for testing a Poisson model. *Canadian Journal of Statistics*, 21, 149–156.

Novoa-Muñoz, F. and Jiménez-Gamero, M. D. (2014). Testing for the bivariate Poisson distribution. *Metrika*, 77, 771–793.

Papageorgiou, H. and Loukas, S. (1988). Conditional even point estimation for bivariate discrete distributions. *Communications in Statistics–Theory and Methods*, 17, 3403–3412.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rayner, J.C. W. and Best, D.J. (1995). Smooth Tests for the bivariate Poisson distribution. *Australian & New Zealand Journal of Statistics*, 37, 233–245.

Sahai, H. and Khurshid, A. (1993). Confidence intervals for the mean of a Poisson distribution: A review. *Biometrical Journal*, 35(7), 857–867.

Self, S.G. and Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82, 605–610.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

SORT 40 (1) January-June 2016, 139-152

Exploring Bayesian models to evaluate control procedures for plant disease

Danilo Alvares[1], Carmen Armero[1], Anabel Forte[1] and Luis Rubio[2]

Abstract

Tigernut tubers are the main ingredient in the production of *orxata* in Valencia, a white soft sweet popular drink. In recent years, the appearance of black spots in the skin of tigernuts has led to important economic losses in *orxata* production because severely diseased tubers must be discarded. In this paper, we discuss three complementary statistical models to assess the disease incidence of harvested tubers from selected or treated seeds, and propose a measure of effectiveness for different treatments against the disease based on the probability of germination and the incidence of the disease. Statistical methods for these studies are approached from Bayesian reasoning and include mixed-effects models, Dirichlet-multinomial inferential processes and mixed-effects logistic regression models. Statistical analyses provide relevant information to carry out measures to palliate the black spot disease and achieve a high-quality production. For instance, the study shows that avoiding affected seeds increases the probability of harvesting asymptomatic tubers. It is also revealed that the best chemical treatment, when prioritizing germination, is disinfection with hydrochloric acid while sodium hypochlorite performs better if the priority is to have a reduced disease incidence. The reduction of the incidence of the black spots syndrome by disinfection with chemical agents supports the hypothesis that the causal agent is a pathogenic organism.

MSC: 62C10, 62F15, 62J12, 62K15, 62P12, 92B15.

Keywords: Dirichlet-multinomial model, logistic regression, measures of effectiveness, tigernuts tubers.

1. Introduction

Tigernuts growing has become an important sector of agribusiness in Valencia (Spain). Tigernut tubers are mostly used for the production of *orxata*, a white soft sweet drink

[1] University of Valencia, Spain. daldasil@alumni.uv.es, carmen.armero@uv.es, anabel.forte@uv.es

[2] Valencian Institute for Agricultural Research, Spain. lrubio@ivia.es

Received: November 2015
Accepted: April 2016

highly appreciated in Spain (Morell and Barber, 1983). The trade around *orxata* has expanded over the past few years but it has been also strongly affected by the appearance of black spots in the skin of tigernuts, making it crucial to figure out how to palliate its negative effects.

Epidemiological data and some greenhouse experiments (unpublished) suggest that the syndrome of black spots could be a disease caused by an unidentified pathogenic organism, which prevent to devise specific strategies to eradicate or control the disease. This lack of information and the difficulties in identifying the aetiology of the disease suggest assaying several general methods of disease control based on selecting pathogen-free seeds or removing the pathogens from seeds by thermal or chemical treatments which have been successfully applied to known pathogenic species of viruses, bacteria and fungi and different crops (Shepard and Claftin, 1975; Sauer and Burroughs, 1986; Grondeau et al., 1994), including tigernuts (García-Jiménez et al., 2004). Moreover, the possible adverse effects in humans are not known so severely diseased tubers are automatically discarded, and only asymptomatic and mildly affected tubers are marketable.

Chemical or thermal treatments can have a detrimental effect on seed germination, which causes yield reduction. Thus, the effectiveness of these methods must be assessed not only by considering the effect on the disease incidence but also on the germination. We evaluated two methods of disease control: i) selection of non-infected tubers used as seeds, and ii) chemical or/and thermal treatments (alone or combined) of infected seeds in order to remove or kill the pathogen. However, for method i), since the pathogen cannot be properly identified and it is not possible to detect pathogen-free seeds, the selection is based on the use of asymptomatic (without black spots) tubers. These seeds could contain the pathogen, although in lower quantities than those severely affected (with black spots covering the whole surface).

The aim of this paper is to gain insight into the transmission of the disease from the tubers used as seed to the progeny of tubers, as well as the procedures for disease control. To the best of our knowledge, this is the first paper devoted to study the black spot disease in tigernuts from a statistical point of view. The structure of this paper is as follows. Section 2 presents two experimental studies specifically designed for the problem. Section 3 discusses the statistical modelling of the data. In particular, Subsections 3.1 and 3.2 analyse the weight of tubers harvested from symptomatic and asymptomatic seeds, and the disease transmission through a mixed-effects model based on the lognormal distribution and the Dirichlet-multinomial inferential process, respectively. Subsection 3.3 deals with effectiveness of different treatments against disease in terms of mixed-effects logistic regression models and a measure of effectiveness which takes into account the germination process and the level of affection of the disease. Conclusions and further remarks are given in Section 4.

2. Experiments and data

Two different greenhouse experiments were designed and carried out with the objective of learning about the black spot disease in tigernuts. *Experiment 1* was aimed at studying the transmission of the disease from seed to the harvested tubers, and *Experiment 2* at analysing the consequences of different treatments against disease.

Experiment 1: Asymptomatic and severely affected seeds were selected and sowed in seven separated pots with five seeds per pot. Five months later, sowing tubers were harvested. The average weight of a tuber (in grams) and the number of asymptomatic tubers (no black spots), with mild symptoms (few small black spots), and severe symptoms (tuber almost completely covered by black spots) in each pot were recorded (see Figure 1).

Experiment 2: Chemical and thermal treatments were applied to severely affected seeds following a balanced two-factor factorial design. The specific combination of both types of treatments is denoted by T_{qt}, where subscripts q and t represent chemical and thermal treatment, respectively. Chemical treatments tested were: no treatment ($q = 1$); disinfection with sodium hypochlorite ($q = 2$); disinfection with hydrochloric acid ($q = 3$); treatment with trifloxystrobin, an active fungicide against a wide range of fungal plant pathogens ($q = 4$); application of a plant defence activator ($q = 5$); and disinfection with trisodium phosphate ($q = 6$). In the case of thermal treatments: no treatment ($t = 1$); incubation in water at 55°C for 30 min ($t = 2$); and 60°C for 30 min ($t = 3$). Eight pots were sowed with five seeds each for every T_{qt} treatment. Germination rate was estimated from the number of seedlings emerged in each pot during the next two weeks. About five months later tubers were harvested and the number of marketable (asymptomatic and mildly diseased) and severely diseased tubers in each pot were registered.

Table 1 summarizes the data from *Experiment 1* together with those from *Experiment 2* corresponding to the absence of chemical and thermal treatment. We joined the data from both experiments because they were independent and shared a common scenario. There seems to be no great differences in the mean and standard deviation of the unit weight of tubers of both groups. However, there are considerable differences in the proportion of tubers in each level of disease infection. It is important to emphasise the strong relationship between tubers and seeds severely affected.

Figure 1: *Tigernuts tubers with different levels of symptoms: asymptomatic, mild, and severe.*

Table 1: *Mean and standard deviation of the unit weight of tubers (in grams) and proportion of asymptomatic tubers, with mild and severe symptoms from asymptomatic and severely affected seeds.*

Seeds	**Tubers**				
	Unit-weight		Asymptomatic	Mild	Severe
	Mean	Sd	Proportion		
Asymptomatic	0.424	0.0846	0.698	0.228	0.074
Severe	0.409	0.0689	0.003	0.176	0.821

Table 2 shows the proportion of marketable tubers harvested and of germinated seeds with regard to each particular chemical and thermal treatment. The thermal treatment at 60°C for 30 min, independently of the chemical treatment, dramatically reduced the germination rate. Notice that no data were collected for T_{33} since no seeds germinated.

Table 2: *Proportion of marketable harvested tubers and of germinated seeds (in brackets) for each treatment.*

Chemical treatment	**Thermal treatment**		
	No treatment	55°C for 30 min	60°C for 30 min
No treatment	0.151 (0.425)	0.423 (0.275)	0.474 (0.025)
Disinf. with sodium hypochlorite	0.403 (0.450)	0.373 (0.275)	0.552 (0.075)
Disinf. hydrochloric acid	0.228 (0.625)	0.136 (0.225)	—
Fungicide	0.209 (0.475)	0.148 (0.275)	0.552 (0.050)
Activator plant defense	0.388 (0.375)	0.359 (0.375)	0.191 (0.025)
Disinf. with trisodium phosphate	0.257 (0.325)	0.404 (0.150)	0.123 (0.050)

3. Statistical modelling

Bayesian inference always expresses uncertainty about the quantities of interest and experimental results in probabilistic terms. Bayes' theorem combines the prior distribution and the likelihood function of the data to obtain the posterior distribution, which contains all relevant information of the problem. This distribution was not analytical in all studies of the paper except for the analysis in Subsection 3.2. In those studies the subsequent posterior distribution was approximated by Markov chain Monte Carlo (MCMC) methods (Gelman et al., 2013) using the software WinBUGS (Lunn et al., 2000). In all these inferences, the MCMC algorithm was run for three Markov chains with 100 000 iterations each after a burn-in period of 1 000. The chains were thinned by only storing every 5th iteration in order to reduce auto-correlation in the saved sample. Trace plots of the simulated values of the three chains always appeared overlapping one another indicating stabilization. Convergence of the chains to the posterior distribution was assessed trough the potential scale reduction factor, *Rhat*, and the effective sample size, *neff* (Kass et al., 1998). In all cases, the *Rhat* values were equal to or near 1 and *neff* > 100, thus indicating that the distributions of the simulated values between and within

the three chains were practically identical, and also that sufficient MCMC samples had been obtained, respectively.

3.1. Weight of tubers

The average unit-weight, in grams, of tubers harvested in each pot from asymptomatic and from severely affected seeds is a positive and continuous variable. It would be better for the statistical analysis to have known the individual weight of each tuber of each pot, but this information was not recorded in the experiment.

That variable (Y from now on) can be approached by several simple models (Ntzoufras, 2009) which share the common structure $(Y \mid \boldsymbol{\theta}) \sim f_y(\boldsymbol{\theta})$, where f_y can vary among different distributions – chi-squared, exponential, gamma, inverse-gamma, log-normal, and Weibull – with parametric vector $\boldsymbol{\theta}$ that may depend on both covariates or factors, as for example the treatment group in our study, and random effects for assessing the specific pot individual effect.

We used the deviance information criterion (DIC) for selecting the most appropriate model, the smaller the DIC the better the fit. Table 3 shows the DIC value for each model pointing out the lognormal (LN) model as slightly better than the rest.

Table 3: *Deviance information criterion values of various models for the mean tuber-weight.*

Model	Chi-squared	Exponential	Gamma	Inverse-Gamma	Lognormal	Weibull
DIC	36.49	12.22	−48.19	23.45	−50.83	−49.40

The selected mixed-effects model was $(Y_{ij} \mid \mu_{ij}, \sigma^2) \sim \text{LN}(\mu_{ij}, \sigma^2)$, defined as

$$\begin{array}{rcl} Y_{ij} & = & \exp(\mu_{ij} + \sigma Z_{ij}),\ i = 1, \ldots, 7 \\ \mu_{ij} & = & \alpha + \beta I_{SD}(i) + b_{ij} \\ Z_{ij} & \sim & \text{N}(0,1), \end{array} \tag{1}$$

where Y_{ij} is the average unit-weight of the tubers harvested in pot i of the seed group j, where $j = 1$ stands for asymptomatic seeds and $j = 2$ for severe diseased seeds; α is the common term in μ_{ij} corresponding to asymptomatic seeds and β the additional effect for severely diseased tubers. The indicator function $I_{SD}(i)$ is 1 when tubers from pot i are harvested from severely diseased seeds, and 0 otherwise. Random effects b_{ij} are conditional $i.i.d.$ random variables normally distributed with mean zero and variance σ_j^2, $j = 1, 2$. To complete the Bayesian model we needed to elicit a prior distribution for the parameters and hyperparameters of the model, $(\alpha, \beta, \sigma, \sigma_1, \sigma_2)$. We assumed prior independence among all them as a default scenario, and considered flat prior distributions $\text{N}(0.0, 10\,000)$ for α and β, where variability in this normal distribution is expressed in terms of the variance, and a gamma distribution, $\text{Ga}(0.01, 0.01)$, for σ, as well as the hyperprior distribution for σ_j.

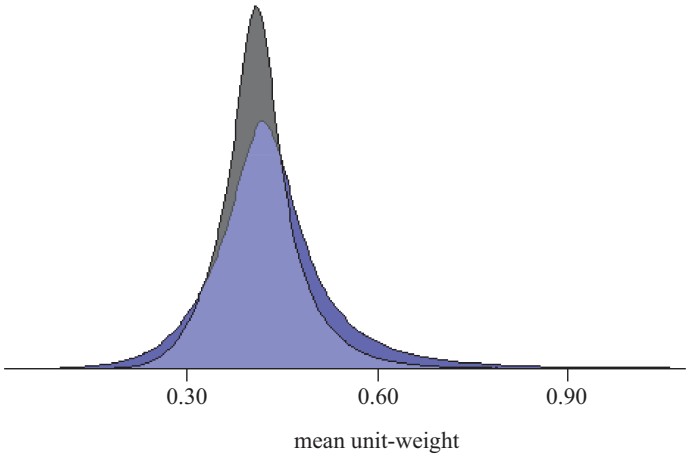

Figure 2: *Posterior distribution for the mean of the average unit-weight of tubers harvested from asymptomatic (blue) and severely diseased seeds (dark gray).*

The approximated posterior mean of the regression coefficients α and β are negative, in particular -0.877 and -0.031 respectively, with $P(\beta > 0 \mid \mathscr{D}) = 0.3701$. In addition, the posterior mean for the standard deviations σ, σ_1 and σ_2 were $\mathrm{E}(\sigma \mid \mathscr{D}) = 0.144$, $\mathrm{E}(\sigma_1 \mid \mathscr{D}) = 0.194$, and $\mathrm{E}(\sigma_2 \mid \mathscr{D}) = 0.151$, with 95% credible intervals (0.070, 0.238), (0.069, 0.421) and (0.072, 0.273), respectively. Figure 2 shows the posterior distribution for the mean of the average unit-weight, in grams, of tubers harvested from asymptomatic and severely diseased seeds. Notice that a great part of both posterior distributions overlap, which could indicate a non-substantial difference in the weight of the tubers harvested from asymptomatic and affected seeds.

3.2. Seed transmission of the disease

We continue with the analysis of the data from *Experiment 1* together with the ones from *Experiment 2* corresponding to the absence of chemical and thermal treatment. We focused on the probability of obtaining tubers with severe, mild or no symptoms of the disease in each pot with regard to the type of seed, asymptomatic or severely affected, planted. For each type of seed, the response variable was the number of harvested tubers in each level of affection in the different pots harvested, which was modelled in terms of a multinomial distribution. Of course, other modelling would be acceptable, for instance the proportional odds models (Liu and Agresti, 2005) to explore the ordinality of the variable of interest. However, we opted to follow a simplified approach that also captures the experimental goals.

The multinomial distribution, $\mathrm{Multin}(n, \theta_1, \ldots, \theta_K)$ (Agresti, 2013) is the probability distribution of the outcomes from a multinomial experiment based on n independent trials, in which each of them can result in one of K mutually exclusive and exhaustive categories. The probability θ_k for each category k does not vary with the data

and $\sum_{k=1}^{K} \theta_k = 1$. In the same way that the multinomial distribution is a generalization of the binomial distribution, the conjugate prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\intercal$ is a multivariate generalization of the beta distribution, known as the Dirichlet distribution, Dirichlet$(\alpha_1, \ldots, \alpha_K)$, with joint density function

$$f(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_K)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \theta_1^{(\alpha_1 - 1)} \cdots \theta_K^{(\alpha_K - 1)}, \ \alpha_k > 0, \ k = 1, \ldots, K, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\intercal$ is the subsequent parametric vector. The combination of a multinomial model and a Dirichlet prior distribution (known as the Dirichlet-multinomial model) was proposed by Lindley (1964) and Good (1965) and results on a Dirichlet posterior distribution for $\boldsymbol{\theta}$ with updated hyper-parameters $\alpha_k + y_k$, $k = 1, \ldots, K$, where y_k is the number of trials in category k.

The literature on Bayesian statistics includes various proposals for prior distributions of $\boldsymbol{\alpha}$ with minimum information (Alvares, 2015). Our choice here is $\alpha_k = 1/K$ because it has been shown to be an objective prior (Berger et al., 2015) with the reference distance approach (see also Perks, 1947).

Figure 3 shows the 95% posterior credible intervals for the probability associated to asymptomatic, mild and severe symptoms tubers depending on the health of the seed from which have grown. Notice that for asymptomatic seeds the probability of harvesting asymptomatic tubers (posterior mean 0.698) is greater than the probabilities of collecting tubers with mild (posterior mean 0.228) or severe symptoms (posterior mean 0.074). However, in the group of diseased seeds the situation is the opposite, and the probability of harvesting tubers with severe symptoms (posterior mean 0.821) is greater than the probabilities corresponding to tubers with mild symptoms (posterior mean 0.176) and no symptoms (posterior mean 0.003). It was clear that the selection

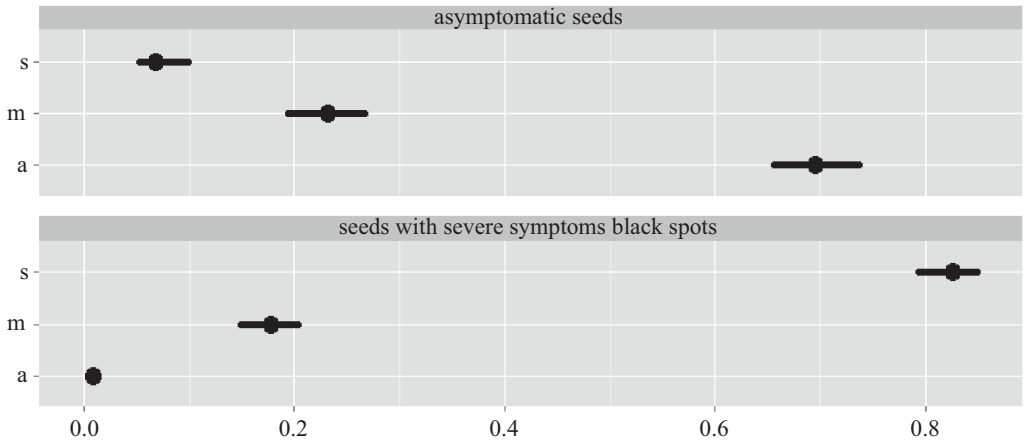

Figure 3: *95% credible interval for the probability associated to asymptomatic (a), mild (m) and severe (s) symptoms tubers harvested from asymptomatic seeds and from seeds with severe symptoms black spots.*

of asymptomatic seeds seemed to be beneficial to reduce the prevalence of tubers with black spots.

As an alternative modelling, we have also examined the three-stage hierarchical multinomial model proposed by (Nandram, 1998). It added as a new level in the model the assumption that the hyperparameters from the prior distribution are unknown and come from a general Dirichlet hyperdistribution, formulated in terms of a parametrization based on the marginal mean of each probability and a common weight. As results were practically equal as those obtained from the reference distance approach prior distribution (the only relevant differences occurred in the fourth decimal place), the most simple Dirichlet-multinomial model was preferable to its hierarchical modelling counterpart.

3.3. Comparison of treatments

We discuss the possible benefits of applying a specific treatment to affected tubers before using them as seeds. We used data from *Experiment 2* for analysing the number of marketable tubers harvested and of germinated seeds from each pot through two marginal mixed-effects logistic regression models for each combination of chemical and thermal treatment. Next, we combined both results into a single measure that quantifies the effectiveness of each treatment.

3.3.1. Germination and disease

Let $Y_{1i}^{(qt)}$ the binomial variable that describes for pot i, $i = 1,\ldots,8$, the number of marketable tubers from a total of $N_{1i}^{(qt)}$ collected from severely affected seeds previously treated with chemical treatment q and thermal treatment t, and represent by $\theta_1^{(qt)}$ the subsequent binomial probability. This probability is modelled through the mixed-effects logistic regression model

$$\begin{aligned} Y_{1i}^{(qt)} &\sim \text{Bin}(N_{1i}^{(qt)}, \theta_1^{(qt)}), \\ \text{logit}(\theta_{1i}^{(qt)}) &= \alpha_1 + \beta_1^{(q)} + \lambda_1^{(t)} + \phi_1^{(qt)} + b_{1i}, \end{aligned} \tag{3}$$

where parameter α_1 indicates the effect of neither chemical nor thermal treatment and $\beta_1^{(q)}$, $\lambda_1^{(t)}$, and $\phi_1^{(qt)}$ include the marginal effect of each treatment, chemical or thermal, and its interaction, respectively. Random effects, b_{1i}, associated to pot i are conditional $i.i.d.$ random variables, $(b_{1i} \mid \sigma_{b1}) \sim \text{N}(0, \sigma_{1b}^2)$. It is worth mentioning that the number of tubers collected in the different pots have a great level of variability: from 8 to 466, mean 201, median 193.5, and standard deviation 77.36 tubers.

The probability of germination with regard to each treatment T_{qt} considered is also analysed through the mixed-effects logistic regression model

$$\begin{aligned} Y_{2i}^{(qt)} &\sim \text{Bin}(N_{2i}^{(qt)}, \theta_2^{(qt)}), \\ \text{logit}(\theta_{2i}^{(qt)}) &= \alpha_2 + \beta_2^{(q)} + \lambda_2^{(t)} + \phi_2^{(qt)} + b_{2i}, \end{aligned} \tag{4}$$

where now $Y_{2i}^{(qt)}$ is the number of germinated seeds in the ith pot from a total $N_{2i}^{(qt)} = 5$ sowed, $\theta_2^{(qt)}$ the probability of germination, parameters α_2, $\beta_2^{(q)}$, $\lambda_2^{(t)}$ and $\phi_2^{(qt)}$, and random effects b_{2i} with the same interpretation as in (3) and standard deviation σ_{2b}.

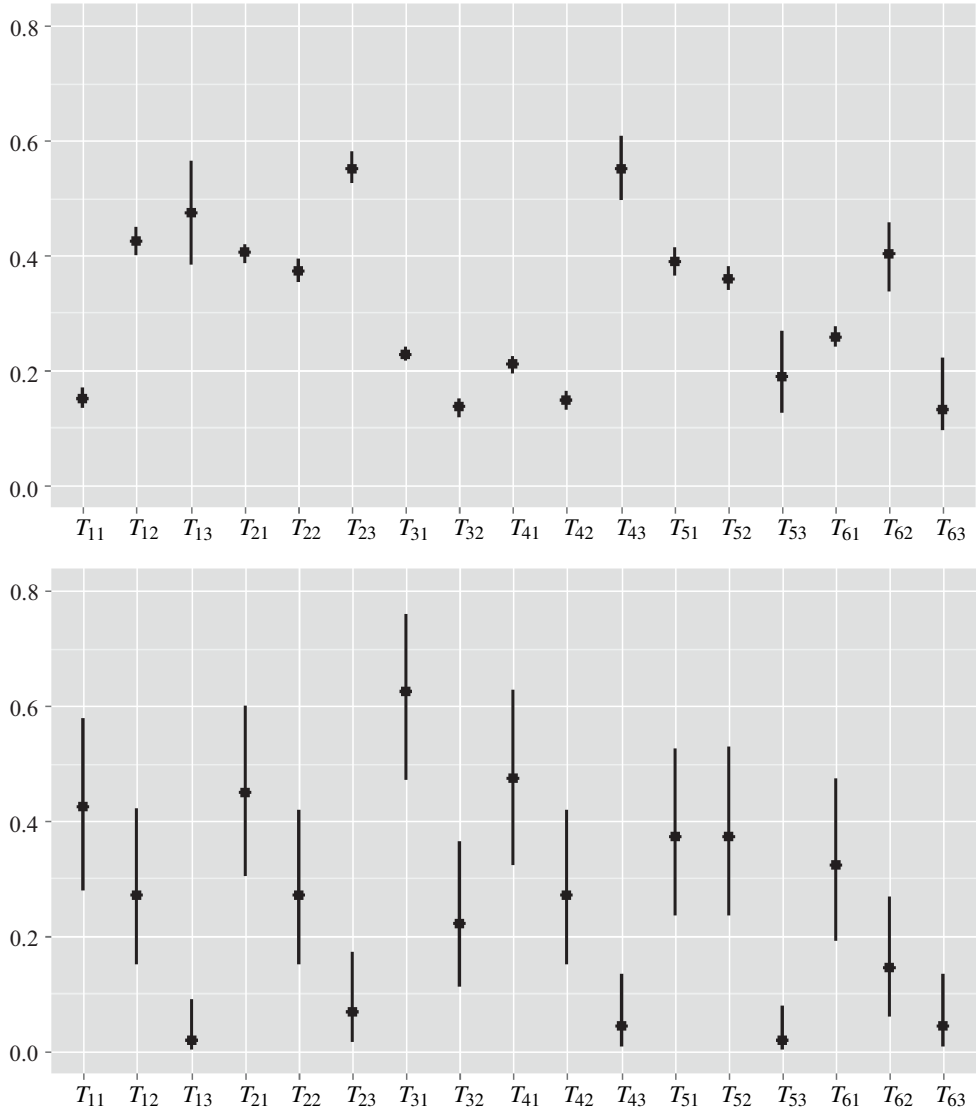

Figure 4: *Posterior mean and* 95% *credible interval for the probability of harvesting asymptomatic tubers from diseased seeds (on top) and for the probability of germination (below) with regard to the previous seed treatment.*

We assume prior independence and non-informative normal distributions for the subsequent marginal priors. In particular, we choose N(0, 10 000) prior distributions for the α's, β's, λ's, and ϕ's, and Ga$(0.01, 0.01)$ as the hyperprior for the standard deviation of the random effects. Posterior mean for the standard deviation of the random effects are 2.845 (model 3) and 0.236 (model 4). Figure 4 shows the posterior mean and a 95% credible interval for the probability of harvesting asymptomatic tubers from severely affected seeds (on top) and for the probability of germination (below). Information in both figures is with regard to the different chemical and thermal treatments considered.

Treatments T_{23} and T_{43}, both based on a temperature of 60°C, achieve the best results with regard to the probability of harvesting asymptomatic tubers. Treatment T_{31} and, to a lesser extent, T_{11}, T_{21}, and T_{41} achieve the greatest values for the probability of seed germination. None of them included thermal treatment. It is important to note the great difference between the precision of both types of intervals, as a result of the different number of trials in the binomial variables defined in models 3 and 4.

Table 4: *Posterior mean of the measure of effectiveness θ_{eqt} for thermal and chemical treatments and some given values of v (values for the best and worst treatments are in blue and red, respectively).*

v $1-v$	0.2 0.8	0.3 0.7	0.4 0.6	0.5 0.5	0.6 0.4	0.7 0.3	0.8 0.2
T_{11}	0.370	0.344	0.318	0.292	0.266	0.239	0.213
T_{12}	0.309	0.327	0.344	0.362	0.379	0.397	0.414
T_{13}	0.116	0.161	0.206	0.251	0.296	0.341	0.386
T_{21}	0.455	0.460	0.465	**0.470**	**0.475**	**0.480**	**0.485**
T_{22}	0.300	0.314	0.328	0.341	0.355	0.369	0.382
T_{23}	0.176	0.226	0.277	0.327	0.378	0.428	0.479
T_{31}	**0.549**	**0.509**	**0.470**	0.431	0.392	0.352	0.313
T_{32}	0.205	0.194	0.183	0.172	0.161	0.151	**0.140**
T_{41}	0.430	0.406	0.383	0.360	0.337	0.314	0.291
T_{42}	0.254	0.245	0.236	0.227	0.218	0.209	0.201
T_{43}	0.156	0.208	0.261	0.313	0.366	0.418	0.471
T_{51}	0.374	0.375	0.375	0.376	0.377	0.378	0.378
T_{52}	0.362	0.357	0.352	0.348	0.343	0.338	0.333
T_{53}	**0.059**	**0.075**	**0.092**	**0.108**	**0.125**	**0.141**	0.158
T_{61}	0.317	0.312	0.308	0.304	0.299	0.295	0.291
T_{62}	0.188	0.206	0.225	0.243	0.262	0.280	0.299
T_{63}	0.087	0.105	0.123	0.142	0.160	0.178	0.196

3.3.2. Dealing with effectiveness

Chemical and thermal treatments provide antagonistic outputs. Thermal treatments produce good results regarding the incidence of the disease in exchange for a considerable reduction of the probability of germination. Chemical results are not so evident. Follow-

ing the spirit of mixture models (Marin et al., 2005), we define a measure of effectiveness associated to a given combination of treatments (q,t) that weights the incidence of the disease $\theta_1^{(qt)}$ and the probability of germination $\theta_2^{(qt)}$

$$\theta_e^{(qt)} = v\,\theta_1^{(qt)} + (1-v)\,\theta_2^{(qt)}, \tag{5}$$

where v, $0 \leq v \leq 1$, is the weighting constant. This measure of effectiveness $\theta_e^{(qt)}$ is simple, sensible, easy to understand, and apply to take decisions in disease management programs.

Table 4 shows the posterior mean of $\theta_e^{(qt)}$ for each treatment and some elicited values of v. When priority is germination ($v \leq 0.5$), the most effective treatment is T_{31}. If priority is achieving a great proportion of asymptomatic tubers ($v \geq 0.5$), the best option will be T_{21}. The worst results (no matter the value of v) are for T_{53}. Another important information is that thermal treatments, at 55 °C and 60 °C, drastically reduced germination. In the case of a balanced decision ($v = 0.5$), the best and worst options are treatments T_{21} and T_{53}, respectively.

4. Conclusions

We have used data from two experimental studies designed to analyse the transmission of black spot disease in tigernuts and the effectiveness of different chemical and thermal treatments to control its incidence. Statistical methods include linear mixed models, Dirichlet-multinomial inferential processes and logistic mixed regression models.

The disease seems not to affect the size of the harvested tubers. In addition, it seems practically impossible to harvest asymptomatic tubers from severely affected seeds and highly likely to obtain severely affected tubers. In the case of asymptomatic seeds, about 70% of the tubers remained symptomless, whereas the rest were distributed between mild and severe symptoms with 23% and 7% approximately. It seems important to select asymptomatic seeds to minimize the disease incidence.

Germination and transmission of the disease from seeds to tubers have been discussed for several procedures which combine chemical and thermal treatments in seeds before they are sown. We propose a measure of effectiveness for treatments which allow to balance probability of germination and disease incidence. The results indicate the bad performance of thermal treatments for germination. This is probably due to the high temperature levels considered, thus suggesting the need to perform other experiments with a larger range of temperature levels.

The study also showed that the best chemical treatments when prioritizing germination is hydrochloric acid while sodium hypochlorite performs better if the priority is to have a reduced disease incidence. The low efficacy of the broad-spectrum fungicide treatment suggests that the causal agent of the black spot disease is not a fungus (al-

though some fungi can be resistant to this fungicide). However, the hypothesis that the syndrome of black spots is caused by a pathogenic organism is supported by the disease incidence reduction after seed disinfection with several chemical agents. This is an interesting result that could address future experimental studies about the subject.

5. Acknowledgments

Alvares's work was supported by Coordination for the Improvement of Higher Level Personnel (BEX: 0047/13–9), Brazil. Armero and Forte's work was supported by Grant MTM2013–42323–P from the Spanish Ministry of Economy and Competitiveness, and ACOMP/2015/202 from the Generalitat Valenciana. Experimental work was supported by grant 5425 from Valencian Institute for Agricultural Research to L. Rubio. The authors are very grateful to the editor and referees for their valuable suggestions and insights.

References

Agresti, A. (2013). *Categorical Data Analysis.* 3rd edition. John Wiley and Sons.

Alvares, D. (2015). Distribuciones previas objetivas para el modelo Dirichlet-multinomial: una aplicación en la agricultura. *Master's thesis*, University of Valencia, Spain.

Berger, J. O., Bernardo, J. M. and Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, 10, 189–221.

García-Jiménez, J., Busto, J., Vicent, A. and Armengol, J. (2004). Control of *Dematophora necatrix* on *Cyperus esculentus* tubers by hot-water treatment. *Crop Protection*, 23, 619–623.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013). *Bayesian Data Analysis.* Texts in Statistical Science, 3th edition. Chapman and Hall/CRC.

Good, I. (1965). *The Estimation of Probabilities: an Essay on Modern Bayesian Methods.* 1st edition. MIT Press.

Grondeau, C., Samson, R. and Sands, D. (1994). A review of thermotherapy to free plant materials from pathogens, especially seeds from bacteria. *Critical Reviews in Plant Sciences*, 13, 57–75.

Kass, R.E., Carlin, B.P., Gelman, A. and Neal, R.M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. *The American Statistician*, 52, 93–100.

Lindley, D. (1964). The Bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, 35, 1622–1643.

Liu, I. and Agresti, A. (2005). The analysis of ordered categorical data: an overview and a survey of recent developments. *Test*, 14, 1–73.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Marin, J.M., Mengersen, K.L. and Robert, C. (2005). *Bayesian Modelling and Inference on Mixtures of Distributions.* D. Dey and C. Rao (editors), Handbook of Statistics, volume 25. Elsevier.

Morell, J. and Barber, S. (1983). Chufa y horchata: características físicas, químicas y nutritivas. *Technical report*, Institute of Agrochemistry and Food Technology, Valencia, Spain.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97–126.

Ntzoufras, I. (2009). *Bayesian Modeling using WinBUGS.* John Wiley and Sons.

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, 73, 285–334.

Sauer, D.B. and Burroughs, R. (1986). Disinfection of seeds surfaces with sodium hypochlorite. *Phytopathology*, 76, 745–749.

Shepard, J.F. and Claftin, L.E. (1975). Critical analyses of the principles of seed potato certification. *Annual Review of Phytopathology*, 13, 271–293.

SORT 40 (1) January-June 2016, 153-176

Transmuted geometric distribution with applications in modelling and regression analysis of count data

Subrata Chakraborty[1,*] and Deepesh Bhati[2]

Abstract

A two-parameter transmuted geometric distribution is proposed as a new generalization of the geometric distribution by employing the quadratic transmutation techniques of Shaw and Buckley. The additional parameter plays the role of controlling the tail length. Distributional properties of the proposed distribution are investigated. Maximum likelihood estimation method is discussed along with some data fitting experiments to show its advantages over some existing distributions in literature. The tail flexibility of density of aggregate loss random variable assuming the proposed distribution as primary distribution is outlined and presented along with a illustrative modelling of aggregate claim of a vehicle insurance data. Finally, we present a count regression model based on the proposed distribution and carry out its comparison with some established models.

MSC: 62E15

Keywords: Aggregate claim, count regression, geometric distribution, transmuted distribution.

1. Introduction

A random variable (rv) X follows the geometric distribution with parameter q, denoted by $\mathscr{GD}(q)$ (see Johnson et al., 2005), pp. 210, equation (5.8)) if its probability mass function(pmf) is given by

$$P(X=t) = pq^t, \qquad t = 0,1,2,\cdots, 0 < q < 1, p = 1-q \tag{1}$$

* Corresponding Author: subrata_arya@yahoo.co.in

[1] Department of Statistics, Dibrugarh University, Dibrugarh-786004, Assam, India.

[2] Department of Statistics, Central University of Rajasthan, Ajmer-305817, Rajasthan, India, deepesh.bhati@curaj.ac.in

Received: August 2015

Accepted: April 2016

For the geometric distribution in (1) the cumulative distribution function (cdf) and survival function (sf) are respectively given by

$$F_X(t) = 1 - q^{t+1} \quad \text{and} \quad S_X(t) = P(X \geq t) = q^t.$$

In last few decades, many generalizations of geometric distribution were attempted by researchers by using different methods, for example, see Jain and Consul (1971), Philippou et al. (1983), Tripathi et al. (1987), Makčutek (2008), Gómez (2010), Chakraborty and Gupta (2015), Sastry et al. (2014) and references therein.

The transmutation, in particular the quadratic rank transmutation(QRT) method first introduced by Shaw and Buckley in 2007 has been used by many researchers to generate a large number of new distributions staring with suitable continuous baseline distributions (see Owoloko et al., 2015, Oguntunde and Adejumo, 2015 and Yousof et al., 2015 for details). It is an interesting way of generating a new and more flexible distribution by adding an additional parameter (α) to a baseline distribution. The QRT method produces a new family distribution that can be seen as a mixture of the maximum and minimum order statistics for a sample of size two from the baseline distribution and also as a mixture of the baseline distribution and its exponentiated version with power parameter two. The new family allows a continuum of distributions in the range of the additional parameter $(-1 < \alpha < 1)$. This method is applicable to any type of baseline distribution like symmetric, centred, and defined over $\mathbb{Z}$; provides explicit expression of the cdf, moments for new distribution through those of baseline distribution; and is suitable for simulation through the quantile function of the baseline distribution. Because of the many properties possessed by the method a significant amount of work to develop new flexible continuous distributions by transmutation method has been published in the last few years. The motivation of the present article is to derive a more flexible extension of the geometric distribution by application of the QRT method. The choice of QRT method is not just for its many attractive properties but also due to the fact that so far there is no evidence of any attempt to use transmutation method to generate new discrete distribution.

Accordingly, in this article an attempt is made to derive a new generalization of geometric distribution with two parameters $0 < q < 1$ and $-1 < \alpha < 1$ by using the QRT method of Shaw and Buckley (2007), which is presented in Section 2. Some distributional properties like unimodality, generating function, moments, quantile function are discussed in Section 3. A discussion on the maximum likelihood estimation (MLE) of parameters is presented in Section 4. Finally, in Section 5, applications of the proposed distribution in modelling aggregate claim size data, claim frequency data and in count data regression are presented.

2. A new generalization of geometric distribution

Here we first briefly discuss the QRT method and then propose the new transmuted geometric distributon.

2.1. Quadratic rank transmutation

The general rank transmutation mapping proposed by Shaw and Buckley (2007) for given pair of cdfs F_1 and F_2 having same support is defined as $G_{R12}(u) = F_2\left(F_1^{-1}(u)\right)$ and $G_{R21}(u) = F_1\left(F_2^{-1}(u)\right)$ where $F^{-1}(u)$ is the quantile function corresponding to the cdf $F(u)$. Both $G_{R12}(u)$ and $G_{R21}(u)$ map the unit interval in to itself. In particular, the quadratic rank transmutation (QRT) mapping is defined by $G_{R12}(u) = u + \alpha u(1-u)$. This implies

$$F_2\left(F_1^{-1}(u)\right) = u + \alpha u(1-u) = (1+\alpha)u - \alpha u^2 \Rightarrow F_2(x) = (1+\alpha)F_1(x) - \alpha F_1(x)^2$$

A discrete rv Y with cdf $F_Y(.)$ and pmf $P(Y=y)$ is said to be constructed by the QRT method of Shaw and Buckley (2007) by transmuting another discrete rv X with cdf $F_X(.)$ and pmf $P(X=x)$, if

$$\begin{aligned} F_Y(y) &= (1+\alpha)F_X(y) - \alpha F_X(y)^2 \text{and} \\ P(Y=y) &= (1+\alpha-2\alpha F_X(y))P(X=y) + \alpha\left(P(X=y)\right)^2 \end{aligned} \tag{2}$$

The distribution F_Y is then refereed to as the transmuted-F_X. In particular, for $\alpha = 0$ it gives the parent distribution function $F_X(y)$, for $\alpha = -1$, $F_X(y)^2$ the distribution of the maximum of two iid rvs with cdf $F_X(x)$, and for $\alpha = 1$, $2F_X(y) - F_X(y)^2$ the distribution of the minimum of two iid rvs with cdf $F_X(x)$.

Mirhossaini and Dolati (2008), expressing the cdf in (2) as $F_Y(y) = F_X(y)(1+$ $+\alpha\bar{F}_X(y))$ where $\bar{F}_X(y) = 1 - F_X(y)$, viewed it as a univariate counterpart of the Farlie-Gumbel-Morgenstern family (see Drouet-Mari and Kotz (2001)) of bivariate cdf $H_{XY}(x,y)$ generated from two independent univariate cdfs $F_X(x)$ and $F_Y(y)$ by the formula $H_{XY}(x,y) = F_X(x)F_Y(y)\left(1+\alpha\bar{F}_X(x)\bar{F}_Y(y)\right), -1 < \alpha < 1$.

Kozubowski and Podgórski (2016) in a very recent paper have shown that the transmuted-F_X distribution can be seen as the distribution of maxima(or minima) of a random number N of iid rvs with the base distribution $F_X(x)$, where N has a Bernoulli distribution shifted up by one.

More over by rewriting the cdf in (2) as

$$F_Y(y) = \frac{1+\alpha}{2}\left(2F_X(y) - F_X(y)^2\right) + \frac{1-\alpha}{2}\left(F_X(y)\right)^2$$

it can be seen as a convex combination (finite mixture) of the cdfs of the maximum and minimum of two iid rv following $F_X(.)$. This implies $(F_X(y))^2 \leq F_X(y) \leq 2F_X(y) - (F_X(y))^2$ since $(F_X(y))^2 \leq 2F_X(y) - (F_X(y))^2$. Therefore the transmuted-F_X family provides a continuum of distributions over the range of the additional parameter $\alpha \in (-1,1)$.

2.2. Transmuted geometric distribution

Suppose an rv X has $\mathscr{GD}(q)$ in (1). Then the cdf of the transmuted geometric rv Y will be constructed as

$$\begin{aligned} F_Y(y) &= (1+\alpha)\left(1-q^{y+1}\right) - \alpha\left(1-q^{y+1}\right)^2 \\ &= 1-(1-\alpha)q^{y+1} - \alpha q^{2(y+1)}, \quad y = 0,1,2,\cdots; 0 < q < 1, -1 < \alpha < 1. \end{aligned}$$

and the corresponding pmf will then be given by

$$p_y = P(Y=y) = (1-\alpha)\,q^y(1-q) + \alpha(1-q^2)q^{2y}, \quad y = 0,1,2,\cdots. \tag{3}$$

where $0 < q < 1, -1 < \alpha < 1$. The distribution in (3) will henceforth be referred to as the *transmuted geometric distribution* $(\mathscr{TGD})$ with two parameters q and α. In short, $\mathscr{TGD}(q,\alpha)$.

Particular cases:

1. For $\alpha = 0$, (3) reduces to $\mathscr{GD}(q)$ in (1).
2. For $\alpha = -1$, (3) reduces to a special case of the exponentiated geometric distribution of Chakraborty and Gupta (2015) with power parameter equal to 2. This is the distribution of the maximum of two iid $\mathscr{GD}(q)$ rvs.
3. For $\alpha = 1$, (3) reduces to $\mathscr{GD}(q^2)$ with pmf $(1-q^2)q^{2y}$, which is the distribution of the minimum of two iid $\mathscr{GD}(q)$ rvs.

Remark 1 $\mathscr{TGD}(q,\alpha)$ forms a continuous bridge between the distributions of the minimum to maximum in a sample of size two from $\mathscr{GD}(q)$.

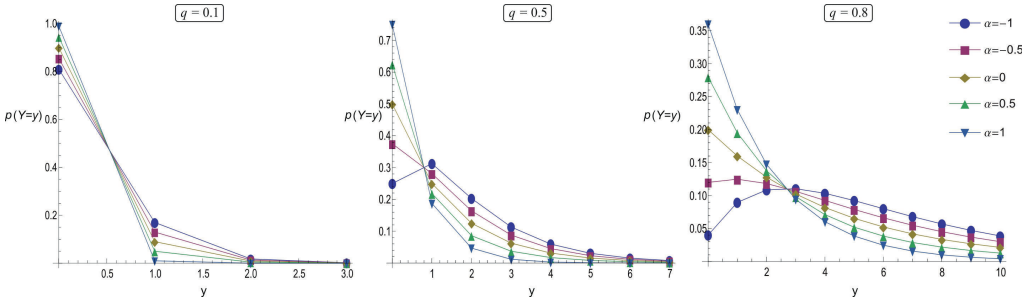

Figure 1: *PMF plot of $\mathscr{TGD}(q,\alpha)$ for different value of parameter.*

3. Distributional properties

3.1. Shape of the $\mathscr{TGD}(q,\alpha)$

The graphs of the pmf of $\mathscr{TGD}(q,\alpha)$ are plotted for various combinations of the values of the two parameters q and α in Figure 1. When $-1 < \alpha < 0$, the pmf is unimodal with either zero or non-zero mode, while for $0 \leq \alpha < 1$, the pmf is always a decreasing function with unique mode at $Y = 0$. The above assertions are mathematically established later in Section 3.3. Moreover, the spread of $\mathscr{TGD}(q,\alpha)$ increases with q and decreases with α.

Furthermore, $\mathscr{TGD}(q,\alpha)$ has at most a tail as long as $\mathscr{GD}(q)$ can be seen from the pmf plots in the Figure 1 and also from the monotonicity of the ratio of the successive probabilities(see theorem 1). The shortest tail occurs when $\alpha = 1$.

3.2. Monotonicity

Here we briefly discuss some useful monotonic properties of $\mathscr{TGD}(q,\alpha)$ and its direct consequences.

Theorem 1 *For $0 < \alpha < 1$ the $\mathscr{TGD}(q,\alpha)$ distribution with pmf given in (3), the ratio p_y/p_{y-1}, $y = 1,2,\cdots,$ forms a monotone increasing sequence.*

Proof. Firstly, we have $p_0 \neq 0, p_1 \neq 0$ and $0 < \alpha < 1$. Now

$$\mathscr{Q}(y) = p_y/p_{y-1} = \frac{(1-\alpha)(1-q)q^y + \alpha(1-q^2)q^{2y}}{(1-\alpha)(1-q)q^{y-1} + \alpha(1-q^2)q^{2(y-1)}}$$

$$= q\left(1 + \frac{\alpha(1+q)q^y}{(1-\alpha)} \Big/ 1 + \frac{\alpha(1+q)q^{y-1}}{(1-\alpha)}\right)$$

further,

$$\triangle\mathscr{Q}(y) = \mathscr{Q}(y+1) - \mathscr{Q}(y) = \frac{(1-q)^2 q^{y+1}(1+q)(1-\alpha)\alpha}{(q^2(1-\alpha)+\alpha(1+q)q^y)\,(q(1-\alpha)+\alpha(1+q)q^y)}$$

Since, for $0 < \alpha < 1$, $\mathscr{Q}(y) > 0$, therefore p_y/p_{y-1} forms a monotone increasing sequence for $0 < \alpha < 1$. ■

The following results follow as a consequence of Theorem 1. For $0 < \alpha < 1$, $\mathscr{TGD}(q,\alpha)$

i. is infinitely divisible (see Warde and Katti, 1971).

ii. pmf is a decreasing sequence (see Johnson and Kotz, 2005 p.75), which in turn indicates that, $\mathscr{TGD}$ has a zero vertex (see Warde and Katti, 1971). This fact was also mentioned in Remark 3.

iii. is DFR(decreasing failure rate), which in turn implies IMRL(increasing mean residual life).

iv. an upper bound for the variance of the $\mathscr{TGD}(q,\alpha)$ can be obtained for $0 < \alpha < 1$ as

$$Var(Y) \geq \frac{p_1}{p_0} = \frac{q(1-\alpha)+\alpha q^2(1+q)}{1-\alpha+\alpha(1+q)}$$

Corollary 1 *For $-1 < \alpha < 0$, $\mathscr{TGD}(q,\alpha)$ distribution with pmf given in (3) is log-concave.*

Proof. The result follows from that fact that p_y/p_{y-1}, $y = 1,2,\ldots,$ forms a monotone decreasing sequence for $-1 < \alpha < 0$ that is $p_{y+1}/p_y < p_y/p_{y-1} \Rightarrow p_y^2 > p_{y-1}p_{y+1}\ \forall y.$ ■

The following results follow as a consequence of corollary 1: For $-1 < \alpha < 0$, $\mathscr{TGD}(q,\alpha)$ distribution is

i. IFR (increasing failure rate), which in turn implies DMRL (decreasing mean residual life).

ii. Strongly unimodal.

iii. At most has a geometric tail.

3.3. Mode

Theorem 2 *$\mathscr{TGD}(q,\alpha)$ is unimodal with a nonzero mode for $-1<\alpha<-(q(2+q))^{-1}$ provided that $q>0.414$.*

Proof. A pmf $P(Y=y)$ with support $y=0,1,2,\cdots$, is uni modal if there exists a unique point $M(\neq 0)$, in the support of Y such that $P(Y=y)$ is increasing on $(0,1,\ldots,M)$ and decreasing on $(M,M+1,\ldots)$. M is then the unique mode of $P(Y=y)$. Thus $\mathscr{TGD}(q,\alpha)$ will have a non zero mode if,

$$\begin{aligned}
&P(Y=1)>P(Y=0)\\
&\Rightarrow (1-\alpha)(1-q)q+\alpha q^2(1-q^2)>(1-\alpha)(1-q)+\alpha(1-q^2)\\
&\Rightarrow (1-\alpha)(1-q)^2+\alpha(1-q^2)(1-q^2)<0\\
&\Rightarrow \alpha<-(1-q)^2/\left((1-q^2)^2-(1-q)^2\right)=-1/(q(2+q)
\end{aligned}$$

But the condition $-1<\alpha<-(q(2+q)^{-1}$ makes sense only if $q(2+q)>1$ which implies $q>\sqrt{2}-1\cong 0.414$. ∎

For example, with $q=0.8$ non zero modes occur when $-1<\alpha<-0.4464$ as can be clearly seen in the third plot of the pmfs in the Figure 1.

Remark 2 For $q<0.414$, the condition of non-zero unimodality leads to α outside its permissible range of $-1<\alpha$

Remark 3 For $0\leq\alpha\leq 1$, the pmf is decreasing, and the mode occurs at the point 0. This indicates the suitability of the proposed distribution for count data which feature, relatively, a large number of zeros. Moreover the proportion of zeros in $\mathscr{TGD}(q,\alpha)$ is more(less) than that of $\mathscr{GD}(q)$ depending on $\alpha>(<)0$.

3.4. An alternative derivation of the $\mathscr{TGD}(q,\alpha)$

Theorem 3 *$\mathscr{TGD}(q,\alpha)$ is the discrete analogue of the skew exponential distribution of Shaw and Buckley (2007).*

Proof. The pdf and cdf of the skew exponential distribution derived using the quadratic rank transmutation (Shaw and Buckley, 2007) are respectively given by

$$f_X(x)=(1-\alpha)\beta e^{-\beta x}+2\alpha\beta e^{-2\beta x},\quad x>0,\beta>0,-1<\alpha<1$$

and

$$F_X(x)=(1+\alpha)(1-e^{-\beta x})-\alpha(1-e^{-2\beta x})^2,\quad x>0,\beta>0,-1<\alpha<1.$$

Hence, the pmf of the discrete analogue (see Chakraborty, 2015, for a detail review of various methods of construction of discrete analogues of continuous distributions.) of X, $Y = \lfloor X \rfloor$, where $\lfloor X \rfloor$ is the floor function, is given by the formula $P(Y = y) = S_X(y) - S_X(y+1) = F_X(y+1) - F_X(y)$. On simplification, this reduces to the pmf of $\mathscr{TGD}(q = e^{-\beta}, \alpha)$. ■

3.5. Generating functions

Theorem 4 *The probability generating function (PGF) of $\mathscr{TGD}(q,\alpha)$ is given by*

$$G_Y(z) = \frac{(1-q)(1-\alpha q(1-z)-q^2 z)}{(1-qz)(1-q^2 z)}, \qquad |q^2 z| < 1$$

Proof. It is known that the pgf $\mathbb{E}(z^X)$ of $X \sim \mathscr{GD}(q)$ is equal to $\frac{1-q}{1-qz}$ (see p. 215, Johnson et al., 2005).

Therefore pgf of $Y \sim \mathscr{TGD}(q,\alpha)$ is given by

$$G_Y(z) = \mathbb{E}(z^Y) = \sum_{y=0}^{\infty} z^y P(Y=y) = \sum_{y=0}^{\infty} z^y \left((1-\alpha)(1-q)q^y + \alpha(1-q^2)q^{2y}\right)$$

$$= \frac{(1-q)(1-\alpha)}{1-qz} + \frac{\alpha(1-q^2)}{1-q^2 z}$$

The result follows on simplification. ■

Remark 4 The other generating functions like characteristic function, moment generating function and cumulant generating function can be easily derived from the PGF by using the results $\Phi_Y(z) = G_Y(e^{iz})$, $\mathbb{M}_Y(z) = G_Y(e^z)$ and $\mathbb{K}_Y(z) = \log(G_Y(e^z))$ respectively.

3.6. Moments and related measures

Here we derive various moments and related measures of $\mathscr{TGD}(q,\alpha)$.

Theorem 5 *The r^{th} factorial moment of $Y \sim \mathscr{TGD}(q,\alpha)$ is given by*

$$\mathbb{E}\left(Y_{(r)}\right) = (1-\alpha)r!\left(\frac{q}{1-q}\right)^r + \alpha r!\left(\frac{q^2}{1-q^2}\right)^r.$$

where $Y_{(r)} = Y(Y-1)\cdots(Y-r+1)$

Table 1: *Expressions for various measures of $\mathscr{TGD}(\alpha,q)$.*

S.No.	Measures	Expression
1	Mean $\mathbb{E}(Y)$	$\frac{q(1-\alpha)+q^2}{1-q^2}$
2	Variance $\mathbb{V}(Y)$	$\frac{q\left(1-\alpha^2+q(1-\alpha^2+q(1-\alpha)+2)\right)}{(1-q^2)^2}$
3	Index of Dispersion (ID)	$\frac{q\left(1-\alpha^2+q(1-\alpha^2+q(1-\alpha)+2)\right)}{(1-q^2)(q(1-\alpha)+q^2)}$
4	$\gamma^{th}quantile\ (y_\gamma)$	$\left\lfloor \frac{\log\left(\alpha-1+\sqrt{\alpha^2-2\alpha(1-2\gamma)+1}\right)-\log(2\alpha)}{\log q} \right\rfloor - 1$
5	Median $(y_{0.5})$	$\left\lfloor \frac{\log\left(\alpha-1+\sqrt{\alpha^2+1}\right)-\log(2\alpha)}{\log q} \right\rfloor - 1$

Proof. It is known that the r^{th} factorial moment $\mathbb{E}(X_{(r)})$ of $X \sim \mathscr{GD}(q)$ is given by

$$\mathbb{E}\left(X_{(r)}\right) = r!\left(\frac{q}{1-q}\right)^r \tag{4}$$

Therefore the r^{th} factorial moment of $Y \sim \mathscr{TGD}(q,\alpha)$ using equation (3) is given by

$$\mathbb{E}\left(Y_{(r)}\right) = (1-\alpha)(1-q)\sum_{y=r}^{\infty} y_{(r)}q^y + \alpha(1-q^2)\sum_{y=r}^{\infty} y_{(r)}q^{2y} \tag{5}$$

The result then follows upon using (4). ∎

Note 1. Alternatively, the above theorem can also be proved using the result $\mathbb{E}(Y_{(r)}) = \frac{d^r}{dz^r}G_Y(z)|_{z=1}$.

By using Theorem 5, the descriptive statistics mean, variance, index of dispersion quantile functions as well as median are given in Table 1. However, we do not present the expressions for skewness as well as kurtosis as they are quite gigantic, instead we present 3-D surface plot of these two measures in Figure 2(a) and 2(b). In Figure 2(a), the q-α surface cuts the skewness surface at zero indicated in blue, hence $\mathscr{TGD}(\alpha,q)$ possess positive skewness above q-α surface and negative skewness below q-α surface. Moreover, if we look in Figure 2(b) horizontal q-α surface drawn at value 3 which never intersect the kurtosis surface, indicating leptokurtic nature of $\mathscr{TGD}(\alpha,q)$. Further, Figure 2(c) shows that the horizontal q-α surface cuts the ID surface at 1 indicating under or

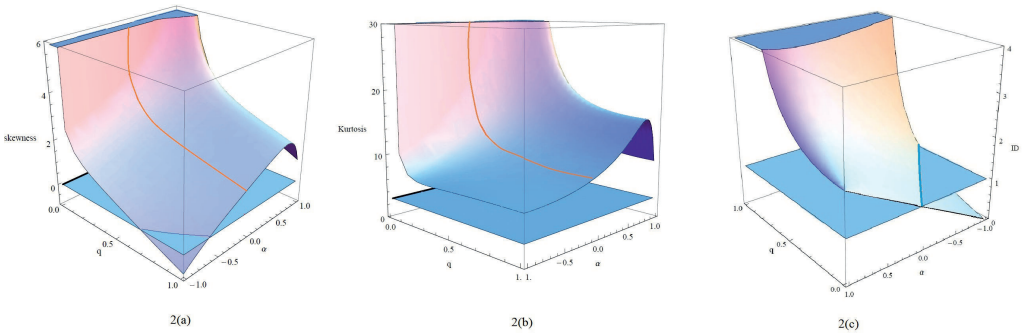

Figure 2: *q-α surface plot of 2(a) Skewness, 2(b) Kurtosis and 2(c) Index of Dispersion of $\mathscr{TGD}(q,\alpha)$.*

over-dispersion for $\alpha \in (-1,0)$ or $(0,1)$ respectively (see Remark 3). Finally skewness and kurtosis of $\mathscr{GD}(q)$ is depicted in red curve on respective surfaces.

Remark 5 A random number $Y \sim \mathscr{TGD}(q,\alpha)$ can be drawn by first generating a uniform random number U in $(0,1)$ and then using the method of inversion to get a sampled observation Y by using result 4 of Table 1.

4. Maximum likelihood estimator

In this section, we focus on the maximum likelihood estimator (MLE), though other estimators can also be derived easily, such as (i) sample proportion of 1's and 0's, (ii) sample quantiles, (iii) method of moments.

For a sample $(y_1, y_2, \cdots, y_n)$ of size n drawn from $\mathscr{TGD}(q,\alpha)$, the likelihood function is given by $L = \prod_{i=1}^{n} \left((1-\alpha)q^{y_i}(1-q) + \alpha q^{2y_i}(1-q^2)\right)$. Taking logarithms on both sides gives the log-likelihood function as

$$l = \log L = n\log(1-q) + n\bar{y}\log(q) + \sum_{i=1}^{n} \log\left((1-\alpha) + \alpha q^{y_i}(1+q)\right) \tag{6}$$

By differentiating (6) with respect to q and α and equating to 0, the following likelihood equations are obtained.

$$\frac{\partial l}{\partial q} = -\frac{n}{1-q} + \frac{n\bar{y}}{q} + \sum_{i=1}^{n} \frac{\alpha q^{y_i} + \alpha y_i(1+q)q^{y_i-1}}{1-\alpha+\alpha(1+q)q^{y_i}} = 0$$

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \frac{(1+q)q^{y_i} - 1}{1-\alpha+\alpha(1+q)q^{y_i}} = 0$$

Since the likelihood equations have no closed form solution, the MLEs $\hat{q}$ and $\hat{\alpha}$ of the parameters q and α can be obtained by maximizing the log-likelihood function using global numerical maximization techniques. Further, the second order partial derivatives of the log-likelihood function are given by

$$\frac{\partial^2 l}{\partial q^2} = -\frac{n}{(1-q)^2} - \frac{n\bar{y}}{q^2} - \sum_{i=1}^{n}\left(\frac{\alpha(1+q)(y_i-1)y_i q^{y_i-2} + 2\alpha y_i q^{y_i-1}}{1-\alpha+\alpha(1+q)q^{y_i}} - \left(\frac{\alpha(1+q)y_i q^{y_i-1} + \alpha q^{y_i}}{1-\alpha+\alpha(1+q)q^{y_i}}\right)^2\right)$$

$$\frac{\partial^2 l}{\partial q \partial \alpha} = \sum_{i=1}^{n}\left(\frac{(1+q)y_i q^{y_i-1} + q^{y_i}}{1-\alpha+\alpha(1+q)q^{y_i}} - \frac{\left(\alpha(1+q)y_i q^{y_i-1} + \alpha q^{y_i}\right)\left((1+q)q^{y_i}-1\right)}{\left(1-\alpha+\alpha(1+q)q^{y_i}\right)^2}\right)$$

$$\frac{\partial^2 l}{\partial \alpha^2} = -\sum_{i=1}^{n}\left(\frac{\left((1+q)q^{y_i}-1\right)^2}{1-\alpha+\alpha(1+q)q^{y_i}}\right)$$

The approximate Fisher information matrix can then be obtained as

$$\left(\begin{array}{cc} \frac{\partial^2 l}{\partial q^2} & \frac{\partial^2 l}{\partial q \partial \alpha} \\ \frac{\partial^2 l}{\partial q \partial \alpha} & \frac{\partial^2 l}{\partial \alpha^2} \end{array}\right)_{q=\hat{q},\alpha=\hat{\alpha}} \tag{7}$$

where $\hat{q}$ and $\hat{\alpha}$ are the MLEs of q and α respectively.

5. Application and data analysis

5.1. An actuarial application

In an actuarial context, non-life insurance companies are often interested in modelling the aggregate claim of a portfolio of policies. Let Z_j, $j = 1,2,\cdots$ be the rv denoting the size or amount of the j^{th} claim and Y be the rv denoting the number of claims. Then the aggregate claim of that portfolio is defined as $S = \sum_{j=1}^{Y} Z_j$. Assuming that the claim amounts Z_j to be identically and independently distributed among themselves as well as with claim frequency Y, the pdf of S can be obtained as $g_S(s) = \sum_{j=1}^{\infty} p_j f^{*j}(s)$ where p_j denotes the probability of the jth claim (called the primary distribution) and $f^{*j}(s)$ is the j-fold convolution of $f(s)$, the pdf of the claim amount (the secondary distribution). For more details one can see Rolski et al. (1999), Antzoulakos and Chadjiconstantinidis (2004), Klugman et al. (2008)) and the references therein.

In the following theorem, we present the distribution of aggregate claim when the primary distribution is $\mathscr{TGD}(q,\alpha)$ and the secondary distribution is exponential with mean $(1/\theta)$.

Theorem 6 *If $\mathscr{TGD}(q,\alpha)$ distribution is the primary distribution and the exponential distribution with parameter $\theta > 0$ is the secondary distribution, then the pdf of rv $S = \sum_{j=1}^{Y} Z_j$ is given by*

$$g_S(s) = \begin{cases} (1-\alpha)(1-q) + \alpha\left(1-q^2\right) & \text{for} \quad s = 0 \\ (1-q)q\theta\left((1-\alpha)e^{-(1-q)s\theta} + q(1+q)\alpha e^{-\left(1-q^2\right)s\theta}\right) & \text{for} \quad s > 0 \end{cases} \tag{8}$$

Proof. Since the claim severity distribution follows an exponential distribution with parameter $\theta > 0$, the j-fold convolution of the exponential distribution is a gamma distribution with parameter j and θ, having density function

$$f^{*j}(z) = \frac{\theta^j}{(j-1)!} z^{j-1} e^{-\theta z}, \qquad j = 1,2,\cdots,$$

Hence, the pdf of the rv S is given by

$$\begin{aligned} g_S(s) = \sum_{j=1}^{\infty} p_j f^{*j}(s) &= \sum_{j=1}^{\infty} \frac{\theta^j}{(j-1)!} s^{j-1} e^{-\theta s}\left((1-\alpha)(1-q)q^j + \alpha\left(1-q^2\right)q^{2j}\right) \\ &= (1-q)q\theta\left((1-\alpha)e^{-(1-q)s\theta} + q(1+q)\alpha e^{-\left(1-q^2\right)s\theta}\right) \end{aligned}$$

where $g_S(s)$ has a jump at $s = 0$ with probability $(1-\alpha)(1-q) + \alpha(1-q^2)$. ∎

Henceforth, we denote the distribution of S with $\mathscr{TGD}(q,\alpha)$ as primary and exponential as secondary distribution as $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$. Further, it is also well-known that the mean of the aggregate rv is the product of the respective means of the primary and secondary rvs, hence in our proposed model

$$\mathbb{E}(S) = \frac{q(1-\alpha)+q^2}{1-q^2}\frac{1}{\theta}$$

We now compare the aggregate loss model as defined in (8) with the aggregate loss model obtained by considering the geometric distribution as the primary distribution and exponential as the secondary distribution for claim severity, hence the density of

the compound geometric-exponential distribution $\mathscr{CG}$-$\mathscr{ED}$ (see pp.152 of Tse, 2009) is given as

$$g_S(s) = \begin{cases} 1-q_1 & \text{for} \quad s=0 \\ (1-q_1)\,q_1\theta e^{-(1-q_1)s\theta} & \text{for} \quad s>0 \end{cases} \tag{9}$$

with mean $\mathbb{E}(X) = \frac{1-q_1}{q_1}\frac{1}{\theta}$.

It is a well known that in the case of reinsurance, the reinsurance company will be interested in those aggregate claim models that are suitable for modelling extreme value. In the following theorem we show that with the same mean and different parameter values, $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$ captures heavy tail values as compared to $\mathscr{CG}$-$\mathscr{ED}(q_1,\theta)$.

Theorem 7 *With the same mean, $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$ has thinner (thicker) tail as compared to $\mathscr{CG}$-$\mathscr{ED}(q_1,\theta)$ for $-1<\alpha<0 (0<\alpha<1)$.*

Proof. Without loss of generality, we consider $\theta = 1$. By equating the means of $\mathscr{CTG}$-$\mathscr{ED}$ with $\mathscr{CG}$-$\mathscr{ED}$, we get

$$\frac{q(1-\alpha)+q^2}{1-q^2} = \frac{1-q_1}{q_1} \qquad \text{which gives} \qquad q_1 = \frac{1-q^2}{1+q(1-\alpha)}.$$

We now compare the tail behaviour of two distributions by taking the limiting ratio (LR) of their sf (see pp. 60, Tse, 2009):

$$LR = \lim_{t\to\infty} \frac{\bar{G}_{CTG\text{-}ED}(t)}{\bar{H}_{CG\text{-}ED}(t)}$$

where $\bar{G}_{CTG\text{-}ED}(t) = q\left((1-\alpha)e^{-(1-q)t} + \alpha q e^{-(1-q^2)t}\right)$ and $\bar{H}_{CG\text{-}ED}(t) = \frac{q(q+1-\alpha)}{1+q(1-\alpha)}$ $\exp[-\frac{(1-q^2)t}{1+q(1-\alpha)}]$ are respectively the sf of $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$ and $\mathscr{CG}$-$\mathscr{ED}(q_1,\theta)$.

Substituting these values in LR, we obtain

$$LR = \lim_{t\to\infty}\left((1-\alpha)e^{\frac{\alpha q(1-q)t}{1+q(1-\alpha)}} + \alpha q e^{-\frac{(1-\alpha)q(1-q^2)t}{1+q(1-\alpha)}}\right)$$

Now observe that for $-1<\alpha<0$, $LR = \lim_{t\to\infty}\frac{\bar{G}_{CTG\text{-}ED}(t)}{\bar{H}_{CG\text{-}ED}(t)} = 0$.

$\Rightarrow \mathscr{CTG}$-$\mathscr{ED}$ has thinner tail then $\mathscr{CG}$-$\mathscr{ED}$.

whereas for $0<\alpha<1$, $LR = \lim_{t\to\infty}\frac{\bar{G}_{CTG\text{-}ED}(t)}{\bar{H}_{CG\text{-}ED}(t)} = \infty$.

$\Rightarrow \mathscr{CTG}$-$\mathscr{ED}$ has thicker tail then $\mathscr{CG}$-$\mathscr{ED}$. ∎

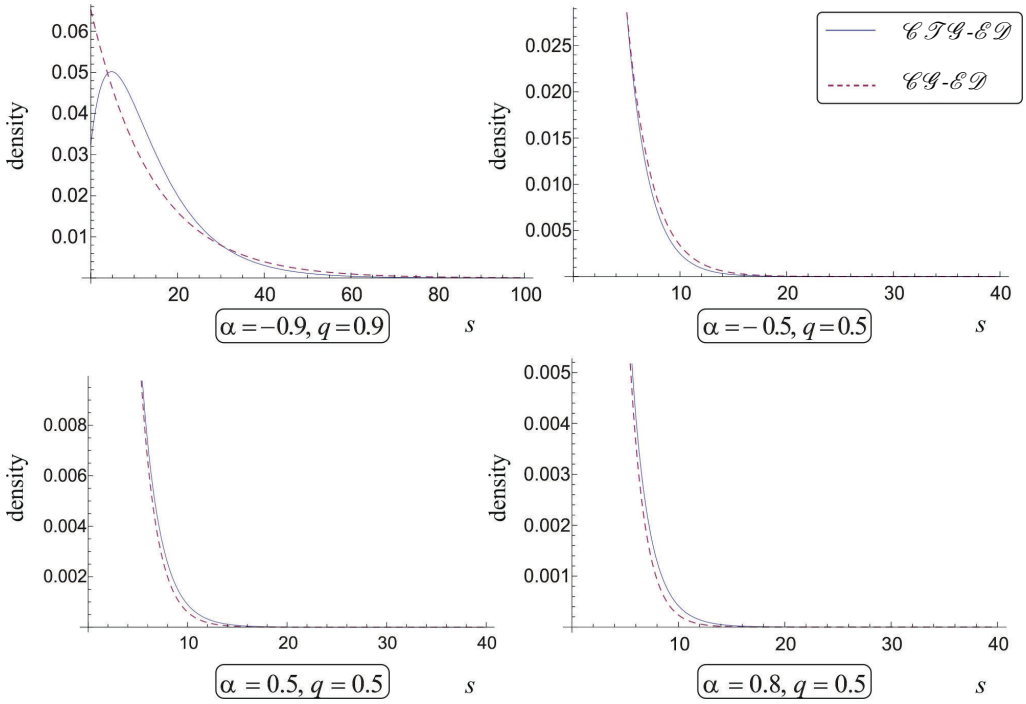

Figure 3: *PDF of aggregate loss rv (compound geometric-exponential distribution in red dashed lines and compound transmuted geometric-exponential distribution in blue lines) for different values of parameter q and α.*

Tail behaviour of $\mathscr{CTG}$-$\mathscr{ED}$ and $\mathscr{CG}$-$\mathscr{ED}$ distributions for different parameter values are presented in Figure 3.

5.1.1. Illustration: aggregate loss modelling

To illustrate the applicability and superiority of the proposed aggregate model compared to other existing aggregate models such as Poisson-exponential, negative binomial-exponential and geometric-exponential, in short X-exponential models having densities indicated in Table 2, we consider a vehicle insurance data set of one-year vehicle insurance policies taken out in 2004 or 2005. There are 67856 policies of which 4624 (6.8%) had at least one claim. Table 3 gives some in-depth information about the claims frequency (X) and total claim(S) for the data set. Full access to this dataset is available on the webiste of the Faculty of Business and Economics, Macquarie University, Australia – see also Jong and Heller (2008). As the variability in total claim data is very high, we scale these observations by scale factor 0.001, remembering the fact that scaling will not effect the comparison, and apply the maximum likelihood method to estimate the parameters of aggregate model. The log-likelihood function for proposed $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$ model is given as

$$l = (n-m)\log(\theta(1-q)q) + m\log\left(\alpha\left(1-q^2\right) + (1-\alpha)(1-q)\right) + \sum_{s_i>0}\log\left((1-\alpha)e^{-\theta(1-q)s_i} + \alpha q(q+1)e^{-\theta(1-q^2)s_i}\right)$$

where m is the number of policies having no claim, $(n-m)$ is the number of policies having at least one claim and n be the total number of policies. As we can see the log-likelihood equations obtained from the log-likelihood function cannot help in determining the estimates of parameter, hence we make use of numerical techniques to search global maximum of log-likelihood surface. We make use of `FindMaximum` function of Mathematica software package v.10.0. The estimates and other comparative measures such as log-likelihood value(LL), Akaike Information Criteria(AIC) are shown in Table 4. Based on the AIC value it can be claimed that the proposed $\mathscr{CTG}$-$\mathscr{ED}(q,\alpha,\theta)$ model gives the best fit for the vehicle insurance data among all the models considered.

Table 2: *Density of X-exponential models.*

S.No.	distribution of X	Density of aggregate rv.
1	Poisson	$g_S(s) = \begin{cases} e^{-\lambda} & \text{for} \quad s=0 \\ \sqrt{\frac{\theta\lambda}{s}}e^{-\theta s-\lambda}\,\mathscr{I}_1\left(2\sqrt{\lambda\theta s}\right) & \text{for} \quad s>0 \end{cases}$ where, $\mathscr{I}_1(.)$ is the modified Bessel function of first kind
2	Negative binomial	$g_S(s) = \begin{cases} (1-q)^r & \text{for} \quad s=0 \\ q\theta r(1-q)^r e^{-\theta s}\,{}_1F_1(r+1;2;\theta q s) & \text{for} \quad s>0 \end{cases}$ where ${}_1F_1(.;.;.)$ is the confluent hypergeometric function
3	Geometric	$g_S(s) = \begin{cases} 1-q & \text{for} \quad s=0 \\ (1-q)\,q\theta e^{-(1-q)s\theta} & \text{for} \quad s>0 \end{cases}$

Table 3: *Descriptive statistics of the vehicle insurance dataset.*

	Number of claims	Total claim amount
Mean	0.072	137.27
variance	0.077	1115769.69
Index of Dispersion	1.0734	8128.29
min	0	0
max	4	55922.1

Table 4: *Estimated value of parameters of X-exponential models.*

S.No.	Distribution of X	Estimated parameter	LL	AIC
1	Poisson	$\hat{\lambda} = 0.12057$, $\hat{\theta} = 0.87832$	-25699.3	51402.6
2	Negative binomial	$\hat{r} = 0.51168$, $\hat{q} = 0.1291$, $\hat{\theta} = 0.55250$	-24740.6	49487.2
3	Geometric	$\hat{q} = 0.06814$, $\hat{\theta} = 0.53273$	-24745.7	49495.4
4	Transmuted Geometric	$\hat{q} = 0.2313$, $\hat{\alpha} = 0.9147$, $\hat{\theta} = 0.5693$	-24702.0	49410.0

5.2. Count data modelling

In this section we demonstrate the utility of $\mathscr{TGD}(q,\alpha)$ in count data modelling considering a real data set on the number of automobile insurance claims per policy in portfolios from Great Britain and Zaire (Willmot, 1987). This data set contain 87% of zeros as well as with variance to mean ratio 1.051 indicating the presence of over-dispersion in the data set. Hence the proposed model is expected to provide adequate fit. Here $\mathscr{TGD}(q,\alpha)$ is compared with the following existing ones.

i. Negative binomial ($\mathscr{NB}$) (Johnson et al., 2005).

ii. Poisson inverse Gaussian (Willmot, 1987) ($\mathscr{P}-\mathscr{IG}$) with pmf defined as

$$P(X=x) = \frac{1}{x!}\sqrt{\frac{2\phi}{\pi}}e^{\phi/\mu}\phi^{-\frac{1}{4}+\frac{x}{2}}\left(2+\frac{\phi}{\mu^2}\right)^{\frac{1-2x}{4}}K_{\frac{1}{2}-x}\left(\sqrt{2\phi+\frac{\phi^2}{\mu^2}}\right)$$

where $x = 0,1,2,\ldots$, $\phi,\mu > 0$ and $K_a(.)$ is modified Bessel function of the third kind.

iii. New discrete distribution (Gómez et al., 2011) ($\mathscr{ND}$) with pmf

$$P(X=x) = \frac{\log(1-\alpha\theta^x) - \log(1-\alpha\theta^{x+1})}{\log(1-\alpha)}$$

where $x = 0,1,2,\ldots$, $\alpha < 1, 0 < \theta < 1$, and

iv. Zero distorted generalized geometric (Sastry et al., 2014) ($\mathscr{ZDGGD}$) with pmf

$$P(X=x) = \begin{cases} 1-q^{\alpha+1} & \text{if} \quad x=0 \\ (1-q)q^{\alpha+x+1} & \text{if} \quad x>0 \end{cases}$$

where $0 < q < 1, -1 < \alpha < 1$.

Table 5: *Fit of automobile claim data in Great Britain, 1968 (Willmot, 1987).*

# claims	Observed	Expected frequency				
	Frequency	$\mathscr{NB}$	$\mathscr{P}-\mathscr{IG}$	$\mathscr{ND}$	$\mathscr{ZDGGD}$	$\mathscr{TGD}$
0	370412	370438.99	370435	370413	370412	370412
1	46545	46451.28	46476.4	46538.3	46555.16	46546.7
2	3935	4030.50	3995.76	3942.39	3913.70	3929.19
3	317	297.82	307.67	318.57	329.00	323.23
4	28	20.09	23.12	25.64	27.76	26.53
5	3	1.28	1.74	2.06	2.38	2.38
Total	421240	421240	421240	421240	421240	421240
estimated		$\hat{p}=0.338$	$\hat{\phi}=0.338$	$\hat{\alpha}=-1.349$	$\hat{q}=0.0845$	$\hat{q}=0.0821$
parameter		$\hat{r}=0.131$	$\hat{\mu}=0.131$	$\hat{\theta}=0.080$	$\hat{\alpha}=-0.146$	$\hat{\alpha}=-0.5121$
χ^2-statistic		9.15	2.74	0.71	0.72	**0.31**
df		3	3	3	3	3
p-value		0.03	0.43	0.87	0.87	**0.96**
$l_{\max}$		–171136.9	–171134.4	–171133.0	–171134.1	**–171133.0**

Table 6: *SE, CI, and CL of estimated parameters for the data sets in Table 5.*

Models	Parameters	ML Estimate	S.E.	CI	CL
$\mathscr{NB}$	$\hat{r}$	0.131	0.5684	(–0.983, 1.255)	0.2228
	$\hat{p}$	0.338	0.0011	(0.336, 0.340)	0.0039
$\mathscr{P}-\mathscr{IG}$	$\hat{\phi}$	0.338	0.0188	(0.3017, 0.3756)	0.0739
	$\hat{\nu}$	0.131	0.0005	(0.1306,0.1328)	0.0022
$\mathscr{ND}$	$\hat{\alpha}$	–1.349	0.1120	(–1.5686, –1.1295)	0.4390
	$\hat{\theta}$	0.080	0.0018	(0.0768, 0.0840)	0.0071
$\mathscr{ZDGGD}$	$\hat{q}$	0.0845	0.0011	(0.0817, 0.0863)	0.0046
	$\hat{\alpha}$	–0.146	0.0051	(–0.1160, –0.1359)	0.0200
$\mathscr{TGD}$	$\hat{q}$	0.0821	0.0011	(0.079, 0.0844)	0.0046
	$\hat{\alpha}$	–0.5121	0.0236	(–0.558, –0.465)	0.0920

The data fitting results for the above four distributions in (i) to (iv) presented in Table 5 are taken from the respective papers. From the findings of the data fitting presented in Table 5, to assess the fit of the competing models we first compare the expected frequencies with the observed one for each model, which reveals that the $\mathscr{TGD}(q,\alpha)$ predicts most of the observed counts more closely than the other models. The χ^2 statistics and its p-values implies lack of fit for NB and also for PIG. The rest of the models provides good fit, with $\mathscr{TGD}(q,\alpha)$ being the best among the lot with highest with p-value of 0.96. Moreover, we also compute standard error (SE), confidence interval (CI) and confidence length (CL) for the parameter estimates. It can be clearly seen from Table 6, that the SE of the estimates of proposed distribution is smaller compared to

other distributions. Hence, it is envisaged that the proposed distribution may serve as an alternative model for modelling data with a large proportion of zeros and over-dispersion.

5.3. Count regression modelling including covariates

In this section, we present the count regression modelling assuming the discrete response variable (Y) as a function of a set of independent (exogenous) variables. Furthermore, we also consider that the mean (θ) of response variable is related with the set of exogenous variables by the positive valued function $\theta = \theta(\boldsymbol{x})$. There are several possible choices for the selection of function $\theta(\boldsymbol{x})$ and thus to ensure the non-negativity of the mean of the response variable, we consider the log-link function as $\theta_i(\boldsymbol{x}) = e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}$, where $\boldsymbol{x}_i^{\mathsf{T}} = (x_{i1}, x_{i2}, \cdots, x_{ip})$ and $\boldsymbol{\beta}^{\mathsf{T}} = (\beta_1, \beta_2, \cdots, \beta_p)$ be the set of covariates and their coefficients. This selection of log-link function includes both random and fixed effects on the same exponential scale. Further, to estimate the parameters, we use following reparametrization

$$\nu = 1 - \alpha \quad \text{and} \qquad q = \left(-\nu + \sqrt{4\theta + 4\theta^2 + \nu^2}\right)/2(1+\theta)$$

where $\theta_i(\boldsymbol{x}) = e^{\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta}}$. The above re parametrization enable us to bring the regression coefficients $(\boldsymbol{\beta})$ and parameters of the response variable into the log-likelihood functions. The log-likelihood function for a random sample $(y_i, \boldsymbol{x}_i)$ of size n with count y_i and a vector $\boldsymbol{x}_i$ of covariates for $i = 1, 2, \cdots, n$ can be written as

$$l(\nu, \theta | y, \boldsymbol{x}) = \sum_{i=1}^{n} \log \left(\nu \left(1 - \frac{-\nu + \sqrt{4\theta_i + 4\theta_i^2 + \nu^2}}{2(1+\theta_i)} \right) \left(\frac{-\nu + \sqrt{4\theta_i + 4\theta_i^2 + \nu^2}}{2(1+\theta_i)} \right)^{y_i} \right.$$
$$\left. + (1-\nu) \left(1 - \left(\frac{-\nu + \sqrt{4\theta_i + 4\theta_i^2 + \nu^2}}{2(1+\theta_i)} \right)^2 \right) \left(\frac{-\nu + \sqrt{4\theta_i + 4\theta_i^2 + \nu^2}}{2(1+\theta_i)} \right)^{2y_i} \right)$$

The parameters $(\nu, \beta_1, \beta_2, \cdots, \beta_p)$ in the above log-likelihood function can be estimated by maximizing the log-likelihood function for a given data set using the `optim()` function in R (for more details one can browse `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/optim.html`), where the initial values of the parameters were chosen from Poisson regression model.

In the next section we present an application of the proposed count regression model to a real life data set and compare its performance with following popular regression models:

i. Poisson regression model

$$P(Y_i = y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \qquad y_i = 0,1,2,... \tag{10}$$

where $\mu_i > 0$. The regression model is obtained by putting $\mu_i = e^{x_i^\top \beta}$.

ii. Generalized Poisson model ($\mathscr{GP}$-2): The pmf of a generalized Poisson ($\mathscr{GP}$-2) regression model (Consul and Famoye , 1992, Yang et al., 2009) is given as

$$P(Y_i = y_i|\theta_i,\nu_i) = \frac{\mu_i(\mu_i+\phi\mu_i y_i)^{y_i-1}}{(1+\phi\mu_i)^{y_i}y_i!}e^{\left(-\frac{\mu_i+\phi\mu_i y_i}{1+\phi\mu_i}\right)}, \quad y_i = 0,1,2,... \tag{11}$$

where $\phi > 0$ is dispersion parameter and $\mu_i = e^{x_i^\top \beta}$ in (11). For more details refer Yang et al. (2009) and finally with

iii. Generalized Negative Binomial ($\mathscr{NB}$-2) (Greene, 2008): The pmf of a generalized negative binomial ($\mathscr{NB}$-2) regression model is given as

$$P(Y_i = y_i|\theta, r_i) = \frac{\Gamma(\theta+y_i)r_i^\theta(1-r_i)^{y_i}}{y_i!\Gamma(\theta)} \tag{12}$$

where $y_i = 0,1,2,\ldots$ and $r_i = \theta/(\theta+\lambda_i)$ and $\lambda_i = e^{x_i^\top \beta}$.

Table 7: *Exploratory data description.*

Variable	Nature of variable	Measurement	Mean	Variance
OFP	Response	Number of physician visits	6.046	57.169
HOSP	Explanatory	Number of days of hospital stays	0.297	0.513
POORHLTH		Self-perceived health status, poor =1, else =0.	0.13	0.113
EXCLHLTH		Self-perceived health status, excellent =1, else 0	0.071	0.066
NUMCHRON		Number of chronic conditions	1.533	1.788
MALE		Gender; male = 1, else =0	0.408	0.241
SCHOOL		Number of year of education	10.355	13.25
PRIVINS		Private insurance indicator, yes =1, no = 0	0.794	0.164

Table 8: *Maximum likelihood estimates of the parameters of different regression models.*

Parameter	Poisson			$\mathscr{GP}$-2			$\mathscr{NB}$-2			$\mathscr{TGM}$		
	est. (s.e)	t-statistic	p-value	est. (s.e.)	t-statistic	p-value	est. (s.e.)	t-statistic	p-value	est.(s.e.)	t-statistic	p-value
Intercept(β_1)	0.99(0.04)	27.51	$<$ 0.00001	0.82 (0.09)	9.12	0.00001	0.88(0.09)	10.17	$<$ 0.00001	0.205 (0.17)	1.207	0.22758
HOSP(β_2)	0.19(0.01)	20.39	$<$ 0.00001	0.28 (0.04)	6.22	0.00001	0.24(0.03)	7.14	$<$ 0.00001	0.090 (0.04)	2.0397	0.04249
POORHLTH(β_3)	0.21(0.03)	8.28	$<$ 0.00001	0.39 (0.09)	4.53	0.00001	0.31(0.07)	4.32	0.000015	–1.174 (0.65)	–1.8106	0.07035
EXCLHLTH(β_4)	–0.21(0.04)	-4.84	$<$ 0.00001	–0.17 (0.09)	–1.8	0.07201	–0.18(0.01)	-1.92	0.054857	1.132 (0.11)	10.1784	0.00001
NUMCHRON(β_5)	0.16(0.01)	24.48	$<$ 0.00001	0.20 (0.02)	9.54	0.00001	0.19(0.02)	10.04	$<$ 0.00001	0.166 (0.03)	5.7383	0.00001
MALE(β_6)	–0.1(0.02)	–5.31	$<$ 0.00001	–0.13 (0.05)	-2.54	0.01116	–0.12(0.05)	–2.52	0.01181	0.073 (0.10)	0.7439	0.45702
SCHOOL(β_7)	0.03(0.001)	12.19	$<$ 0.00001	0.04 (0.01)	5.29	0.00001	0.04(0.01)	5.29	$<$ 0.00001	0.069 (0.02)	3.6983	0.00022
PRIVINS(β_8)	0.15(0.02)	5.85	$<$ 0.00001	0.15 (0.07)	2.27	0.02331	0.14(0.06)	2.28	0.02271	0.471 (0.11)	4.2646	0.00001
Dispersion parameter	—	—		0.28 (0.01)	30.08	0.00001	1.16 (0.04)	25.03	$<$ 0.00001	0.066(0.034)	1.9064	0.05674
V		10.6736			–1.188642			–2.28383			—	
$l_{\max}$		–8813.74			–5614.72			–5607.2			**– 4973.1**	
AIC ($-2l_{\max}+2k$)		17643			11247.44			11232.39			**9964.202**	

5.3.1. A numerical illustration of count regression

We examine the US National Medical Expenditure Survey 1987/88 (NMES) data obtained from Journal of Applied Econometrics 1997 Data Archive at `http://qed.econ.queensu.ca/jae/1997-v12.3/ deb-trivedi/`, which were originally employed by Deb and Trivedi (1997) in their analysis of various measures of health-care utilization. For illustration purpose we consider the first 2000 observations for fitting the regression model. The exploratory data description of the response variable as well as the set of explanatory variables is given in Table 7, from where it can be seen that the mean and variance of the number of physician visit (OFP) variable indicates presence of the over-dispersion as well as existence of large number of zeros. Hence it seems appropriate to apply our model for the present data set with the number of physician visits (OFP) as the response variable and remaining seven as explanatory variables.

Table 8 presents the maximum likelihood estimates of the parameters of the models Poisson($\mathscr{P}$), negative binomial($\mathscr{NB}$-2), generalized Poisson ($\mathscr{GP}$-2), and transmuted geometric ($\mathscr{TGM}$), their standard errors, t-statistics and $p-$values.

For comparison between the different fitted models, we have used the value of the maximum of the log-likelihood function ($l_{\max}$) and the Akaike information criterion (AIC). The model with the lowest AIC value is considered to be the best. It can be observed that the estimates of all parameters except the parameters of POORHLTH, MALE and dispersion parameter are found significant at 5% level of significance. Unlike the other models considered here the number of physician visit has not been influenced by the gender profile and poor health status of the patient. Most of the estimated parameters values under the $\mathscr{TGM}$ model differs in values obtained under other competitive models. The estimate of dispersion parameter for $\mathscr{TGM}$ found significant at 5% level of significance as opposed to $\mathscr{GP}$-2 and $\mathscr{NB}$-2 models which gives an indication of capturing dispersion of data. Moreover, with respect to the values of $l_{\max}$ and consequently *AIC*, our proposed model turns out to be the best. Hence, we conclude that proposed $\mathscr{TGM}$ regression model gives satisfactory fit and can be considered suitable for count data regression analysis.

Since the models under consideration namely $\mathscr{P}$, $\mathscr{NB}$-2, $\mathscr{GP}$-2, are not nested within $\mathscr{TGM}$, it may of interest to employ the Vuong test (see Vuong (1989)) for non-nested models to discriminate among these models. The Vuong statistic is given by

$$V = \frac{1}{\zeta\sqrt{n}}\left(l_{\mathscr{TGM}}(\hat{\Theta}_1) - l_g(\hat{\Theta}_2)\right) \tag{13}$$

where

$$\zeta^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\frac{f_{\mathscr{TGM}}(\hat{\Theta}_1|y_i,x_i)}{g(\hat{\Theta}_2|y_i,x_i)}\right)\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f_{\mathscr{TGM}}(\hat{\Theta}_1|y_i,x_i)}{g(\hat{\Theta}_2|y_i,x_i)}\right)\right)^2$$

where $f_{\mathscr{TGM}}$ and g represent $\mathscr{TGM}$ and the other competing model respectively.

As statistic V ia asymptotically standard normal, the rejection of test in favour of $\mathscr{TGM}$ occurs if $V > 1.96$, at the 5% level of significance . From our findings in Table 8, it is seen that the proposed $\mathscr{TGM}$ regression model is preferred over Poisson (since $V > 1.96$), but do not distinguish between $\mathscr{GP}$-2 model (since $-1.96 < V < 1.96$). However the test rejects the $\mathscr{TGM}$ model when compared with $\mathscr{NB}$-2 (since $V < -1.96$).

6. Concluding remarks

In this paper the transmutation technique is used to offer a new flexible generalization of the geometric distribution as a viable alternative to some existing models. Different distributional properties of the distribution are found to be simple and attractive. The theoretical result regarding possibility of applying this new distribution to model aggregate claim in the actuarial context is presented and its suitability for modelling large aggregate claims is established and complimented with a real life data set. Illustrative data fitting with the proposed model for a popular data set from automobile insurance sector having over-dispersion turned out to be very useful. Finally, a count regression model based on the proposed distribution provided best fit in terms of the AIC value when compared with some existing models for analysing a data set from the health sector. Based on these findings, it is envisaged that the transmuted geometric distribution with two parameters can be very useful in modelling and analysis of count data of different types. Further, this idea of applying transmutation to discrete distribution may be applied to construct new generalizations of other distributions.

Acknowledgments

The authors gratefully acknowledge the suggestions of the editor-in-chief and three anonymous referees on earlier versions of the manuscript which resulted in much improved presentation.

References

Antzoulakos, D. and Chadjiconstantinidis, S. (2004). On mixed and compound mixed Poisson distributions. *Scandinavian Actuarial Journal*, 3, 161–188.

Chakraborty, S. and Gupta, R. D. (2015). Exponentiated geometric distribution: another generalization of geometric distribution. *Communication in Statistics-Theory and Methods*, 44, 1143–1157.

Chakraborty, S.(2015). Generation of discrete analogues of continuous distributions-a survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2, 6.

Consul, P. C. and Famoye, F. (1992). Generalized Poisson regression model. *Communication in Statistics-Theory and Methods*, 2, 89–109.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the Elderly: a finite mixture approach. *Journal of Applied Econometrics*, 12, 313–336.

Drouet-Mari, D, Kotz, S. (2001). *Correlation and Dependence*. Imperial College press, London.

Gómez-Déniz, E. (2010). Another generalization of the geometric distribution. *Test*, 19, 399–415.

Gómez-Déniz, E., Sarabia, J. M. and Odeja, E. C. (2011). A new discrete distribution with actuarial applications. *Insurance: Mathematics and Economics*, 48, 406–412.

Greene, W. (2008). Functional forms for the negative binomial model for count data. *Economics Letter*, 99, 585–590.

Jain, G. C. and Consul, P. C. (1971). A generalized negative binomial distribution. *SIAM Journal of Applied Mathematics*, 21, 501–513.

Johnson, N. L. Kemp, A. W. and Kotz, S. (2005). *Univariate Discrete Distributions*. 2nd ed. Wiley, New York.

Jong, D. P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, Cambridge.

Klugman, S. A., Panjer, H. H. and Willmot, G. E. (2008). *Loss Models: From Data to Decisions*. 3rd ed. John Wiley and Sons, New York, 3, 161–188.

Kozubowski, T.J. and Podgórski, K. (2016). Transmuted distributions and random extrema. *Statistics and Probability Letters*, available online 13 April 2016, in press.

Makčutek, J. (2008). A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Rep Phys*, 60, 501–509.

Mirhossaini, S. M. and Dolati, A. (2008). On a new generalization of the exponential distribution. *Journal of Mathematical Extension*, 3, 27–42.

Oguntunde, P. E. and Adejumo, A. O. (2015). The transmuted inverse exponential distribution. *International Journal of Advanced Statistics and Probability*, 3, 1–7.

Owoloko, E. A., Oguntunde, P. E. and Adejumo, A. O. (2015). Performance rating of the transmuted exponential distribution: an analytical approach. *Scandinavian Actuarial Journal*, 4, 1–15.

Philippou, A. N., Georghiou, C. and Philippou, G. N. (1983). A generalized geometric distribution and some of its properties. *Statistics and Probability Letters*, 1, 171–175.

Rolski, T., Schmidli, H., Schmidt, V. and Teugel, J. (1999). *Stochastic Processes for Insurance and Finance*. John Wiley and Sons, New York.

Sastry, D. V. S., Bhati, D., Rattihalli, R. N. and Gómez-Déniz, E.(2004). On zero distorted generalized geometric distribution. *Communication in Statistics-Theory and Methods*, accepted.

Shaw, W. and Buckley, I. (2007). The alchemy of probability distributions: beyond Gram- Charlier expansions and a skew-kurtotic-normal distribution from a rank transmutation map. *Research Report*. Available in arXiv:0901.0434v1 [q-fin.ST].

Tripathi, R. C., Gupta, R. C. and White, T. J.(1987). Some generalizations of the geometric distribution. *Sankhyā*, Series B, 49, 218–223.

Tse, Yiu-Kuen (2009). *Non-life Actuarial Models Theory, Methods and Evaluation*. Cambridge University Press, UK.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57, 307–333.

Warde, W. D. and Katti, S. K.(1971). Infinite divisibility of discrete distributions II. *Annual of Mathematical Statistics*, 42, 1088–1090.

Willmot, G. E. (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative binomial. *Scandinavian Actuarial Journal*, 3-4, 113–127.

Yang, Z., Hardin, J. W. and Addy, C. L. (2009). A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inference*, 139, 1514–1521.

Yousof, H. M., Afify, A. Z., Alizadeh, M., Butt, N. S., Hamedani, G. G. and Ali, M. M. (2015). The Transmuted exponentiated generalized-G family of distributions. *Pakistan Journal of Statistics and Operations Research*, DOI: 10.18187/pjsor.v11i4.1164.

SORT 40 (1) January-June 2016, 177-200

Compound distributions motivated by linear failure rate

Narjes Gitifar[1], Sadegh Rezaei[1] and Saralees Nadarajah[2]

Abstract

Motivated by three failure data sets (lifetime of patients, failure time of hard drives and failure time of a product), we introduce three different three-parameter distributions, study basic mathematical properties, address estimation by the method of maximum likelihood and investigate finite sample performance of the estimators. We show that one of the new distributions provides a better fit to each data set than eight other distributions each having three parameters and three distributions each having two parameters.

MSC: 62E15.

Keywords: Linear failure rate distribution, maximum likelihood estimation, Poisson distribution.

1. Introduction

Systems or components having linear failure rates are common in real life. Examples include concrete under multiaxial states of stress (Donida and Mentrasti, 1982), composite laminates with transverse shear (Reddy and Reddy, 1992) and load-sharing systems (Sutar and Naik-Nimbalkar, 2014). There are also many real data sets that exhibit approximately linear failure rates at least in the upper tails. We present three examples.

The first data set, due to Dispenzieri et al. (2012), consists of the number of days from visit to clinic until death of 100 patients. The data result from a study of the relationship between serum free light chain and mortality. The 100 patients were selected randomly from a total of 7874 patients, including patients who had not died. The patients who had died were diagnosed with monoclonal gammapothy.

[1] Amirkabir University of Technology, Tehran, IRAN, email: srezaei@aut.ac.ir

[2] University of Manchester, Manchester M13 9PL, UK

Received: July 2015

Accepted: April 2016

Table 1: *Summary statistics of the three data sets.*

Statistic	Data set 1	Data set 2	Data set 3
minimum	0.0054	0.0053	0.0035
first quartile	0.3368	0.3977	0.318
median	0.4774	0.7770	0.4211
third quartile	0.7412	0.9304	0.5581
maximum	0.9514	1.4040	0.6878

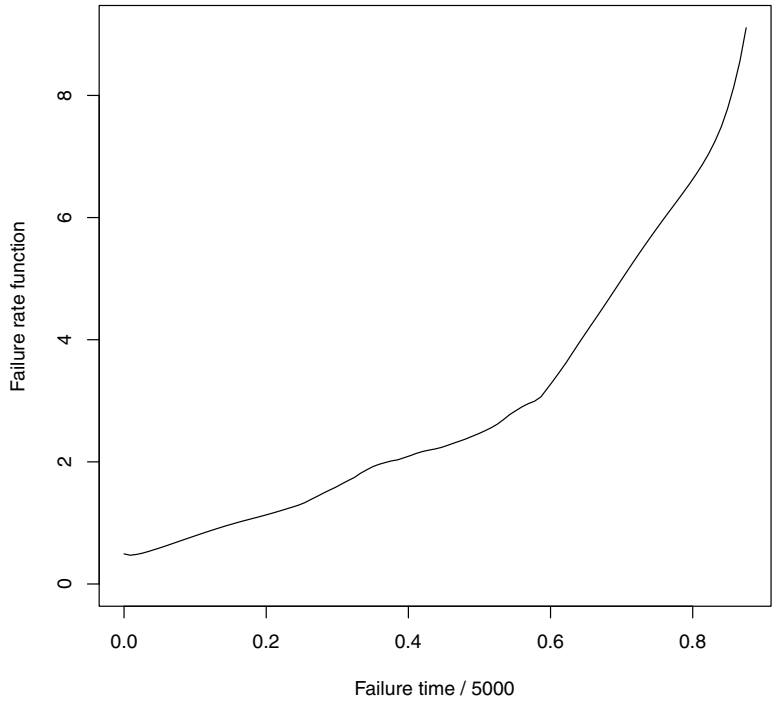

Figure 1: *Kaplan-Meier estimate of the failure rate function of the patient data of Dispenzieri et al. (2012).*

The second data set from https://www.backblaze.com/hard-drive-test-data.html is one hundred failure times in days of hard drives. The data were selected randomly from a total of 52422 hard drives, which included hard drives which had not failed. The data were collected by a large backup storage provider over two years. On each day, the Self-Monitoring, Analysis, and Reporting Technology (SMART) statistics of operational drives were recorded. When a hard drive was no longer operational, it was marked as a failure and removed.

The third data set due to Hong and Meeker (2013) is one hundred failure data in weeks of a product called Product D2 that is used in offices or residences. Product D2 is "similar to a high-end copying machine connected to the Internet and installed with a

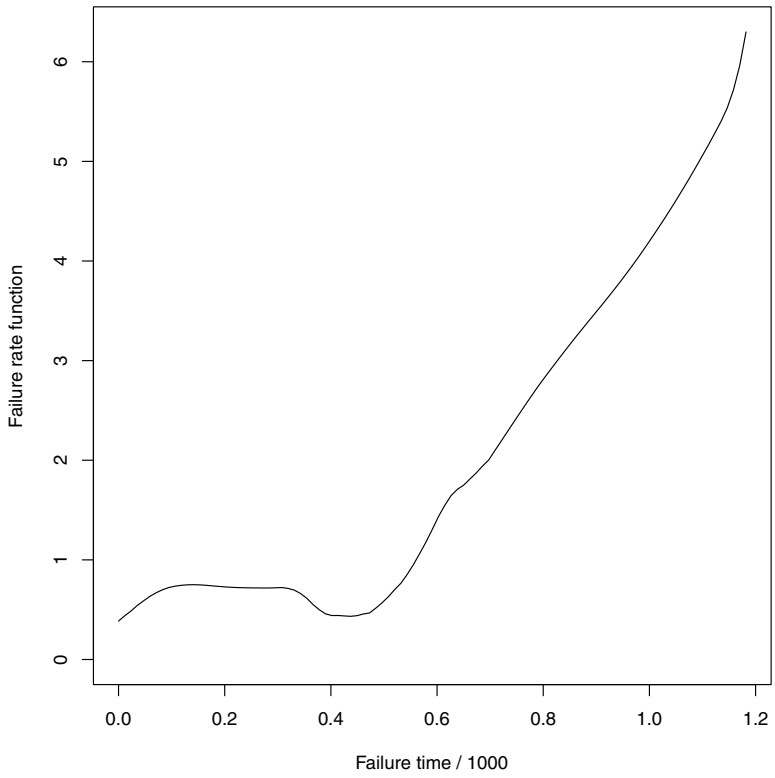

Figure 2: *Kaplan-Meier estimate of the failure rate function of the hard drive failure data.*

smart chip to record the number of pages that have been printed, as a function of time" (Hong and Meeker, 2013, page 136). The one hundred data were selected randomly from a total of 1800 observations.

All three data sets are presented in the appendix.

Kaplan-Meier estimates of the failure rate function (FRF) of the three data sets are shown in Figures 1, 2 and 3. We can see that the FRFs are approximately linear at least in the upper tails. The histogram of the three data sets are shown in Figures 8, 9 and 10. Some summary statistics of the three data sets are shown in Table 1.

We suppose that the patient's body or the hard drive or the product D2 is made of a number of components say N working independently in series. The assumption of the series structure is more reasonable than a parallel structure because it is unlikely that a patient's body will fail if and only if all its components fail or that a hard drive will break if and only if all its components break or that a product will fail if and only if all its components fail. It is more likely that a patient's body will fail if and only if any of its components fails or that a hard drive will break if and only if any of its components breaks or that a product will fail if and only if any of its components fails. However, in practice the components may not work independently. The distribution of the failure

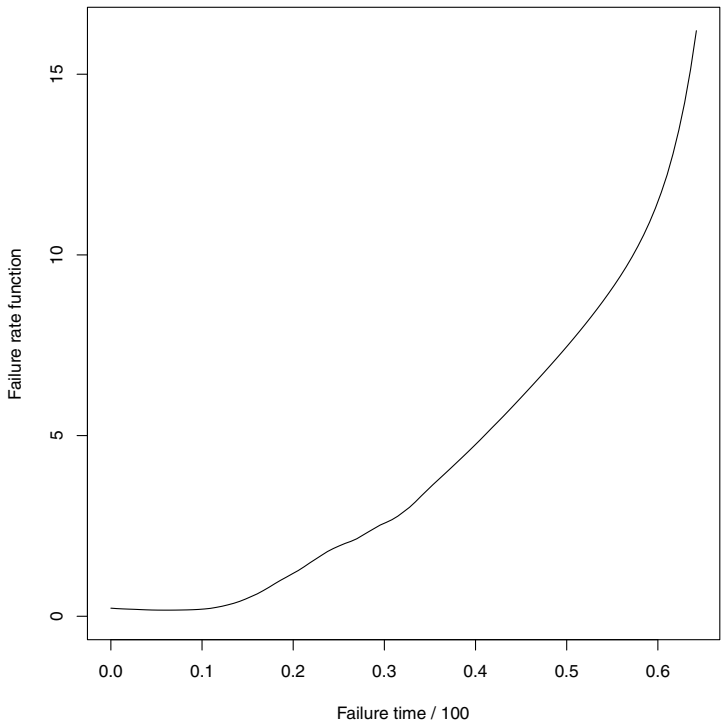

Figure 3: *Kaplan-Meier estimate of the failure rate function of the failure data of Hong and Meeker (2013).*

time may not have a closed form if we assume that the components are dependent, see (2) below and its discussion. We shall suppose independence for simplicity.

The number N may vary from one patient to another or one hard drive to another or one product to another. It may depend on the type of hard drive, type of patient, type of product, weight, length, and so on. So, we may take N as a random variable. The failure time can be written as $X = \min(Y_1, Y_2, \ldots, Y_N)$, where $Y_1, Y_2, \ldots, Y_N$ denote the failure times of the N components.

Standard models for N are the geometric, zero truncated Poisson, logarithmic, zero truncated negative binomial and zero truncated binomial distributions. For simplicity, we shall consider only the first three since each of them has one parameter. The last two distributions have two parameters each. That is, we take N to have one of the following probability mass functions (PMFs):

$$\Pr(N = n) = (1 - \lambda)\lambda^{n-1}$$

for $0 < \lambda < 1$ and $n = 1, 2, \ldots$;

$$\Pr(N = n) = \frac{\lambda^n}{(e^\lambda - 1)\, n!}$$

for $\lambda > 0$ and $n = 1, 2, \ldots$; or

$$\Pr(N = n) = -\frac{1}{\ln(1-\lambda)} \frac{\lambda^n}{n}$$

for $0 < \lambda < 1$ and $n = 1, 2, \ldots$.

Since the failure rate for the three data sets is approximately linear at least in the upper tail (see Figures 1, 2 and 3), we shall suppose $Y_1, Y_2, \ldots$ too follow a distribution that has a linear FRF. The distribution characterized by a linear failure rate is actually known as the linear failure rate (LFR) distribution due to Bain (1974). Its probability density function (PDF) and cumulative distribution function (CDF) are specified by

$$f_Y(y; \gamma, \beta) = (\beta + \gamma y) \exp\left(-\beta y - \frac{\gamma}{2} y^2\right)$$

and

$$F_Y(y; \gamma, \beta) = 1 - \exp\left(-\beta y - \frac{\gamma}{2} y^2\right),$$

respectively, for $y > 0$, $\beta \geq 0$, $\gamma \geq 0$ and $\beta + \gamma > 0$. It is easy to see that the FRF is $h_Y(y; \gamma, \beta) = \beta + \gamma y$, a linear function of y. Both parameters, β and γ, are referred to as scale parameters.

The distribution of $X = \min(Y_1, Y_2, \ldots, Y_N)$ can now be derived given the assumptions that N is either geometric, Poisson or logarithmic and $Y_1, Y_2, \ldots$ are independent LFR random variables independent of N. In the general case, the CDF and the PDF of X can be derived as

$$\begin{aligned} F_X(x) &= \Pr[\min(Y_1, Y_2, \ldots, Y_N) < x] = 1 - \Pr[\min(Y_1, Y_2, \ldots, Y_N) > x] \\ &= 1 - \sum_{n=1}^{\infty} \Pr[\min(Y_1, Y_2, \ldots, Y_n) > x \mid N = n] \Pr(N = n) \\ &= 1 - \sum_{n=1}^{\infty} \Pr[Y_1 > x, Y_2 > x, \ldots, Y_n > x] \Pr(N = n) \\ &= 1 - \sum_{n=1}^{\infty} \Pr^n[Y > x] \Pr(N = n) = 1 - \sum_{n=1}^{\infty} [1 - F_Y(x)]^n \Pr(N = n) \end{aligned}$$

and

$$f_X(x) = f_Y(x) \sum_{n=1}^{\infty} n [1 - F_Y(x)]^{n-1} \Pr(N = n),$$

respectively. In the case N is geometric, we obtain

$$f_X(x;\lambda,\gamma,\beta) = \frac{(1-\lambda)(\beta+\gamma x)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)}{\left[1-\lambda\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\right]^2},$$

which we shall refer to as the linear failure rate geometric (LFRG) distribution and write $X \sim \text{LFRG}(\lambda,\gamma,\beta)$ for $0<\lambda<1$, $\beta\geq 0$, $\gamma\geq 0$ and $\beta+\gamma>0$. In the case N is zero truncated Poisson, we obtain

$$f_X(x;\lambda,\gamma,\beta) = \lambda\left(1-e^{-\lambda}\right)^{-1}(\beta+\gamma x)\exp\left(-\lambda-\beta x-\frac{\gamma}{2}x^2\right)\exp\left[\lambda\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\right], \tag{1}$$

which we shall refer to as the linear failure rate Poisson (LFRP) distribution and write $X \sim \text{LFRP}(\lambda,\gamma,\beta)$ for $\lambda>0$, $\beta\geq 0$, $\gamma\geq 0$ and $\beta+\gamma>0$. In the case N is logarithmic, we obtain

$$f_X(x;\lambda,\gamma,\beta) = -\frac{\lambda(\beta+\gamma x)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)}{\ln(1-\lambda)\left[1-\lambda\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\right]},$$

which we shall refer to as the linear failure rate logarithmic (LFRL) distribution and write $X \sim \text{LFRL}(\lambda,\gamma,\beta)$ for $0<\lambda<1$, $\beta\geq 0$, $\gamma\geq 0$ and $\beta+\gamma>0$. These distributions do not have linear failure rates. But $h_X(y;\lambda,\gamma,\beta) \sim h_Y(y;\gamma,\beta) \sim \gamma y$ as $y\to\infty$. So, the assumption of linear failure rate for $Y_1, Y_2, \ldots$ guarantees that linear failure rate holds for X too at least in the upper tail.

The limiting cases of the LFRG, LFRP and LFRL distributions as $\lambda \downarrow 0$ is the LFR distribution. The LFRG and LFRL distributions limit to a degenerate distribution as $\lambda \uparrow 1$.

If $Y_1, Y_2, \ldots$ are dependent random variables then the CDF of X can only be expressed as

$$F_X(x) = 1 - \sum_{n=1}^{\infty} \Pr[Y_1 > x, Y_2 > x, \ldots, Y_n > x] \Pr(N=n). \tag{2}$$

This cannot be reduced to a closed form unless the joint dependence of $(Y_1, Y_2, \ldots, Y_n)$ takes a very simple form.

In the rest of this section, Section 2 and Section 3, we shall focus on the LFRP distribution. The details for the LRFG and LRFL distributions can be derived similarly. One of the most popular models for counts is the zero truncated Poisson distribution. Some of its recent applications can be found in van der Heijden et al. (2003), Elhai et al. (2008), Ginebra and Puig (2010) and Xu and Hu (2011).

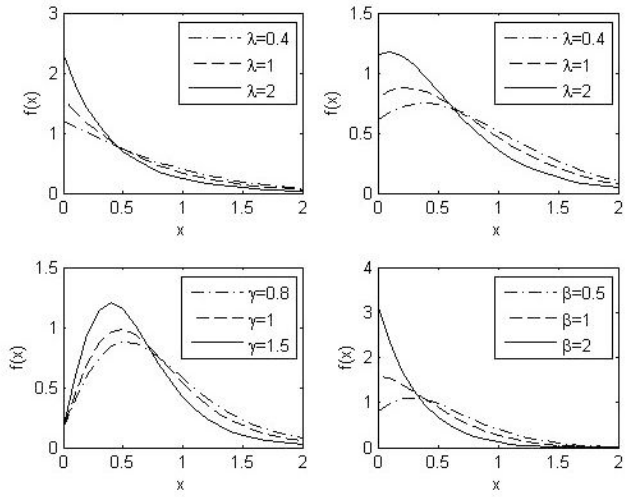

Figure 4: *Probability density function of the LFRP distribution for (a)* $\gamma = 0.5$ *and* $\beta = 1$*, (b)* $\gamma = 1$ *and* $\beta = 0.5$*, (c)* $\beta = 0.05$ *and* $\lambda = 3$*, (d)* $\gamma = 2$ *and* $\lambda = 1$*.*

Possible shapes of (1) are shown in Figure 4. We see that both monotonically decreasing and unimodal shapes are possible. The mode of (1) is the root of

$$\frac{\gamma}{\beta+\gamma x} - \beta - \gamma x = \lambda(\beta+\gamma x)\exp\left(-\beta x - \frac{\gamma}{2}x^2\right).$$

Furthermore, $f_X(0) = \lambda\beta/\left(1-e^{-\lambda}\right)$ and

$$f_X(x) \sim \lambda\gamma\left(1-e^{-\lambda}\right)^{-1} x\exp\left(-\lambda - \beta x - \frac{\gamma}{2}x^2\right)$$

as $x \to \infty$. The lower tail of the PDF has a fixed point while its upper tail decays exponentially.

The CDF and FRF of $X \sim \text{LFRP}(\lambda,\gamma,\beta)$ are

$$F_X(x) = \frac{1}{e^{\lambda}-1}\left\{e^{\lambda} - \exp\left[\lambda\exp\left(-\beta x - \frac{\gamma}{2}x^2\right)\right]\right\}$$

and

$$h_X(x) = \frac{(\beta+\gamma x)\lambda\exp\left(-\beta x - \frac{\gamma}{2}x^2\right)}{1-\exp\left[-\lambda\exp\left(-\beta x - \frac{\gamma}{2}x^2\right)\right]}, \tag{3}$$

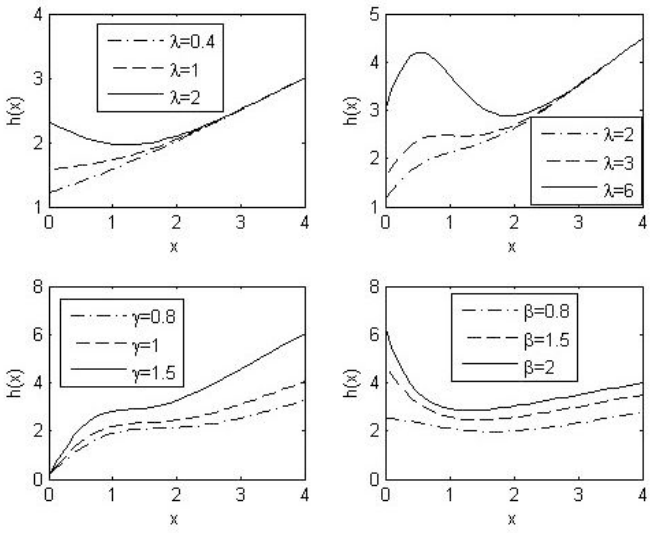

Figure 5: *Failure rate function of the LFRP distribution for (a)* $\gamma = 0.5$ *and* $\beta = 1$*, (b)* $\gamma = 1$ *and* $\beta = 0.5$*, (c)* $\beta = 0.05$ *and* $\lambda = 3$*, (d)* $\lambda = 3$ *and* $\gamma = 0.5$*.*

respectively, for $x > 0$, $\lambda > 0$, $\beta \geq 0$, $\gamma \geq 0$ and $\beta + \gamma > 0$. Figure 5 shows possible shapes of (3) for different parameter values. We see that the LFRP distribution can exhibit increasing, decreasing and upside down bathtub shapes for the failure rate. The LFR distribution can exhibit only increasing or constant failure rates.

Reliability and survival analysis often encounter upside down bathtub failure rates. Examples can be found in redundancy allocations in systems (Singh and Misra, 1994) and mortality modelling (Silva et al., 2010).

The mode or the anti-mode of (3) is the root of

$$\frac{\gamma}{\beta+\gamma x} - \beta - \gamma x = -\lambda(\beta+\gamma x)\exp\left(-\beta x - \frac{\gamma}{2}x^2\right)\left\{\exp\left[\lambda\exp\left(-\beta x - \frac{\gamma}{2}x^2\right)\right] - 1\right\}^{-1}.$$

Furthermore, $h_X(0) = \lambda\beta/\left(1-e^{-\lambda}\right)$ and $h_X(x) \sim \gamma x$ as $x \to \infty$. The lower tail of the FRF has a fixed point. As already noted, the upper tail of the FRF of the LFRP distribution behaves in the same manner as that of the LFR distribution. Yet the former does exhibit upside down bathtub failure rates while the latter does not.

The qth quantile of $X \sim \text{LFRP}(\lambda, \gamma, \beta)$ say x_q defined by $F_X(x_q) = q$ is

$$x_q = -\frac{\beta}{\gamma} + \sqrt{\frac{\beta^2}{\gamma^2} - \frac{2}{\gamma}\ln\left\{\ln\left[e^{\lambda} - q\left(e^{\lambda}-1\right)\right]^{\frac{1}{\lambda}}\right\}}.$$

In particular, the median of X is

$$\text{Median}(X) = -\frac{\beta}{\gamma} + \sqrt{\frac{\beta^2}{\gamma^2} - \frac{2}{\gamma}\ln\left\{\ln\left[e^{\lambda} - \frac{1}{2}(e^{\lambda}-1)\right]^{\frac{1}{\lambda}}\right\}}.$$

Quantiles are useful for estimation and simulation.

Several other distributions have been introduced in the literature by taking $X = \min(Y_1, Y_2, \ldots, Y_N)$, where N is a geometric, zero truncated Poisson or a logarithmic random variable: By taking N to be a geometric random variable and $Y_1, Y_2, \ldots$ to be independent and identical Weibull random variables, Barreto-Souza et al. (2011) introduced the three-parameter Weibull geometric (WG) distribution given by the PDF

$$f(x) = \frac{(1-\lambda)\beta\gamma^{-\beta}x^{\beta-1}\exp\left[-(x/\gamma)^{\beta}\right]}{\left\{1-\lambda\exp\left[-(x/\gamma)^{\beta}\right]\right\}^2}$$

for $x > 0$, $0 < \lambda < 1$, $\beta > 0$ and $\gamma > 0$; By taking N to be a zero truncated Poisson random variable and $Y_1, Y_2, \ldots$ to be independent and identical Weibull random variables, Lu and Shi (2012) introduced the three-parameter Weibull Poisson (WP) distribution given by the PDF

$$f(x) = \frac{\lambda\beta\gamma^{-\beta}x^{\beta-1}\exp\left\{-(x/\gamma)^{\beta} + \lambda\exp\left[-(x/\gamma)^{\beta}\right]\right\}}{\exp(\lambda)-1}$$

for $x > 0$, $\lambda > 0$, $\beta > 0$ and $\gamma > 0$; By taking N to be a logarithmic random variable and $Y_1, Y_2, \ldots$ to be independent and identical Weibull random variables, Ciumara and Preda (2009) introduced the three-parameter Weibull logarithmic (WL) distribution given by the PDF

$$f(x) = -\frac{(1-\lambda)\beta\gamma^{-\beta}x^{\beta-1}\exp\left[-(x/\gamma)^{\beta}\right]}{\ln\lambda\left\{1-(1-\lambda)\exp\left[-(x/\gamma)^{\beta}\right]\right\}}$$

for $x > 0$, $0 < \lambda < 1$, $\beta > 0$ and $\gamma > 0$; By taking N to be a geometric random variable and $Y_1, Y_2, \ldots$ to be independent and identical generalized exponential random variables, Mahmoudi and Jafari (2012) introduced the three-parameter generalized exponential geometric (GEG) distribution given by the PDF

$$f(x) = \frac{(1-\lambda)\beta\gamma\exp(-\gamma x)\left[1-\exp(-\gamma x)\right]^{\beta-1}}{\left\{\lambda\left[1-\exp(-\gamma x)\right]^{\beta}-1\right\}^2}$$

for $x > 0$, $0 < \lambda < 1$, $\beta > 0$ and $\gamma > 0$; By taking N to be a zero truncated Poisson random variable and $Y_1, Y_2, \ldots$ to be independent and identical generalized exponential random variables, Mahmoudi and Jafari (2012) introduced the three-parameter generalized exponential Poisson (GEP) distribution given by the PDF

$$f(x) = \frac{\lambda\beta\gamma \exp(-\gamma x)\left[1-\exp(-\gamma x)\right]^{\beta-1} \exp\left\{\left[1-\exp(-\gamma x)\right]^{\beta}\right\}}{\exp(\lambda)-1}$$

for $x > 0$, $\lambda > 0$, $\beta > 0$ and $\gamma > 0$; By taking N to be a logarithmic random variable and $Y_1, Y_2, \ldots$ to be independent and identical generalized exponential random variables, Mahmoudi and Jafari (2012) introduced the three-parameter generalized exponential logarithmic (GEL) distribution given by the PDF

$$f(x) = \frac{\lambda\beta\gamma \exp(-\gamma x)\left[1-\exp(-\gamma x)\right]^{\beta-1}}{\ln(1-\lambda)\left\{\lambda\left[1-\exp(-\gamma x)\right]^{\beta}-1\right\}}$$

for $x > 0$, $0 < \lambda < 1$, $\beta > 0$ and $\gamma > 0$.

A final motivation for the LFRP distribution is that it provides better fits for the three data sets than at least eight other distributions each having three parameters and at least three distributions each having two parameters. The eight distributions are the LFRG, LFRL, WG, WP, WL, GEG, GEP and GEL distributions.

The rest of this paper is organized as follows: estimation of the parameters of the LFRP distribution by the method of maximum likelihood is considered in Section 2; finite sample performance of the maximum likelihood estimators is assessed by simulation in Section 3; application of the LFRP distribution to the three data sets is illustrated in Section 4; some conclusions are noted in Section 5.

We have given above simple expressions for the PDF, its shape, FRF, its shape, quantiles and median of $X \sim \text{LFRP}(\lambda, \gamma, \beta)$. Simple expressions for further mathematical properties of $X \sim \text{LFRP}(\lambda, \gamma, \beta)$ do not appear to be possible; for example, using the series expansions

$$(1-z)^{-2} = \sum_{k=0}^{\infty} \binom{-2}{k} (-z)^k,$$

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!},$$

$$(1-z)^{-1} = \sum_{k=0}^{\infty} z^k,$$

and equation (2.3.15.3) in Prudnikov et al. (1986), one can express the nth moments of LFRG, LFRP and LFRL distributions as

$$E(X^n) = (1-\lambda)\sum_{k=0}^{\infty}\binom{-2}{k}(-\lambda)^k A(n,k),$$

$$E(X^n) = \frac{\lambda e^{-\lambda}}{1-e^{-\lambda}}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}A(n,k)$$

and

$$E(X^n) = -\frac{1}{\ln(1-\lambda)}\sum_{k=0}^{\infty}\lambda^{k+1}A(n,k),$$

respectively, where

$$A(n,k) = \frac{n!\exp\left[\frac{(k+1)\beta^2}{4\gamma}\right]}{\gamma^{\frac{n+1}{2}}(k+1)^{\frac{n+2}{2}}}\left[\beta\sqrt{k+1}D_{-n-1}\left(\frac{\beta\sqrt{k+1}}{\sqrt{\gamma}}\right)+(n+1)\sqrt{\gamma}D_{-n-2}\left(\frac{\beta\sqrt{k+1}}{\sqrt{\gamma}}\right)\right],$$

where $D_\nu(\cdot)$ denotes the parabolic cylinder function of order ν. These expressions are not simple. They are infinite sums of terms involving a special function which is defined in terms of an integral. So, the moments could be computed more efficiently by numerical integration, i.e., by

$$E(X^n) = \int_0^{\infty} x^n \frac{(\beta+\gamma x)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)}{\left[1-(1-\lambda)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\right]^2}dx,$$

$$E(X^n) = \lambda e^{-\lambda}\left(1-e^{-\lambda}\right)^{-1}\int_0^{\infty} x^n(\beta+\gamma x)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\exp\left[\lambda\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)\right]dx$$

and

$$E(X^n) = -\frac{1}{\ln(1-\lambda)}\int_0^{\infty} x^n \frac{(\beta+\gamma x)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)}{1-(1-\lambda)\exp\left(-\beta x-\frac{\gamma}{2}x^2\right)}dx.$$

Hence, we shall not consider further mathematical properties.

2. Estimation

We suppose $x_1, x_2, \ldots, x_n$ is a random sample from LFRP(β, γ, λ) with β, γ, λ unknown. Then the log-likelihood function of β, γ, λ can be expressed as

$$\ln L = n \ln \lambda - n \ln \left(e^{\lambda} - 1\right) + \sum_{i=1}^{n} \ln (\beta + \gamma x_i) - \beta \sum_{i=1}^{n} x_i + \frac{\gamma}{2} \sum_{i=1}^{n} x_i^2 + \\ + \lambda \sum_{i=1}^{n} \exp\left(-\beta x_i - \frac{\gamma}{2} x_i^2\right). \tag{4}$$

The associated normal equations are

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \frac{n e^{\lambda}}{e^{\lambda} - 1} + \sum_{i=1}^{n} \exp\left(-\beta x_i - \frac{\gamma}{2} x_i^2\right),$$
$$\frac{\partial \ln L}{\partial \gamma} = \sum_{i=1}^{n} \frac{x_i}{\beta + \gamma x_i} - \frac{1}{2} \sum_{i=1}^{n} x_i^2 - \lambda \sum_{i=1}^{n} \frac{x_i^2}{2} \exp\left(-\beta x_i - \frac{\gamma}{2} x_i^2\right),$$
$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{\beta + \gamma x_i} - \sum_{i=1}^{n} x_i + \lambda \sum_{i=1}^{n} x_i \exp\left(-\beta x_i - \frac{\gamma}{2} x_i^2\right).$$

The maximum likelihood estimates of (λ, γ, β) say $\left(\widehat{\lambda}, \widehat{\gamma}, \widehat{\beta}\right)$ are the simultaneous solutions of $\partial \ln L / \partial \lambda = 0$, $\partial \ln L / \partial \gamma = 0$ and $\partial \ln L / \partial \beta = 0$. These equations being non-linear, some quasi-Newton algorithm will be needed to solve them simultaneously. An alternative is to obtain $\left(\widehat{\lambda}, \widehat{\gamma}, \widehat{\beta}\right)$ by direct numerical maximization of (4). We shall pursue this simpler approach. Numerical maximization of (4) was performed by using optim in R (R Development Core Team, 2014). Extensive numerical calculations showed that the surface of (4) was reasonably smooth. optim was able to locate the maximum for a wide range of starting values. The solution returned by optim was unique for all starting values.

Reasonable starting values for the parameters are useful to ease optimization. The method of moments can be used to obtain them. Equating the sample moments $m_1 = (1/n) \sum_{i=1}^{n} x_i$, $m_2 = (1/n) \sum_{i=1}^{n} x_i^2$ and $m_3 = (1/n) \sum_{i=1}^{n} x_i^3$ with the theoretical versions given by

$$E\left(X^i\right) = \lambda \left(1 - e^{-\lambda}\right)^{-1} \int_0^{\infty} x^i (\beta + \gamma x) \exp\left(-\lambda - \beta x - \frac{\gamma}{2} x^2\right) \exp\left[\lambda \exp\left(-\beta x - \frac{\gamma}{2} x^2\right)\right] dx,$$

we have $m_1 = \mathrm{E}(X)$, $m_2 = \mathrm{E}\left(X^2\right)$ and $m_3 = \mathrm{E}\left(X^3\right)$. These equations were solved using a quasi-Newton algorithm.

The distribution of $\left(\widehat{\lambda},\widehat{\gamma},\widehat{\beta}\right)$ as $n\to\infty$, under certain regularity conditions (see, for example, Ferguson, 1996 and pages 461-463 in Lehmann and Casella, 1998), is trivariate normal with mean (λ,β,γ) and covariance given by the inverse of

$$\mathbf{I}=\begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{pmatrix}=\begin{pmatrix} \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\lambda^2}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\lambda\partial\gamma}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\lambda\partial\beta}\right) \\ \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\gamma\partial\lambda}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\gamma^2}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\gamma\partial\beta}\right) \\ \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\beta\partial\lambda}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\beta\partial\gamma}\right) & \mathrm{E}\left(-\frac{\partial^2 \ln L}{\partial\beta^2}\right) \end{pmatrix}.$$

I is referred to as the expected information matrix.

In practice, n is finite. Cox and Hinkley (1979) recommended that the distribution of $\left(\widehat{\lambda},\widehat{\gamma},\widehat{\beta}\right)$ be approximated by a trivariate normal distribution with mean (λ,β,γ) and covariance taken to be the inverse of

$$\mathbf{J}=\begin{pmatrix} J_{11} & J_{12} & J_{13} \\ J_{21} & J_{22} & J_{23} \\ J_{31} & J_{32} & J_{33} \end{pmatrix}=\left.\begin{pmatrix} -\frac{\partial^2 \ln L}{\partial\lambda^2} & -\frac{\partial^2 \ln L}{\partial\lambda\partial\gamma} & -\frac{\partial^2 \ln L}{\partial\lambda\partial\beta} \\ -\frac{\partial^2 \ln L}{\partial\gamma\partial\lambda} & -\frac{\partial^2 \ln L}{\partial\gamma^2} & -\frac{\partial^2 \ln L}{\partial\gamma\partial\beta} \\ -\frac{\partial^2 \ln L}{\partial\beta\partial\lambda} & -\frac{\partial^2 \ln L}{\partial\beta\partial\gamma} & -\frac{\partial^2 \ln L}{\partial\beta^2} \end{pmatrix}\right|_{\lambda=\widehat{\lambda},\gamma=\widehat{\gamma},\beta=\widehat{\beta}}.$$

J is referred to as the observed information matrix. Cox and Hinkley (1979)'s approximation is known to be a better approximation than one based on the expected information matrix.

The elements of the observed information matrix are

$$J_{11}=\frac{n}{\widehat{\lambda}^2}-\frac{ne^{\widehat{\lambda}}}{\left(e^{\widehat{\lambda}}-1\right)^2},$$

$$J_{22}=\sum_{i=1}^{n}\frac{x_i^2}{\left(\widehat{\beta}+\widehat{\gamma}x_i\right)^2}-\frac{\widehat{\lambda}}{4}\sum_{i=1}^{n}x_i^4\exp\left(-\widehat{\beta}x_i-\frac{\widehat{\gamma}}{2}x_i^2\right),$$

$$J_{33}=\sum_{i=1}^{n}\frac{1}{\left(\widehat{\beta}+\widehat{\gamma}x_i\right)^2}-\widehat{\lambda}\sum_{i=1}^{n}x_i^2\exp\left(-\widehat{\beta}x_i-\frac{\widehat{\gamma}}{2}x_i^2\right),$$

$$J_{12}=J_{21}=\frac{1}{2}\sum_{i=1}^{n}x_i^2\exp\left(-\widehat{\beta}x_i-\frac{\widehat{\gamma}}{2}x_i^2\right),$$

$$J_{13} = J_{31} = \sum_{i=1}^{n} x_i \exp\left(-\widehat{\beta}x_i - \frac{\widehat{\gamma}}{2}x_i^2\right),$$

$$J_{23} = J_{32} = \sum_{i=1}^{n} \frac{x_i}{\left(\widehat{\beta}+\widehat{\gamma}x_i\right)^2} - \frac{\widehat{\lambda}}{2}\sum_{i=1}^{n} x_i^3 \exp\left(-\widehat{\beta}x_i - \frac{\widehat{\gamma}}{2}x_i^2\right).$$

The regularity conditions referred to hold as $n \to \infty$. In practice, n is finite. So, it is natural to ask: how large n should be for the maximum likelihood estimates to perform well? We answer this question in Section 3.

3. Simulation

Here, we assess the performance of the maximum likelihood estimates with respect to sample size n. The assessment is based on a simulation study:

1. generate ten thousand samples of size n from (1). The inversion method was used to generate samples.
2. compute the maximum likelihood estimates for the ten thousand samples, say $\left(\widehat{\lambda}_i, \widehat{\beta}_i, \widehat{\gamma}_i\right)$ for $i = 1, 2, \ldots, 10000$.
3. compute the biases and mean squared errors given by

$$\text{bias}_h(n) = \frac{1}{10000}\sum_{i=1}^{10000}\left(\widehat{h}_i - h\right),$$

and

$$\text{MSE}_h(n) = \frac{1}{10000}\sum_{i=1}^{10000}\left(\widehat{h}_i - h\right)^2$$

for $h = \lambda, \beta, \gamma$.

We repeated these steps for $n = 10, 11, \ldots, 100$ with $\lambda = 1$, $\beta = 1$ and $\gamma = 1$, so computing $\text{bias}_\lambda(n)$, $\text{bias}_\beta(n)$, $\text{bias}_\gamma(n)$ and $\text{MSE}_\lambda(n)$, $\text{MSE}_\beta(n)$, $\text{MSE}_\gamma(n)$ for $n = 10, 11, \ldots, 100$.

Figures 6 and 7 show how the three biases and the three mean squared errors vary with respect to n. The broken lines in Figure 6 correspond to the biases being zero. The broken lines in Figure 7 correspond to the mean squared errors being zero. The following observations can be made:

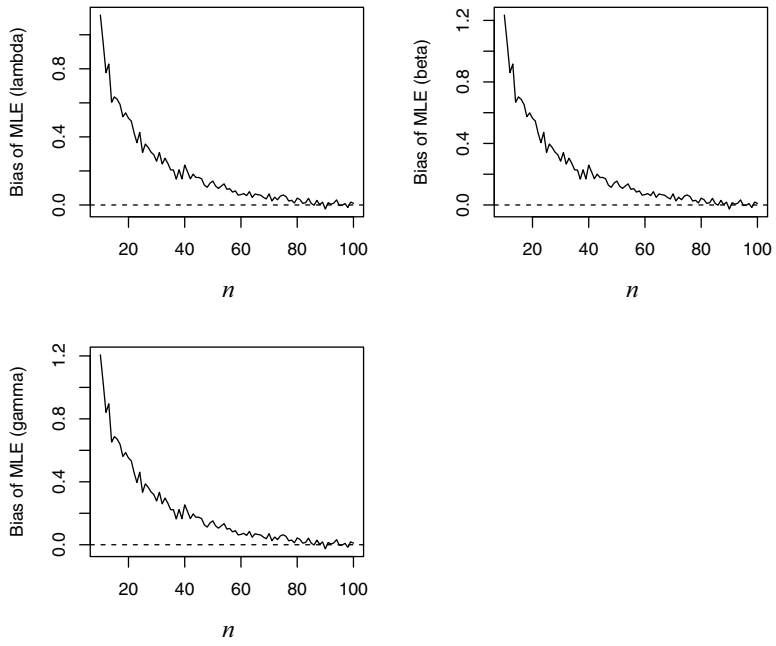

Figure 6: *From top to bottom and from left to right:* $\text{bias}_\lambda(n)$, $\text{bias}_\beta(n)$ *and* $\text{bias}_\gamma(n)$ *versus* $n = 10, 11, \ldots, 100$.

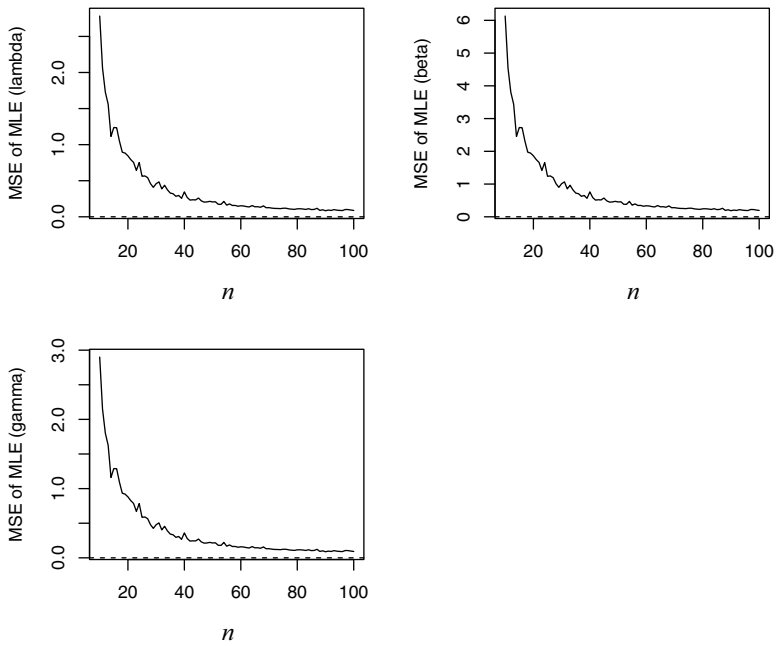

Figure 7: *From top to bottom and from left to right:* $\text{MSE}_\lambda(n)$, $\text{MSE}_\beta(n)$ *and* $\text{MSE}_\gamma(n)$ *versus* $n = 10, 11, \ldots, 100$.

1. the biases for each parameter are generally positive;
2. the biases for each parameter decrease to zero as $n \to \infty$;
3. the biases appear smallest for the parameter, λ;
4. the mean squared errors for each parameter decrease to zero as $n \to \infty$;
5. the mean squared errors appear smallest for the parameter, λ;
6. the mean squared errors appear largest for the parameter, β;
7. the biases and mean squared errors for each parameter appear reasonably small for all $n \geq 60$.

We have presented results for only one choice for (λ, β, γ), namely that $(\lambda, \beta, \gamma) = (1, 1, 1)$. But the results were similar for a wide range of other choices. In particular, the biases and mean squared errors for each parameter appeared reasonably small for all $n \geq 60$.

The three real data sets in Section 4 each has a sample size greater than or equal to sixty. So, we can expect the estimates in Section 4 to be reasonable.

4. Real data applications

Here, we return to the three data sets to illustrate the applicability of the LFRP distribution. The following distributions were fitted to each data: the LFR, LFRG, LFRP, LFRL, WG, WP, WL, GEG, GEP and GEL distributions. We also fitted the Weibull and gamma distributions given by the PDFs

$$f(x) = \frac{\beta x^{\beta-1}}{\gamma^{\beta}} \exp\left[-\left(\frac{x}{\gamma}\right)^{\beta}\right]$$

and

$$f(x) = \frac{x^{\beta-1}}{\gamma^{\beta}\Gamma(\beta)} \exp\left(-\frac{x}{\gamma}\right),$$

respectively, for $x > 0$, $\alpha > 0$ and $\beta > 0$. Each distribution was fitted by the method of maximum likelihood. The parameter estimates, standard errors, $-\ln L$, AIC values and BIC values are given in Tables 2, 3 and 4. The standard errors were computed by inverted the observed information matrices.

We see that the LFRP distribution yields the smallest $-\ln L$, the smallest AIC and the smallest BIC for each data set. It provides a significantly better fit than the LFR distribution for each data set, as judged by the likelihood ratio test. The standard errors for the LFRP distribution appear reasonable, as they are smaller than the parameter estimates.

Table 2: *Parameter estimates, standard errors, log-likelihood, AIC and BIC for the twelve distributions fitted to the patient data of Dispenzieri et al. (2012).*

Distribution	$\widehat{\lambda}$	SE	$\widehat{\beta}$	SE	$\widehat{\gamma}$	SE	$-\ln L$	AIC	BIC
LFR			0.348	0.176	5.071	0.739	7.747	19.494	24.704
LFRG	0.001	0.000	0.348	0.176	5.069	0.738	7.751	21.503	29.318
LFRP	1.894	0.851	1.132	0.661	5.591	1.212	4.960	15.921	23.736
LFRL	0.001	0.000	0.342	0.173	5.063	0.734	7.750	21.500	29.315
WG	0.999	0.000	1.839	0.156	0.561	0.032	11.838	29.676	37.491
WP	2.230	0.910	1.434	0.204	0.394	0.065	8.818	23.637	31.452
WL	0.001	0.000	1.848	0.157	0.563	0.032	11.841	29.682	37.498
GEG	0.999	0.000	2.012	0.285	2.925	0.307	21.182	48.365	56.180
GEP	3.850	1.032	1.095	0.326	3.947	0.377	11.689	29.377	37.193
GEL	0.001	0.000	2.011	0.285	2.923	0.307	21.184	48.368	56.183
Weibull			1.839	0.156	0.561	0.032	11.837	27.674	32.885
Gamma			2.068	0.272	0.245	0.036	19.387	42.774	47.985

Table 3: *Parameter estimates, standard errors, log-likelihood, AIC and BIC for the twelve distributions fitted to the hard drive failure data.*

Distribution	$\widehat{\lambda}$	SE	$\widehat{\beta}$	SE	$\widehat{\gamma}$	SE	$-\ln L$	AIC	BIC
LFR			0.296	0.138	2.530	0.393	42.043	88.087	93.297
LFRG	0.000	0.000	0.292	0.135	2.465	0.384	42.072	90.143	97.959
LFRP	1.753	0.691	0.776	0.375	2.841	0.572	38.849	83.698	91.514
LFRL	0.000	0.000	0.296	0.138	2.530	0.393	42.044	90.088	97.904
WG	0.999	0.000	1.751	0.152	0.774	0.046	46.993	99.986	107.801
WP	1.831	0.738	1.484	0.189	0.584	0.079	43.848	93.695	101.511
WL	0.001	0.000	1.772	0.154	0.774	0.045	46.984	99.967	107.783
GEG	0.999	0.000	1.842	0.257	2.017	0.216	55.885	117.769	125.585
GEP	3.569	1.080	1.016	0.327	2.664	0.256	47.407	100.814	108.629
GEL	0.000	0.000	1.876	0.263	2.035	0.217	55.887	117.774	125.589
Weibull			1.772	0.154	0.775	0.045	46.982	97.964	103.174
Gamma			1.902	0.249	0.369	0.055	54.304	112.608	117.818

The parameter estimates and the log-likelihood values of the LFRG and LFRL distributions are very close for all three data sets. This suggests that the likelihood surfaces for the LFRG and LFRL distributions attain their maximum points along the border corresponding to $\lambda = 0$. We noted earlier LFRG and LFRL distributions reduce to the LFR distribution as $\lambda \downarrow 0$. So, the fits of LFRG and LFRL distributions do not improve on the fit of the LFR distribution for the three data sets.

Table 4: *Parameter estimates, standard errors, log-likelihood, AIC and BIC for the twelve distributions fitted to the failure data of Hong and Meeker (2013).*

Distribution	$\widehat{\lambda}$	SE	$\widehat{\beta}$	SE	$\widehat{\gamma}$	SE	$-\ln L$	AIC	BIC
LFR			0.028	0.061	9.349	0.968	−32.148	−60.296	−55.086
LFRG	0.000	0.000	0.047	0.081	9.523	1.000	−32.069	−58.138	−50.323
LFRP	5.023	1.719	1.361	1.458	15.188	3.681	−48.555	−91.111	−83.295
LFRL	0.000	0.000	0.019	0.052	9.389	0.967	−32.133	−58.267	−50.451
WG	0.999	0.000	3.149	0.256	0.482	0.016	−44.743	−83.485	−75.670
WP	4.940	1.837	1.703	0.313	0.287	0.054	−46.938	−87.876	−80.061
WL	0.000	0.000	3.146	0.255	0.483	0.016	−44.745	−83.489	−75.674
GEG	0.003	0.000	4.552	0.476	0.753	0.133	−29.354	−52.708	−44.893
GEP	8.160	1.966	1.859	0.584	7.352	0.588	−42.532	−79.064	−71.249
GEL	2.082×10^{-5}	0.000	5.546	0.918	5.304	0.442	−25.126	−44.253	−36.437
Weibull			3.146	0.255	0.483	0.016	−44.745	−85.489	−80.279
Gamma			5.371	0.735	0.081	0.012	−31.814	−59.629	−54.418

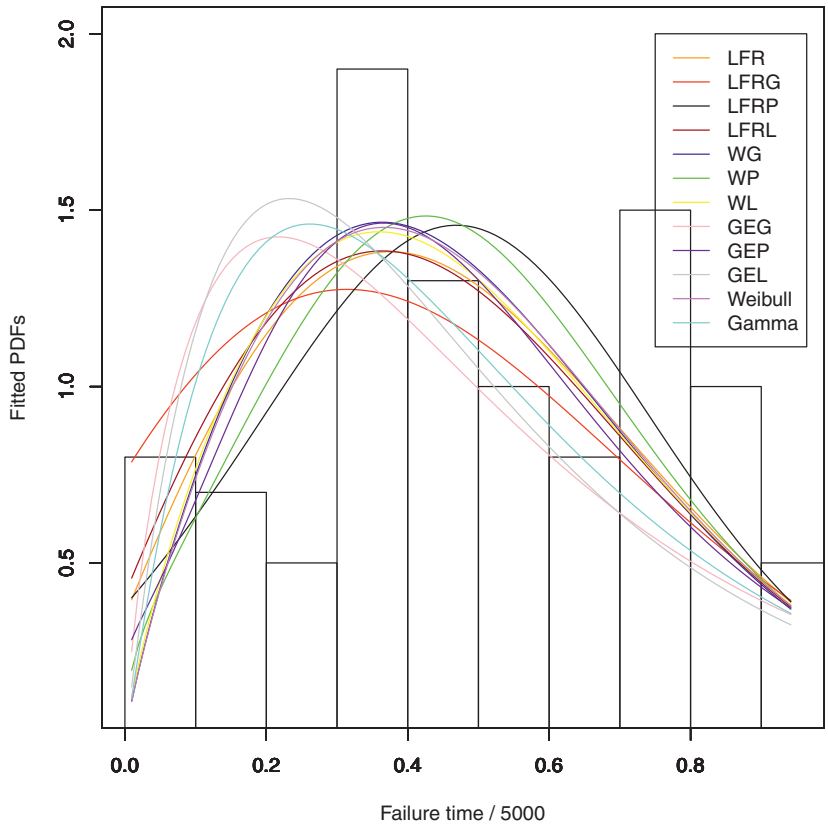

Figure 8: *Density plots for the twelve distributions fitted to the patient data of Dispenzieri et al. (2012).*

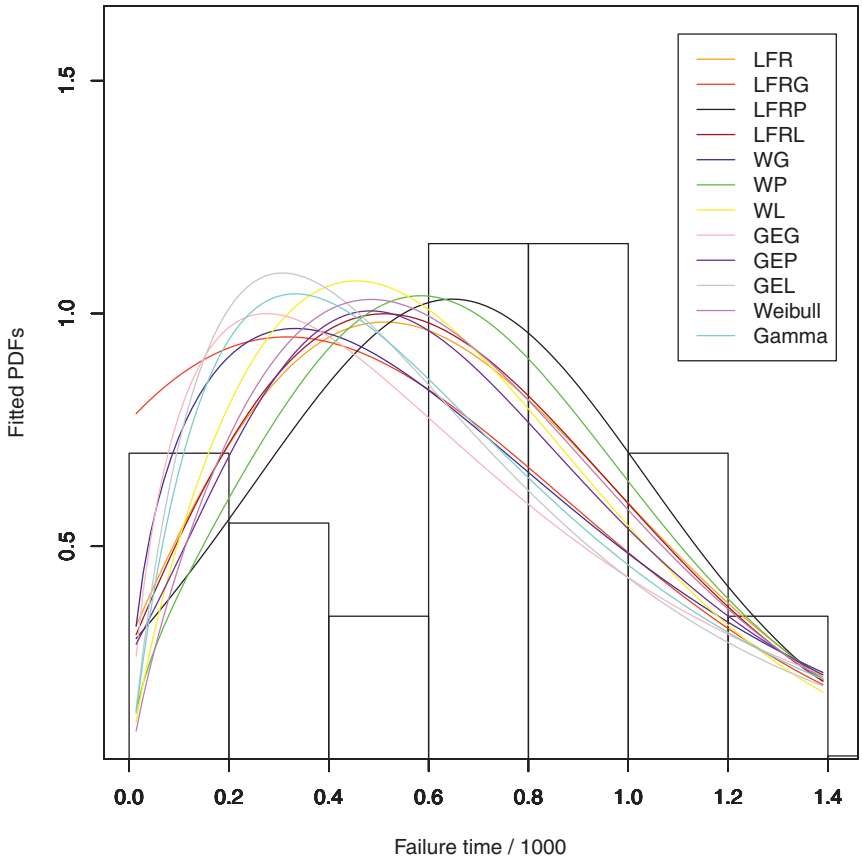

Figure 9: *Density plots for the twelve distributions fitted to the hard drive failure data.*

The density plots for the fit of the distributions for the three data sets are shown in Figures 8 to 10. The fitted PDFs of the LFRP distribution captures the observed histograms better than others. Hence, we can say that the LFRP distribution provides the best fit for at least three real data sets.

The parameter estimates of the best fitting LFRP distribution for the three data sets can be interpreted as follows:

- the patient's body can be modelled as a series system having an average of $\widehat{\lambda}/\left[1-e^{-\widehat{\lambda}}\right]=2.2$ components with the 95 percent confidence interval $(0.37,4.09)$, where the failure rate of each component is linear with an intercept of 1.132 and a slope of 5.591. That is, the failure rate of each component at time zero is 1.132 and the failure rate increases by 5.591 for every unit increase in time;
- the hard drive can be modelled as a series system having an average of $\widehat{\lambda}/\left[1-e^{-\widehat{\lambda}}\right]=2.1$ components with the 95 percent confidence interval $(1.26,2.98)$, where the failure rate of each component is linear with an intercept of 0.776 and

a slope of 2.841. That is, the failure rate of each component at time zero is 0.776 and the failure rate increases by 2.841 for every unit increase in time;

- the product D2 can be modelled as a series system having an average of $\widehat{\lambda}/\left[1-e^{-\widehat{\lambda}}\right]=5.1$ components with the 95 percent confidence interval $(-1.97, 12.08)$, where the failure rate of each component is linear with an intercept of 1.361 and a slope of 15.188. That is, the failure rate of each component at time zero is 1.361 and the failure rate increases by 15.188 for every unit increase in time.

Note that $\lambda/\left[1-e^{-\lambda}\right]$ is the expected value of a zero truncated Poisson random variable. The stated confidence intervals were obtained by the delta method.

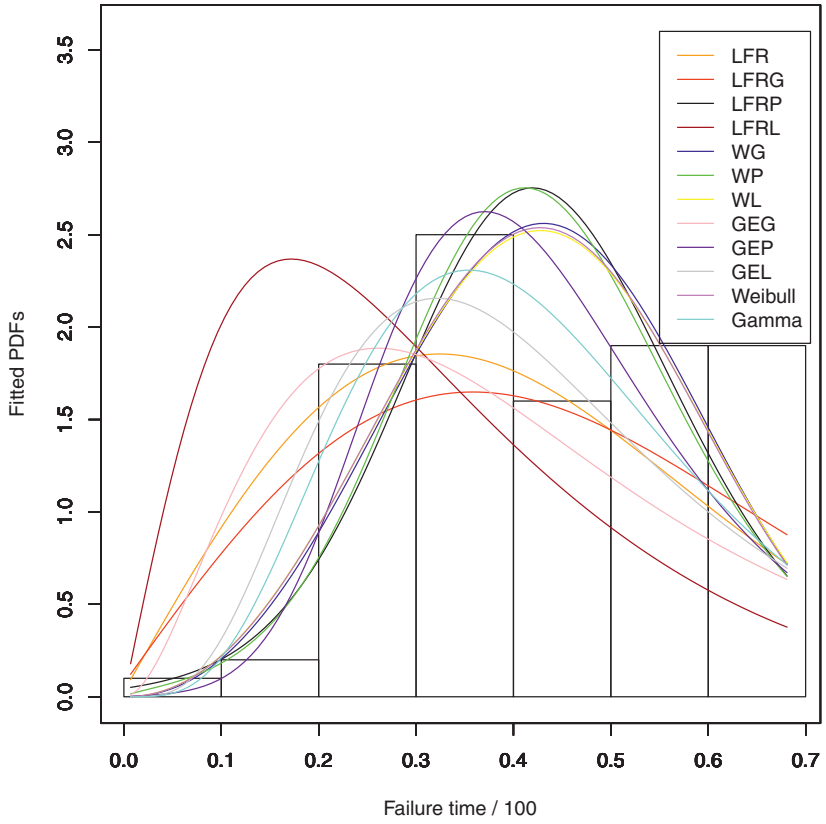

Figure 10: *Density plots for the twelve distributions fitted to the failure data of Hong and Meeker (2013).*

5. Conclusions

We have proposed three distributions motivated by three failure data sets: the linear failure rate geometric, linear failure rate Poisson and linear failure rate logarithmic distributions. Each of these distributions has three parameters.

We have studied mathematical properties and estimation issues for the linear failure rate Poisson distribution. We have shown in particular that its failure rate function can be decreasing, increasing and upside down bathtub shaped, more varied than the failure rate function of the linear failure rate distribution.

Among the twelve distributions fitted to the three data sets, the linear failure rate Poisson distribution gave the best fit. The adequacy of fits was assessed in terms AIC values, BIC values and density plots.

A future work is to estimate the parameters of the linear failure rate Poisson distribution by the method of percentiles, the method of probability weighted moments, the method of least squares, the method of weighted least squares, the method of generalized moments, and other methods. Another future work is to propose bivariate and multivariate generalizations of the linear failure rate Poisson distribution.

Appendix: Three data sets

The first data is

```
0.1102 0.2390 0.4598 0.7146 0.2608 0.0838 0.8746 0.1578
0.3358 0.0198 0.7192 0.7916 0.4486 0.4080 0.6048 0.3686
0.4686 0.5418 0.3760 0.8684 0.1572 0.4860 0.0118 0.4732
0.5450 0.8982 0.5674 0.2602 0.4330 0.3608 0.3648 0.5124
0.1360 0.7548 0.8960 0.4816 0.0818 0.3268 0.9514 0.8650
0.3372 0.5438 0.5392 0.5750 0.3672 0.6694 0.3068 0.2536
0.3756 0.3962 0.4690 0.3416 0.6430 0.9104 0.4426 0.7280
0.7370 0.7666 0.6420 0.2000 0.3588 0.6632 0.8752 0.8934
0.6526 0.1370 0.5222 0.7746 0.9230 0.6422 0.3298 0.7286
0.0054 0.3754 0.2448 0.9466 0.3256 0.3726 0.0516 0.4496
0.7850 0.8670 0.0758 0.5174 0.7742 0.5464 0.6152 0.7594
0.8310 0.4036 0.8954 0.7970 0.3638 0.0142 0.7998 0.1658
0.4572 0.7540 0.9220 0.3688
```

For computational stability with fitting distributions, we have divided each observation by 5000.

The second data is

```
1.293458333 0.251375000 1.265458333 1.404000000
1.280416667 1.201500000 1.193458333 0.340333333
1.101166667 1.059250000 1.360541667 1.245125000
1.098041667 1.049875000 1.167875000 1.271500000
1.182000000 0.925916667 0.963333333 1.119666667
0.867791667 0.845375000 0.803416667 0.323500000
1.165083333 1.065958333 1.103583333 1.035583333
1.173958333 0.886916667 0.789958333 0.671791667
0.782666667 0.534125000 0.691000000 0.813750000
0.773416667 0.629291667 0.520291667 0.635000000
```

```
0.695041667 0.712625000 0.428000000 0.423208333
0.615541667 0.254416667 0.160791667 0.125083333
0.416791667 0.215416667 0.214958333 0.185375000
0.228458333 0.206958333 0.228833333 0.190083333
0.205000000 0.007458333 0.192750000 0.227666667
0.155916667 0.179791667 0.018625000 0.169458333
0.066416667 0.005333333 0.115416667 0.080375000
0.495833333 0.854916667 0.498750000 0.902875000
0.967958333 0.786916667 0.920583333 0.943875000
0.807666667 0.761708333 0.733583333 1.043833333
0.893583333 0.746500000 0.736583333 0.880500000
0.889708333 0.780666667 0.668041667 0.861291667
0.711916667 0.718500000 0.863041667 0.908000000
0.833791667 0.671416667 0.826083333 0.823000000
0.784375000 0.667833333 0.669750000 0.835750000
```

For computational stability with fitting distributions, we have divided each observation by 1000.

The third data is

```
0.222673061 0.257639905 0.328155859 0.515672484
0.583401130 0.642256077 0.621521735 0.587506929
0.594755485 0.316753044 0.550884304 0.312962380
0.516646945 0.546445582 0.600493703 0.297813235
0.332441913 0.333245894 0.364800151 0.429097225
0.627439232 0.313363071 0.579554283 0.391397547
0.125167305 0.541816854 0.665764686 0.398880874
0.402492151 0.423982077 0.428143776 0.341767913
0.514537781 0.686683383 0.333088363 0.249962985
0.226748439 0.286643595 0.645490088 0.584664074
0.397377064 0.609634794 0.353187577 0.536304985
0.406031202 0.586163204 0.648786836 0.516497130
0.318475607 0.494774308 0.436782434 0.245923132
0.618409876 0.255245760 0.464312202 0.454133994
0.387982016 0.218311879 0.526363495 0.418258490
0.272839591 0.151997829 0.492728139 0.290973052
0.471553883 0.363069573 0.668371780 0.501805967
0.600306622 0.477109810 0.515188714 0.283784543
0.600625759 0.299420135 0.368553098 0.653382502
0.687845701 0.379423961 0.279504337 0.407995757
0.685695223 0.259685231 0.514854899 0.501119729
0.003522425 0.672089253 0.630145059 0.310811342
0.384073475 0.388312955 0.268080935 0.437408445
0.634243302 0.239656858 0.391844012 0.347107733
0.499160234 0.325770026 0.290634387 0.371908794
```

For computational stability with fitting distributions, we have divided each observation by 100.

Acknowledgments

The authors would like to thank the Editor and the three referees for careful reading and comments which greatly improved the paper.

References

Bain, L.J. (1974). Analysis for the linear failure rate life testing distribution. *Technometrics*, 16, 551–559.

Barreto-Souza, W., de Morais, A.L. and Cordeiro, G.M. (2011). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation*, 81, 645–657.

Ciumara, R. and Preda, V. (2009). The Weibull-logarithmic distribution in lifetime analysis and its properties, In *Proceedings of the XIIIth International Conference "Applied Stochastic Models and Data Analysis" (ASMDA-2009)*, editors L. Sakalauskas, C. Skiadas, and E.K. Zavadskas, Institute of Mathematics and Information/Vilnius Gediminas Technical University, Vilnius, Lithuania, 395–399.

Cox, D.R. and Hinkley, D.V. (1979). *Theoretical Statistics*. London: Chapman and Hall.

Dispenzieri, A., Katzmann, J., Kyle, R., Larson, D., Therneau, T., Colby, C., Clark, R., Mead, G., Kumar, S., Melton III, L.J. and Rajkumar, S.V. (2012). Use of monclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clinic Proceedings*, 87, 512–523.

Donida, G. and Mentrasti, L. (1982). A linear failure criterion for concrete under multiaxial states of stress. *International Journal of Fracture*, 19, 53–66.

Elhai, J.D. Calhoun, P.S. and Ford, J.D. (2008). Statistical procedures for analyzing mental health services data. *Psychiatry Research*, 160, 129–136.

Ferguson, T.S. (1996). *A Course in Large Sample Theory*. London: Chapman and Hall.

Ginebra, J. and Puig, X. (2010). On the measure and the estimation of evenness and diversity. *Computational Statistics and Data Analysis*, 54, 2187–2201.

Hong, Y. and Meeker, W.Q. (2013). Field-failure predictions based on failure-time data with dynamic covariate information. *Technometrics*, 55, 135–149.

Lehmann, L.E. and Casella, G. (1998). *Theory of Point Estimation*, second edition. New York: Springer Verlag.

Lu, W. and Shi, D. (2012). A new compounding life distribution the Weibull-Poisson distribution. *Journal of Applied Statistics*, 39, 21–38.

Mahmoudi, E. and Jafari, A.A. (2012). Generalized exponential-power series distributions. *Computational Statistics and Data Analysis*, 56, 4047–4066.

Prudnikov, A.P., Brychkov, Y.A. and Marichev, O.I. (1986). *Integrals and Series*, volume 1. Amsterdam: Gordon and Breach Science Publishers.

R Development Core Team. (2014). *A Language and Environment for Statistical Computing: R Foundation for Statistical Computing*. Vienna, Austria.

Reddy, Y.S.N. and Reddy, J.N. (1992). Linear and non-linear failure analysis of composite laminates with transverse shear. *Composites Science and Technology*, 44, 227–255.

Silva, R.B. Barreto-Souza, W. and Cordeiro, G.M. (2010). A new distribution with decreasing, increasing and upside-down bathtub failure rate. *Computational Statistics and Data Analysis*, 54, 935–944.

Singh, H. and Misra, N. (1994). On redundancy allocations in systems. *Journal of Applied Probability*, 31, 1004–1014.

Sutar, S.S. and Naik-Nimbalkar, U.V. (2014). Accelerated failure time models for load sharing systems. *IEEE Transactions on Reliability*, 63, 706–714.

van der Heijden, P.G.M., Bustami, R., Cruyff, M.J.L., Engbersen, F.G. and van Houwelingen, H.C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, 3, 305–322.

Xu, S. and Hu, Z. (2011). Mapping quantitative trait loci using the MCMC procedure in SAS. *Heredity*, 106, 357–369.

SORT 40 (1) January-June 2016, 201-224

A statistical learning based approach for parameter fine-tuning of metaheuristics

Laura Calvet[1,*], Angel A. Juan[1], Carles Serrat[2] and Jana Ries[3]

Abstract

Metaheuristics are approximation methods used to solve combinatorial optimization problems. Their performance usually depends on a set of parameters that need to be adjusted. The selection of appropriate parameter values causes a loss of efficiency, as it requires time, and advanced analytical and problem-specific skills. This paper provides an overview of the principal approaches to tackle the Parameter Setting Problem, focusing on the statistical procedures employed so far by the scientific community. In addition, a novel methodology is proposed, which is tested using an already existing algorithm for solving the Multi-Depot Vehicle Routing Problem.

MSC: 90-08, 62-07.

Keywords: Parameter fine-tuning, metaheuristics, statistical learning, biased randomization.

1. Introduction

Mathematical optimization plays an important role both in research and in our everyday lives. Management of portfolios, vehicle routing or DNA sequence assembly, are only some of the fields in which optimization techniques are employed.

Most of the existing proposals to solve optimization problems can be classified into exact methods or heuristic/metaheuristic approaches (Talbi, 2009). The former guarantee the optimality of the solution found. Unfortunately, a number of relevant problems are particularly complex, and tackling them with state-of-the-art exact methods would

[*] *Corresponding author:* Laura Calvet, lcalvetl@uoc.edu

[1] Department of Computer Science, Open University of Catalonia, IN3, 08018 Barcelona, Spain, lcalvetl@uoc.edu, ajuanp@uoc.edu

[2] Department of Mathematics, Universitat Politècnica de Catalunya, 08028 Barcelona, Spain, carles.serrat@upc.edu

[3] Portsmouth Business School, University of Portsmouth, PO1 3DE, UK, jana.ries@port.ac.uk

Received: March 2015
Accepted: May 2016

require substantial computer memory and time. Problems of this kind are known to be NP-hard (Bovet and Crescenzi, 1994). The Facility Location Problem, the Knapsack Problem and the Multi-Depot Vehicle Routing Problem (MDVRP) are some examples of NP-hard problems. In these cases, heuristics present some experience-based techniques that implement strategies to obtain a sufficiently good solution in a reasonable amount of time. Although they do not provide any theoretical guarantee, they are a popular choice when solving NP-hard problems. Owing to its nature, any heuristic is problem-dependent, which restricts its application to one particular class of problems. Also, heuristics usually provide sub-optimal solutions. These factors have led to the introduction of metaheuristics.

Birattari and Kacprzyk (2009) defines metaheuristics as "general algorithmic templates that can be easily adapted to solve the most different optimization problems". A number of them are nature-inspired, include stochastic components and have several parameters (Boussaïd et al., 2013). They are present in a large number of research areas such as telecommunications (Martins and Ribeiro, 2006), machine learning (Carvalho et al., 2011), and vehicle routing (Gendreau et al., 2008), among others.

Although the performance of metaheuristics is known to depend on its parameter values, the scientific community has not formally addressed the so-called Parameter Setting Problem (PSP) until the end of the last century. According to Eiben et al. (1999), during the first decades of metaheuristics research, many scientists based their choices on tuning the parameters "by hand", i.e. experimenting with different values and selecting the ones that provide the best outputs, or "by analogy", applying settings that have been proven successful for similar problems. More recently, the need for a systematic approach towards setting of metaheuristic parameters has been increasingly outlined in the literature (Hooker, 1995; Johnson, 2002). Subsequently, researchers employ a scientific approach to tackle the PSP more frequently. It is important to highlight that the selection of a systematic methodology leads to a gain of efficiency, as in general, less time is required to fine-tune the parameters while the performance of the metaheuristic is the same if not improved. However, there is no methodology commonly accepted by the scientific community and there is also a lack of publications that compare, in an exhaustive and objective manner, the main approaches and the techniques used so far. Moreover, some of the proposed methodologies are not easily reproducible or are highly metaheuristic and problem dependent. These are some of the reasons why, in spite of the amount of parameter fine-tuning works, many practitioners go on tuning by hand or designing algorithms without parameters (or with a very low number of them), even in the case when more parameterized algorithms could lead to better performances.

This article aims to contribute to the literature by proposing a general and automated statistical learning based procedure to tackle the PSP. It is accompanied by some methodological guidelines to validate the results. In order to test the methodology and illustrate its application, the approach is employed to fine-tune a hybrid algorithm implemented to solve the MDVRP.

The remainder of this article is organized as follows. Section 2 presents a formal definition of the PSP, the existing approaches, and their main contributions. Our methodology is outlined in Section 3, followed by Section 4, which shows its application on a hybrid algorithm. A discussion of the results is reported in Section 5. Finally, Section 6 presents concluding remarks.

2. Related work on the Parameter Setting Problem

Ries et al. (2012) define the PSP as the search for a set of parameter values θ^* in the parameter space Θ such that $\forall\theta \in \Theta : \theta^* \succeq \theta$ (where $\succeq$ denotes a relation of preference), for a given metaheuristic m in the metaheuristic space M, and a given instance x or group of them X in the instance space I. In practice, the amount of time available for experimenting T may be a restriction. In this case, the solution is approximate $(\hat{\theta})$. With regards to the difficulty of this problem, Montero et al. (2014) states that: (a) it is time consuming; (b) the best set of parameter values depends on the problem at hand; and (c) the parameters can be interrelated.

During the last decades, a large number of methodologies have been put forward to solve the PSP. These proposals can be classified in two groups (Birattari and Kacprzyk, 2009): Parameter Control Strategies (PCS), and Parameter Tuning Strategies (PTS). This classification is extended by Instance-specific Parameter Tuning Strategies (IPTS), which include features of the aforementioned groups.

This section provides a brief description of each approach and some of the most cited works. We refer the interested reader to more specific publications such as Eiben et al. (1999), De Jong (2007) and Battiti and Brunato (2010) for an expanded review of PCS, Birattari and Kacprzyk (2009) in the case of PTS, and Ries (2009) for IPTS.

2.1. Parameter Control Strategies (PCS)

These methodologies aim for a dynamic fine-tuning of the parameters by controlling and adapting their values while solving a problem instance. They follow two basic steps: firstly, an initial set of parameter values is chosen; secondly, an adaptation mechanism is integrated which changes relevant parameter values. Most of these strategies apply Adaptive Parameter Control, which means that their adaptation mechanism is based on the assessment of particular information that is stored during the iterative process of a metaheuristic. This information is usually related to the goodness of intermediate solutions. Figure 1 outlines the main instructions of a PCS based on Adaptive Parameter Control. The main drawbacks of this approach are the potentially high computational effort required and the lack of acquired understanding about good parameter values each time an instance is solved.

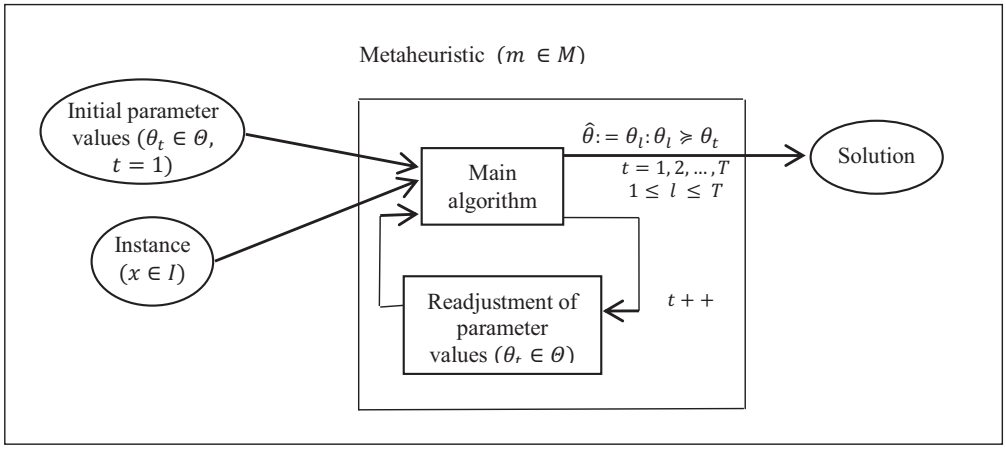

Figure 1: *Scheme of PCS applying an Adaptive Parameter Control.*

Eiben et al. (1999) addressed the PSP in Evolutionary Algorithms (EAs). Three categories were defined to classify the PCS. The first one, Deterministic Parameter Control, alters the value of a parameter by some deterministic rule, which is usually time based. The second category, Adaptive Parameter Control, does employ feedback to determine the direction and/or magnitude of a parameter change. This is the most used kind of control. Consequently, we will focus on it. The third, Self-Adaptive Parameter Control (Smith, 2008), encodes the parameters to be adapted into the chromosomes of an EA. De Jong (2007) described the main motivations to use dynamic parameter setting strategies in EAs: first, as the running proceeds, information about the fitness landscape is generated, which may be used to improve the performance; also, changing the parameters is needed as an EA "evolves from a more diffuse global search process to a more focused converging local search process".

Table 1: *Representative works employing PCS.*

Work	Main techniques	Metaheuristic	Optimization problem
Battiti and Tecchiolli (1994) and Battiti and Brunato (2005)	Reactive Scheme	Tabu Search (TS)	Quadratic Assignment Problem (QAP), and Maximum Clique Problem
Zennaki and Ech-Cherif (2010)	Support Vector Machines	TS	TSP
Lessmann et al. (2011)	Regression Models	Particle Swarm Optimization (PSO)	Water Supply Network Planning Problem

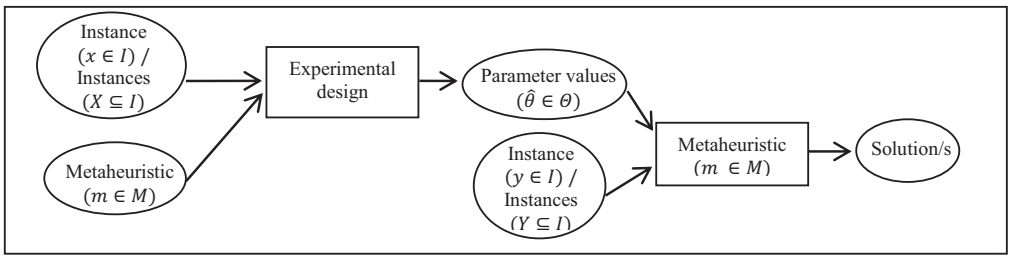

Figure 2: *Scheme of PTS.*

Table 1 gathers a few representative works following this approach. Nowadays, it constitutes a popular choice, mostly in EAs. From the literature, it can be concluded that the parameter fine-tuning is a difficult task, partly due to the potential interactions between parameters (Eiben et al., 1999; De Jong, 2007 and Smith, 2008). The worth of applying PCS is sometimes doubted (Beasley et al., 1993) or not recommended for static optimization problems (De Jong, 2007). However, most authors agree that this approach has a long way to go.

2.2. Parameter Tuning Strategies (PTS)

This approach relies on the concept of robustness (Viana et al., 2005). A robust algorithm provides good results for a given set of instances of a problem using a fixed set of parameter values. The basic procedure (Figure 2) involves finding a set of parameter values providing satisfactory results for a set of instances, usually using statistical and/or optimization techniques. Some authors analyse only a representative subset of instances and apply the set of parameter values found to solve all the instances. This approach also includes the case of solving one instance.

The work of Czarn et al. (2004) is an outstanding contribution from a statistical point of view. It addresses the issues of blocking when using design of experiments (DOE) for variation or noise due to seed, testing individual parameters and interactions, and performing power analyses, among others.

Table 2 shows some works relying on this approach. Many authors focus on minimizing the number of runs, presenting simple models without interactions (e.g., Coy et al., 2001; Pongcharoen et al. 2007 and Xu et al., 1998). DOE and regression analysis are the most employed techniques. The main criticism these works may receive is that most need an initialization of methodology-specific parameters that in some cases is not fully reported. Fortunately, the number of papers that report applications of their methodology in more than one problem or in real-world problems is increasing.

Table 2: *Representative works implementing PTS.*

Work	Main techniques	Metaheuristic	Optimization problem
Park and Kim (1998)	Simplex method	SA	Graph Partitioning Problem, Permutation Flow Shop Scheduling Problem, and Short-term Production Scheduling Problem
Xu et al. (1998)	Tree growing and pruning method based on statistical tests	TS	Steiner Tree-Star Problem
Coy et al. (2001)	DOE and Linear Regression	Routing heuristics	Vehicle Routing Problem
Bartz-Beielstein et al. (2004)	DOE, Classification and Regression Trees, and Design and Analysis of Computer Experiments	PSO and Nelder-Mead Simplex Algorithm	Elevator Group Controller Problem
Ramos et al. (2005)	Logistic Regression	EA	TSP
Birattari and Kacprzyk (2009), Birattari et al. (2010)	Racing Algorithm (Maron and Moore, 1993) and the Friedman's two-way analysis of variance by ranks (Conover, 1999)	Iterated Local Search (ILS) and Ant Colony Optimization (ACO)	QAP and TSP
Adenso-Díaz and Laguna (2006)	DOE and Local Search	Neighbourhood structure, TS, SA, TS, Heuristic based on the SA and the TS, and TS	Steiner Problem, Part-Machine Grouping Problem, Part-Machine Grouping Problem, Single-Machine Scheduling, Proportionate Flowshops, and Bandwidth Packing
Pongcharoen et al. (2007)	DOE	GA	TSP
Ridge and Kudenko (2007)	DOE and Desirability Functions	ACO	TSP
Gunawan et al. (2013)	DOE, Response Surface Methodology and ParamILS (Hutter et al., 2009)	SA	Industry Spares Inventory Optimization Problem

2.3. Instance-specific Parameter Tuning Strategies (IPTS)

As in the case of PCS, IPTS aim for an instance-specific tailoring of the parameters. At the same time, these strategies use a fixed set of parameter values, as the PTS, avoiding the need of modifying the metaheuristic algorithm and reducing the potential computational effort required to adapt parameter values during the algorithmic run. In order to implement these strategies the relation between the parameter values and the performance of the metaheuristic has to be analysed, taking into account instance features. The next step consists in developing a mechanism able to use the features of a new instance to recommend a set of parameter values. The key element is the selection of instance features that are easy and fast to compute, and good at discriminating instances on the shape of their fitness landscapes. These landscapes represent the relationship between the objective function values and the parameters. This learning may take a non-negligible amount of time, but it is assumed that this approach requires less computational time than the PCS approach does. The procedure is shown in Figure 3.

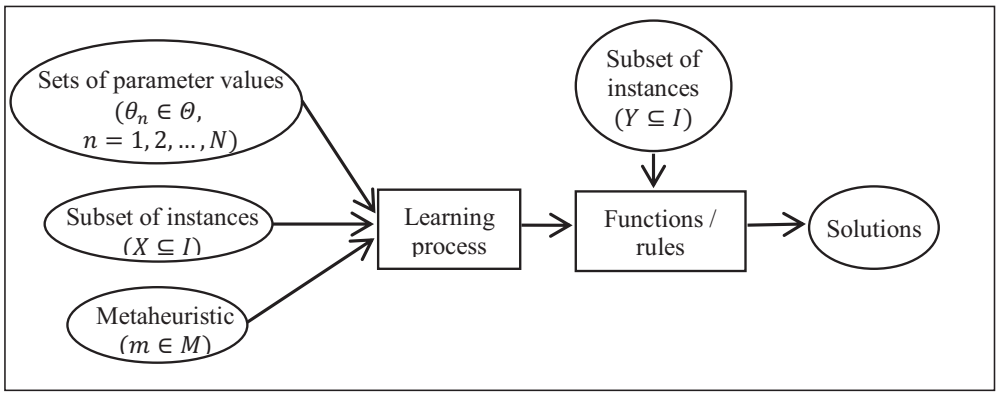

Figure 3: *Scheme of IPTS.*

Some contributions are included in Table 3. The number of works is low since it is relatively new. As in the previous cases, they employ a variety of techniques and analyse several problems.

Table 3: *Representative works implementing PTS.*

Work	Main techniques	Metaheuristic	Optimization problem
Ries (2009)	DOE and Fuzzy Logic	Guided Local Search and GA	TSP
Pavón et al. (2009)	Case-Based Reasoning and Bayesian Networks	GA	Root Identification Problem
Dobslaw (2010)	DOE and Artificial Neural Networks	PSO	TSP

It has been seen that the literature on the PSP is relatively diverse. However, more research is needed to fully explore and compare the performance of different techniques from statistics and operations research (OR), and to achieve that researchers and practitioners become aware of the relevant effect that an adequate parameter-fine tuning may have. In this paper we mainly focus on the parameter fine-tuning of metaheuristic algorithms from an OR perspective. Notice, however, that the literature on parameter fine-tuning of general algorithms is much more extensive, and it has been mainly developed by the computer science community. This community addresses a larger variety of problems (not only of optimization nature), and tends to employ algorithms with a larger number of parameters and to consider more complex and/or time-consuming approaches for setting the parameters of different types of algorithms, including searching and classification algorithms, etc. Thus, for example, Ansótegui et al. (2015) or Hutter et al. (2011) describe general but complex methods that can be used in the fine-tuning process of several types of algorithms. These general approaches are rarely considered by the OR community. Accordingly, one of the main contributions of this paper is to provide the OR community with an alternative methodology, which is easier to use and faster, and that can be employed to simplify and make more agile the fine-tuning process of metaheuristic algorithms.

2.4. Approaches comparison

All approaches have different advantages. The dynamic adaptation of the parameter values that characterizes PCS usually provides better results. However, the computational effort tends to be higher. On the other hand, the PTS approach is the easiest and fastest to use, once a set of parameter values is selected. Although the code of the algorithm is not changed, finding an adequate set may be also time-consuming. The last group of strategies represents a compromise solution: it takes less computational time than the PCS approach, but requires implementing a learning mechanism, for which statistical learning skills are needed.

Therefore, there is no approach that stands out from the others. Probably, the most adequate depends on the specific problem to tackle, the instances to solve, the available time and the skills of the researcher. Despite this fact, some general guidelines can be formulated. PTS can be considered as the best option when working with robust algorithms. Regarding IPTS, they are more complex than PTS but provide better results when the algorithm is not robust. In case of prioritizing the algorithm performance, PCS usually constitute the most recommendable approach.

3. Our approach

We propose a methodology that follows the PTS approach. There are several reasons for choosing it. Firstly, it is not computationally intensive, since it may focus on a subset of instances. The inference from a representative sample of benchmark instances to the whole set usually provides good results, specifically if the analysed algorithm is robust. There are two conditions that imply robustness. First, the algorithm has to be little sensitive to small changes in the parameter values, and second, the fitness landscapes for different instances have to be similar. These conditions guarantee that the best set of parameter values for one instance will probably provide good results for the others. The high number of works following this approach, which cover several metaheuristics and optimization problems, shows that many metaheuristic algorithms can be considered robust. Another reason for focusing on PTS is that there is no methodology based on this approach and widely employed, but at the same time, there are plenty of techniques that can be used. Some of them have been intensively tested as DOE and regression analysis. However, others remain to be investigated.

Our methodology is based on clustering (Hastie et al., 2009) and DOE (Montgomery, 2012). These are two well-established techniques that can be easily implemented using free statistical software. Clustering groups instances that have a similar fitness landscape. It facilitates the selection of representative instances and also provides information that can be used to perform a more flexible fine-tuning if each group is treated independently, i.e. exploring the fitness landscape of an instance to find a good set of parameter values and applying it to solve the instances assigned to the same group. Regarding DOE, it enables experimenters to identify and quantify the effects of several parameters and their interactions on the objective function value.

The remainder of this section presents a statistical learning based methodology to obtain a list of sets of parameter values, and a more global procedure to validate and assess its goodness.

3.1. General methodology

A four-step procedure is exposed herein. It is assumed that the experimenter has described and modelled a problem, and has chosen the metaheuristic to tackle it and a set of benchmark instances.

- The first step involves choosing a subset of the instances. Their fitness landscapes will be analysed in order to obtain sets of parameter values that provide good results for them. The subset has to be representative as these sets of parameter values will be used to solve the whole set of instances. An approach to select a representative subset is, firstly, to determine the instance features that have a major influence on which set of parameter values is the most adequate, and then, choose the instances in such a way that the feature values of the subset are representative of

those of the entire set of instances. For example, if we have a parameter for which its optimum value is known to depend on the instance size, a representative subset of the instances will present the same proportion of instances of a given size that the whole set does. This approach can be particularly difficult when there are several non-independent parameters. A possible simplification for feature selection consists of choosing those that are commonly used to discriminate instances of a specific problem. Several examples can be found in the literature. Coy et al. (2001) considered, when addressing the Capacitated Vehicle Routing Problem (CVRP), the distribution of customers, the distribution of demand and the location of the depot. Ries et al. (2012) studied the size, the distance metric, a ratio to describe the shape of the area within which a set of cities is distributed and a measure of clustering for the TSP.

In contrast, a problem-independent approach is proposed here. Initially, for a given number of randomly generated sets of parameter values, each instance is solved several times using different seeds for the random number generator of the algorithm (or only once if the algorithm is deterministic). Alternatively, the sets could also be generated using more advanced statistical techniques such as DOE. We consider the median of the objective function values found with the same parameter values but different seeds. The median is a robust measure to aggregate data, but many others could be employed. It is essential to remark the importance that a seed may have in the performance of an algorithm (Juan et al., 2015 and Czarn et al., 2004). Afterwards, feature scaling is applied to the values obtained for each instance. Then, this data is used to cluster instances and select a representative one from each cluster. These instances form the subset to analyse.

Although it is a computationally intensive approach, we think it is effective to assess which instances show a similar relation between parameter values and the performance of an algorithm.

For each instance of the subset, the steps ranging from the second to the fourth are implemented as follows.

- The second step requires selecting the range over which each parameter can be set. Some experience or knowledge about the problem and the metaheuristic may be highly valuable. The ranges should be large enough to cover at least one set of parameter values that can provide a sufficiently good solution with a high probability. On the other hand, a smaller range would allow the experimenter to describe more accurately, with the same resources, the relationship between the parameter values and the objective function value. If there is no a priori information about which are the best regions of the parameter space, a suitable procedure is to perform a rough and fast landscape analysis. Specifically, some possible combinations of parameter values can be selected and utilised to run the algorithm. The best results will identify promising regions. There are several ways of choosing the combina-

tions, as equally-spaced or randomly generated sets. This analysis holds a trade-off between the computational time required and the reliability of the conclusions.

- The third step consists of designing an experiment. A Central Composite Design is studied. Each metaheuristic parameter is considered a factor and the extreme values of its range define the levels. According to this design, the algorithm is executed also several times for each combination of factor values, each one with a different seed.
- In the fourth step, a procedure is developed to search the neighbourhood of the best set of parameter values found. Specifically, another Central Composite Design centred on this set is applied.

Finally, the upshot is a list of recommended sets of parameter values, one per cluster; in particular, those that reported the best results on the last step. The procedure is shown in Figure 4.

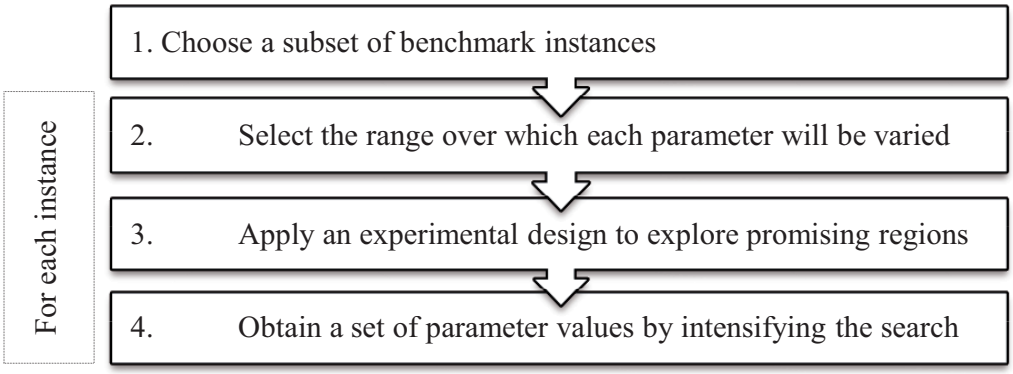

Figure 4: *Outline of the procedure for parameter fine-tuning.*

An extended proceeding (Figure 5) is described below in order to validate the list of sets of parameter values obtained and analyse the results provided by it.

Before all else, a list of sets of parameter values, $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_K)$ where K is the number of clusters, is chosen as has been explained in the precedent section. Later on, each instance of the subset used to select $\hat{\theta}$ is solved with the corresponding set of $\hat{\theta}$, and with different sets, $\bar{\theta}_j$ $(j = 1, 2, \ldots, J)$ (equally spaced, randomly selected or relatively close to the set of $\hat{\theta}$ according to some distance measure). To assess the performance of a set of $\hat{\theta}$ in a specific instance regarding the other sets, the associated solutions are compared. Given a decision level parameter r $(1 \leq r \leq J+1)$, if the rank of the objective function value provided by the proposed set is equal or lower than r, then it is considered a good set for that instance. Once all the instances of the subset are examined, the proportion of them in which the corresponding set has been classified as good can be calculated. $\hat{\theta}$ is validated by comparing this proportion with a predefined parameter

p $(0 < p < 1)$; if the proportion is higher, then the experimenter has enough evidence of the quality of $\hat{\theta}$ to go on to test it with other instances in the next step.

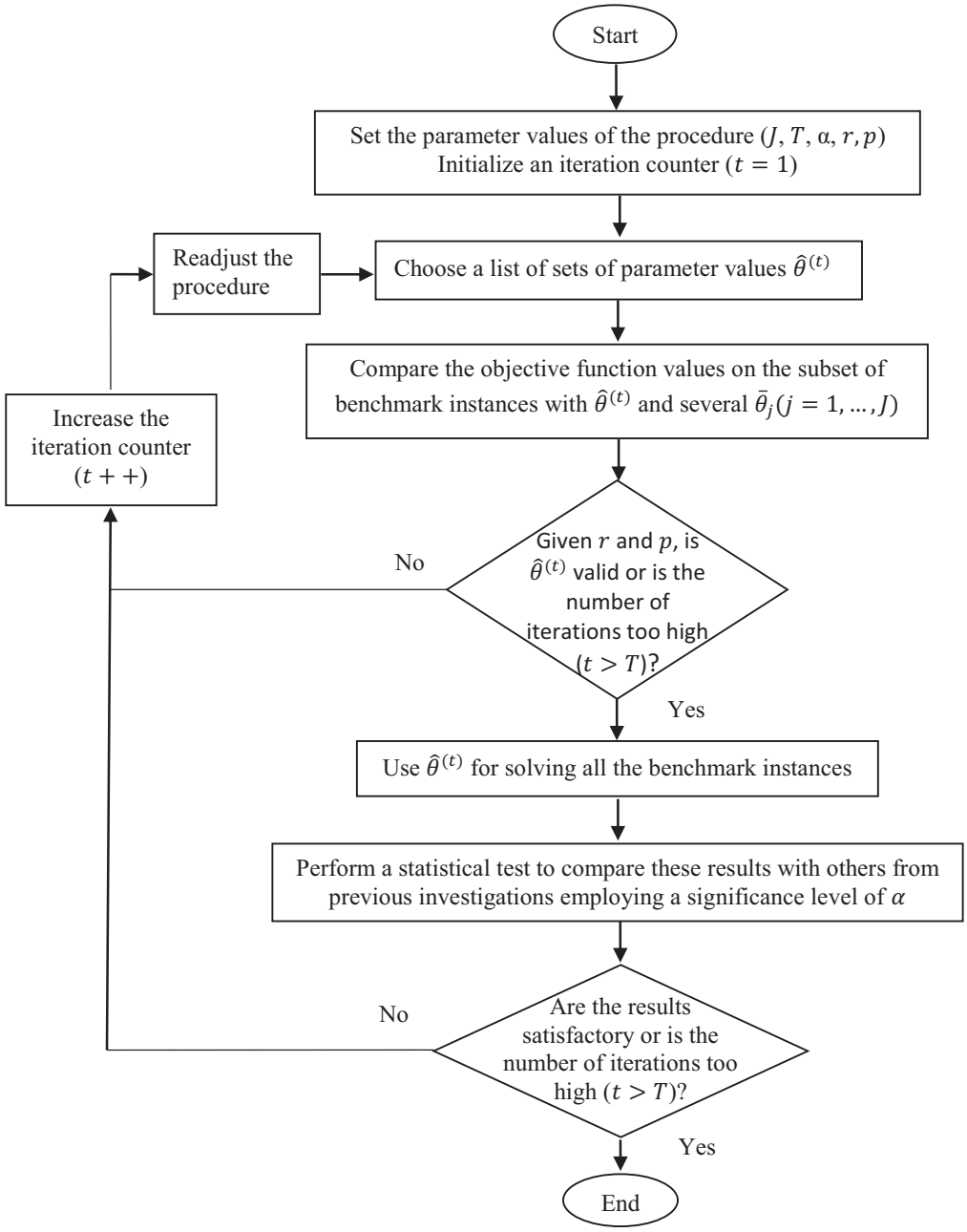

Figure 5: *Flowchart representing the proposed methodology.*

If $\hat{\theta}$ is not validated, the process has to be readjusted and restarted. This readjustment may be done in several ways, some options are: checking the robustness and the adequacy of the clustering, adapting the ranges, dedicating more resources to the search, etc. The best strategy is problem-dependent. As a consequence, the choice should rely on the opinion of the experimenter, who will have acquired valuable information from the outputs observed.

Once the list of sets of parameter values has been labelled as valid, it is applied for solving the other instances (each one with the set proposed for the representative instance of the cluster where it has been assigned). To examine the effectiveness of the procedure, it is desirable to compare the solutions with others reported in the literature for the same instances, by performing the t-test for paired samples if data are normal, or the Wilcoxon signed rank test otherwise. If the means (or the mean ranks if data are not normal) do not differ significantly, it may be classified as a satisfactory outcome as it will mean that the proposed methodology, automated and general, has been proven to be competitive. If the results are unsatisfactory, the procedure should be modified and reinitiated.

It is useful to consider that, since the available resources are usually limited, the possible readjustments should be also limited (T represents this limit). Consequently, the process may end without a satisfactory list of sets of parameter values. In this case, the list which provides on average the best solutions will be accepted.

4. Experimental results

4.1. Case study: Biased randomization and ILS for solving the Multi-Depot Vehicle Routing Problem (MDVRP)

In order to test our methodology, it was implemented to fine-tune the parameters of the hybrid algorithm described in Juan et al. (2015), which combines biased randomization and the ILS metaheuristic to address the MDVRP. A brief introduction to both the problem and the algorithm are presented in this subsection.

The MDVRP is a variant of the well-known CVRP that consists in planning routes to service a number of customers with a homogeneous fleet of vehicles that have a maximum capacity. All routes begin and end at one depot, where all resources are initially located. The objective is to find a solution (Figure 6) that minimizes the total cost while satisfying the associated constraints. Typically, these constraints imply that a single vehicle supplies each customer and it cannot stop twice at the same customer. The MDVRP integrates an allocation problem, in which the customers are assigned to one depot, with several CVRPs, one per depot. In the test case, there is also a maximum number of vehicles per depot and a maximum route length. It is considered a challenging problem as allocation and routing issues are interrelated.

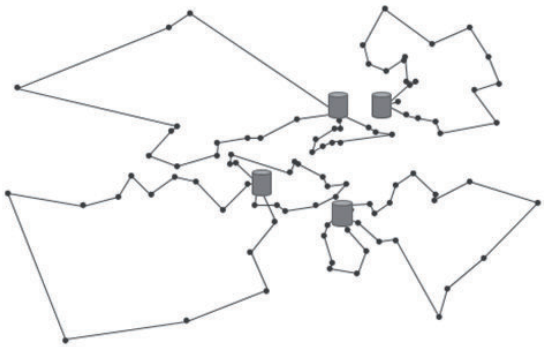

Figure 6: *Solution for a medium-size MDVRP with 4 depots (cylinders).*

The algorithm follows several steps. Initially, a priority list of potentially eligible customers is computed for each depot. The lists are sorted according to a distance-based criterion. Then, they are randomized based on a geometric distribution and used to allocate customers to depots. Afterwards, an initial solution is built by solving each routing problem independently with a version of the Clarke & Wright's Savings (CWS) heuristic (Clarke and Wright, 1964). In short, CWS starts building an initial solution in which each route includes just one customer. Following that, the heuristic considers the possibility of merging two routes if the total cost is reduced. This operation is repeated until no more merges are possible. For this project, the authors developed a biased-randomized version (Juan et al., 2011); while the original seeks always the best possible merging, this version applies biased randomization to select one merging (i.e., multiple solutions can be obtained). In the next phase, an ILS procedure is implemented. A new solution is computed by perturbing the current solution, which implies the reallocation of a given percentage of customers. The new solution replaces the current solution if the former is better. If it is also better than the best solution found so far, the latter is updated. On the other hand, if the new solution is worse than the current one, an acceptance criterion is applied and, consequently, the current base solution can still be modified. This phase ends after a fixed number of iterations. Finally, a post-optimization process is applied to the five best solutions.

This algorithm has three main parameters:

- bM: the parameter of the distribution assigning nodes to depots.
- bR: the parameter of the distribution selecting edges in the CWS heuristic.
- $p*$: the percentage of nodes that are reallocated in the ILS phase.

Note that these parameters take values between 0 and 1.

5. Implementation details

The first step is the selection of a representative subset of instances. Initially, 10 randomly generated sets of parameter values, 7 seeds and the 33 benchmark instances solved in Juan et al. (2015) were selected. Therefore, information from 2310 runs was stored. Data from different seeds was aggregated by computing the median; then feature scaling was applied. The instances that were considered easy-to-solve, those that presented no variation in the results, were separated. This was done to focus the analysis on the instances for which results could be improved by fine-tuning the parameters. Afterwards, a clustering using the *k*-medoids algorithm (Theodoridis and Koutroumbas, 2009) was performed. The range of values considered for setting the value of *k* was 2-12. The final value was selected employing the average silhouette criteria (Rousseeuw, 1987). The composition of the clusters and the representative instances (or medoids) can be observed in Table 4.

Table 4: *Clustering of the benchmark instances.*

Medoids	Clusters
p01	p01
p07	p04, p07, p11, p18, pr02, pr05, pr09
p09	p03, p09, pr04, pr10
p17	p17
p19	p19
p22	p22
p23	p20, p23
pr06	p05, p06, p08, p10, p15, pr01, pr03, pr06, pr07, pr08

Once the subset of instances was formed, the second step, setting the ranges of the parameters, was carried out. After a statistical analysis, it was concluded that just two parameters, *bM* and *bR*, did significantly affect the performance of the algorithm. Therefore, only those two parameters were studied. Five equally spaced values ranging from 0 to 1 were analysed for each parameter. Each instance was solved seven times (considering different seeds) for each possible combination of parameter values. The objective function values were aggregated as before. Then, the values for other possible combinations were estimated by linear interpolation.

The ranges were set to cover the smallest rectangular area of the parameter space that included the lowest objective function values. In particular, the values labelled as the lowest were those meeting the following condition:

$$\text{Objective solution} \leq \text{minimum value} + \beta \cdot (\text{maximum value} - \text{minimum value})$$

The value of β was set at a different value for each instance. More precisely, it was the minimum value that encompassed, at least, 5% of the search space. Figure 7 shows the contour plot and the area in which the search was intensified for each instance.

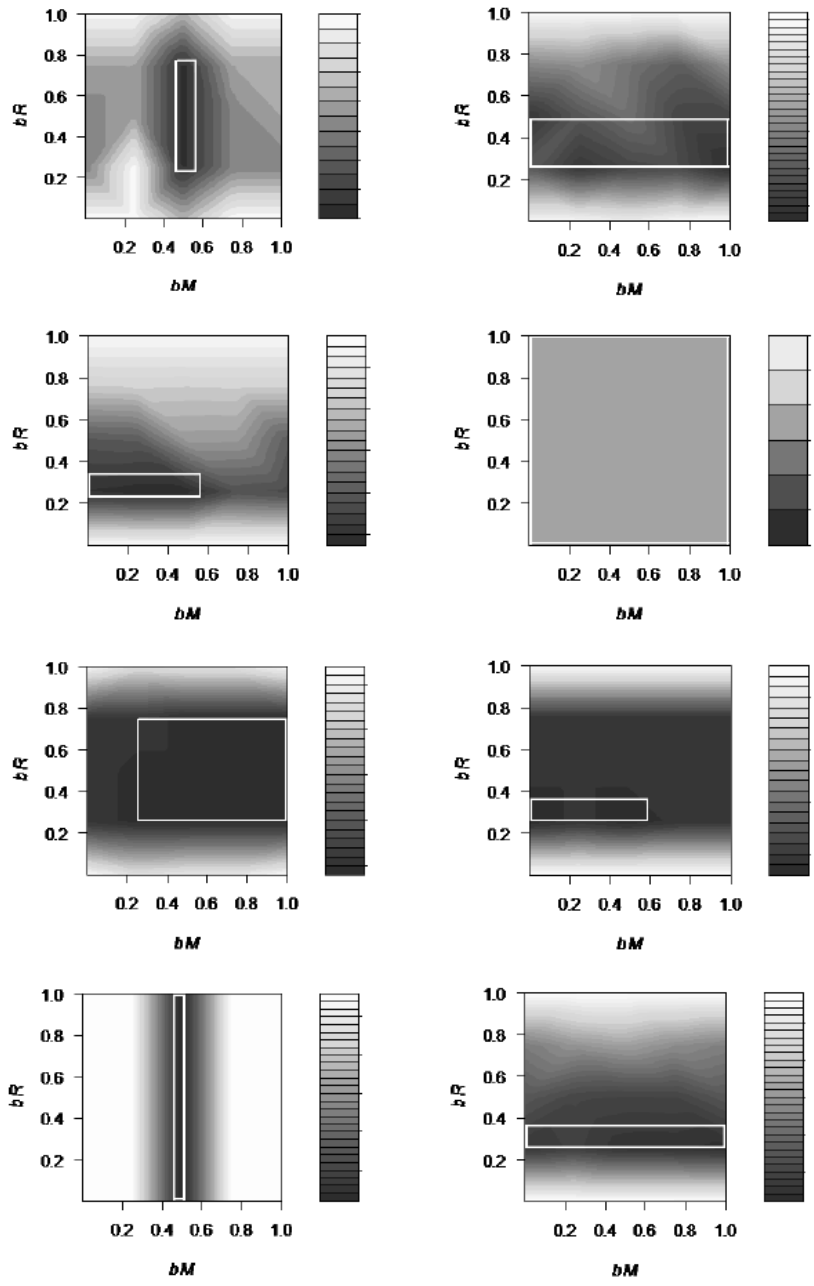

Figure 7: *Contour plots of the medoids sorted from left to right, and top to bottom.*

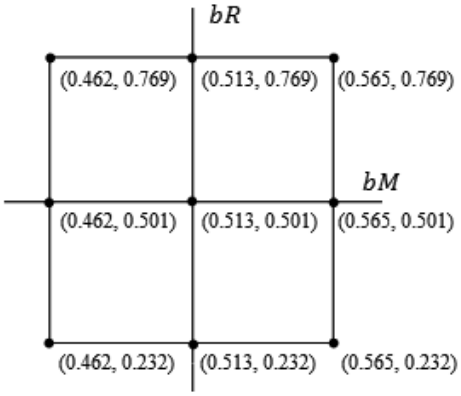

Figure 8: *Scheme of the FCC Design applied to the instance p01.*

The next step was applying a design for each instance of the subset. It was performed to better analyse the relation between the metaheuristic performance and the parameter values. A Face-Centred Central Composite (FCC) Design was selected, as in most of the cases the space parameter could not be expanded (since all parameters could only take values between 0 and 1). Figure 8 displays the scheme for instance p01. The objective function values for the same instance are represented in Figure 9.

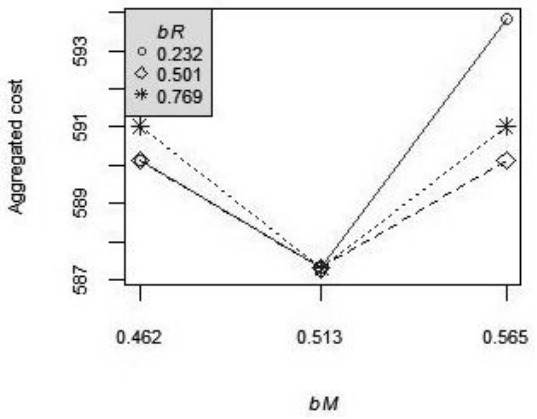

Figure 9: *Solutions of the instance p01.*

Then, the neighbourhood of each set that provided the best solution for an instance was explored applying another FCC Design, centred on that set and covering half of the area analysed with the previous design. The sets that finally presented the best performance were stored. They are outlined in Table 5. Random values were assigned to the instances that did not present variations in the results when changing the parameter values.

Table 5: *Proposed list of sets of parameter values.*

Medoids	Clusters	*bM*	*bR*
p01	p01	0.513	0.501
p07	p04, p07, p11, p18, pr02, pr05, pr09	0.001	0.372
p09	p03, p09, pr04, pr10	0.283	0.283
	p17	random	random
p19	p19	0.443	0.378
p22	p22	0.001	0.231
p23	p20, p23	0.449	0.250
pr06	p05, p06, p08, p10, p15, pr01, pr03, pr06, pr07, pr08	0.500	0.231
	p02, p12, p13, p14, p16, p21	random	random

5.1. Results

The following parameters were chosen to validate the list of sets: $J = 10$, $T = 3$, $\alpha = 0.05$, $r = 6$, $p = 0.7$. The number of sets randomly generated was fixed considering the trade-off between the reliability of our comparisons and the computational time required. The number of iterations was set considering only the time available. The significance level is the one most commonly used in the literature. The value of the fourth parameter is the mean rank that could be expected due to randomness with 11 solutions (1 set proposed and 10 randomly generated). The last parameter was calibrated to force the algorithm to provide good results at most of the instances.

The algorithm was run 7 times with different seeds for each combination of parameter values, the medians and the minimum values were stored. The ranks of the results obtained are detailed in Table 6. Ties receive a rank equal to the average of the ranks they span, shown inside the parentheses.

Table 6: *Ranks of the results provided by our list and by* 10 *random sets.*

Medoids	Rank (medians)	Rank (minimum values)
p01	1	3.5 (1-6)
p07	5	3.5 (1-6)
p09	2	2
p17	2 (1-3)	1
p19	6.5 (2-11)	10.5 (10-11)
p22	11	11
p23	1.5 (1-2)	1
pr06	5	1.5 (1-2)
Valid instances	0.75	0.75

Table 7: *Sets of parameter values for comparison.*

bM	bR	p*
Uniform (0.5, 0.8)	Uniform (0.1, 0.2)	Uniform (0.1, 0.5)

Table 8: *Instances experimental results.*

Inst.	OR medians (1)	OR, minimum values (2)	JR, medians (3)	JR, minimum values (4)	% Gap (1)-(3)	% Gap (2)-(4)
p01	585.000	576.866	593.829	576.866	−1.509	0.000
p02	480.261	476.660	480.261	476.660	0.000	0.000
p03	644.464	641.186	649.229	641.186	−0.739	0.000
p04	1022.085	1019.570	1024.473	1024.062	−0.234	−0.441
p05	760.341	756.281	764.325	754.882	−0.524	0.185
p06	882.827	879.072	880.418	879.763	0.273	−0.079
p07	899.709	897.974	906.395	897.974	−0.743	0.000
p08	4440.534	4434.552	4438.407	4426.747	0.048	0.176
p09	3920.743	3906.561	3923.248	3900.274	−0.064	0.161
p10	3706.763	3667.344	3705.012	3687.054	0.047	−0.537
p11	3598.972	3584.691	3592.891	3585.690	0.169	−0.028
p12	1318.955	1318.955	1318.955	1318.955	0.000	0.000
p13	1318.955	1318.955	1318.955	1318.955	0.000	0.000
p14	1360.115	1360.115	1360.115	1360.115	0.000	0.000
p15	2573.393	2556.846	2573.393	2557.528	0.000	−0.027
p16	2605.565	2585.373	2605.565	2600.099	0.000	−0.570
p17	2720.231	2714.663	2725.799	2725.799	−0.205	−0.410
p18	3831.996	3806.783	3835.388	3806.783	−0.089	0.000
p19	3883.686	3883.686	3883.686	3881.427	0.000	0.058
p20	4080.348	4074.779	4091.482	4091.482	−0.273	−0.410
p21	5706.530	5692.789	5701.902	5692.789	0.081	0.000
p22	5808.738	5806.370	5806.480	5786.288	0.039	0.346
p23	6134.441	6128.873	6145.576	6123.306	−0.182	0.091
pr01	861.319	861.318	861.319	861.318	0.000	0.000
pr02	1330.495	1310.679	1331.543	1314.364	−0.079	−0.281
pr03	1813.634	1813.634	1814.452	1813.634	−0.045	0.000
pr04	2084.843	2077.582	2089.785	2079.832	−0.237	−0.108
pr05	2379.075	2359.947	2379.797	2368.525	−0.030	−0.363
pr06	2709.792	2693.680	2713.593	2696.504	−0.140	−0.105
pr07	1109.235	1109.235	1109.235	1109.235	0.000	0.000
pr08	1680.896	1674.930	1678.872	1674.594	0.120	0.020
pr09	2148.216	2147.192	2153.317	2142.650	−0.237	0.212
pr10	3016.255	3008.129	3028.606	3014.874	−0.409	−0.224

According to our methodology, the list of sets can be considered valid as it presents a rank equal to or below 6 in 75% of the analysed instances, both considering medians and minimum values. In order to test our results, the algorithm was executed with the parameter values suggested in Juan et al. (2015). Both series of results are comparable as were obtained using the same computer and stopping criteria based on the number of iterations. Table 7 presents the parameter values used in the aforementioned paper. Instead of setting fixed values, the authors introduced randomness by employing uniform distributions. The lower and upper bounds were selected after some tests.

Table 8 shows the results obtained solving all instances with the proposed list of sets (our results, OR), and with the set proposed in Juan et al. (2015) (indicated as JR in the table).

6. Discussion of the results

The comparison of the solutions shows that our procedure achieves better results in most of the instances. Table 9 presents the average and the standard deviation of the differences, and the p-values of the test to compare the mean ranks of the results. It is a non-parametric test as the null hypothesis of the Shapiro-Wilk test, a test of normality, was rejected in all cases. The means are negatives, indicating that our methodology provides better solutions. The p-values reveal that the differences of the mean ranks are not statistically significant. Even though, the magnitude of the mean difference can be considered relevant in the context of the MDVRP.

Table 9: *Means and standard deviations of the differences and statistical tests.*

		Mean of the differences	Standard deviation of the differences	P-value of the comparison of mean ranks
All instances	Medians	−0.149	0.330	0.954
	Minimum values	−0.070	0.219	0.980
All instances except the studied subset and those not analysed	Medians	−0.117	0.247	0.942
	Minimum values	−0.100	0.217	0.942

Results on all instances except the subset of representative instances selected initially and those not analysed because of the null variation of their results allow us to demonstrate the good performance of our methodology, which is not directly attributed to the instances deeply studied but to their representativeness, without considering the changes in the instances that where discarded, which are due to randomness.

7. Conclusions

This paper has addressed the Parameter Setting Problem which, due to the relevance of metaheuristics in a number of fields, is increasingly getting more attention.

We have presented an overview of the main approaches: Parameter Control Strategies (PCS), Parameter Tuning Strategies (PTS), and Instance-specific Parameter Tuning Strategies (IPTS). While PCS dynamically adapt the parameter values during the resolution of an instance, PTS leave the parameter values fixed and employ them to solve several instances. IPTS represent a compromise solution, the parameter values are not modified during the search but they can be different for each instance, depending on its features. The benefits and pitfalls of each approach have been discussed. In addition, a new methodology which stands out for being automated and, problem- and metaheuristic-independent, has been presented. It incorporates techniques of clustering, which allows splitting the set of instances and, as a consequence, gives more flexibility to the fine-tuning by analysing each subset independently, and design of experiments. As a result, we have developed a methodology that avoids the strictness of common PTS, which present only a set of parameter values, and the need of modifying the main algorithm and spending more time on the resolution of instances that characterizes PCS. At the same time, our methodology is simpler than IPTS as it does not require a learning procedure able to recommend an instance-specific set of parameter values. In order to illustrate and test our methodology, it has been applied to a hybrid algorithm. The case study provides promising results.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (TRA2013-48180-C3-P, MTM2015-64465-C2-1-R (MINECO/FEDER), TRA2015-71883-REDT), and FEDER. Likewise, we want to acknowledge the support received by the Department of Universities, Research and Information Society of the Catalan Government (2014-CTP-00001).

References

Adenso-Diaz, B. and Laguna, M. (2006). Fine-tuning of algorithms using fractional experimental designs and local search. *Operations Research*, 54 , 99–114.

Ansótegui, C., Malitsky, Y., Samulowitz, H., Sellmann, M. and Tierney, K. (2015). Model-based genetic algorithms for algorithm configuration. In *Proceedings of the 24th International Conference on Artificial Intelligence* IJ-CAI'15 (pp. 733-739). AAAI Press.
URL: http://dl.acm.org/citation. cfm?id=2832249.2832351.

Bartz-Beielstein, T., Parsopoulos, K. E. and Vrahatis, M. N. (2004). Design and analysis of optimization algorithms using computational statistics. *Applied Numerical Analysis & Computational Mathematics*, 1, 413–433.

Battiti, R. and Brunato, M. (2005). *Reactive search: machine learning for memory-based heuristics*. Technical Report Teofilo F. Gonzalez (Ed.), Approximation Algorithms and Metaheuristics, Taylor Francis Books (CRC Press.

Battiti, R. and Brunato, M. (2010). Reactive search optimization: learning while optimizing. In *Handbook of Metaheuristics* (pp. 543–571). Springer.

Battiti, R. and Tecchiolli, G. (1994). The reactive tabu search. *ORSA journal on computing*, 6, 126–140.

Beasley, D., Bull, D.R., Martin, R.R. et al. (1993). An overview of genetic algorithms: Part 2, research topics. *University computing*, 15, 170–181.

Birattari, M. and Kacprzyk, J. (2009). *Tuning metaheuristics: a machine learning perspective*, volume 197. Springer.

Birattari, M., Yuan, Z., Balaprakash, P. and Stützle, T. (2010). F-race and iterated f-race: An overview. In *Experimental methods for the analysis of optimization algorithms* (pp. 311–336). Springer.

Boussaïd, I., Lepagnot, J. and Siarry, P. (2013). A survey on optimization metaheuristics. *Information Sciences*, 237 , 82–117.

Bovet, D.P. and Crescenzi, P. (1994). *Introduction to the Theory of Complexity*. Hertfordshire, UK, UK: Prentice Hall International (UK) Ltd.

Carvalho, A.R., Ramos, F.M. and Chaves, A.A. (2011). Metaheuristics for the feedforward artificial neural network architecture optimization problem. *Neural Computing and Applications*, 20 , 1273–1284.

Clarke, G. and Wright, J.W. (1964). Scheduling of vehicles from a central depot to a number of delivery points. *Operations research*, 12, 568–581.

Conover, W.J. (1999). *Practical Nonparametric Statistics*. (3rd ed.). John Wiley & Sons.

Coy, S. P., Golden, B.L., Runger, G.C. and Wasil, E.A. (2001). Using experimental design to find effective parameter settings for heuristics. *Journal of Heuristics*, 7, 77–97.

Czarn, A., MacNish, C., Vijayan, K., Turlach, B. and Gupta, R. (2004). Statistical exploratory analysis of genetic algorithms. *Evolutionary Computation, IEEE Transactions on*, 8, 405–421.

De Jong, K. (2007). Parameter setting in eas: a 30 year perspective. In *Parameter setting in evolutionary algorithms* (pp. 1–18). Springer.

Dobslaw, F. (2010). A parameter tuning framework for metaheuristics based on design of experiments and artificial neural networks. In *International Conference on Computer Mathematics and Natural Computing*. WASET.

Eiben, A.E., Hinterding, R. and Michalewicz, Z. (1999). Parameter control in evolutionary algorithms. *Evolutionary Computation, IEEE Transactions on*, 3, 124–141.

Gendreau, M., Potvin, J.-Y., Bräumlaysy, O., Hasle, G. and Løkketangen, A. (2008). *Metaheuristics for the vehicle routing problem and its extensions: A categorized bibliography*. Springer.

Gunawan, A., Lau, H.C. and Wong, E. (2013). Real-world parameter tuning using factorial design with parameter decomposition. In *Advances in Metaheuristics* (pp. 37–59). Springer.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning*. (2nd ed.). Springer.

Hooker, J.N. (1995). Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1 , 33–42.

Hutter, F., Hoos, H.H. and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *Learning and Intelligent Optimization* (pp. 507–523). Springer.

Hutter, F., Hoos, H.H., Leyton-Brown, K. and Stützle, T. (2009). Paramils: an automatic algorithm configuration framework. *Journal of Artificial Intelligence Research*, 36, 267–306.

Johnson, D.S. (2002). A theoreticians guide to the experimental analysis of algorithms. *Data structures, near neighbor searches, and methodology: fifth and sixth DIMACS implementation challenges*, 59, 215–250.

Juan, A.A., Faulin, J., Jorba, J., Riera, D., Masip, D. and Barrios, B. (2011). On the use of monte carlo simulation, cache and splitting techniques to improve the clarke and wright savings heuristics. *Journal of the Operational Research Society*, 62 , 1085–1097.

Juan, A.A., Pascual, I., Guimarans, D. and Barrios, B. (2015). Combining biased randomization with iterated local search for solving the multidepot vehicle routing problem. *International Transactions in Operational Research*, 22, 647–667.

Lessmann, S., Caserta, M. and Arango, I.M. (2011). Tuning metaheuristics: A data mining based approach for particle swarm optimization. *Expert Systems with Applications*, 38, 12826–12838.

Maron, O. and Moore, A.W. (1993). Hoeffding races: Accelerating model selection search for classification and function approximation. *Robotics Institute*, (p. 263).

Martins, S.L. and Ribeiro, C.C. (2006). Metaheuristics and applications to optimization problems in telecommunications. In *Handbook of optimization in telecommunications* (pp. 103–128). Springer.

Montero, E., Riff, M.-C. and Neveu, B. (2014). A beginner's guide to tuning methods. *Applied Soft Computing*, 17, 39–51.

Montgomery, D.C. (2008). *Design and analysis of experiments*. (8th ed.). John Wiley & Sons.

Park, M.-W. and Kim, Y.-D. (1998). A systematic procedure for setting parameters in simulated annealing algorithms. *Computers & Operations Research*, 25, 207–217.

Pavón, R., Díaz, F., Laza, R. and Luzón, V. (2009). Automatic parameter tuning with a bayesian case-based reasoning system. a case of study. *Expert Systems With Applications*, 36, 3407–3420.

Pongcharoen, P., Chainate, W. and Thapatsuwan, P. (2007). Exploration of genetic parameters and operators through travelling salesman problem. *Science Asia*, 33, 215–222.

Ramos, I.C., Goldbarg, M.C., Goldbarg, E.G. and Neto, A.D.D. (2005). Logistic regression for parameter tuning on an evolutionary algorithm. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on* (pp. 1061-1068). IEEE volume 2.

Ridge, E. and Kudenko, D. (2007). Analyzing heuristic performance with response surface models: prediction, optimization and robustness. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation* (pp. 150–157). ACM.

Ries, J. (2009). *Instance-based flexible parameter tuning for meta-heuristics using fuzzy-logic*. Ph.D. thesis University of Portsmouth.

Ries, J., Beullens, P. and Salt, D. (2012). Instance-specific multi-objective parameter tuning based on fuzzy logic. *European Journal of Operational Research*, 218, 305–315.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 , 53–65.

Smith, J.E. (2008). Self-adaptation in evolutionary algorithms for combinatorial optimisation. In *Adaptive and Multilevel Metaheuristics* (pp. 31–57). Springer.

Talbi, E.-G. (2009). *Metaheuristics: from design to implementation*, volume 74. John Wiley & Sons.

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*, volume 74. John Wiley & Sons.

Viana, A., Sousa, J.P. and Matos, M.A. (2005). Constraint oriented neighbourhoodsa new search strategy in metaheuristics. In *Metaheuristics: progress as real problem solvers* (pp. 389–414). Springer.

Xu, J., Chiu, S.Y. and Glover, F. (1998). Fine-tuning a tabu search algorithm with statistical tests. *International Transactions in Operational Research*, 5, 233–244.

Zennaki, M. and Ech-Cherif, A. (2010). A new machine learning based approach for tuning metaheuristics for the solution of hard combinatorial optimization problems. *Journal of Applied Sciences*, 10, 1991–2000.

Information for authors and subscribers

Author Guidelines

SORT accepts for publication only original articles that have not been submitted simultaneously to any other journal in the areas of statistics, operations research, official statistics or biometrics.

Articles should be preferably applied and may include computational or educational elements. Publication will be exclusively in English. All articles will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the first corresponding author if changes are requested in form or content.

To submit an article, the author must upload it in **PDF format**.

The article should be prepared in double-spaced **format**, using a 12-point typeface.

The **title page** must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (75100 words) followed by the keywords and MSC2010 classification of the American Mathematical Society.

Before submitting an article, the author(s) would be well advised to ensure that the text uses **correct English**.

Bibliographic references within the text must follow this format: author surname followed by the year of publication in parentheses [i.e., Mahalanobis (1936), Rao (1982b)]. The complete reference citations should be listed alphabetically at the end of the article, with multiple publications by a single author listed chronologically. Examples of reference formats are as follows:

- Article: Casella, G. and Robert, C. (1998). Post-processing accept-reject samples: recycling and rescaling. Journal of Computational and Graphical Statistics, 7, 139–157.
- Book: Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). Bayesian Data Analysis, 2nd Ed. Chapman & Hall / CRC, New York.
- Chapter in book: Engelmann, B. (2006). Measures of a rating's discriminative power-applications and limitations. In: Engelmann, B. and Rauhmeier, R. (eds), The Basel Risk Parameters: Estimation, Validation, and Stress Testing. Springer, New York.
- Online article (put issue or page numbers and last accessed date): Marek, M. and Lesaffre, E. (2011). Hierarchical Generalized Linear Models: The R Package HGLMMM. Journal of Statistical Software, 39 (13). http://www.jstatsoft.org/v39/i13. Last accessed 28 March 2011.

Explanatory footnotes should be numbered sequentially and placed at the bottom of the corresponding page. **Tables and figures** should also be numbered sequentially.

Once the article has been accepted, the journal editorial office will **contact the author** with instructions about this final version, which should be submitted using **LaTeX**. The journal secretary will provide authors with LaTeX templates and appropriate references to the MSC2010 classification of the American Mathematical Society. All the submissions should be handled by RACO (Revistes Catalanes en Accs Obert) website.

New Authors: please register at: http://www.raco.cat/index.php/SORT/user/register. Upon successful registration you will be sent an e-mail with instructions to verify your registration.

Submission Preparation Checklist

As part of the submission process, authors are required to check off their submission's compliance with all of the following items, and submissions may be returned to authors that do not adhere to these guidelines.

1. The submitted manuscript follows the guidelines to authors published by SORT
2. Published articles are under a Creative Commons License BY-NC-ND
3. Font size is 12 point
4. Text is double-spaced
5. Title page includes title, name(s) of author(s), professional affiliation(s), complete address of corresponding author
6. Abstract is 75100 words and contains no notation, no references and no abbreviations
7. Keywords and MSC2010 classification have been provided
8. Bibliographic references are according to SORT's prescribed format
9. English spelling and grammar have been checked
10. Manuscript is submitted in PDF format

Copyright notice and author opinions

SORT Statistics and Operations Research Transactions
Institut d'Estadstica de Catalunya (Idescat)
Via Laietana, 58 08003 Barcelona. SPAIN
Tel. +34-93.557.30.76 – Fax +34-93.557.30.01
sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form

SORT *(Statistics and Operations Research Transactions)*

Name _____

Organisation _____

Street Address _____

Zip/Postal code _____ City _____

State/Country _____ Tel. _____

Fax _____ NIF/VAT Registration Number _____

E-mail _____

Date _____

Signature

I wish to subscribe to **SORT *(Statistics and Operations Research Transactions)*** for the year 2016 (volume 40)

Annual subscription rates:

– Spain: €42 (4 % VAT included)
– Other countries: €46 (4 % VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4 % VAT included)
– Other countries: €17/issue (4 % VAT included)

Please send this subscription form (or a photocopy) to:

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-557 30 01

Or by e-mail to:

sort@idescat.cat