Austrian Journal of Statistics

AUSTRIAN STATISTICAL SOCIETY

Volume 47, Number 2, 2018

ISSN: 1026597X, Vienna, Austria

Österreichische Zeitschrift für Statistik

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Austrian Journal of Statistics; Information and Instructions

GENERAL NOTES

The Austrian Journal of Statistics is an open-access journal with a long history and is published approximately quarterly by the Austrian Statistical Society. Its general objective is to promote and extend the use of statistical methods in all kind of theoretical and applied disciplines. Special emphasis is on methods and results in official statistics.

Original papers and review articles in English will be published in the Austrian Journal of Statistics if judged consistently with these general aims. All papers will be refereed. Special topics sections will appear from time to time. Each section will have as a theme a specialized area of statistical application, theory, or methodology. Technical notes or problems for considerations under Shorter Communications are also invited. A special section is reserved for book reviews.

All published manuscripts are available at

http://www.ajs.or.at

(old editions can be found at http://www.stat.tugraz.at/AJS/Editions.html)

Members of the Austrian Statistical Society receive a copy of the Journal free of charge. To apply for a membership, see the website of the Society. Articles will also be made available through the web.

PEER REVIEW PROCESS

All contributions will be anonymously refereed which is also for the authors in order to getting positive feedback and constructive suggestions from other qualified people. Editor and referees must trust that the contribution has not been submitted for publication at the same time at another place. It is fair that the submitting author notifies if an earlier version has already been submitted somewhere before. Manuscripts stay with the publisher and referees. The refereeing and publishing in the Austrian Journal of Statistics is free of charge. The publisher, the Austrian Statistical Society requires a grant of copyright from authors in order to effectively publish and distribute this journal worldwide.

OPEN ACCESS POLICY

This journal provides immediate open access to its content on the principle that making research freely available to the public supports a greater global exchange of knowledge.

ONLINE SUBMISSIONS

Already have a Username/Password for Austrian Journal of Statistics?
Go to http://www.ajs.or.at/index.php/ajs/login

Need a Username/Password?
Go to http://www.ajs.or.at/index.php/ajs/user/register

Registration and login are required to submit items and to check the status of current submissions.

AUTHOR GUIDELINES

The original LaTeX-file guidelinesAJS.zip (available online) should be used as a template for the setting up of a text to be submitted in computer readable form. Other formats are only accepted rarely.

SUBMISSION PREPARATION CHECKLIST

- The submission has not been previously published, nor is it before another journal for consideration (or an explanation has been provided in Comments to the Editor).
- The submission file is preferable in LaTeXfile format provided by the journal.
- All illustrations, figures, and tables are placed within the text at the appropriate points, rather than at the end.
- The text adheres to the stylistic and bibliographic requirements outlined in the Author Guidelines, which is found in About the Journal.

COPYRIGHT NOTICE

The author(s) retain any copyright on the submitted material. The contributors grant the journal the right to publish, distribute, index, archive and publicly display the article (and the abstract) in printed, electronic or any other form.

Manuscripts should be unpublished and not be under consideration for publication elsewhere. By submitting an article, the author(s) certify that the article is their original work, that they have the right to submit the article for publication, and that they can grant the above license.

Austrian Journal of Statistics

Volume 47, Number 2, 2018

Editor-in-chief: Matthias TEMPL

http://www.ajs.or.at

Published by the AUSTRIAN STATISTICAL SOCIETY

http://www.osg.or.at

Österreichische Zeitschrift für Statistik

Jahrgang 47, Heft 2, 2018

ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT

Impressum

Published approximately quarterly by the Austrian Statistical Society, C/o Statistik Austria
Guglgasse 13, A–1110 Wien

Contents

Editorial

This is the last editorial that is visible in the hard copy version of an issue. The Austrian Journal of Statistics is free and open access. Due to this spirit we are moving from hard copies to electronical online versions only. Future issues will be a collection of articles without any special formatting on page numbers and table of contents. However, such information is visible in the online version of the issue.

This current issue includes five scientific papers, accessible online at http://www.ajs.or.at.

The first article is related to compositional data analysis and deals with the interpretation of regression coefficients for compositional regression. The second article investigates in censoring schemes for a life time distribution. The third contribution enhance previous work from the authors for a multivariate setting. The aim is to find an optimal allocation given box constraints in survey sampling. Distribution characteristics for fuzzy sets are considered in the fourth paper. The last paper again deals with a life time distribution.

Matthias Templ
(Editor-in-Chief)

Institute of Data Analysis and Process Design
Zurich University of Applied Sciences
Rosenstrasse 3, CH–8400 Winterthur,
Switzerland
E-mail: matthias.templ@gmail.com

Winterthur, 30. Januar 2018

Austrian Journal of Statistics
February 2018, Volume 47, 3–19.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i2.652

Interpretation of Compositional Regression with Application to Time Budget Analysis

Ivo Müller
Palacký University
Karel Hron
Palacký University
Eva Fišerová
Palacký University
Jan Šmahaj
Palacký University
Panajotis Cakirpaloglu
Palacký University
Jana Vančáková
Prostor Plus

Abstract

Regression with compositional response or covariates, or even regression between parts of a composition, is frequently employed in social sciences. Among other possible applications, it may help to reveal interesting features in time allocation analysis. As individual activities represent relative contributions to the total amount of time, statistical processing of raw data (frequently represented directly as proportions or percentages) using standard methods may lead to biased results. Specific geometrical features of time budget variables are captured by the logratio methodology of compositional data, whose aim is to build (preferably orthonormal) coordinates to be applied with popular statistical methods. The aim of this paper is to present recent tools of regression analysis within the logratio methodology and apply them to reveal potential relationships among psychometric indicators in a real-world data set. In particular, orthogonal logratio coordinates have been introduced to enhance the interpretability of coefficients in regression models.

Keywords: regression analysis, compositional data, time budget structure, orthogonal logratio coordinates, interpretation of regression parameters.

1. Introduction

Regression analysis becomes challenging when compositional data as observations carrying relative information (Aitchison 1986; Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015) occur in the role of response or explanatory variables. Although this might frequently seem to be a purely numerical problem, compositional data in any form inducing a constant sum constraint (proportions, percentages) rather represent a conceptual feature. In fact, compositional data may not necessarily be expressed with a constant sum of components (parts). The decision whether data at hand are compositional or not depends on the purpose of analysis - whether it is absolute values of components, or rather their relative structure, that is of primary interest.

One of most natural examples of compositional data are time budget (time allocation) data, discussed already in the seminal book on compositional data analysis (Aitchison 1986, p. 365). Apart from the compositional context, due to its psychological, social, and economic

impacts, time allocation and its statistical analysis receives attention in many publications. The distribution of the total amount of time among productive-, maintenance-, and leisure activities reflects the current status and soundness of economy, with its labour-saving inventions, communication technologies, means of transportation, information and mass media channels, and level of consumption (Becker 1965; Garhammer 2002; Gershuny 2000; Juster and Stafford 1991; Robinson and Godbey 1997). The economy is usually closely linked to political arrangement, which through welfare state institutions (including child-care facilities) relieve citizens of many obligations, thus opening possibilities for loosening and restructuring their daily schedules (Korpi 2000; Gershuny and Sullivan 2003; Crompton and Lyonette 2006). Leisure time service is further provided for by various sports programs, holiday resorts, outdoor activities and the like, for both adolescents and adults. Moreover, frequently also supplementary qualitative/quantitative variables (age, gender, variables resulting from psychometric scales) are of simultaneous interest, which calls for the use of regression modelling.

When considering the problem of time allocation from the statistical point of view, the individual activities represent relative contributions to the overall time budget. Particularly, although the input data can be obtained either in the original time units, or directly in proportions or percentages, the relevant information is conveyed by ratios between the parts (time activities). Consequently, also differences between relative contributions of an activity should be considered in ratios instead of absolute differences as they better reflect relative scale of the original observations.

Both scale invariance and relative scale issues are completely ignored when the raw time budget data or any representation thereof (like proportions or percentages) are analysed using standard statistical methods. Although there do exist methods whose aim is to solve purely numerical problems resulting from the nature of observations carrying relative information (being of one dimension less than the actual number of their parts), these methods usually do not represent a conceptual solution to the problem of compositional data analysis. Instead, any reasonable statistical methodology for this kind of observations should be based on ratios between parts, or even *logratios* (logarithm of ratios), which are mathematically much easier to handle (Aitchison 1986; Pawlowsky-Glahn *et al.* 2015). Logratios as a special case of a more general concept of logcontrasts are used to construct coordinates with respect to the Aitchison geometry that captures all the above mentioned natural properties of compositions. Nevertheless, possibly due to apparent complexity of the logratio methodology, logratio methods haven't still convincingly entered applications in social sciences, specifically psychological applications; methods to analyse time budget, mentioned in the seminal book of Van den Ark (van den Ark 1999) and resulting from fixing the unit-sum constraint of compositional data, were mostly overcome during the last 15 years of intensive development in the field of compositional data. Very recently statistical analysis of psychological (ipsative) data seems to attract attention (Batista-Foguet, Ferrer-Rosell, Serlavós, Coenders, and Boyatzis 2015; van Eijnatten, van der Ark, and Holloway 2015). Nevertheless, still rather specific methods are used without providing a concise data analysis, particularly concerning regression modelling that frequently occurs in psychometrics.

For this reason, the aim of this paper is to perform a comprehensive regression analysis of time budget structure of college students by taking real-world data from a large psychological survey at Palacký University in Olomouc (Czech Republic). With that view, relations with other response/explanatory variables (as well as those within the original composition) will be analysed using proper regression modelling.

The structure of the paper is as follows. In the next section, the orthonormal logratio coordinates are introduced first, and then regression modelling is discussed in more detail in Section 3. In order to achieve better interpretability of regression parameters while preserving all important features of regression models for compositional data, orthogonal coordinates (instead of orthonormal ones) are introduced as an alternative in Section 4. Section 5 is devoted to logratio analysis of the concrete time budget data set and the final Section 6 (Discussion) concludes.

2. Orthonormal logratio coordinates for compositional data

For a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)'$, considering all possible logratios $\ln(x_i/x_j)$, $i, j = 1, \ldots, D$, for statistical analysis means to take into account $D(D-1)/2$ variables (up to sign of the logarithm). This would lead to a complex ill-conditioned problem already for data sets with moderate number of variables. Moreover, information related to the original parts (although expressed possibly in logratios) is usually of primary interest. For this reason, a natural choice is to aggregate logratios meaningfully to logcontrasts (variables of type $\sum_{i=1}^{D} c_i \ln x_i$, where $\sum_{i=1}^{D} c_i = 0$), that are able to capture all the relative information about single compositional parts (time activities). In other words, when x_1 plays the role of such a part, we proceed to variable $\ln(x_1/x_2) + \ldots + \ln(x_1/x_D) = (D-1)\ln(x_1/\sqrt[D-1]{\prod_{i=2}^{D} x_i})$, i.e. to logcontrast that highlights the role of x_1 (**?**). In order to build a system of orthonormal coordinates, this variable needs to be further scaled and also the remaining $D-2$ coordinates, orthonormal log-contrasts, are constructed consequently (we refer to isometric logratio (ilr) coordinates (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, and Barceló-Vidal 2003)). One possible choice of ilr coordinates that fulfil the above requirements (for any of parts x_l, $l = 1, \ldots, D$, in place of x_1) is $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$,

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \tag{1}$$

The case of x_1 would be obtained by choosing $l = 1$. In a more general setting, the composition $(x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})'$ stands for such a permutation of the parts $(x_1, \ldots, x_D)'$ that always the l-th compositional part fills the first position, $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$. In such a configuration, the first ilr coordinate $z_1^{(l)}$ explains all the relative information (merged into the corresponding logcontrast) about the original compositional part x_l, the coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ then explain the remaining logratios in the composition. Note that the only important position is that of $x_1^{(l)}$ (because it can be fully explained by $z_1^{(l)}$), the other parts can be chosen arbitrarily, because different ilr coordinates are orthogonal rotations of each other (Egozcue *et al.* 2003). Although this particular choice of ilr coordinates has been used successfully in many geological and chemometrical applications (Buccianti, Egozcue, and Pawlowsky-Glahn 2014; Filzmoser, Hron, and Reimann 2012; Kalivodová, Hron, Filzmoser, Najdekr, Janečková, and Adam 2015), no experiences are recorded in the psychometrical context.

3. Regression analysis within the logratio methodology

Regression analysis is an important tool for analysing the relationships between the response variable Y and known explanatory variables $\mathbf{x}$, see, e.g. (Montgomery, Peck, and Vining 2006). Although in the psychometrical context it is often difficult to distinguish whether the covariates are driven by an error as well, or not, we will follow the assumption of fixed covariates in order to enable estimation of regression parameters using the standard least squares (LS) method, resulting in easy-to-handle statistical inference (hypotheses testing). When the response variables or explanatory variables are compositional, special treatment in regression is necessary. A natural way for introducing regression with compositional explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ is to perform a standard multiple regression where the explanatory variables $\mathbf{z}_i = (1, z_{i,1}, z_{i,2}, \ldots, z_{i,D-1})'$ represent the ilr coordinates of $\mathbf{x}_i$ and 1 for the intercept (Hron, Jelínková, Filzmoser, Kreuziger, Bednář, and Barták 2012). Using a special choice of ilr coordinates $\mathbf{z}^{(l)}$ given by (1), we can consider the lth ilr basis, for $l = 1, 2, \ldots, D$, and we obtain D different multiple regression models in the form

$$Y_i = \beta_0 + \beta_1^{(l)} z_{i,1}^{(l)} + \cdots + \beta_{D-1}^{(l)} z_{i,D-1}^{(l)} + \varepsilon_i^{(l)}, \quad i = 1, 2, \ldots, n, \tag{2}$$

where $\beta_0, \beta_1^{(l)}, \ldots \beta_{D-1}^{(l)}$ are unknown regression parameters and $\varepsilon_i^{(l)}$ are random errors in the lth model. Due to the orthogonality of different ilr bases, the intercept term β_0 is the same for all D models (similarly as the index of determination R^2 or the F statistic to test the overall significance of the covariates) (Hron *et al.* 2012). The regression parameters can be estimated in the standard way by the least squares (LS) method. Using the notation $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ for the observation vector, $\mathbf{Z}^{(l)} = (\mathbf{z}_1^{(l)}, \mathbf{z}_2^{(l)}, \ldots, \mathbf{z}_n^{(l)})'$ for $n \times D$ design matrix, $\boldsymbol{\beta}^{(l)} = (\beta_0, \beta_1^{(l)}, \ldots, \beta_{D-1}^{(l)})'$ for regression parameters, and $\boldsymbol{\varepsilon}^{(l)} = (\varepsilon_1^{(l)}, \varepsilon_2^{(l)}, \ldots, \varepsilon_n^{(l)})'$ for the error term, models (2) can be rewritten in the matrix form

$$\mathbf{Y} = \mathbf{Z}^{(l)}\boldsymbol{\beta}^{(l)} + \boldsymbol{\varepsilon}^{(l)}, \ l = 1, 2, \ldots, D. \tag{3}$$

We can consider that random errors in the lth model are not correlated with the same variance $\sigma_{(l)}^2$. Then the best linear unbiased estimators of regression parameters $\boldsymbol{\beta}^{(l)}$ by the LS method are

$$\widehat{\boldsymbol{\beta}}^{(l)} = (\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}\mathbf{Z}'^{(l)}\mathbf{Y}, \ l = 1, 2, \ldots, D. \tag{4}$$

From the practical point of view, only the parameter $\beta_1^{(l)}$ is important, since it corresponds to the first ilr coordinate $z_1^{(l)}$ that explains all the relative information about the part $x_1^{(l)}$. The other parameters $\beta_2^{(l)}, \ldots, \beta_{D-1}^{(l)}$ do not have such straightforward interpretation. So, we can say, e.g., that the absolute change of the conditional mean of Y with respect to coordinate $z_1^{(l)}$ is about $\beta_1^{(l)}$, if other coordinates $z_j^{(l)}$, $j = 2, 3, \ldots, D-1$ (representing subcomposition $(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)'$), are fixed.

The unbiased estimator of $\sigma_{(l)}^2$ in the lth model (3) is

$$\widehat{\sigma}_{(l)}^2 = (\mathbf{Y} - \mathbf{Z}^{(l)}\widehat{\boldsymbol{\beta}}^{(l)})'(\mathbf{Y} - \mathbf{Z}^{(l)}\widehat{\boldsymbol{\beta}}^{(l)})/(n - D), \tag{5}$$

that can be used to estimate the variance-covariance matrix of the estimator of regression parameters,

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}^{(l)}) = \widehat{\sigma}_{(l)}^2(\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}. \tag{6}$$

Under assumption of normality of random errors we can perform any standard statistical inference, e.g. test the significance of regression parameters, or to construct confidence intervals for them. The significance of the individual regression parameters in the lth model, $l = 1, 2, \ldots, D$, can be tested by the following statistics:

$$T_0 = \frac{\widehat{\beta}_0}{\widehat{\sigma}_{(l)}\sqrt{\{(\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}\}_{1,1}}}; \quad T_i^{(l)} = \frac{\widehat{\beta}_i^{(l)}}{\widehat{\sigma}_{(l)}\sqrt{\{(\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}\}_{i+1,i+1}}}, \tag{7}$$

$i = 1, 2, \ldots, D-1$. Here the symbol $\{(\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}\}_{i+1,i+1}$ denotes the $(i+1)$th diagonal element of the matrix $(\mathbf{Z}'^{(l)}\mathbf{Z}^{(l)})^{-1}$. Under the null hypothesis that regression parameters are zeros, the statistics T_0 and $T_i^{(l)}$ each follow a Student t-distribution with $n - D$ degrees of freedom. The statistic T_0 is the same irrespective of the choice of $l = 1, \ldots, D$ in (2), see (Hron *et al.* 2012) for details. Of course, the response variable can have also another distribution than normal, i.e. the methodology of generalized linear models (Dobson and Barnett 2008) can be directly implemented.

Similarly, when the response variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_D)'$ are compositional and explanatory variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)'$ are non-compositional, one can use the regression models where the response variables $Z_1, \ldots, Z_{D-1}$ represent the ilr coordinates of $\mathbf{Y}$ (Egozcue, Daunis-i Estadella, Pawlowsky-Glahn, Hron, and Filzmoser 2011). Using the ilr coordinates (1), where only the first ilr coordinate $Z_1^{(l)}$ is of interest, we obtain D different multiple regression models in the form

$$Z_{i,1}^{(l)} = \gamma_0^{(l)} + x_{i,1}\gamma_1^{(l)} + \cdots + x_{i,k}\gamma_k^{(l)} + \varepsilon_i^{(l)}, \ i = 1, 2, \ldots, n, \ l = 1, 2, \ldots, D. \tag{8}$$

In this case, the interpretation of regression parameters is the following. For example, if $x_2, \ldots, x_k$ are fixed, then for each change of 1 unit in x_1, the conditional mean of $Z_1^{(l)}$ changes $\gamma_1^{(l)}$ units. Nevertheless, similarly as for the case of regression with compositional explanatory variables, because the orthonormal coordinates (1) have to be interpreted in terms of *scaled* logratios under natural logarithm, the interpretation of these "units" and thus also values of regression parameters might get rather complex for practical purposes. Under the usual multiple regression model assumptions, (8) can be expressed in the matrix form

$$\mathbf{Z}_1^{(l)} = \mathbf{X}\boldsymbol{\gamma}^{(l)} + \boldsymbol{\varepsilon}^{(l)},\ l = 1, 2, \ldots, D, \tag{9}$$

where $\mathbf{Z}_1^{(l)} = (Z_{1,1}^{(l)}, Z_{2,1}^{(l)}, \ldots, Z_{n,1}^{(l)})'$ is an observation vector, $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_k)'$ is a vector of regression parameters, and $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$ is $n \times (k+1)$ design matrix. Here $\mathbf{1}$ is a vector of n ones. When the random errors in the lth model are not correlated with the same variance $\sigma^2_{e,(l)}$, the best linear unbiased estimator of regression parameters $\boldsymbol{\gamma}^{(l)}$ by the LS method is

$$\widehat{\boldsymbol{\gamma}}^{(l)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{Z}_1^{(l)},\ l = 1, 2, \ldots, D, \tag{10}$$

with the estimated variance-covariance matrix

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\gamma}}^{(l)}) = \widehat{\sigma}^2_{e,(l)}(\mathbf{X}'\mathbf{X})^{-1}. \tag{11}$$

The unbiased estimator of $\sigma^2_{e,(l)}$ in model (9) is

$$\widehat{\sigma}^2_{e,(l)} = (\mathbf{Z}_1^{(l)} - \mathbf{X}\widehat{\boldsymbol{\gamma}}^{(l)})'(\mathbf{Z}_1^{(l)} - \mathbf{X}\widehat{\boldsymbol{\gamma}}^{(l)})/(n-k-1). \tag{12}$$

Again, under assumption of normality of random errors we can test the significance of regression parameters, or construct confidence intervals for them. In this case, the significance of the individual regression parameters in the lth model, $l = 1, 2, \ldots, D$, can be tested by the statistic:

$$U_i^{(l)} = \frac{\widehat{\gamma}_i^{(l)}}{\widehat{\sigma}_{e,(l)}\sqrt{\{(\mathbf{X}'\mathbf{X})^{-1}\}_{i+1,i+1}}},\quad i = 0, 1, \ldots, k. \tag{13}$$

Under the null hypothesis that regression parameters are zeros, the statistics $U_i^{(l)}$ follow a Student t-distribution with $n-k-1$ degrees of freedom.

Finally, within the logratio methodology we can consider also the case of regression among parts of a composition, in particular, between a part x_0 and the rest of compositional parts, $x_1, \ldots, x_D$, in a $(D+1)$-part composition. Following (Buccianti *et al.* 2014; Hrůzová, Todorov, Hron, and Filzmoser 2016), a natural choice is to consider the case of regression with compositional explanatory variables, where the response is formed by coordinate, carrying the relative information of x_0 (with respect to compositional covariates), i.e.,

$$z_0 = \sqrt{\frac{D}{D+1}}\ln\frac{x_0}{\sqrt[D]{\prod_{i=1}^D x_i}}.$$

By construction, z_0 is orthonormal to the rest of coordinates, assigned to explanatory parts as in (1).

4. Orthogonal coordinates for compositional regression

Although the above regression models in orthonormal logratio coordinates are theoretically well justified, both the normalizing constants to reach orthonormality and the natural logarithm itself result in quite a complex interpretation of the regression parameters. A way out is to move to *orthogonal* coordinates, where nothing from the above properties of regression

modelling in coordinates is lost (in particular, values of $T_i^{(l)}$ and $U_i^{(l)}$ statistics, neither the geometrical features of regression with compositional response (Egozcue *et al.* 2011)), while, at the same time, a substantial simplification in parameter interpretation is gained. Following (1), these considerations lead to orthogonal coordinates

$$z_i^{(l)*} = \log_2 \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ i = 1, \ldots, D-1, \tag{14}$$

for $l = 1, \ldots, D$, where the normalizing constants are omitted and the original natural logarithm is replaced by the binary one. Let's see the effect of using the orthogonal coordinates for all regression models introduced above (parameters of their corresponding versions in orthogonal coordinates (14) are always marked with an asterisk). Considering regression with compositional explanatory variables first, from properties of LS estimation and the relation between logarithms of different bases we get

$$\beta_0^* = \beta_0,\ \beta_1^{(l)*} = \ln(2)\sqrt{\frac{D-1}{D}}\beta_1^{(l)},$$

generally

$$\beta_i^{(l)*} = \ln(2)\sqrt{\frac{D-i}{D-i+1}}\beta_i^{(l)},\ i = 1, \ldots, D-1,$$

and similarly for their estimates and the respective standard errors. Analogously, for models resulting from regression with compositional response we get

$$\gamma_i^{(l)*} = \log_2(\mathrm{e})\sqrt{\frac{D}{D-1}}\gamma_i^{(l)},\ i = 0, \ldots, k.$$

Finally, in regression within composition both the above effects are combined, i.e., for D regression models

$$Z_{i0} = \beta_0 + \beta_1^{(l)} z_{i,1}^{(l)} + \cdots + \beta_{D-1}^{(l)} z_{i,D-1}^{(l)} + \varepsilon_i^{(l)},\ i = 1, 2, \ldots, n, \tag{15}$$

$(l = 1, \ldots, D)$ we obtain

$$\beta_0^* = \log_2(\mathrm{e})\sqrt{\frac{D+1}{D}}\beta_0,\ \beta_i^{(l)*} = \sqrt{\frac{(D+1)(D-i)}{D(D-i+1)}}\beta_i^{(l)},\ i = 1, \ldots, D-1.$$

Indeed, the interpretation of regression coefficients gets simpler now. For regression with compositional regressors and non-compositional response, first note that a unit additive increment in a log-transformed coordinate z is equivalent to a two-fold multiplicative increase in the relative dominance of the original compositional variable x, if the base-2 logarithm is used, that is,

$$\Delta z_1^{(l)*} = \log_2 \frac{x_1^{(l)}}{\sqrt[D-1]{\prod_{i=2}^{D} x_i^{(l)}}} \cdot 2 - \log_2 \frac{x_1^{(l)}}{\sqrt[D-1]{\prod_{i=2}^{D} x_i^{(l)}}} = 1.$$

The coefficient $\beta_1^{(l)*}$ in the regression equation then has the usual meaning of an additive increase in the response y that corresponds to increasing z by one (i.e., increasing the dominance of x twice), while keeping all else fixed. For example, if $\beta_1^{(l)*} = 3$, the value of the response gets higher by 3 units when the relative dominance of the part x_l with respect to the average of the other parts, see the logratio in (14), is doubled, at constant values of the other involved covariates (orthogonal coordinates). Next, in case of regression with compositional response and non-compositional regressors, $\gamma_j^{(l)*}$ is the additive increment of the log-transformed response z when adding one to an explanatory variable x_j, $j = 1, \ldots, k$, (at constant values of the other covariates)

$$\gamma_j^{(l)*} = \Delta Z_1^{(l)*} = \log_2 \frac{Y_1^{(l)}}{\sqrt[D-1]{\prod_{i=2}^{D} Y_i^{(l)}}}\delta_j^{(l)} - \log_2 \frac{Y_1^{(l)}}{\sqrt[D-1]{\prod_{i=2}^{D} Y_i^{(l)}}} = \log_2 \delta_j^{(l)},$$

where $\delta_j^{(l)} = 2^{\gamma_j^{(l)*}}$ is the multiplicative increase in the relative dominance of the original compositional response y. So, for a unit additive change in x_j, the ratio of $Y_1^{(l)}$ to the "mean value" of the other compositional responses grows $\delta_j^{(l)} = 2^{\gamma_j^{(l)*}}$ times. Finally, an analogous interpretation for regression within composition can be obtained, namely, a two-fold multiplicative increase in the relative dominance of x_l (or equivalently, a unit additive increment in coordinate $z_1^{(l)*}$) brings the increase in the relative dominance of the response x_0 of

$$\delta_1^{(l)} = 2^{\beta_1^{(l)*}}, \text{ where } \Delta Z_0^* = \log_2 \delta_1^{(l)}.$$

Note also that the above expression for the proportionality coefficient δ stays the same irrespective of the base to which the logarithm was taken, as factor 2 in the expression now stands for two-fold increase in dominance, not for the logarithmic base.

5. Time budget analysis

Following the previous developments, the decision to admit that the time budget data are by their nature compositional invites one to couch analysis in terms of logratios instead of working with the original observations in percentages; namely, the latter would lead to biased conclusions due to relative character of compositions. The aim of this section is to demonstrate on real-world psychometric data that working with logratios in the regression context is as accessible as dealing with the original observations.

5.1. Data and methods

For this purpose, we employ data from (Vančáková 2013) that were obtained in a large psychometric study, guaranteed and realized by the Department of Psychology, Palacký University in Olomouc, Czech Republic. A questionnaire called "Leisure Time" was distributed among students at the above university, reaching a total of $N = 414$ respondents (347 women, 67 men) who provided complete answers. The items included in the questionnaire tapped three distinct areas: i) personal characteristics (age, gender, faculty and field of study); ii) leisure time (its concept, absolute and relative amount, content); iii) personality traits (self-esteem and attitude to challenges). In terms of current analysis, of particular interest are relationships among the following variables: Daily Time Budget as expressed in seven compositional variables (parts, summing up to 100 percent) *study/work*, *commuting*, *food*, *hygiene&dressing*, *sleep*, *household duties*, and *leisure time*; personality variables *self-esteem* (z-score from a 10-item Rosenberg Self-Esteem Scale (Rosenberg 1965) included in the questionnaire) and *challenge* ("Are you a person who invites challenges, i.e. opportunities to surpass yourself?", originally 4-choice response collapsed into dichotomic and coded as 1 for "always" or "almost always", and 0 for "almost never" or "never"); and covariates of *age* (in years) and *gender* (dichotomic, coded as 1 for men and -1 for women). Distribution of the variables *age* and *self-esteem* is visualized in Figure 1 in the form of EDA-plots using the R package StatDA (Filzmoser 2013).

Although the respondents were asked to enter data on Daily Time Budget in percentages, the obtained range of the sum of parts was $\langle 7, 520 \rangle$ due to misunderstanding the units to use and their prescribed constant sum constraint (of course, most of the row sums were exactly or close to 100). Nevertheless, the important information on relative contributions of parts to the overall time budget was unaffected by using whatever units, which thus emphasizes even further the necessity to apply the logratio methodology in statistical processing. Note once again that for the logratio methodology the constant sum representation of compositional data is not a necessary requirement. However, for the purpose of easier comparisons, in the following the percentage representation was taken for all time budget observations.

Besides paying attention to differences, as well as agreement, in logratio vs. "standard"

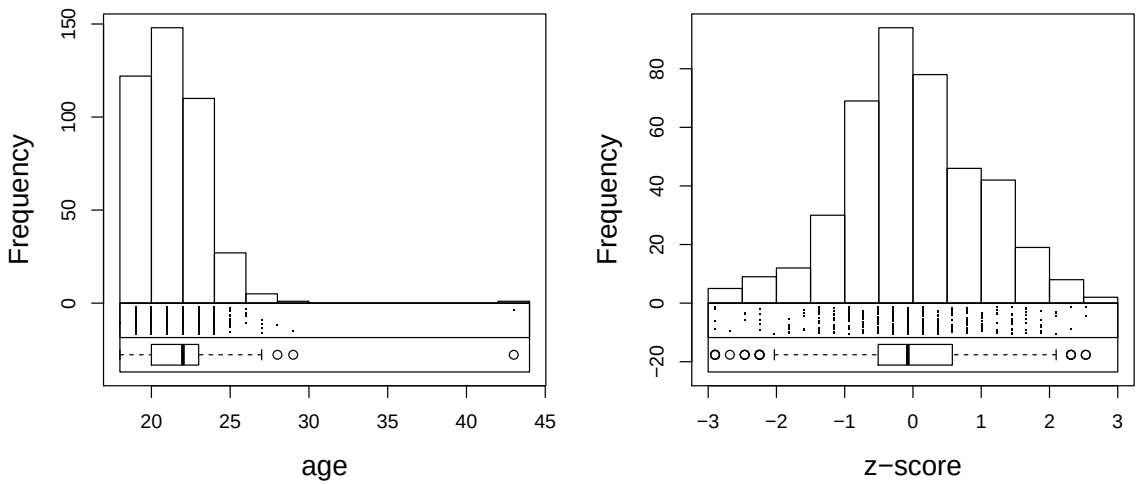

Figure 1: EDA-plots for variables *age* (left) and *self-esteem* (right).

methodologies, we will keep our thoughts focused on some tentative conjectures about interconnections among variables. This data set allows for exploring possible influences among several prominent psychological factors. On the one hand, we have the pair of personality traits of self-esteem and openness to challenge which we expect to be bundled close together and even boost each other if challenges are being tackled successfully, or else restrain each other in a downward spiral. On the other hand, the necessity of time allocation brings about an inevitable interplay of work, active relaxation, and sleep (passive relaxation). And then, of course, these two broad areas come into mutual contact in complex ways.

These considerations lead us, at the outset, to postulate a firm and positive relationship between personality traits of *challenge* and *self-esteem*. Next, within compositional variables, we deem as highly probable a negative relationship between *work/study* and *leisure time*, and between *work/study* and *sleep* on the premise that working/studying takes away time from both these forms of relaxation. *Sleep* is considered loosely associated with *leisure time* on the grounds that the time left after deducting all duties is being distributed between both. If there is more time available, it will add up to both sleep and leisure. If any at all, the relationship between *sleep* and *challenge* is expected to be negative, as the person who is busy taking challenges might have less time for sleep. The association between *sleep* and *self-esteem* is less clear-cut but it can be conceived along the lines that a self-assured person participates in numerous activities and thus sleeps less, while, on the other hand, an insecure person may seek sleep as a welcome escape from reality. As a consequence, *work/study* should be positively related with both *challenge* and *self-esteem*, and *leisure time* negatively related with both. Any effects of gender may be obscured in this dataset as men are seriously underrepresented among respondents.

In the following, the relationships among variables are determined through regression analysis. A logratio approach (which is deemed appropriate whenever a compositional variable out of Daily Time Budget is present) is compared to a standard non-compositional approach, e.g. Linear Model (LM) or Generalized Linear Model (GLM). In the statistical analysis we focused on those relations that are primarily not gender related. Moreover, preliminary exploratory analysis using variation matrix (Aitchison 1986) and compositional biplot (Aitchison and Greenacre 2002), see Figure 2, revealed strong relationship between *food* and *hygiene&dressing* components; because of their rather marginal importance for psychological interpretation, these parts will be excluded from further consideration (but kept as parts of the initial composition). On the other hand, there seems to be no relation between *commuting*

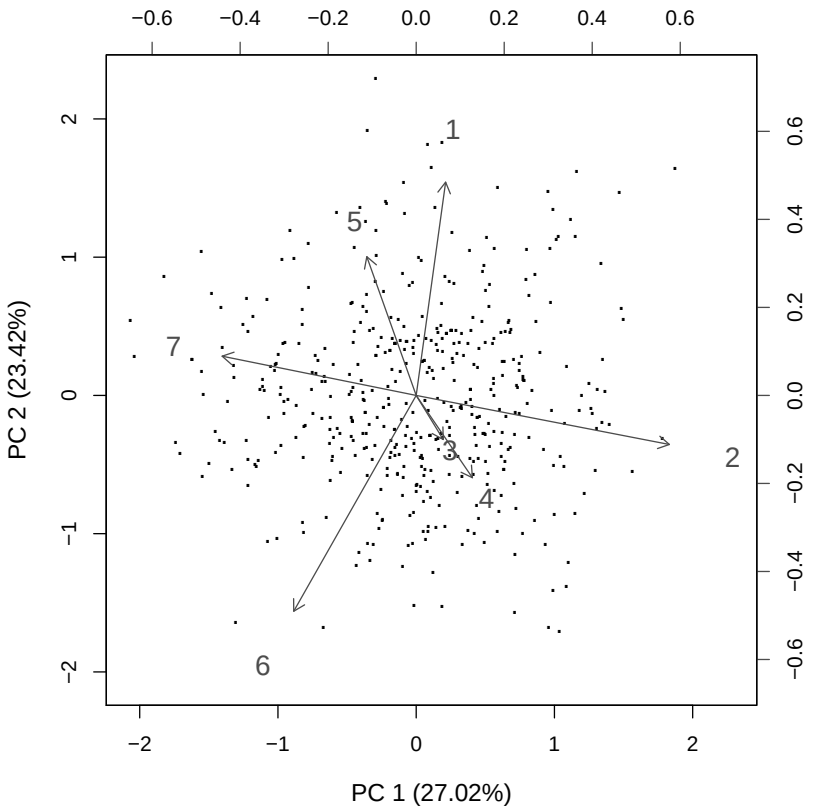

Figure 2: Compositional biplot for time budget data. Number codes correspond to single activities (1 − *study/work*, 2 − *commuting*, 3 − *food*, 4 − *hygiene&dressing*, 5 − *sleep*, 6 − *household duties*, 7 − *leisure time*)

and *leisure time*, or *study/work* and *household duties*, respectively. Interestingly, there is some nearly constant ratio also between *study/work* and *sleep* throughout the sampled population, which goes against the hypothesized association.

5.2. Regression analysis

From the essence of the data set, interconnections among variables (compositional and non-compositional, or even within the time budget composition) are of primary interest. For this purpose, several regression models were applied to data. Accordingly, in addition to Daily Time Budget, non-compositional variables of *challenge*, *self-esteem*, *age*, and *gender* were taken into consideration here. In order to enable direct interpretation of regression output, orthogonal coordinates (as described in Section 4), instead of orthonormal ones, were employed for the compositional variables within logratio approach.

As a first step, let us explore the manner how seeking challenges is determined by Daily Time Budget and other explanatory variables. That is, the response now is non-compositional (binomial), while some of the regressors are compositional and others not. For this purpose, binomial regression (a special case of logistic regression) was applied, first with compositional regressors in logratio coordinates, second with the original variables in percentages; note that any representation of the orthogonal logratio coordinates would lead to the same parameter estimates for the non-compositional covariates. From the time budget variables just those of potential psychological influence were included (*study/work*, *commuting*, *sleep*, *household duties*, and *leisure time*); of course, due to construction of the regression model in coordinates, all parts of the original composition were taken into account for the estimation purposes under logratio approach. On the other hand, perfect collinearity among compositional variables

makes it impossible to include all of them simultaneously as regressors in a standard linear model. Following (Hron *et al.* 2012), common regression output like parameter estimates, their standard errors, values of corresponding statistics and their P-values (using function `glm` from R-package MASS, see (Venables and Ripley 2002) for further details) are collected in Table 1 (all tables with detailed results are included as supplementary material), where names of the original parts stand as notation for the corresponding orthogonal coordinates (14). It can be seen that both the *study/work* coordinate and the *self-esteem* variable are contributing the most (in the positive direction, due to positive sign of their coefficients) in explaining the *challenge* response. The interpretation of coefficients is such that if the relative dominance of *study/work* in time budget doubles (with respect to average contribution of the other parts), the odds for seeking challenges increases $\exp(0.422) = 1.53$-fold (other covariates staying fixed); similarly, a unit increase in *self-esteem* z-score brings increase of the odds for seeking challenges $\exp(0.452) = 1.57$-fold. Note that, in line with the methodology described in the previous section, five regression models were employed to obtain the estimates for the compositional coordinates. By applying orthogonal coordinates (14), the interpretation of regression coefficients gets much easier than with original orthonormal coordinates (1). The tight link between *challenge* and *self-esteem* is thus established. On the other hand, we don't see significance of either *sleep* or *leisure time*, though the direction (sign of coefficient) is as expected.

For all binomial regression models the usual model diagnostics can be done, being the same irrespective which ilr coordinate system for representation of compositional predictors is taken. Specifically, jackknife deviance residuals against linear predictor, normal scores plots of standardized deviance residuals, plot of approximate Cook statistics against leverage/(1-leverage), and case plot of Cook statistic as listed, e.g. in function `glm.diag.plots` from the package `boot` can be obtained. In our case normality of residuals is rather limited, though plots of the Cook statistics do not show a significant amount of influential/leverage points that supports reliability of the results.

Finally, note that it would be also possible to add interactions between single compositional parts (represented by the respective ilr variables) and non-compositional predictors. An example of that would be possible interaction between variable *study/work* changing with *age*, i.e. fresh students and students shortly before finishing the study might have different values on *study/work* than others. Nevertheless, in order to keep simplified level of the modelling, interactions were not allowed; moreover, even when such a promising interaction was added to the model, the resulting parameter was not significant.

Table 1: Logratio approach: Results from regression of *challenge* on orthogonal coordinates of the explanatory composition and further covariates. For explanations see text.

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.69708    1.24154  -0.561  0.57448
study/work         0.42200    0.14164   2.979  0.00289
commuting         -0.06723    0.10961  -0.613  0.53959
sleep             -0.20460    0.17476  -1.171  0.24168
household duties  -0.02904    0.11187  -0.260  0.79519
leisure time      -0.13142    0.12714  -1.034  0.30129
self-esteem        0.45187    0.11105   4.069 4.72e-05
age                0.04298    0.05516   0.779  0.43586
gender             0.19698    0.15494   1.271  0.20360

    Null deviance: 552.4  on 413  degrees of freedom
Residual deviance: 521.4  on 404  degrees of freedom
AIC: 541.4

Number of Fisher Scoring iterations: 4
```

The output of binomial regression with the original compositional variables is shown in Table 2. The interpretation of regression parameters is analogous to standard multiple regression. The exponential function $\exp(\cdot)$ of the estimate of regression parameter corresponding to given covariate (either in percentages or in other units) represents amount by which the odds of *challenge* would increase/decrease if that covariate were one unit higher by constant values of the other covariates. By taking this interpretation into account, there is not much difference from the logratio approach above (also the model fit, expressed by AIC criterion, stays almost the same), which would indicate that the distortion of covariance structure among percentage covariates (see, e.g., (Aitchison 1986) for details) didn't have dramatic influence on regression output. The strength of association between openness to *challenge* and *self-esteem* remains unchanged. Nevertheless, the interesting influence of *study/work* coordinate, which was clearly visible using the logratio coordinates, is now lost.

Table 2: GLM approach: Results from binomial regression of *challenge* on original explanatory composition (in percentages) and further covariates. For explanations see text.

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.44332  1.90962  -0.232    0.816
study/work         0.02123  0.01853   1.146    0.252
commuting         -0.00242  0.03916  -0.062    0.951
sleep             -0.00467  0.02077  -0.225    0.822
household duties   0.00073  0.03098   0.024    0.981
leisure time      -0.01892  0.02253  -0.840    0.401
self-esteem        0.44518  0.11046   4.030 5.57e-05
age                0.03610  0.05418   0.666    0.505
gender             0.16862  0.15198   1.110    0.267

    Null deviance: 552.40  on 413  degrees of freedom
Residual deviance: 524.04  on 405  degrees of freedom
AIC: 542.04

Number of Fisher Scoring iterations: 4
```

As a second step, let us look the other way around and search for possible significant covariates of Daily Time Budget. For this purpose regression with compositional response was employed, the response variables now being the five chosen Daily Time Budget variables. The logratio approach leads to five univariate regression models (with orthogonal coordinates corresponding to individual compositional parts) and the results are displayed in Table 3 (to save space, just regression estimates and possible significance at the usual level $\alpha = 0.05$, marked by asterisk, are provided). The effects of particular covariates on response coordinates are evident. For example, by increasing the value of *self-esteem* by one, the relative dominance of *leisure time* in the composition (with respect to average of parts) increases approximately by 6 percent ($2^{0.088} = 1.06$). Similarly, taking challenges brings the relative dominance of *study/work* 18 percent higher ($2^{0.237} = 1.18$), and one more year of age 2.9 percent higher. The positive association between *study/work* and taking *challenges* is in accordance with our anticipations, but with *self-esteem* and *leisure time* a contrary direction was expected. The connection between *sleep* and both *challenge* and *self-esteem* remained below significance. It is interesting to see also some gender influence on both *sleep* and *leisure time*. Due to coding used (1 for male and -1 for female) it can be concluded that for males *sleep* and *leisure time* play a more important role in the overall time budget than for females. More precisely, the part *sleep* is explained only by gender. Hence, $\widehat{z}_1^{(sleep)*} = 1.644$ is the fitted value of the coordinate $z_1^{(sleep)*}$ for males, while $\widehat{z}_1^{(sleep)*} = 1.357$ for females. It means that the relative dominance of *sleep* in the composition to the "mean value" of the other compositional responses is $2^{1.644} = 3.125$ for males (3.125 times higher relative contribution of sleep than for the averaged rest of components), and $2^{1.357} = 2.562$ for females. Further, it can

be concluded that the relative dominance of *sleep* for males is $2^{2\widehat{\gamma}^*_{(sleep)}} = 1.22$ times greater than for females. Although results for *food* and *hygiene&dressing* variables are in general not discussed in this section, it is worth to note that for *hygiene&dressing* a significant role of gender (in the negative sense) was revealed; accordingly, this compositional part plays a more important role in time budget of females than for males.

Table 3: Logratio approach: Results from regression with compositional response. Significant regression parameters (at $\alpha = 0.05$) marked by asterisk.

	study/work	commuting	sleep	household	leisure time
(Intercept)	0.60673	-1.17704	1.50068*	-1.27186*	0.96194*
challenge	0.23723*	-0.02216	-0.01922	-0.07174	-0.13237
self-esteem	-0.02201	0.02367	0.03083	-0.01959	0.08817*
age	0.04117*	-0.00789	0.00186	0.02723	-0.01801
gender	-0.03217	-0.05116	0.14381*	-0.04412	0.22449*

By way of comparison, the same regression model was analysed under the assumption of Dirichlet distribution for the compositional response that is popular also in psychometric context (Georguieva, Rosenheck, and Zelterman 2008) and, although rather inconsistent with logratio methodology, is still frequently recommended for modelling compositional data. For this purpose function `DirichReg` from the package DirichletReg (Maier 2014) was applied by expressing the input compositions in proportional representation; regression output is collected in Table 4. The interpretation of regression parameters is analogous to standard multiple regression by considering proportional representation of the response and the fact that parameters of the Dirichlet distribution, being not scale invariant, are predicted. Apart from apparent computational complexity of the model, Dirichlet regression does not seem to shed new light into the problem; moreover, some of the potential relationships that have emerged with the logratio approach are lost again.

Table 4: GLM approach: Results from Dirichlet regression with compositional response. Significant regression parameters (at $\alpha = 0.05$) marked by asterisk.

	study/work	commuting	sleep	household	leisure time
(Intercept)	2.05549*	0.94065*	2.59815*	0.98208	2.21778*
challenge	0.08550	-0.05467	-0.06799	-0.08346	-0.14143
self-esteem	0.03210	0.04770	0.06618	0.02641	0.09649*
age	0.00825	-0.01401	-0.01578	-0.00027	-0.02443
gender	-0.03630	-0.03837	0.06755	-0.03792	0.11748*

From the previous analysis, *leisure time* seems to be strongly linked with the non-compositional variables. A natural question thus arises whether regression could reveal also some relations within parts of the time budget composition. Thus, as the third step, the corresponding logratio model from Section 2 was applied, by expressing both the response and regressors in orthogonal logratio coordinates (and with additional non-compositional covariates). Similarly as before, Table 5 collects results from four regression models, each highlighting the role of one of compositional explanatory variables (without influence on the non-compositional covariates). Though the R^2 statistic gives rather low value (as is usual in social science), some patterns stand out. In particular, relative dominance of *leisure time* is positively influenced by *sleep* (increasing the dominance of sleep twice enlarges the dominance of leisure time by 27 percent, as $2^{0.34} = 1.27$) and marginally by *self-esteem* (unit increase in self-esteem increases the dominance of leisure time by 6 percent); negative effects on leisure time are formed by *study/work* (10 percent decrease in dominance, $2^{-0.16} = 0.90$) and *commuting* (decrease in

dominance by 13 percent). Consistency with the previous logratio model (Table 3, regression with compositional response) is underlined by the roles of *self-esteem* and *gender* covariates. Again, a psychological interpretation can be easily derived. Here we are able to pinpoint the significant positive association of *sleep* and *leisure time*, as well as negative association of *work/study* and *leisure time*. Marginally significant is the connection between *leisure time* and *self-esteem* which appeared significant in previous regression (Table 3).

Table 5: Logratio approach: Results from regression of *leisure time* coordinate on orthogonal coordinates of the explanatory composition and further covariates. For explanations see text.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.47988    0.45492   1.055  0.29211
study/work        -0.15646    0.05576  -2.806  0.00526
commuting         -0.20414    0.04377  -4.664 4.22e-06
sleep              0.33976    0.06374   5.330 1.64e-07
household duties   0.05734    0.04304   1.332  0.18351
challenge         -0.09358    0.08937  -1.047  0.29568
self-esteem        0.08353    0.04330   1.929  0.05442
age               -0.01908    0.01991  -0.958  0.33852
gender             0.16760    0.05951   2.817  0.00509

Residual standard error: 0.8544 on 404 degrees of freedom
Multiple R-squared:  0.1619,    Adjusted R-squared:  0.1433
F-statistic: 8.674 on 9 and 404 DF,  p-value: 6.21e-12
```

For the final comparison we consider the standard linear regression model where the original parts in percentages are involved (except for *food* and *hygiene&dressing*), see Table 6 for the regression summary. Although conclusions from this model as regards non-compositional covariates would be pretty similar as with logratio methodology, the situation is different in other respects. By comparing R^2 for these two models and P-values at respective compositional covariates it is easy to see that for the standard regression model these values are very strongly driven by the constant-sum constraint of the original composition. In particular, note that by including all the compositional parts, R^2 would be brought up to 1, i.e., relations between the response and covariates would be completely driven by constant sum constraint of the input data. Of course, as statistical processing of the original compositions violates both scale invariance and relative scale properties of observations, it cannot be concluded that by considering compositional data without a constant sum constraint, the resulting regression model would be relevant. Nevertheless, in percentage representation, which is the case here, the irrelevance of the standard approach is clearly observable.

5.3. Results

The logratio approach to regression analysis supports our hypothesis of strong negative association between *work/study* and *leisure time*, as well as of strong positive association between *challenge* and *self-esteem*. Next, *leisure time* is significantly tied to *self-esteem* but the direction here appeared positive, rather than negative as expected. The reason could be that self-assured people don't feel the urge to work that much and rather take things easy, allowing themselves more leisure. Also, an explanation in keeping with (Šípek 2001) says that people with higher self-esteem may be better prepared to use their free time and it may be easier for them to admit their needs (for rest and reward). The connection between *leisure time* and *challenge* was not born out (remained below significance, though direction was negative as anticipated). The above regression results were agreed on by both logratio and standard linear model approaches. Both approaches also showed a relationship between *sleep* and *leisure time*. However, here the directions differed: logratio showed it to be positive (as hypothesized), linear model negative. On top of that, logratio approach was capable of revealing a

Table 6: Standard LM approach: Results from regression of *leisure time* on other compositional parts (in percentages) and further covariates. For explanations see text.

```
 Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               56.88605    2.96589  19.180  < 2e-16
work/study                -0.60560    0.02698 -22.446  < 2e-16
commuting                 -0.94166    0.07235 -13.016  < 2e-16
sleep                     -0.51838    0.03738 -13.869  < 2e-16
household duties          -0.63007    0.06094 -10.339  < 2e-16
challenge                 -0.41193    0.48376  -0.852  0.39498
self-esteem                0.46017    0.23468   1.961  0.05058
age                        0.05062    0.10826   0.468  0.64032
gender                     1.00488    0.31835   3.157  0.00172

Residual standard error: 4.636 on 405 degrees of freedom
Multiple R-squared:  0.6014,    Adjusted R-squared:  0.5935
F-statistic: 76.37 on 8 and 405 DF,  p-value: < 2.2e-16
```

significant positive connection between *challenge* and *work/study*.

The psychologically relevant variables seem to form a well-defined cluster of *challenge, self-esteem*, and *work/study*. Somewhat in opposition stands the pair of *leisure time* and *sleep*. However, their position with respect to the main cluster is less clearly marked, as *leisure time* is negatively linked to *work/study* but positively (perhaps only marginally) to *self-esteem*. Nevertheless it seems reasonable to assume that working/studying does take time away from both leisure and sleep simultaneously.

Finally, it is also worth noting that standard regression models were presented mostly for the sake of comparison of the logratio approach with alternatives that would be most possibly used instead. While in some cases their output might seem meaningful, it can also happen that by ignoring the relative structure of Daily Time Budget some interesting features are lost, as was the case in Table 2 and Table 4. For some cases, like when percentage representation of the relative contributions is analysed, it is very easy to demonstrate that scale invariance of compositional data leads to clear failure of the standard approach (Table 6).

6. Discussion

Specific habits of time allocation reveal a lot about an individual, a community, a society, or a culture. In each society, options available to individuals for earning their living determine the amount of time they will spend working, or preparing themselves for any such productive activity through study or apprenticeship. In modern times, we have witnessed a continuous reduction in working hours, at least in industrialized countries. At the same time, due to constant total time budget, this development leaves more space for other activities, both necessary (self- and home-maintenance like sleep, eating, hygiene, care for family and house) and discretionary (leisure activities like socializing, culture, sports, reading, idling, etc.). As the time budget data are usually accompanied with other psychometric variables, regression modelling is the first and intuitive choice for a relevant statistical analysis.

Due to relative character of time budget allocation, it seems natural to work with (log-)ratios rather than with observations in the original scale (i.e. represented usually in proportions or percentages). It turned out that logratios meet the scale invariance and relative scale requirements (among others that are important for reasonable processing of compositional data) commonly raised in connection with any observations carrying primarily relative information. The main problem is then how to construct logratio coordinates, both meaningful from the mathematical point of view (guaranteed in particular by orthonormality of coor-

dinates) and at the same time providing easy interpretation. The aim of the paper was to enhance interpretability of regression analysis output by employing orthogonal coordinates in place of the mathematically preferred orthonormal ones, demonstrated for the particular case of time budget data. The reason for the choice of alternative coordinates is that all the beneficial properties of the orthonormal coordinates are maintained also by the orthogonal ones, but the latter enable (by avoiding the scaling constants and changing the logarithmic base) a more straightforward interpretation. We are convinced that better interpretability of the regression models, discussed in the paper, can help with applicability of the logratio methodology in psychological research, and also in general.

Acknowledgements

The authors gratefully acknowledge the support of the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the grant COST Action CRoNoS IC1408.

References

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Aitchison J, Greenacre M (2002). "Biplots of Compositional Data." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 375–392.

Batista-Foguet JM, Ferrer-Rosell B, Serlavós R, Coenders G, Boyatzis RE (2015). "An Alternative Approach to Analyze Ipsative Data. Revisiting Experiential Learning Theory." *Frontiers in Psychology*, **6**(1742).

Becker GS (1965). "A Theory of the Allocation of Time." *The Economic Journal*, **75**(299), 493–517.

Buccianti A, Egozcue JJ, Pawlowsky-Glahn V (2014). "Variation Diagrams to Statistically Model the Behavior of Geochemical Variables: Theory and Applications." *Journal of Hydrology*, **519**, 988–998.

Crompton R, Lyonette C (2006). "Work-life 'Balance' in Europe." *Acta Sociologica*, **49**(4), 379–393.

Dobson AJ, Barnett AG (2008). *An Introduction to Generalized Linear Models*. CRC Press, Boca Raton.

Egozcue JJ, Daunis-i Estadella J, Pawlowsky-Glahn V, Hron K, Filzmoser P (2011). "Simplicial Regression: The Normal Model." *Journal of Applied Probability and Statistics*, **6**(1&2), 87–108.

Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003). "Isometric Logratio Transformations for Compositional Data Analysis." *Mathematical Geology*, **35**(3), 279–300.

Filzmoser P (2013). "StatDA: Statistical Analysis for Environmental Data. R package version 1.6.7." URL http://CRAN.R-project.org/package=StatDA.

Filzmoser P, Hron K, Reimann C (2012). "Interpretation of Multivariate Outliers for Compositional Data." *Computers & Geosciences*, **39**, 77–85.

Garhammer M (2002). "Pace of Life and Enjoyment of Life." *Journal of Happiness Studies*, **3**, 217–256.

Georguieva R, Rosenheck R, Zelterman D (2008). "Dirichlet Component Regression and Its Applications to Psychiatric Data." *Computation Statistics & Data Analysis*, **52**, 5344–5355.

Gershuny J (2000). *Changing Times. Work and Leisure in Postindustrial Society*. Oxford University Press, Oxford.

Gershuny J, Sullivan O (2003). "Time Use, Gender and Public Policy Regimes." *Social Politics*, **10**, 205–228.

Hron K, Jelínková M, Filzmoser P, Kreuziger R, Bednář P, Barták P (2012). "Statistical Analysis of Wines Using a Robust Compositional Biplot." *Talanta*, **90**, 46–50.

Hrůzová K, Todorov V, Hron K, Filzmoser P (2016). "Classical and Robust Orthogonal Regression between Parts of Compositional Data." *Statistics*, **50**(6), 1261–1275.

Juster FT, Stafford FP (1991). "The Allocation of Time: Empirical Findings, Behavioral Models, and Problems of Measurement." *Journal of Economic Literature*, **29**(2), 471–522.

Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T (2015). "PLS-DA for Compositional Data with Application to Metabolomics." *Journal of Chemometrics*, **29**, 21–28.

Korpi W (2000). "Faces of Inequality: Gender, Class and Patterns of Inequalities in Different Types of Welfare States." *Social Politics*, **7**, 127–191.

Maier MJ (2014). "DirichletReg: Dirichlet Regression in R. R package version 0.5-2." URL http://dirichletreg.r-forge.r-project.org/.

Montgomery DC, Peck EA, Vining GG (2006). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. Wiley, Chichester.

Robinson JP, Godbey G (1997). *Time for Life: The Surprising Way Americans Use Their Time*. Pennsylvania University Press, University Park (PA).

Rosenberg M (1965). *Society and the Adolescent Self-image*. Princeton University Press, Princeton.

van den Ark LA (1999). *Contributions to Latent Budget Analysis: A Tool for the Analysis of Compositional Data*. DSWO-press, Leiden.

van Eijnatten FM, van der Ark LA, Holloway SS (2015). "Ipsative Measurement and the Analysis of Organizational Values: an Alternative Approach for Data Analysis." *Quality & Quantity*, **49**, 559–579.

Vančáková J (2013). "Game, Free Time and the Individual (in Czech)." URL http://www.vmonline.cz/cz/hra-volny-cas-a-jedinec-vancakova-2013/.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Springer, New York.

Šípek J (2001). *Introduction to Geopsychology: The World and Wanderings Through It in a Context of Recent Times (in Czech)*. ISV Publishing House, Praha.

Affiliation:

Ivo Müller, Karel Hron, Eva Fišerová
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University
17. listopadu 12
CZ-77146 Olomouc, Czech Republic
E-mail: ivo.muller@upol.cz, hronk@seznam.cz, eva.fiserova@upol.cz

Jan Šmahaj, Panajotis Cakirpaloglu
Department of Psychology
Philosophical Faculty
Palacký University
Vodární 6
771 80 Olomouc, Czech Republic
E-mail: jan.smahaj@upol.cz, panajotis.cakirpaloglu@upol.cz

Jana Vančáková
Prostor Plus
Na Pustině 1068, 280 02 Kolín, Czech Republic
E-mail: jana.vancakova@centrum.cz

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 47
February 2018

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2017-02-23
Accepted: 2017-11-28

Austrian Journal of Statistics
February 2018, Volume 47, 21–32.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i2.575

Bayes Prediction Bound Lengths under Different Censoring Criterion: A Two-Sample Approach

Gyan Prakash
Moti Lal Nehru Medical College, Allahabad, U.P., India

Abstract

The censoring arises when exact lifetimes are known partially only, and it is useful in life testing experiments for time and cost restrictions. In literature, there are several types of censoring plans available. In which three different censoring plans have addressed in the present comparative study. The Burr Type-XII distribution considered here as the underlying model and the comparison made on Two-Sample Bayes prediction bound lengths. The analysis of the present discussion has carried out by a real life example and simulated data both.

Keywords: Burr Type-XII distribution, two-sample plan, Type-II censoring, right censoring, progressive Type-II right censoring, Bayes prediction bound length.

1. Introduction

The cumulative density and probability density function of Burr Type-XII distribution are given as

$$F\left(x;\theta,\sigma\right)=1-\left(1+x^{\sigma}\right)^{-\theta}\ ;\ \theta>0,\sigma>0,x\geq 0 \tag{1}$$

and

$$f\left(x;\theta,\sigma\right)=\sigma\theta x^{\sigma-1}\left(1+x^{\sigma}\right)^{-\theta-1}\ ;\ \theta>0,\sigma>0,x\geq 0. \tag{2}$$

The two-parameter Burr Type-XII distribution has unimodal or decreasing failure rate function

$$\rho(x)=\sigma\theta x^{\sigma-1}\left(1+x^{\sigma}\right)^{-1}\ ;\ \theta>0,\sigma>0,x\geq 0 \tag{3}$$

The shape of the failure rate function $\rho(x)$ does not affected by the parameter θ. The parameter θ and σ both are known as shape parameter. Also, $\rho(x)$ has a unimodal curve when $\sigma>1$ and it has decreased failure rate function when $\sigma\leq 1$. The Burr Type-II distribution is applied in several areas including study of quality control and reliability, duration study and failure time modeling. The analysis of business failure data, the efficacy of analgesics in clinical trials, and the times to failure of electronic components are the other areas of application of the said distribution. Zimmer, Keats, and Wang (1998) discussed at several statistical properties of the underlying distribution based on reliability analysis.

El-Sagheer (2016) discussed in his recent paper, about the point and interval predictions based on general progressive Type-II censored data by using generalized Pareto distribution under Bayesian setup for two-sample prediction approach. Rao, Aslam, and Kundu (2015) discuss about the multi-component stress strength reliability based on ML estimation criteria by assuming Burr Type-XII distribution in his recent paper. Using Koziol-Green model of random censorship Danish and Aslam (2014) deals the Bayes estimation for unknown parameters of the underlying distribution by assuming both the informative and non-informative priors. Jang, Jung, Park, and Kim (2014) discussed some estimation based on Bayesian setup for Burr Type-XII distribution under progressive censoring.

Soliman, Abd-Ellah, Abou-Elheggag, and Modhesh (2012) obtained some Bayes estimation from Burr Type-XII distribution by using progressive first-failure censored data. Lee, Wu, and Hong (2009) obtained Bayes and empirical Bayes estimators of reliability parameters under progressively Type-II Burr censored samples. Many works have done on underlying distribution, a little few of them discussed above, and a few more are Rodriguez (1977), Nigm (1988), Al-Huesaini and Jaheen (1995), Ali-Mousa and Jaheen (1998), Wu and Yu (2005), El-Sagheer and Ahsanullah (2015), Soliman, Abd-Ellah, Abou-Elheggag, and El-Sagheer (2015) and El-Sagheer (2016).

It is not always possible that the experimentally observed the lifetimes of all inspected units in life testing experiments, due to time limitation and/or cost or material resources for data collection. In addition, when some sample values at either or both extremes adulterated, the trimmed samples are useful. There are several types of censoring plans available in literature, in which only three common censoring plans have addressed in the present study.

The article presents a comparative study under Two-Sample Bayes prediction bounds length by using different censoring plans, wiz, Item-Failure, right Item-Failure, and Progressive Type-II censoring. The Bayes prediction bounds lengths have obtained from the underlying model. The properties of the procedures are illustrated by simulated data as well as a real data set.

2. Bayes prediction bound lengths (Two-sample technique)

When sufficient information regarding the past and the present behavior of an observation is available, we predict the nature of the future behavior of an observation in the present section. A Bayesian statistical analysis has applied here for predicting future statistic from the model given in Eq. (2), based on all three considered censoring plans.

Let $x_{(1)}, x_{(2)}, ..., x_{(r)}$ be the first r observed ordered failure items from a sample of size n under considered censoring scheme for the model Eq. (2). If $y_{(1)}, y_{(2)}, ..., y_{(k)}$ is the second (unobserved) items censored data of size k drawn independently from the same model of size N, then the first sample is known as informative sample, while the second sample is referred to as future sample. Our aim is to predict the j^{th} order statistic in the future sample based on an informative sample. This prediction technique is known as the, Two-sample Bayes prediction technique. Recently, Prakash and Singh (2013) discussed about the Bayes prediction limits under two-sample plan for the Pareto model.

2.1. Item-failure censoring

Let us suppose a total of n items from considering model are put under the life test and the test terminates when first $r^{th}(r \leq n)$ item fails. This censoring scheme is known as Item-Failure censoring scheme. In such test situations, the observations usually occurred in ordered of weakest items failed first.

Let us assume that $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be n ordered items from Eq. (2). If $\underline{x} \cong \left(x_{(1)}, x_{(2)}, ..., x_{(r)}\right)$ be first r observed failure items, then the joint probability density function for these order statistics is defined as

$$f_I\left(\underline{x}|\theta,\sigma\right) = \left(\prod_{i=1}^{r} f\left(x_{(i)};\theta,\sigma\right)\right)\left(1-F\left(x_{(r)};\theta,\sigma\right)\right)^{n-r}$$

$$= \left(\prod_{i=1}^{r} \sigma\theta x_{(i)}^{\sigma-1}\left(1+x_{(i)}^{\sigma}\right)^{-\theta-1}\right)\left(1+x_{(r)}^{\sigma}\right)^{-\theta(n-r)}$$

$$\Rightarrow f_I\left(\underline{x}|\theta,\sigma\right) \propto \theta^r exp\left(-\theta T_I\left(\underline{x};\theta,\sigma\right)\right)\ ; \tag{4}$$

where $T_I\left(\underline{x};\theta,\sigma\right) = \sum_{i=1}^{r} log\left(1+x_{(i)}^{\sigma}\right) + (n-r)log\left(1+x_{(r)}^{\sigma}\right)$.

There is no honest way to define, which prior probability estimate is better. Based on personal beliefs, one may choice a flexible family of priors, and chosen one from that family, which matches best. In the present study, Gamma distribution $G(1,\alpha)$ taken as the conjugate family of prior for unknown parameter θ, with the probability density function

$$\pi_\theta = \alpha e^{-\alpha\theta}\ ;\ \alpha > 0, \theta > 0. \tag{5}$$

Based on Bayes theorem, the posterior density about the parameter θ under considered censoring plan is defined as

$$\pi^*_{I\theta} = \frac{f_I\left(\underline{x}|\theta,\sigma\right)\cdot\pi_\theta}{\int_\theta f_I\left(\underline{x}|\theta,\sigma\right)\cdot\pi_\theta\, d\theta}. \tag{6}$$

Using Eq. (4) and Eq. (5) in Eq. (6), the posterior density is now obtained as

$$\pi^*_{I\theta} \propto \frac{\theta^r exp\left(-\theta T_I\left(\underline{x};\theta,\sigma\right)\right)\cdot e^{-\alpha\theta}}{\int_\theta \theta^r exp\left(-\theta T_I\left(\underline{x};\theta,\sigma\right)\right)\cdot e^{-\alpha\theta}\, d\theta}$$

$$\Rightarrow \pi^*_{I\theta} = \frac{\left(T^*_I\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\Gamma(r+1)}\theta^r exp\left(-\theta T^*_I\left(\underline{x};\theta,\sigma\right)\right)\ ;\ T^*_I\left(\underline{x};\theta,\sigma\right) = T_I\left(\underline{x};\theta,\sigma\right) + \alpha. \tag{7}$$

The Bayes predictive density of future observation Y is denoted by $h_I\left(Y|\underline{x}\right)$ and obtained by simplifying the following relation

$$h_I\left(Y|\underline{x}\right) = \int_\theta f_I\left(y;\theta,\sigma\right)\cdot\pi^*_{I\theta}\, d\theta$$

$$\Rightarrow h_I\left(Y|\underline{x}\right) = (r+1)\sigma y^{\sigma-1}\left(1+y^\sigma\right)^{-1}\frac{\left(T^*_I\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\left(T^*_I\left(\underline{x};\theta,\sigma\right)+log\left(1+y^\sigma\right)\right)^{r+2}}. \tag{8}$$

Based on predictive density Eq. (8) of the future observation Y, the cumulative predictive density function is denoted as $G_I\left(Y|\underline{x}\right)$ and obtained as

$$G_I\left(Y|\underline{x}\right) = Pr\left(Y \le y\right)$$

$$= \left(T^*_I\left(\underline{x};\theta,\sigma\right)\right)^{r+1}(r+1)\sigma\int_0^y \frac{y^{\sigma-1}\left(1+y^\sigma\right)^{-1}}{\left(T^*_I\left(\underline{x};\theta,\sigma\right)+log\left(1+y^\sigma\right)\right)^{r+2}}dy$$

$$G_I\left(Y|\underline{x}\right) = 1 - \left(\frac{T_I^*\left(\underline{x};\theta,\sigma\right)}{T_I^*\left(\underline{x};\theta,\sigma\right) + log\left(1+y^{\sigma}\right)}\right)^{r+1}. \tag{9}$$

Now, if Y_j denote the j^{th} order statistic in future sample of size $k; 1 \leq j \leq k$, then from k future observations, the probability density function of the j^{th} ordered future observation is given as

$$\Phi_I\left(y_j\right) = j\left({}^kC_j\right)\left(G_I\left(Y_j|\underline{x}\right)\right)^{j-1}\left(1 - G_I\left(Y_j|\underline{x}\right)\right)^{k-j} h_I\left(Y_j|\underline{x}\right)$$

$$\Rightarrow \Phi_I\left(Y_j\right) = j\left({}^kC_j\right)\left(1 - \left(\frac{T_I^*\left(\underline{x};\theta,\sigma\right)}{T_I^*\left(\underline{x};\theta,\sigma\right) + log\left(1+y_j^{\sigma}\right)}\right)^{r+1}\right)^{j-1}$$

$$\cdot\left(\left(\frac{T_I^*\left(\underline{x};\theta,\sigma\right)}{T_I^*\left(\underline{x};\theta,\sigma\right) + log\left(1+y_j^{\sigma}\right)}\right)^{r+1}\right)^{k-j}$$

$$\cdot(r+1)\sigma y_j^{\sigma-1}\left(1+y_j^{\sigma}\right)^{-1}\frac{\left(T_I^*\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\left(T_I^*\left(\underline{x};\theta,\sigma\right) + log\left(1+y_j^{\sigma}\right)\right)^{r+2}};\ y_j > 0. \tag{10}$$

Let us assume the transformation

$$Z = 1 - \left(\frac{T_I^*\left(\underline{x};\theta,\sigma\right)}{T_I^*\left(\underline{x};\theta,\sigma\right) + log\left(1+y_j^{\sigma}\right)}\right)^{r+1}$$

then the probability density function for the j^{th} ordered future observation becomes

$$\Phi_I\left(Z\right) = j\left({}^kC_j\right)\left(Z\right)^{j-1}\left(1-Z\right)^{k-j}; Z > 0. \tag{11}$$

Now, we say that (l_1, l_2) is a $100(1-\epsilon)\%$ prediction limits for a future random variable Y, if

$$Pr\left(l_1 \leq Y \leq l_2\right) = 1 - \epsilon. \tag{12}$$

Here l_1 and l_2 be the lower and upper Bayes prediction limits of the random variable Y, and $1-\epsilon$ is called the confidence prediction coefficient. To find the prediction limits under the two-sample plan for Y_j, j^{th} observation from a set of k future observations, we rewrite the Eq. (12) under the equal tail limits, as

$$Pr\left(Y_j \leq l_{1j}\right) = \frac{\epsilon}{2} = Pr\left(Y_j \leq l_{2j}\right) \forall\, j = 1, 2, ..., k. \tag{13}$$

Using the Eq. (11) and Eq. (13), the expressions of the limits for the j^{th} future observation are obtained by solving following equations

$$j\left({}^kC_j\right)\int_0^{\hat{l}_1} Z^{j-1}\left(1-Z\right)^{k-j} dZ = \frac{\epsilon}{2}$$

and

$$j\left({}^kC_j\right)\int_0^{\hat{l}_2} Z^{j-1}\left(1-Z\right)^{k-j} dZ = 1 - \frac{\epsilon}{2}, \tag{14}$$

where $\hat{l}_i = 1 - \left(\frac{T_I^*(\underline{x};\theta,\sigma)}{T_I^*(\underline{x};\theta,\sigma)+log\left(1+l_{ij}^{\sigma}\right)}\right)^{r+1}; i = 1, 2.$

Solving Eq. (14) for $j = 1$, the lower and upper Bayes prediction limits for the first future observation are given as

$$l_{11I} = \left\{exp\left(\left(\epsilon^* - 1\right) T_I^*\left(\underline{x};\theta,\sigma\right)\right) - 1\right\}^{1/\sigma} ;\ \epsilon^* = \left(\frac{2-\epsilon}{2}\right)^{-1/k(r+1)}$$

and

$$l_{21I} = \left\{exp\left(\left(\epsilon^{**} - 1\right) T_I^*\left(\underline{x};\theta,\sigma\right)\right) - 1\right\}^{1/\sigma} ;\ \epsilon^{**} = \left(\frac{\epsilon}{2}\right)^{-1/k(r+1)}.$$

Similarly, solving the Eq. (14) for $j = k$, the prediction limits for the last future observation is

$$l_{1kI} = \left\{exp\left(\left(\tau^* - 1\right) T_I^*\left(\underline{x};\theta,\sigma\right)\right) - 1\right\}^{1/\sigma} ;\ \tau^* = \left(1 - \left(\frac{\epsilon}{2}\right)^{\frac{1}{k}}\right)^{-1/(r+1)}$$

and

$$l_{2kI} = \left\{exp\left(\left(\tau^{**} - 1\right) T_I^*\left(\underline{x};\theta,\sigma\right)\right) - 1\right\}^{1/\sigma} ;\ \tau^{**} = \left(1 - \left(\frac{2-\epsilon}{2}\right)^{\frac{1}{k}}\right)^{-1/(r+1)}.$$

Hence, the Bayes prediction lengths for the smallest (first) and the largest (last) future observations are obtained as

$$L_{(IS)} = l_{21I} - l_{11I}$$

and

$$L_{(IL)} = l_{2kI} - l_{1kI}. \tag{15}$$

2.2. Right item-failure censoring

Since all n items from the considered model are put under the life test without replacement. In which only $r(\leq n)$ ordered items are measurable, while the remaining $(n-r)$ items are censored. These $(n-r)$ censored lifetimes will be ordered distinctly. This process is known as the right Item failure-censoring scheme (Prakash (2014)).

Now, let us consider a sequence of independent random sample from Burr Type-XII distribution of size n such as $x_{(1)}, x_{(2)}, ..., x_{(r-1)}, x_{(r)}, x_{(r+1)}, ..., x_{(n)}$. All n items are put to test without replacement and the first r items $\underline{x} \cong \left(x_{(1)}, x_{(2)}, ..., x_{(r-1)}, x_{(r)}\right)$ are fully measured while remaining $(n-r)$ items $\left(x_{(r+1)}, x_{(r+2)}, ..., x_{(n)}\right)$ are censored. Based on above the joint probability density function of these order statistics is defined as

$$f_R\left(\underline{x}|\theta,\sigma\right) \propto \left(\prod_{i=1}^{r} f\left(x_{(i)};\theta,\sigma\right)\right) \cdot \left(\prod_{i=r+1}^{n} \left(1 - F\left(x_{(i)};\theta,\sigma\right)\right)\right)$$

$$\Rightarrow f_R\left(\underline{x}|\theta,\sigma\right) \propto \theta^r exp\left(-\theta T_R\left(\underline{x};\theta,\sigma\right)\right) ;\ T_R\left(\underline{x};\theta,\sigma\right) = \sum_{i=1}^{n} log\left(1 + x_{(i)}^{\sigma}\right). \tag{16}$$

Using Eq. (5) and Eq. (16) in Eq. (6), the posterior density for unknown parameter θ under right item-failure censoring is obtained as

$$\pi_{R\theta}^* = \frac{\left(T_R^*\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\Gamma(r+1)} \theta^r exp\left(-\theta T_R^*\left(\underline{x};\theta,\sigma\right)\right) ;\ T_R^*\left(\underline{x};\theta,\sigma\right) = T_R\left(\underline{x};\theta,\sigma\right) + \alpha. \tag{17}$$

On similar lines, the Bayes predictive density, cumulative predictive density functions of future observation Y and probability density function of the j^{th} ordered future observation are obtained respectively as

$$h_R\left(Y|\underline{x}\right) = (r+1)\sigma y^{\sigma-1}\left(1+y^{\sigma}\right)^{-1}\frac{\left(T_R^*\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\left(T_R^*\left(\underline{x};\theta,\sigma\right)+log\left(1+y^{\sigma}\right)\right)^{r+2}},$$

$$G_R\left(Y|\underline{x}\right) = 1-\left(\frac{T_R^*\left(\underline{x};\theta,\sigma\right)}{T_R^*\left(\underline{x};\theta,\sigma\right)+log\left(1+y^{\sigma}\right)}\right)^{r+1}$$

and

$$\Phi_R\left(Z\right) = j\left({}^kC_j\right)\left(Z\right)^{j-1}\left(1-Z\right)^{k-j}; Z>0 \tag{18}$$

where $Z = 1-\left(\frac{T_R^*(\underline{x};\theta,\sigma)}{T_R^*(\underline{x};\theta,\sigma)+log\left(1+y_j^{\sigma}\right)}\right)^{r+1}$.

Solving Eq. (18) for $j=1$ and $j=k$, the lower and upper Bayes prediction bound limits for first and last future observation are given respectively as

$$l_{11R} = \left\{exp\left(\left(\epsilon^*-1\right)T_R^*\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma},$$

$$l_{21R} = \left\{exp\left(\left(\epsilon^{**}-1\right)T_R^*\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma},$$

$$l_{1kR} = \left\{exp\left(\left(\tau^*-1\right)T_R^*\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma}$$

and

$$l_{2kR} = \left\{exp\left(\left(\tau^{**}-1\right)T_R^*\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma}.$$

Now, the Bayes prediction intervals for first and last future observations are obtained similarly as

$$L_{(RS)} = l_{21R}-l_{11R}$$

and

$$L_{(RL)} = l_{2kR}-l_{1kR}. \tag{19}$$

2.3. Progressive Type-II censoring

The progressive censoring seems to be a great importance in strategic interval experiments. In many industrial experiments involving lifetimes of machines or units, it is required to dismiss the experiments early with failures must be limited for various reasons. This censoring criterion plays a significant role in such lifetime studies, in which the experiments terminate early.

Let us suppose an experiment in which n independent and identical units $x_{(1)}, x_{(2)}, ..., x_{(n)}$ are placed on a live test at beginning time and first $r; (1 \le r \le n)$ failure items are observed. At the time of each failure occurring prior to termination point, one (or more) enduring units detached from the test. The experiment is terminated at the time of r^{th} failure, and all remaining surviving units are removed from the test. See Prakash (2015) for more details on Progressive censoring.

Let $\underline{x} \cong \left(x_{(1)}, x_{(2)}, ..., x_{(r)}\right)$ are the lifetimes of completely observed units to fail and $R_1, R_2, ..., R_r$ are the numbers of units withdrawn at these failure times. Here, $R_1, R_2, ..., R_r$ all are predefined integers following the relation

$$\sum_{i=1}^{r} R_i + r = n.$$

Based on progressively type-ii censoring scheme the joint probability density function of order statistics $\underline{x} \cong \left(x_{(1)}, x_{(2)}, ..., x_{(r)}\right)$ is defined as

$$f_p\left(\underline{x}|\theta,\sigma\right) = C_p \prod_{i=1}^{r} f\left(x_{(i)};\theta,\sigma\right)\left(1-F\left(x_{(i)};\theta,\sigma\right)\right)^{R_i}; \tag{20}$$

Here, C_p is known as progressive normalizing constant. Simplifying Eq. (20), we get

$$\Rightarrow f_P\left(\underline{x}|\theta,\sigma\right) \propto \theta^r exp\left(-\theta T_P\left(\underline{x};\theta,\sigma\right)\right)\ ;\ T_P\left(\underline{x};\theta,\sigma\right) = \sum_{i=1}^{r}\left(1+R_i\right) log\left(1+x_{(i)}^{\sigma}\right).$$

The posterior density about the parameter θ under progressive censoring plan is

$$\pi_{P\theta}^{*} = \frac{\left(T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\Gamma(r+1)}\theta^r exp\left(-\theta T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)\ ;\ T_P^{*}\left(\underline{x};\theta,\sigma\right) = T_P\left(\underline{x};\theta,\sigma\right)+\alpha.$$

Similarly, the Bayes predictive density, cumulative predictive density functions of future observation Y and probability density function of the j^{th} ordered future observation under progressive censoring are obtained and given respectively as

$$h_P\left(Y|\underline{x}\right) = (r+1)\sigma y^{\sigma-1}\left(1+y^{\sigma}\right)^{-1}\frac{\left(T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)^{r+1}}{\left(T_P^{*}\left(\underline{x};\theta,\sigma\right)+log\left(1+y^{\sigma}\right)\right)^{r+2}},$$

$$G_P\left(Y|\underline{x}\right) = 1-\left(\frac{T_P^{*}\left(\underline{x};\theta,\sigma\right)}{T_P^{*}\left(\underline{x};\theta,\sigma\right)+log\left(1+y^{\sigma}\right)}\right)^{r+1}$$

and

$$\Phi_P\left(Z\right) = j\left({}^{k}C_{j}\right)\left(Z\right)^{j-1}\left(1-Z\right)^{k-j}; Z>0 \tag{21}$$

where $Z = 1-\left(\frac{T_P^{*}\left(\underline{x};\theta,\sigma\right)}{T_P^{*}\left(\underline{x};\theta,\sigma\right)+log\left(1+y_j^{\sigma}\right)}\right)^{r+1}$.

Substituting $j=1$ and $j=k$ in Eq. (21). The lower and upper Bayes prediction bound limits for first and last future observation are given as

$$l_{11P} = \left\{exp\left(\left(\epsilon^{*}-1\right)T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma},$$

$$l_{21P} = \left\{exp\left(\left(\epsilon^{**}-1\right)T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma},$$

$$l_{1kP} = \left\{exp\left(\left(\tau^{*}-1\right)T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma}.$$

and

$$l_{1kP} = \left\{exp\left(\left(\tau^{**}-1\right)T_P^{*}\left(\underline{x};\theta,\sigma\right)\right)-1\right\}^{1/\sigma}.$$

Thus, the Bayes prediction intervals for the smallest and the largest future observation are obtained and given as

$$L_{(PS)} = l_{21P}-l_{11P}$$

and

$$L_{(PL)} = l_{2kP}-l_{1kP}.$$

3. Numerical analysis

The performance of the proposed procedures is studied by a numerical illustration based on a real data set for a clinical trial describe a relief time (in hours) for 30 arthritic patients considered here form data provided by Wingo (1993) and used recently by Wu, Wu, Chen, Yu, and Lin (2010). The data are given in the Table (1).

Table 1: Relief time (in hours) for 30 arthritic patients

0.70	0.58	0.54	0.59	0.71	0.55	0.63	0.84	0.49	0.87
0.73	0.72	0.62	0.82	0.84	0.29	0.51	0.61	0.57	0.29
0.36	0.46	0.68	0.34	0.44	0.75	0.39	0.41	0.46	0.66

We fit the Burr Type-XII distribution to the given data in Table (1). The Kolmogorov-Smirnov (K-S) distances between the fitted and the empirical distribution functions is 0.0675 with p-value is > 0.05. Based on the K-S test statistic, Burr Type-XII distribution provides an adequate fit the data sets. In addition, the graph for both the empirical survival function and the estimated survival functions is given in Figure (3.3). (El-Sagheer (2015))

We carry out this comparison by considering the given data of size $n(= 30)$ with $\sigma(= 1.00)$ and $\alpha(= 0.50)$. The selected values of level of significance are $\epsilon = 99\%, 95\%, 90\%$.

3.1. Item-failure censoring scheme

Let the test is terminated when $r(= 5, 10, 15)$, as it is supposed from $n = 30$. Help of a considered set of parametric values, obtains the one-sided two-sample Bayes prediction bound lengths with the data given in Table (1) and presented in Table (3).

It is noted that when confidence level ϵ increases the length of intervals tends to be wider. A decreasing trend has been seen in bound lengths when censored sample size increases.

3.2. Right item-failure censoring scheme

The one-sided two-sample Bayes prediction bound lengths have been obtained under similar set of considered parametric set of values as discussed above and presented in Table (3) for right item-failure censoring data.

All properties have seen similar for the bound lengths obtained under item-failure censoring criterion. However, the bound lengths become narrower as compared to item-failure censoring criterion for all considered parametric set of values

3.3. Progressive censoring scheme

The Bayes prediction bound length under two-sample criterion have been obtained and presented in Table (3) for a similar set of parametric values as discussed above in censoring plan $R_i; i = 1, 2, ..., r$, given in (2).

Again, all the behaviors have seen similar as discussed above when compared with both censoring criteria. Further, it is noted that the magnitude of bound lengths under progressive censoring criteria are wider than compared to item-failure or right item-failure censoring criterion. It is also remarkable that for small confidence level, the bound length for largest observation is narrower as compared to the item-failure censoring criterion.

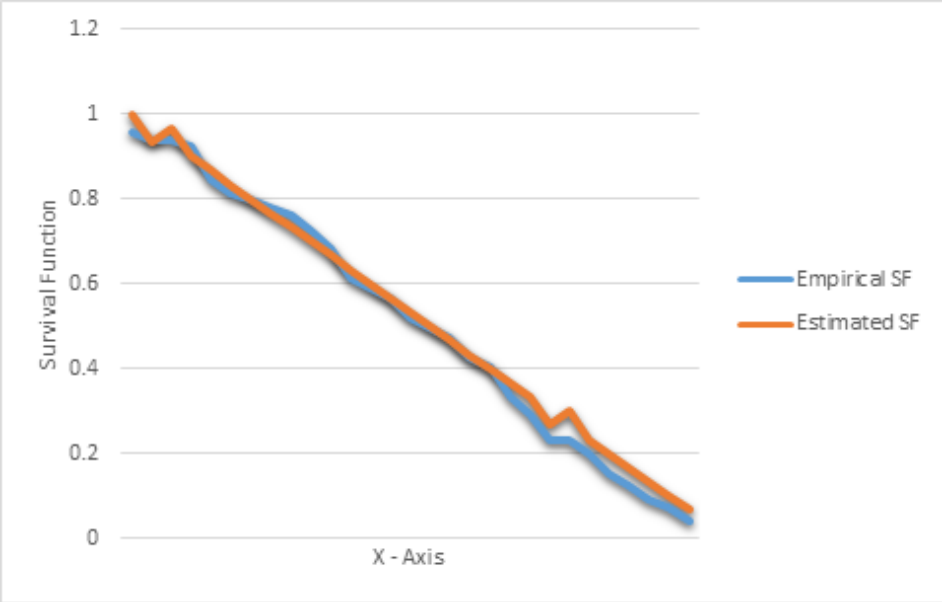

Figure 1: Empirical and estimated survival functions

Table 2: Different progressive censoring plan

Case	m	$R_i; i = 1, 2, ..., m$
1	5	1 2 1 0 1
2	10	1 0 0 3 0 0 1 0 0 1
3	15	1 0 2 0 0 1 0 2 0 0 0 1 0 0 1

4. Simulation study

Based on simulation, the performances of the procedures are studied in the present section.

Using Eq. (5), the values of shape parameter θ have been generated by using $\alpha(= 0.25, 0.50, 1.00)$. Using these three generated values of θ with a known set of values of parameter $\sigma(= 0.50, 1.00, 2.00)$, generates $10,000$ random samples, each of size $n = 30$.

All desired censored samples are generated by using following relation $x_i = \left\{(1 - U_i)^{-\frac{1}{\theta}} - 1\right\}^{\frac{1}{\sigma}}$. Here, U_i are independently distributed $U(0, 1)$. The one-sided two-sample Bayes prediction bound lengths based on simulated data are presented in the Tables 04-06 for item-failure, right item-failure, and progressive censored data respectively.

The bound length becomes wider as combination of prior parameter increase. However, a decreasing trend has seen for higher set of prior values ($\alpha = 1.00, \sigma = 2.00$). All other properties have seen similar as discussed in the previous section.

Conclusion

The properties of Bayes prediction bound lengths based on two-sample technique are the main aim of the present discussion. The underlying model is assumed here as the Burr Type-XII distribution and the analysis presented by simulated data set and a real data set provided by Wingo (1993). The item-failure, right item-failure, and progressive Type-II censoring is used for the present comparative study.

Table 3: Two-sample Bayes prediction bound lengths under different censoring plans

$\alpha = 0.50$	Item-Failure Censoring Plan					
$\sigma = 1.00$	The First Future Observation			The Last Future Observation		
$r \downarrow \epsilon \rightarrow$	99%	95%	90%	99%	95%	90%
5	3.3742	3.3501	3.3237	4.7201	4.6645	4.5772
10	2.5178	2.4556	2.3996	3.3748	3.3128	3.1849
15	2.1914	1.9791	1.9152	2.8201	2.6846	2.5436
Right Item-Failure Censoring Plan						
5	3.2511	3.2078	3.1423	4.5177	4.2942	4.1967
10	2.4158	2.3359	2.2312	3.2116	3.1018	3.0186
15	2.0414	1.9268	1.8753	2.6171	2.5366	2.4507
Progressive Type-II Censoring Plan						
5	3.8061	3.7488	3.6002	5.1837	4.8177	3.9006
10	3.1998	3.0739	2.8952	3.5538	3.1084	2.9082
15	2.7939	2.5121	2.4109	3.1664	3.0008	2.4610

Table 4: Bound lengths under item-failure censoring plan

$n = 30$		The First Future Observation			The Last Future Observation		
$(\alpha, \sigma) \downarrow$	$r \downarrow \epsilon \rightarrow$	99%	95%	90%	99%	95%	90%
0.25, 0.50	5	2.9547	2.9031	2.6749	3.0745	3.0616	3.0194
	10	2.1815	2.1215	2.0149	2.3124	2.1822	2.0974
	15	1.6953	1.4341	1.1327	1.9057	1.6216	1.5998
0.50, 1.00	5	3.1231	3.0178	2.9104	3.3538	3.3297	3.2607
	10	2.3081	2.0387	1.8196	2.5024	2.4504	2.2652
	15	1.9128	1.8114	1.6418	2.1488	1.9871	1.9333
1.00, 2.00	5	3.0445	3.0193	2.9522	3.3124	3.2786	3.0501
	10	2.2696	2.1615	2.1527	2.4615	2.4101	2.2336
	15	1.8892	1.7889	1.7203	2.1235	1.9426	1.9294

Based on selected parametric values, the one-sided two-sample Bayes prediction bound lengths are wider under the Progressive censoring scheme as compared to other censoring patterns. It is also remarkable that for small confidence level, the bound length for largest observation is narrower under Progressive censoring criterion as compared to the item-failure censoring criterion.

References

Al-Huesaini EK, Jaheen ZF (1995). "Bayesian Prediction Bounds for the Burr Type-XII Failure Model." *Communications in Statistics-Theory and Methods*, **24**(7), 1829–1842.

Ali-Mousa MAM, Jaheen ZF (1998). "Bayesian Prediction for the Two-Parameter Burr Type-XII Model Based on Doubly Censored Data." *Journal of Applied Statistical Science*, **7**(2-3), 103–111.

Danish MY, Aslam M (2014). "Bayesian Analysis of Censored Burr-XII Distribution." *Electronic Journal of Applied Statistical Analysis*, **7**(2), 326–342.

El-Sagheer RM (2015). "Estimation of the Parameters of Life for Distributions Having Power Hazard Function Based on Progressively Type-II Censored Data." *Advances and Applications in Statistics*, **45**(1), 1–27.

Table 5: Bound lengths under right item-failure censoring plan

$n = 30$		The First Future Observation			The Last Future Observation		
$(\alpha, \sigma) \downarrow$	$r \downarrow \epsilon \rightarrow$	99%	95%	90%	99%	95%	90%
0.25, 0.50	5	2.7179	2.7089	2.3797	3.0415	3.0316	3.0104
	10	2.1152	1.9841	1.8189	2.2224	2.1212	2.0714
	15	1.6138	1.3508	1.1612	1.8570	1.5206	1.4778
0.50, 1.00	5	3.0320	2.8185	2.3913	3.3438	3.2297	3.2107
	10	2.1233	1.9133	1.6351	2.3524	2.2314	2.1782
	15	1.7911	1.5373	1.2124	2.1187	1.9711	1.8333
1.00, 2.00	5	2.9029	2.8089	2.4484	3.2124	3.1860	3.0310
	10	2.0641	1.9005	1.7893	2.3315	2.1401	2.0836
	15	1.7804	1.3134	1.2128	2.0925	1.9006	1.7294

Table 6: Bound lengths under progressive type-II censoring plan

$n = 30$		The First Future Observation			The Last Future Observation		
$(\alpha, \sigma) \downarrow$	$r \downarrow \epsilon \rightarrow$	99%	95%	90%	99%	95%	90%
0.25, 0.50	5	3.2003	2.9086	2.6928	3.3415	3.3165	3.2606
	10	2.3002	2.2605	2.1892	2.4124	2.3305	2.1975
	15	1.7975	1.7801	1.6168	2.1306	1.9155	1.6792
0.50, 1.00	5	3.3233	3.0875	2.9622	3.6139	3.5078	3.2754
	10	2.5178	2.0889	1.9303	2.7289	2.6124	2.2698
	15	2.1783	1.8681	1.7838	2.3047	2.1159	2.0706
1.00, 2.00	5	3.3079	3.2105	3.0276	3.4599	3.2622	3.0314
	10	2.4659	2.1485	1.9389	2.6144	2.3186	2.1355
	15	2.0126	1.9037	1.4691	2.1072	2.0107	1.9163

El-Sagheer RM (2016). "Bayesian Prediction Based on General Progressive Censored Data from Generalized Pareto Distribution." *Journal of Statistics Applications and Probability*, **5**(1), 43–51.

El-Sagheer RM, Ahsanullah M (2015). "Statistical Inference for a Step-Stress Partially Accelerated Life Test Model Based on Progressively Type - II Censored Data from Lomax Distribution." *Journal of Applied Statistical Science*, **21**, 307–323.

Jang DO, Jung M, Park JH, Kim C (2014). "Bayesian Estimation of Burr Type-XII Distribution Based on General Progressive Type-II Censoring." *Applied Mathematical Sciences*, **69**(8), 3435–3448.

Lee WC, Wu JW, Hong CW (2009). "Assessing the Lifetime Performance Index of Products from Progressively Type-II Right Censored Data Using Burr-XII Model." *Mathematics and Computers in Simulation*, **79**(7), 2167–2179.

Nigm AM (1988). "Prediction Bounds for the Burr Model." *Communications in Statistics - Theory and Methods*, **17**(1), 287–297.

Prakash G (2014). "Right Censored Bayes Estimator for Lomax Model." *Statistics Research Letters*, **3**(1), 23–28.

Prakash G (2015). "Progressively Censored Rayleigh Data under Bayesian Estimation." *The International Journal of Intelligent Technologies and Applied Statistics*, **8**(3), 257–373.

Prakash G, Singh DC (2013). "Bayes Prediction Intervals for the Pareto Model." *Journal of Probability and Statistical Science*, **11**(1), 109–122.

Rao GS, Aslam M, Kundu D (2015). "Burr-XII Distribution Parametric Estimation and Estimation of Reliability of Multicomponent Stress-strength." *Communications in Statistics - Theory and Methods*, **44**, 4953–4961.

Rodriguez RN (1977). "A Guide to the Burr Type-XII Distributions." *Biometrika*, **64**(1), 129–134.

Soliman AA, Abd-Ellah AH, Abou-Elheggag NA, El-Sagheer RM (2015). "Inferences for Burr-X Model Using Type-II Progressively Censored Data with Binomial Removals." *Arabian journal of Mathematics*, **4**(2), 127–139.

Soliman AA, Abd-Ellah AH, Abou-Elheggag NA, Modhesh AA (2012). "Estimation from Burr Type-XII Distribution Using Progressive First-failure Censored Data." *Journal of Statistical Computation and Simulation*, **1**(1), 1–21.

Wingo DR (1993). "Maximum Likelihood Methods for Fitting the Burr Type-XII Distribution to Life Test Data." *Metrika*, **40**(1), 203–210.

Wu JW, Yu HY (2005). "Statistical Inference about the Shape Parameter of the Burr Type-XII Distribution under the Failure-censored Sampling Plan." *Applied Mathematics and Computation*, **163**(1), 443–482.

Wu SF, Wu CC, Chen YL, Yu YR, Lin YP (2010). "Interval Estimation of a Two-parameter Burr-XII Distribution under Progressive Censoring." *Statistics*, **44**(1), 77–88.

Zimmer WJ, Keats JB, Wang FK (1998). "The Burr XII Distribution in Reliability Analysis." *Journal of Quality Technology*, **30**(4), 386–394.

Affiliation:

Gyan Prakash
Department of Community Medicine
Moti Lal Nehru Medical College,
Allahabad, U. P., India.
E-mail: ggyanji@yahoo.com

Austrian Journal of Statistics
February 2018, Volume 47, 33–52.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i2.764

Multivariate Optimal Allocation with Box-Constraints

Ulf Friedrich
Trier University

Ralf Münnich
Trier University

Martin Rupp
Trier University

Abstract

Modern surveys aim at fostering accurate information on demographic and other variables. The necessity for providing figures on regional levels and on a variety of subclasses leads to fine stratifications of the population. Optimizing the accuracy of stratified random samples requires incorporating a vast amount of strata on various levels of aggregation. Accounting for several variables of interest for the optimization yields a multivariate optimal allocation problem in which practical issues such as cost restrictions or control of sampling fractions have to be considered. Taking advantage of the special structure of the variance functions and applying Pareto optimization, efficient algorithms are developed which allow solving large-scale problems. Additionally, integrality- and box-constraints on the sample sizes are considered. The performance of the algorithms is presented comparatively using an open household dataset illustrating their advantages and relevance for modern surveys.

Keywords: stratified random sampling, multi-criteria optimization, linear constraints, integer optimization, Pareto optimality, semismooth Newton.

1. Introduction

Accurate population figures provide an important basis for political and economic decision processes. In light of urban audits and regional policies, these figures, however, have to be made available in sufficient regional detail as well as for many sub-classes, which requires introducing a vast number of strata by regions and content. Censuses, registers, or adequate surveys can provide the information necessary for such research. Using surveys, stratified random sampling provides an adequate basis that allows integrating further optimization techniques while considering practical settings with various constraints. Additionally, several variables of interest may be incorporated in the optimization process which either contains complementary or conflictory information. This finally leads to a multivariate optimal allocation problem under constraints regarding regional as well as context-specific stratifications.

For the stratified random sampling problem we assume a finite population $\mathcal{U}$ of size N with disjoint cross-classification strata $h = 1, \ldots, H$. Let τ_Y denote the total of a variable of interest Y. In stratified random sampling, an unbiased estimator is $\hat{\tau}_Y^{\text{StrRS}} = \sum_{h=1}^{H} N_h \bar{y}_h$, where $\bar{y}_h$ is the sample mean of variable Y and N_h is the population size in stratum h. Its

variance is

$$\mathrm{Var}(\hat{\tau}_Y^{\mathrm{StrRS}}) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \tag{1}$$

with known stratum-specific variances S_h^2 of variable Y and stratum-specific sample sizes n_h for all strata $h = 1, \ldots, H$ (Lohr 2010, chapter 4). In practice, earlier surveys or highly correlated variables yield the necessary information for S_h^2. In our presentation, we tacitly assume that adequate proxies are available as a discussion of proxy quality and its implications is beyond the scope of this work.

Minimizing the variance (1) with respect to the stratum-specific sample sizes n_h while respecting a given total sample size $n_{\max}$ leads to the (univariate) optimal allocation introduced by Tschuprow (1923) and Neyman (1934). In contrast to the equal and proportional allocation, see Cochran (1977), the optimal allocation depends on the variable of interest. The resulting optimal allocation is given in closed form by

$$n_h^* = \frac{N_h S_h}{\sum_{k=1}^{H} N_k S_k} \cdot n_{\max}. \tag{2}$$

This allocation method is extended in Gabler, Ganninger, and Münnich (2012) and Münnich, Sachs, and Wagner (2012b), such that for each stratum-specific sample size n_h box-constraints m_h, M_h with

$$2 \leq m_h \leq n_h \leq M_h \leq N_h \tag{3}$$

are added to the optimization problem. As zero sample sizes in single strata lead to biased estimates and variance estimation of the total estimate requires stratum-specific sample sizes of at least two, a lower constraint $m_h \geq 2$ is applied. Upper constraints $M_h \leq N_h$ have to be introduced to avoid overallocation in strata where n_h given by (2) exceeds N_h. A further reduction of M_h allows to control sample fractions, for example to avoid highly different response burdens in various regions or strata. In addition, M_h prevents a stratum-specific full census which is prohibited by law in specific surveys, for example by judgment of the German Federal Administrative Court (BVerwG, 03/15/2017, 8 C 6.16).

Altogether, the optimal allocation problem under box-constraints is given by

$$\begin{aligned} \min_{n \in \mathbb{R}_+^H} \quad & \mathrm{Var}(\hat{\tau}_Y^{\mathrm{StrRS}}) \\ \text{s.t.} \quad & \sum_{h=1}^{H} n_h = n_{\max} \\ & m_h \leq n_h \leq M_h \quad \forall h = 1, \ldots, H. \end{aligned} \tag{4}$$

The problem can equivalently be stated with the inequality constraint $\sum_{h=1}^{H} n_h \leq n_{\max}$, but equality holds at every optimal solution.

Friedrich, Münnich, de Vries, and Wagner (2015) provide a further extension that ensures integrality of the solution of the optimal allocation. Using Gabler *et al.* (2012) and by separating different sub-regions, the method can be rewritten to a simultaneous optimal allocation for multiple areas. This is achieved by using a quadratic separable decision function.

In the multivariate generalization of the optimal allocation problem, K different variances $\mathrm{Var}(\hat{\tau}_1^{\mathrm{StrRS}})$, ..., $\mathrm{Var}(\hat{\tau}_K^{\mathrm{StrRS}})$ are considered simultaneously. Dalenius (1953) discusses this problem in detail and distinguishes two solution strategies. In the first, one or more of the variances $\mathrm{Var}(\hat{\tau}_k^{\mathrm{StrRS}})$, $k = 1, \ldots, K$, are bounded from above and treated as constraints of an optimization problem in which the total sample size (or the cost of the survey) is minimized. This leads to a univariate optimization problem with non-linear constraints. In the second, the variances are minimized simultaneously subject to linear size (or cost) constraints. This

perspective leads to a multi-objective optimization problem with conflicting objectives that requires an appropriate mathematical theory. In particular, an adequate notion of optimality, such as Pareto optimality, is essential and the problem has to be transformed into a form that is solvable by optimization algorithms. Most of the literature dealing with multivariate allocation splits up depending on which of these two formulations is used.

Indeed, Chatterjee (1968), Chatterjee (1972), and Huddleston, Claypool, and Hocking (1970) use the first variant. Multivariate optimal allocation problems are addressed in the same way in Kokan (1963) and supplemented by existence and uniqueness results in Kokan and Khan (1967). Introducing overhead costs, Ahsan and Khan (1982) discuss the problem with variance constraints for a more general objective function. More recently, Bankier (1988), Hohnhold (2009a), and Hohnhold (2009b) have published allocation techniques with more than one level of strata. These techniques are based on a compensation of the accuracy of regional estimates and population total estimates and, hence, also belong to the first class of methods. Falorsi and Righi (2015) present a generalized framework for defining the optimal inclusion probabilities in multivariate and multi-domain surveys. Falorsi and Righi (2008) and Falorsi and Righi (2016) introduce a solution method using a balanced sampling design.

Combining aspects from both strategies, Kish (1976) proposes to combine aspects of variance and cost minimization in a non-linear model with the help of loss functions and discusses various choices for the objective function within his model.

Turning to the second solution strategy introduced by Dalenius (1953), the multivariate optimal allocation is threated in Folks and Antle (1965) as a multi-objective optimization problem with linear constraints. They discuss the mathematical theory of scalarization and the relationship between the multi-objective problem and the scalarized problem. Moreover, they prove a sufficiency result for the set of efficient (or Pareto optimal) solutions for the simple problem without box-constraints and neglecting the integrality of the solutions. Díaz-García and Ramos-Quiroga (2014) solve the multivariate allocation as a multi-objective problem as well but with the help of stochastic programming. Khan, Ali, Raghav, and Bari (2012) use stochastic programming on another model. Both methods lead to non-linear integer optimization problems which are hard to solve even for small instances. Khan, Khan, and Ahsan (2003) solve multivariate allocation problems by exploiting the separability of the objective function and applying dynamic programming. While dynamic programming is a classical solution method for allocation problems (Arthanari and Dodge 1981, chapter 5), it is not very efficient in practice as the computational study of Bretthauer, Ross, and Shetty (1999) shows.

All strategies using the multi-objective perspective on the problem have to use scalarization techniques to combine the variances for the variables in a one single objective function. The selection of a scalarization technique can be interpreted as the choice of a suitable decision-making function (Schaich and Münnich 1993 and Díaz-García and Cortez 2006). The optimal allocation then highly depends on the concrete choice of a scalarization function.

We also take the second of the two perspectives of Dalenius (1953) and treat the multivariate allocation problem as a multi-objective problem. We extend the theoretical result in Folks and Antle (1965) by giving a (necessary and sufficient) characterization of *all* Pareto optimal points. Moreover, in contrast to earlier publications, we solve the problem while respecting integrality and box-constraints. We compute the set of Pareto optimal solutions, the so-called Pareto frontier, for this refined problem formulation which allows decision makers to choose a personally specified preference from this set.

The solution of allocation problems under the box-constraints (3) may yield non-differentiable points and many standard algorithms, such as classical Newton techniques, may fail to provide the correct optimal solution. To avoid convergence issues, we propose using the semismooth Newton method (Münnich, Sachs, and Wagner 2012a and Wagner 2013). Because stratum-specific sample sizes are integer values, we also provide an alternative algorithm to derive a multivariate optimal allocation in which all stratum-specific samples are integer-valued. This strategy avoids rounding and is based on the integer optimal allocation techniques published

in Friedrich *et al.* (2015).

Each scalarization for multivariate optimal allocation contains an additive linking of variances. Due to the scaling of units of the variables of interest, the variances have to be standardized for comparability. We present an alternative solution that extends the techniques published in Schaich and Münnich (1993).

Finally, it is of great importance for practical applications to solve large problem instances in appropriate time. Our methods solve multivariate optimal allocation problems with several thousand strata within seconds and are reliable tools when dealing with real-world data. This stands in contrast with other algorithms for multivariate allocation problems that are generally computationally tractable for only a small number of strata. The computationally efficient solution of large (integer) multivariate optimal allocation problems supplements the theoretical discussion and certainly is another central innovation of our methods.

In Section 2, we use the method of box-constrained optimal allocation presented in Münnich *et al.* (2012b) as a starting point to derive a generalized multivariate box-constrained optimal allocation problem with various decision-making functions and standardization techniques. Moreover, we establish the link between the multivariate allocation problem and the theory of Pareto optimization. In Section 3, we provide efficient numerical algorithms for selected variants of the developed problem. These are fast enough to solve even large problem instances and avoid rounding the solution by finding the globally optimal integer-valued solution. In Section 4, we present selected performance and simulation results based on the open AMELIA household dataset (Alfons, Burgard, Filzmoser, Hulliger, Kolb, Kraft, Münnich, Schoch, and Templ 2011, as well as Merkle, Burgard, and Münnich 2016).

2. Multivariate optimal allocation

2.1. Preliminaries

In a multivariate optimal allocation problem, several variables of interest are considered simultaneously. The resulting optimization problem has several conflicting objective functions. Thereby, the correlation between the variables of interest, the variable types as well as the purpose of the survey are decisive factors. The use of a scalarization technique is mandatory to treat this conflict of objectives and to solve the optimization problem numerically. The choice of a scalarization technique is not clear in advance, depends on the application, and has a considerable influence on the solution of the problem.

The most intuitive scalarization technique is the weighted sum method, for which each objective is weighted and the weighted objectives are cumulated (Jahn 1986). Another widespread technique is the epsilon-constraint method, which corresponds to minimizing the cost while respecting variance restrictions (Ehrgott 2005 and Falorsi and Righi 2015). As we focus on the minimization of the variance, we do not consider the epsilon-constraint method here. Moreover, we propose a p-norm of the objectives ($p = 1, 2, 4, 8, \infty$), which is discussed in Lin (2005). Schaich and Münnich (1993) study the particular case $p = \infty$ which is equivalent to the so-called *min-max* method.

In addition to scalarization, standardization techniques are also important for standardizing variances of various types of variables of interest. Schaich and Münnich (1993) suggest to replace the variance of the estimators by the coefficient of variation to receive additively comparable values. In order to retain the mathematical properties of the variance function, we use the squared coefficient of variation

$$\mathrm{CV}^2(\hat{\tau}_Y^{\mathrm{StrRS}}) := \frac{\mathrm{Var}(\hat{\tau}_Y^{\mathrm{StrRS}})}{\tau_Y^2}$$

with the population total τ_Y of variable Y for the (CV2)-standardization. Although the principal effect is similar, squaring the coefficient of variation may lead to small differences

in some settings. A drawback of using the squared coefficient of variation is the requirement for the population total τ_Y of variable Y, which is generally not given in advance and which, as it is a ratio, is even more demanding than using only the proxies for the stratum-specific variances.

Furthermore, we propose the alternative (opt) standardization, in which the objectives are standardized by the unique univariate optimal allocations as standardization factors. The standardized objective for the variable of interest Y is given by

$$\text{opt}(\hat{\tau}_Y^{\text{StrRS}}) := \frac{\text{Var}(\hat{\tau}_Y^{\text{StrRS}})}{\text{Var}_Y^{\text{opt}}}$$

where $\text{Var}_Y^{\text{opt}}$ is the univariate optimal allocation for the variable Y computed, for example, with the box-constraint optimal allocation by Münnich *et al.* (2012b). This standardization technique reflects the relative loss for each variable under consideration when using the compromise allocation rather than the single variable optimized allocation. In contrast to (CV2), an advantage of this technique is that the total τ_Y of variable Y is not required. Moreover, if S_h^2 has to be estimated, the uncertainty and blur of this estimation is symmetrically present in the numerator and denominator of the objectives, and, thus, eliminated. Hence, a standardization by the univariate optimal variances results in a more robust multivariate optimal allocation.

2.2. Methods of multivariate optimal allocation

The optimal allocation with respect to only one variable of interest Y with box-constraints for stratum-specific sample sizes is given by (4). The simultaneous consideration of several variables of interest $Y_1, \ldots, Y_K$ yields the following multi-criteria optimization problem

$$\begin{aligned} \min_{n \in \mathbb{R}_+^H} \quad & \left(\text{Var}(\hat{\tau}_1^{\text{StrRS}}), \ldots, \text{Var}(\hat{\tau}_K^{\text{StrRS}})\right) \\ \text{s.t.} \quad & \sum_{h=1}^{H} n_h = n_{\max} \\ & m_h \leq n_h \leq M_h \quad \forall h = 1, \ldots, H \end{aligned} \tag{5}$$

where

$$\text{Var}(\hat{\tau}_k^{\text{StrRS}}) = \sum_{h=1}^{H} \frac{N_h^2 (S_h^k)^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

with H cross-classification strata, stratum sizes N_h, and stratum-specific variances $(S_h^k)^2$ given for each stratum $h = 1, \ldots, H$ and variables of interest $k = 1, \ldots, K$. To prove the existence of a solution, we refer to Jahn (1986, Theorem 6.3). The multivariate allocation problem (5) can be reformulated as a single-objective optimization problem with objective function $f : \mathbb{R}_+^H \to \mathbb{R}_+$ by combining the K original objective functions in one scalar expression. In this scalarization the objective functions are also standardized to make them comparable. Next, we explain the standardized scalarization in detail.

Weighted sum scalarization

Using the weighted sum scalarization method, we obtain the objective function f given by

$$\begin{aligned} f(n) &:= \sum_{k=1}^{K} w_k \frac{\text{Var}(\hat{\tau}_k^{\text{StrRS}})}{\alpha_k} \\ &= \sum_{k=1}^{K} w_k \frac{\sum_{h=1}^{H} \frac{N_h^2 (S_h^k)^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)}{\alpha_k}, \end{aligned} \tag{6}$$

which depends on externally given weights $w_1, \ldots, w_K \in \mathbb{R}_+$ with $\sum_{k=1}^K w_k = 1$. Using the squared coefficient of variation as a standardization method, factor $\alpha \in \mathbb{R}^K$ is defined by $\alpha_k := \tau_k^2$ for all $k = 1, \ldots, K$. Alternatively, if we apply the single unique optimal allocations as a standardization technique, we set $\alpha_k := \mathrm{Var}_k^{\mathrm{opt}}$ for all $k = 1, \ldots, K$.

Alternative scalarization techniques

Alternatively, by using the p-norm ($p < \infty$) as scalarization method, we obtain the objective

$$f(n) := \left(\sum_{k=1}^{K} \left(\sqrt{\frac{\mathrm{Var}(\hat{\tau}_k^{\mathrm{StrRS}})}{\alpha_k}} \right)^p \right)^{\frac{1}{p}}. \tag{7}$$

If we define f by the 2-norm, it is equivalent to the weighted sum with equal weights. Finally, using the *min-max* method, f is given by

$$f(n) := \max_{k=1,\ldots,K} \sqrt{\frac{\mathrm{Var}(\hat{\tau}_k^{\mathrm{StrRS}})}{\alpha_k}}. \tag{8}$$

Properties of the objective function

In the case of the weighted sum scalarization, the objective function f in (6) is continuously differentiable, strictly convex, and separable (Münnich *et al.* 2012b). These properties are essential for the fast algorithms presented in Section 3. If the alternative scalarization methods are used, f changes and may loose some of these properties. In particular, the objective f in (7) is continuously differentiable and strictly convex, but only separable if $p = 2$. If f is not separable, as in the case $p \neq 2$, special attention must be paid to the selection of the solution algorithm.

Furthermore, for $p = \infty$ the objective (8) is not continuously differentiable. However, many classical optimization methods, such as the Newton method, rely on differentiability and are not applicable in this case. For more details we refer to Section 3.

2.3. Weighted sum and Pareto optimization

The scalarization by the weighted sum fits in the theory of Pareto optimality. When optimizing competing objectives, the *Pareto frontier* describes the set of all efficient solutions in the sense that for all points in the frontier one objective can only be improved by diminishing another. Therefore, the Pareto frontier gives a very suitable characterization of all those points decision makers should consider in a multi-criteria optimization problem. On the other hand, it is not advisable to choose an allocation which is not on the Pareto frontier, because it could be improved without cost.

Moreover, the Pareto frontier describes the optimal solutions independently from the weighting, that means independently from the ranking of the variables of interest by decision makers. Instead of determining the ranking in advance, our method allows users to select a preferred solution among all Pareto optimal points *after* the optimization step. Advantages of this procedure are the ability to optimize without a known priority ranking of the variables of interest and the possibility to use additional information at the time of decision, for example variance structures or sensitivity, and the robustness of the solution with respect to the weights.

We describe the entire frontier of Pareto optimal solutions to the multivariate allocation problem (5) mathematically in Appendix A and extend the results by Folks and Antle (1965). We prove that each optimal solution of the weighted sum reformulation for an arbitrary choice of weights is a Pareto optimal solution for (5). Moreover, if we solve the weighted sum problem for all possible choices of weights, we obtain all Pareto optimal solutions of the original problem (subject only to the discretization of the weights). This way, we compute the whole

Pareto frontier of the multivariate allocation problem. We refer to Sections 4.2 and 4.3 for computation algorithms of the Pareto frontier, their implementation, and exemplary numerical results.

2.4. Multivariate optimal integer allocation

So far, we have ignored the requirement that the calculated stratum-specific sample sizes in an (univariate or multivariate) optimal allocation problem have to be in the set of non-negative integers for almost all application problems because, for example, a fraction of a person cannot be drawn in a sample. In general, the solution of the allocation problems in continuous variables presented in Section 2.2 is not an integer but a fractional number. In practical applications this problem is commonly solved by a rounding strategy in the post-processing of the results. However, a rounded solution obtained this way is in general *not* an optimal solution in the set of all integral solutions as in the example data presented in Section 4.4. Therefore, we also discuss an algorithm for the computation of the globally optimal solution in integer variables.

3. Algorithmic solution of allocation problems

In this section we present two efficient algorithms for the numerical solution of (5) in continuous and integer variables. The strict convexity and separability of the scalarized objective function f is crucial for the correctness of both algorithms. Concerning the scalarization and standardization techniques presented in Section 2, f is only separable for the weighted sum or the 2-norm but not for the other p-norms or *min-max*.

3.1. Semismooth Newton

The algorithm is based on developments and derivations published in Münnich *et al.* (2012b) who consider a univariate optimal allocation problem with box-constraints. After scalarization and standardization with the techniques described in Section 2, it is also applicable to the multivariate problem. The main characteristic of the algorithm is to express the stratum-specific sample sizes n_h as a function of the Lagrange multiplier $\lambda \in \mathbb{R}$ by transforming the Karush-Kuhn-Tucker optimality conditions. Then, the expression for $n_h(\lambda)$ is substituted into the equality-constraint of the original problem, which leads to a one-dimensional system of equations depending on λ

$$\Phi(\lambda) := \sum_{h=1}^{H} n_h(\lambda) - n_{\max} = 0 \tag{9}$$

with $n_h(\lambda) := \text{Proj}_{[m_h,M_h]}\left(\frac{S_h^2 N_h^2}{\lambda}\right)^{\frac{1}{2}}$, where $\text{Proj}_{[m_h,M_h]}$ denotes the projection into the interval $[m_h, M_h]$. Due to this cut-off, Φ is not continuously differentiable. Nevertheless, Qi and Sun (1993) show semismoothness for Φ. Münnich *et al.* (2012b) suggest a fixed-point iteration to solve (9). We chose a semismooth Newton method because it allows for additional generalizations. For a detailed presentation of the semismooth Newton method in the context of survey statistics, we refer to Münnich *et al.* (2012a).

It is also necessary to solve non-separable settings of the continuous allocation problem for a complete comparison of the methods in Section 4. These instances are solved with the `R` package `nloptr` (Ypma, Borchers, and Eddelbuettel 2014).

3.2. Solution as integer optimization problem

As in the continuous case, the multivariate optimal *integer* allocation problem is algorithmically tractable whenever the objective function f is separable and convex. The problem

reduces to a single-objective optimization problem and algorithms developed for the univariate allocation problem can be applied directly.

Friedrich *et al.* (2015) present three algorithms for the problem that use the fact that the minimization of a separable and convex function is polynomially solvable in integer variables if the feasible set is a so-called *polymatroid*, which is a convex polytope with strong combinatorial properties. An exhaustive discussion of the mathematical background is given in Friedrich (2016). The algorithms are based on so-called *Greedy* strategies and find the globally optimal integer solution.

In the case of convex objective functions that are not necessarily separable, the problem can still be reformulated as a single-objective integer optimization problem, but the fast Greedy algorithms do not find the optimal solution. Nevertheless, it is possible to solve these more general problems with the help of a reformulation as *linear* integer problems (Hochbaum 1995). A reformulation of this type has been solved with the commercial software FICO Xpress Optimization Suite in Friedrich *et al.* (2015) with the result that computation times worsen significantly (many hours instead of seconds). Therefore, we do not solve the integer version of the non-separable problems in Section 4.

4. Simulation study and results

We use the synthetic AMELIA dataset (Merkle *et al.* 2016) for a simulation study to verify and compare the presented methods. It is a household dataset reflecting the household structure of Europe containing 3 781 289 households and 10 012 600 individuals. We use the household structure with stratification levels districts (DIS – 40 strata), household size (HHS – 6 strata), and degree of urbanization (DOU – 3 strata). This results in $40 \cdot 6 \cdot 3 = 720$ cross-classification strata. As variances have to be compared, cross-classification strata with a total size of $N_h < 2$ are eliminated, so that the simulation only contains 676 strata. The size distribution of the 676 strata is shown in Figure 1 clustered by the classes of household size. Classes 1 up to 5 contain households with the respective number of persons, class 6 contains households with more than five persons.

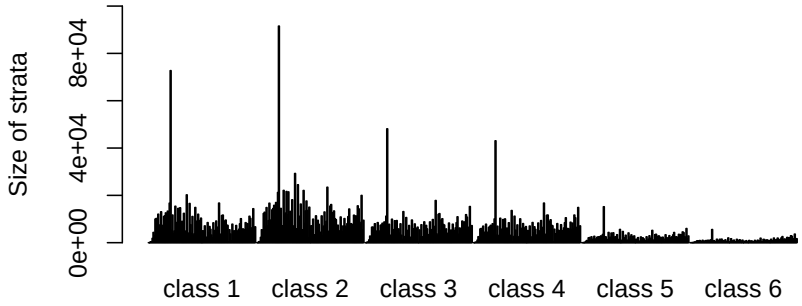

Figure 1: Size distribution of the 676 cross-classification strata clustered by the classes of household size.

We choose the total household income (INC), the social income (SOC), and the age of the main income earner (HAGE) as variables of interest. The social income of a household is defined as the sum of unemployment, old-age, survivors, sickness, disability, and education-related benefits of all people within the household, see Merkle *et al.* (2016). Although the three variables of interest are all continuous, our method is applicable to proportions of categorical variables as well.

The correlations within each district are presented in the boxplots in Figure 2. The correlations over the population are depicted as vertical lines. In particular, concerning variable HAGE, we observe some differences between the overall correlation and the separated district

correlations.

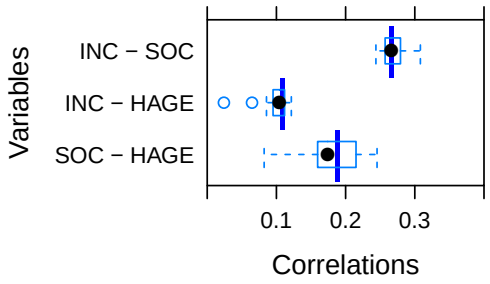

Figure 2: Boxplots of the correlations of the variable of interest per district in contrast to the total population (vertical lines).

We compare the estimates under various settings with the relative mean squared error (RMSE). Because the sampling design is stratified random sampling, the estimates are unbiased and the RMSE comparison is equivalent to the variance comparison (Lohr 2010, chapter 2). Since the AMELIA dataset is used, the true values of the RMSEs can be computed directly for the comparative analysis rather than the Monte Carlo equivalences. Furthermore, most of the following figures and graphs do not contain absolute values of errors, variances or sample sizes, but relative values compared to the case of an independent univariate optimal allocation of the three variables of interest.

4.1. Comparison of variances depending on the decision-making strategy

In the following, we compare the results of the four decision-making functions 2-norm, 4-norm, 8-norm, and min-max as well for the (CV2)-standardization and the alternative (opt)-standardization presented in Section 2.1. The variance functions for each variable of interest have equal weights in these settings and, as pointed out before, the 2-norm is equivalent to the weighted sum with equal weights.

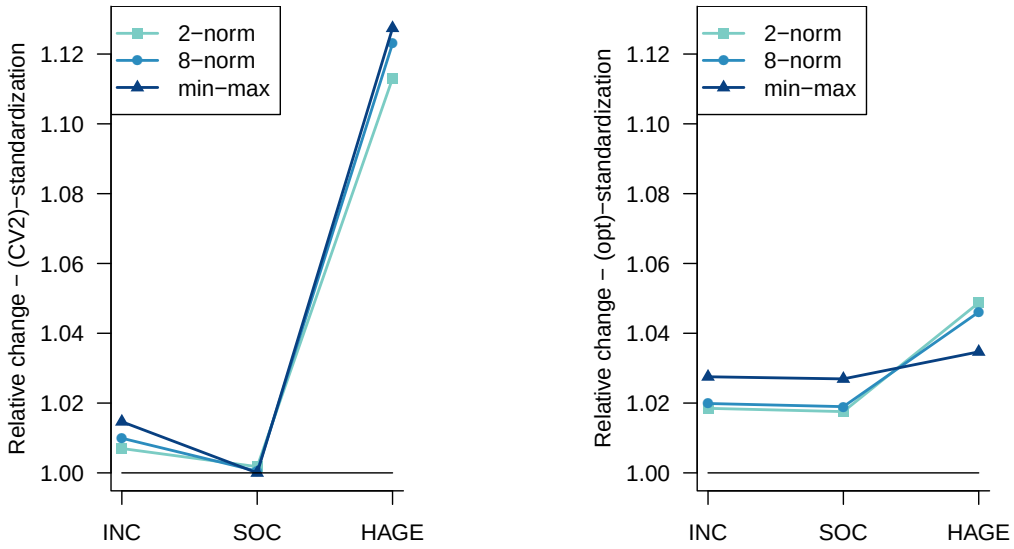

Figure 3: Relative change of RMSE for the estimated population totals for various standardization and scalarization techniques.

In Figure 3 we show the relative increase of the RMSE for the total population estimates of the three variables of interest compared to the optimal univariate allocation computed by the

method of Münnich *et al.* (2012b). For a better visibility, the 4-norm is not displayed in the figure. Its graph is between the graphs of the 2- and 8-norm. As every univariate optimal allocation is optimal, the RMSEs have to be higher or equal in the multivariate case compared to the univariate RMSEs. As a consequence, the graphs in Figure 3 are located on or above the horizontal one-line. In the settings with (CV2)-standardization, the error-increases are not well balanced. Because

$$\mathrm{CV}^2(\hat{\tau}_{\mathrm{SOC}}^{\mathrm{StrRS}}) > \mathrm{CV}^2(\hat{\tau}_{\mathrm{INC}}^{\mathrm{StrRS}}) > \mathrm{CV}^2(\hat{\tau}_{\mathrm{HAGE}}^{\mathrm{StrRS}})$$

for all appropriate allocations, the increase of the RMSE is smallest in variable SOC. In the min-max case ($p = \infty$), the increase for SOC is zero, which means that the multivariate optimal allocation is equal to the univariate optimal allocation with respect to SOC.

In contrast, we observe a well balanced increase of the RMSEs for the (opt)-standardization because the p-norm of the relative change of the variances compared to the univariate optimal allocations is minimized. This results in a well compensated allocation. For $p = \infty$ we obtain an almost equal increase.

In Figure 4 we plot the same settings as in Figure 3, but the RMSEs of the subtotal estimates of each of the 40 districts are presented. As before, the errors are illustrated relative to the errors when using the univariate optimal allocations. Dots which are located to the right of the vertical one-line correspond to subtotal estimates with an increase of the district specific RMSEs. Accordingly, dots to the left of the one-line correspond to estimates with a decrease. Again, the settings with the most compensated errors are those corresponding to the (opt)-standardization. Although the RMSEs in Figure 3 are higher than the univariate RMSEs, the multivariate allocation also leads to error-decreases in some districts, shown as points located to the left of the one-line in the boxplots.

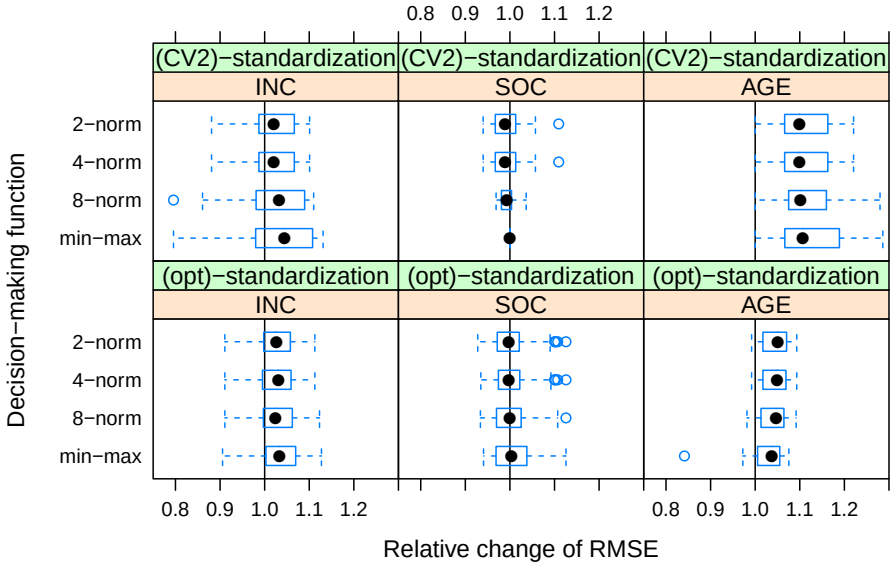

Figure 4: Relative change of RMSE for the estimated district totals for various standardization and scalarization techniques.

Figures 3 and 4 show that the (CV2)-standardization and a scalarization with a larger p accentuate single variables, in particular those with a comparably high CV. This contrasts compensatory methods which may be preferable in cases where no most important variable is obvious.

4.2. Comparison of variances depending on the chosen weights

Predefined weights

Here, we focus on the weighted sum as decision-making function. As illustrated in Section 2.3, this decision-making function facilitates the computation of the whole set of Pareto optimal solutions. We plot the relative increases of the district specific RMSEs for ten combinations of weights for the (CV2)-standardization in Figure 5 and for the (opt)-standardization in Figure 6. The relative error-increases of the total population estimates are comparatively

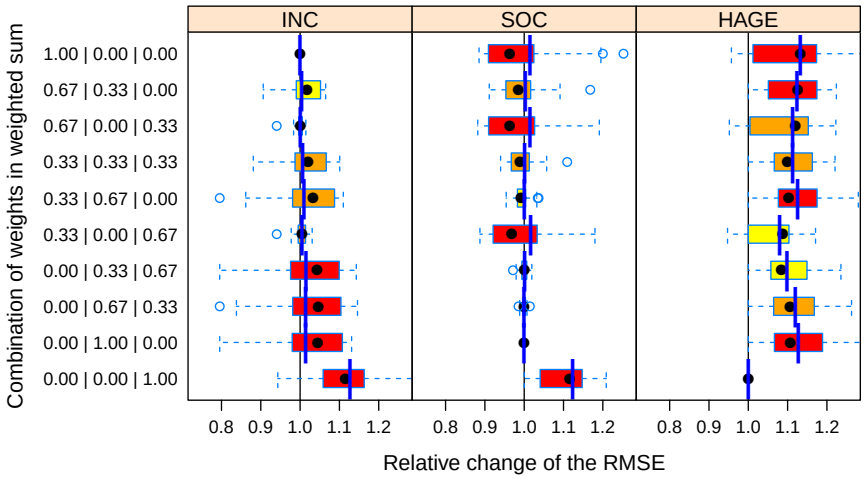

Figure 5: Relative change of the RMSE for the estimated district totals for ten combinations of weights with (CV2)-standardization. Red boxes correspond to a weight of 0.00, orange boxes to a weight of 0.33, and yellow boxes to a weight of 0.67 for the respective variable.

shown as vertical lines. The settings in row one, nine, and ten are equal to the univariate optimal allocations with respect to one of the three variables of interest, which is why the boxplots for the corresponding variables have no spread. In most cases, higher weights coincide with lower estimation errors of the district totals. Nevertheless, this coincidence is not a general statement and depends, among others, on the correlation structure of the variables of interest. Comparing the results of Figures 5 and 6, we observe more compensated error-

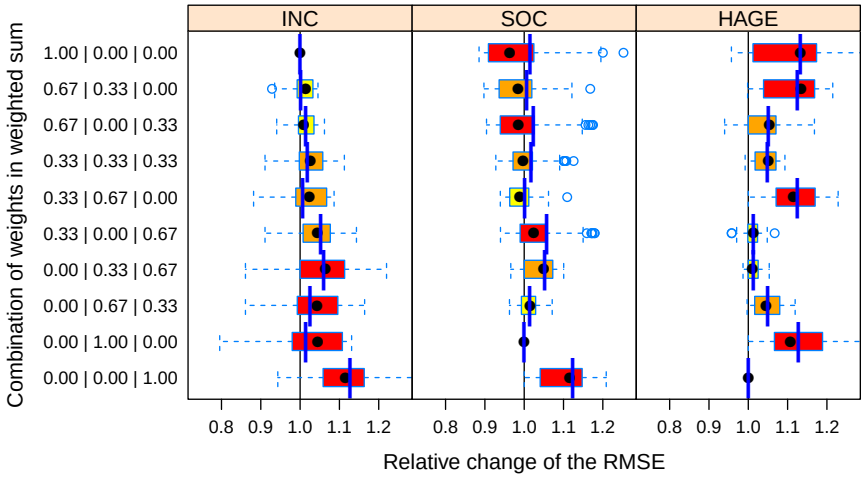

Figure 6: Relative change of the RMSE for the estimated district totals for ten combinations of weights with (opt)-standardization. Red boxes correspond to a weight of 0.00, orange boxes to a weight of 0.33, and yellow boxes to a weight of 0.67 for the respective variable.

increases over all variables and all districts using the (opt)-standardization than using the (CV2)-standardization. For example, in Figure 5 the variable SOC is dominant. If SOC is assigned any weight higher than zero, the increase in the error of the estimates is low for SOC, but high for the other variables of interest. This effect does not occur for the (opt)-standardization and the particular weight combinations $w = (1/3, 1/3, 1/3)$ and $w = (0, 1/3, 2/3)$ in Figure 6.

Pareto optimization

To obtain a characterization of the Pareto frontier, we compute the multivariate optimal allocations for all possible combinations of weights with a resolution of 0.1.

In the heatmaps in Figure 7 we plot the increase of the variances of the total estimates with respect to the univariate optimal allocation variances. Each dot represents one combination of weights. The percentage weight for each separate variable is marked on the related axis. In consequence of the scaling resolution of 0.1, the dots which represent the equal weighting

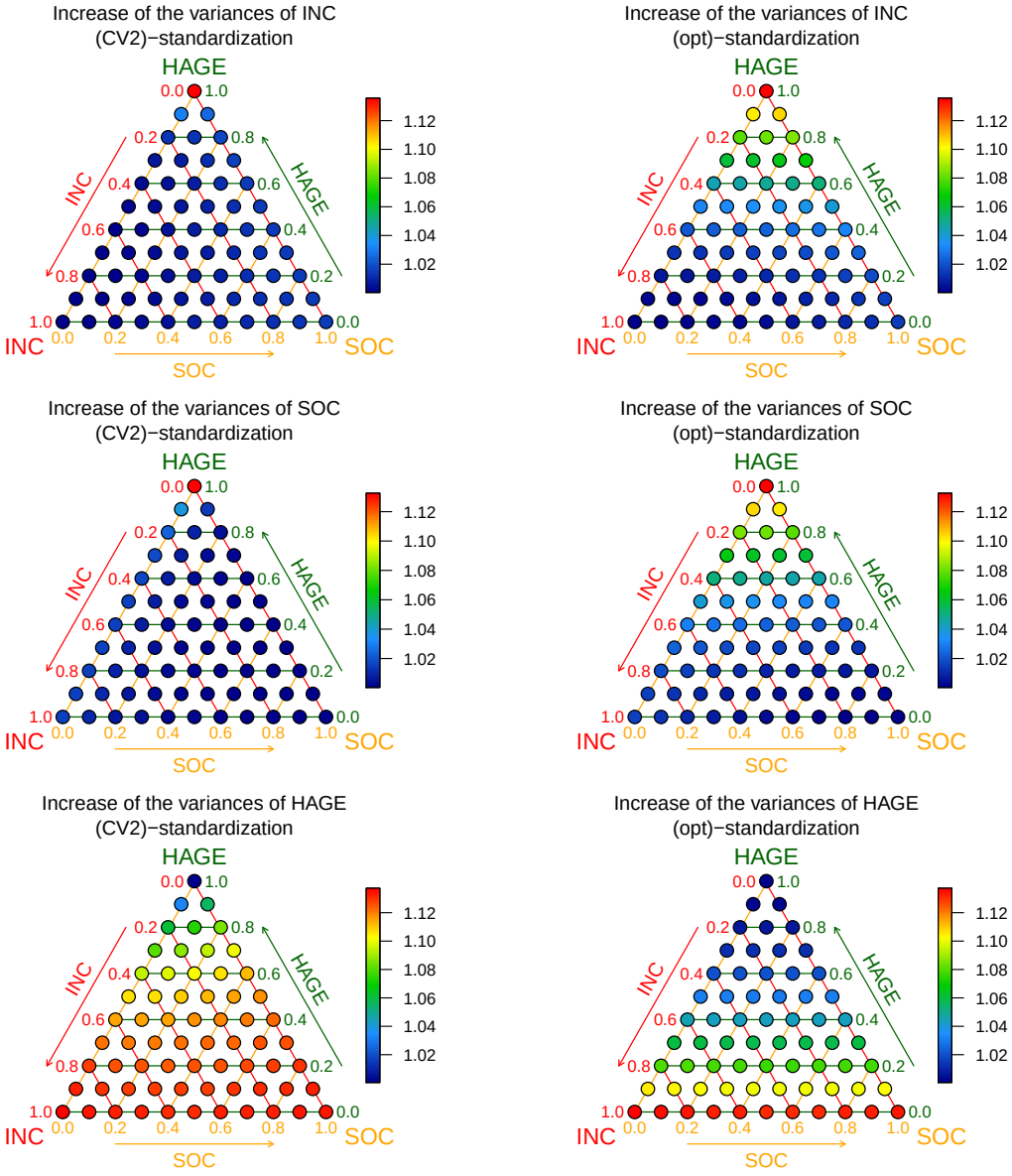

Figure 7: Relative increase of the variances of the population estimates under (CV2)- and (opt)-standardization for 66 combinations of weights for each variable of interest. The percentage weight for each separate variable is marked on the related axis.

$w = (1/3, 1/3, 1/3)$ and the weighting $w = (0, 1/3, 2/3)$ are not contained in the heatmaps.

However, they can be accurately approximated by the surrounding dots. Blue dots are favorable because they represent combinations of weights with a lower increase of the variances. For example, the minimal variance for variable INC is located at the vertex where variable INC is given the full weight 1.00. The variances differ depending on the choice of the standardization strategy.

A similar behavior between the variances of INC and SOC can be observed because of their positive correlation of 0.27. In addition to that, the correlation between HAGE and INC as well as HAGE and SOC is smaller, which results in a higher error-increase of the total estimate of HAGE, even for roughly equal weights. Similarly to Section 4.1, the setting with (opt)-standardization is more balanced in the overall comparison of the heatmaps.

The structure of the heatmaps in Figure 8 is equivalent to Figure 7, but the *cumulated error-increase* of the total estimates of the three variables of interest is plotted. In the case of the (CV2)-standardization, the best choice is an asymmetric weighting. In contrast, the setting with (opt)-standardization is more balanced, so the best choice has roughly equal weights.

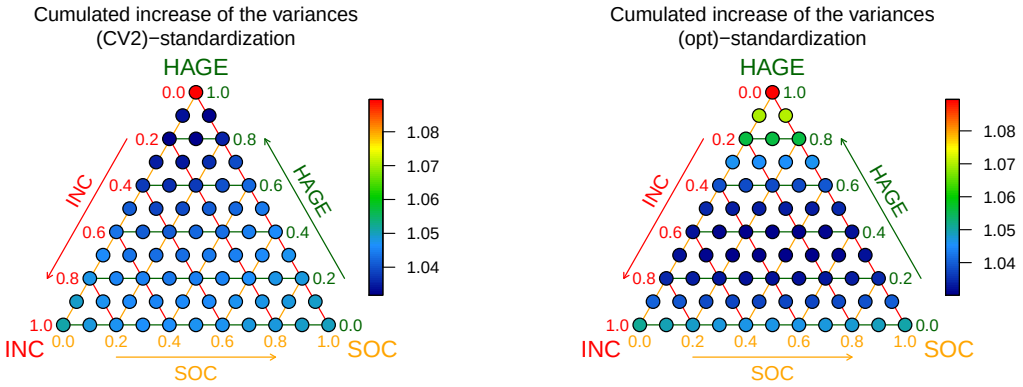

Figure 8: Relative cumulated increase of the variances of the population estimates under (CV2)- and (opt)-standardization for 66 combinations of weights. The percentage weight for each separate variable is marked on the related axis.

By Theorem A.1, each dot in the heatmaps represents the variance of one Pareto optimal solution. To be precise, the dots along the edges (where at least one weight is zero) are weakly Pareto optimal. Combining the heatmaps of the three variables of interest in one plot, we can display the Pareto frontier in Figure 9. Each dot in the three-dimensional space represents one Pareto optimal solution. Each of the three axes represents the error-increase for the corresponding variable. As already observed before, the (opt)-standardization results in a more balanced Pareto frontier.

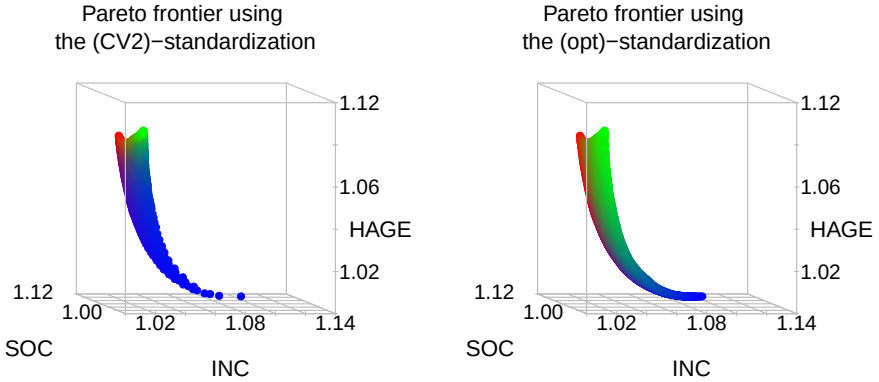

Figure 9: Pareto frontiers for the (CV2)-standardization and (opt)-standardization.

The evaluation of these plots offers valuable support for the decision maker to select the

preferred solution among all efficient solutions. By using the weighted sum and calculating the Pareto frontier, the decision is based on a higher level of reliable information. This contrasts using a p-norm that does not give the user the possibility to choose his preferred solution.

As the computation of the Pareto frontier requires the solution of many optimal allocation problems, it is only realizable in a practical time frame if efficient algorithms are used. We show in Section 4.5 that our algorithms are fast enough to facilitate this analysis of the Pareto frontier for multivariate allocation problems even for large problem instances. Moreover, this finding holds for both the continuous and integer problem.

4.3. Comparison of stratum-specific sample sizes

In Figure 10 the stratum-specific sample sizes are presented on the district level. Each boxplot contains the 40 districts and shows the relative change of the stratum-specific sample sizes compared to the equal weighting $w = (1/3, 1/3, 1/3)$ in line four.

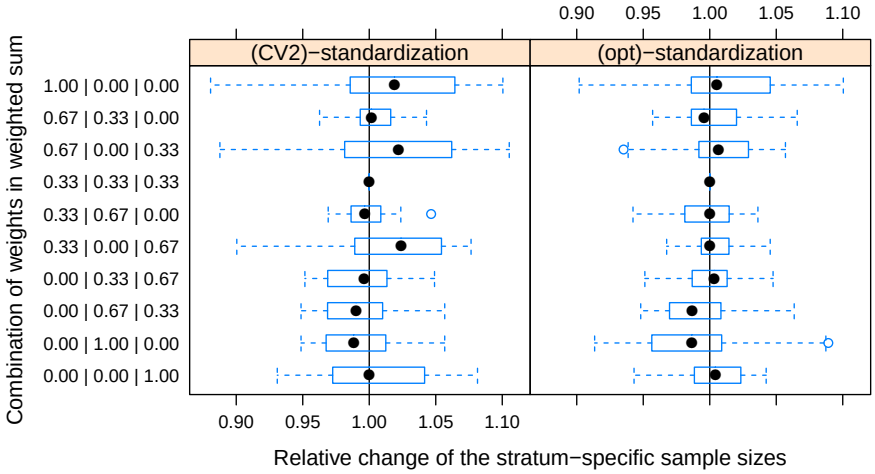

Figure 10: Relative differences in stratum-specific sample sizes of the districts.

On the one hand, there is a spread in the stratum-specific sample sizes depending on the weights (up to 12%), which illustrates the effect of the chosen weighting on the structure of the optimal allocation and the advantage of knowing the Pareto frontier. On the other hand, we recognize clear differences between the (CV2)- and (opt)-standardization. The relative changes of the sample sizes using (opt)-standardization is smaller.

4.4. Comparison of continuous and integer solution

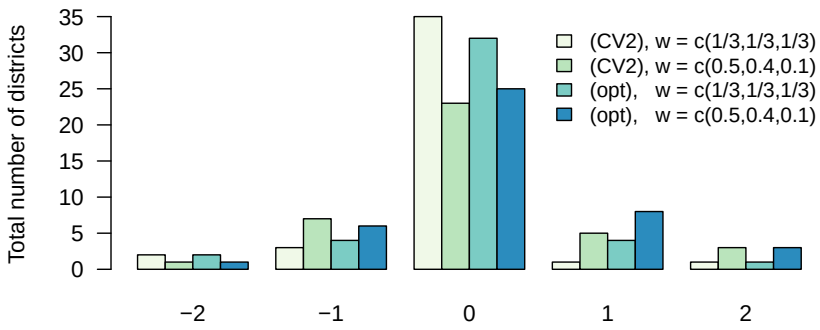

Figure 11: Absolute differences between district-specific sample sizes for rounded and integer allocation.

In Figure 11 we compare the rounded optimal continuous solution of the multivariate allocation problem, computed by the semismooth Newton method, with the optimal integer solution, computed by the Greedy method. We plot the cumulative differences in each district exemplary for two combinations of weights and both standardization techniques.

The rounded sample sizes differ from the optimal integer sample sizes and the differences vary from -2 up to $+2$ per district. The associated RMSEs are also different. This shows the advantage of the integer optimal allocation – especially when considering that the computing times of the continuous and integer solver are of the same magnitude (Friedrich *et al.* 2015).

4.5. Performance of the algorithms

All the numerical results are computed in `R` on a desktop PC with an Intel Core i7-6700 CPU at 3.40GHz $\times$ 8 and an internal memory of 32 GB.

Münnich *et al.* (2012b) and Friedrich *et al.* (2015) show that the fixed-point iteration or semismooth Newton method for the continuous problem as well as the Greedy algorithm for the integer problem have huge advantages in computing time compared to the `R` package `nloptr` which provides an `R` interface to the open-source library `NLopt` for nonlinear optimization. However, as pointed out in Section 2.2, the separability of the objective function is mandatory for these algorithms, so they can only be applied using a weighted sum or 2-norm as standardization technique. The following results are based on the weighted sum setting with equal weights $w = (1/3, 1/3, 1/3)$ and (opt)-standardization. Neither the choice of the weights, nor the selection of the standardization technique changes the numerical performance of the algorithms significantly. The initial point λ^0 for the continuous solvers is calculated as the mean of the three separate univariate optimal stratum-specific sample sizes.

Table 1: Performance of the semismooth Newton algorithm compared to `nloptr`.

	`R` package `nloptr`	Semismooth Newton
Computing time [ms]	6 801	1
Iterations	242	4

The performance of the semismooth Newton algorithm is shown in Table 1 and Table 2. Using similar predefined precisions, we observe distinct improvements in computing time and the number of iterations compared to the `R` package `nloptr`. Moreover, by analyzing column two of Table 2, we can numerically confirm a quadratic convergence rate, which is proved (locally) for the semismooth Newton method in Qi and Sun (1993).

Table 2: Convergence of the semismooth Newton algorithm.

Iteration i	Residual $\|\|\Phi(\lambda^i)\|\|_2$	Objetive $f(n(\lambda^i))$
0	$2.0 \cdot 10^{2}$	$6.0263 \cdot 10^{12}$
1	$6.1 \cdot 10^{0}$	$6.1512 \cdot 10^{12}$
2	$6.0 \cdot 10^{-3}$	$6.1552 \cdot 10^{12}$
3	$5.8 \cdot 10^{-9}$	$6.1553 \cdot 10^{12}$
4	$1.3 \cdot 10^{-11}$	$6.1553 \cdot 10^{12}$

For the computing times of the Greedy methods for the integer allocation problem we refer to the detailed analysis in Friedrich *et al.* (2015), who, in particular, prove a worst-case bound on the running time. The computing times are generally longer than those for the fixed-point iteration in the continuous case, but still well below one second.

5. Conclusion

The optimization of stratified sampling designs must consider many requirements, such as conflicting variables of interest, cost restrictions, or the control of sampling fractions. This results in a multivariate optimal allocation problem under constraints.

We have proposed several scalarization and standardization techniques for the efficient solution of multivariate allocation problems. Whereas the scalarization reflects the decision function when evaluating conflicting goals, the standardization of the variances yields a rescaling of the variables fostering comparability. Furthermore, we have shown how the entire Pareto frontier as the set of all Pareto optimal solutions can be computed. The major benefit is the possibility of an a posteriori choice of a weighting scheme of the variables of interest, so that the decision maker is able to incorporate additional information to achieve the application-specific optimal allocation. As a further advantage, it is not necessary to a priori assess the conflicting goals or rank the variables of interest. Additionally, we have observed considerable differences in estimation errors and stratum-specific sample sizes when varying the weighting schemes. We can underline the importance of the chosen scalarization, standardization, and weighting in multivariate optimal allocation.

We have computed solutions for instances of the continuous and integer allocation problem using the AMELIA dataset. This simulation study presents the algorithms comparatively, underlines their advantages, and allows recommendations for their practical use. In contrast to standard solvers, using the separability and convexity of the given problem yields a substantial increase in the numerical performance, which enables calculating the Pareto frontier in high resolution. The integer algorithm avoids rounding. Moreover, the semismooth Newton method supports extensions with more general restrictions.

Acknowledgements

This research was supported within the project *Research Innovation for Official and Survey Statistics* (RIFOSS), funded by the German Federal Statistical Office, and by the research training group 2126 *Algorithmic Optimization* (ALOP), funded by the German Research Foundation DFG. Finally, we thank the editor for his support and encouragement.

References

Ahsan MJ, Khan SU (1982). "Optimum Allocation in Multivariate Stratified Random Sampling with Overhead Cost." *Metrika*, ***29***, 71–78.

Alfons A, Burgard J, Filzmoser P, Hulliger B, Kolb JP, Kraft S, Münnich R, Schoch T, Templ M (2011). "The AMELI Simulation Study. Research Project Report WP6–D6.1." *Technical report*, AMELI. URL http://ameli.surveystatistics.net.

Arthanari TS, Dodge Y (1981). *Mathematical Programming in Statistics*, volume *341* of *Wiley Series in Probability and Statistics*. Wiley, New York.

Bankier MD (1988). "Power Allocations: Determining Sample Sizes for Subnational Areas." *American Statistical Association*, ***42***(3), 459–472.

Bretthauer KM, Ross A, Shetty B (1999). "Nonlinear Integer Programming for Optimal Allocation in Stratified Sampling." *European Journal of Operational Research*, ***116***(3), 667–680.

Chatterjee S (1968). "Multivariate Stratified Surveys." *Journal of the American Statistical Association*, ***63***(322), 530–534.

Chatterjee S (1972). "A Study of Optimum Allocation in Multivariate Stratified Surveys." *Scandinavian Actuarial Journal*, ***1972***(1), 73–80.

Cochran WG (1977). *Sampling Techniques*. 3rd edition. Wiley, New York.

Dalenius T (1953). "The Multivariate Sampling Problem." *Scandinavian Actuarial Journal*, ***36***, 92–102.

Díaz-García JA, Cortez LU (2006). "Optimum Allocation in Multivariate Stratified Sampling: Multi-Objective Programming." *Technical report*, Centro de Investigación en Matemáticas, Guanajuato, México.

Díaz-García JA, Ramos-Quiroga R (2014). "Optimum Allocation in Multivariate Stratified Random Sampling: A Modified Prékopa's Approach." *Journal of Mathematical Modelling and Algorithms in Operations Research*, ***13***(3), 315–330.

Ehrgott M (2005). *Multicriteria Optimization*. 2nd edition. Springer, Heidelberg.

Falorsi PD, Righi P (2008). "A Balanced Sampling Approach for Multi-Way Stratification Designs for Small Area Estimation." *Survey Methodology*, ***34***(2), 223–234.

Falorsi PD, Righi P (2015). "Generalized Framework for Defining the Optimal Inclusion Probabilities of One-Stage Sampling Designs for Multivariate and Multi-Domain Surveys." *Survey Methodology*, ***41***(1), 215–236.

Falorsi PD, Righi P (2016). "A Unified Approach for Defining Optimal Multivariate and Multi-Domains Sampling Designs." In *Topics in Theoretical and Applied Statistics*, pp. 145–152. Springer, Heidelberg.

Folks JL, Antle CE (1965). "Optimum Allocation of Sampling Units to Strata when there are R Responses of Interest." *Journal of the American Statistical Association*, ***60***(309), 225–233.

Friedrich U (2016). *Discrete Allocation in Survey Sampling and Analytic Algorithms for Integer Programming*. Ph.D. thesis, Trier University.

Friedrich U, Münnich R, de Vries S, Wagner M (2015). "Fast Integer-Valued Algorithms for Optimal Allocations under Constraints in Stratified Sampling." *Computational Statistics and Data Analysis*, ***92***, 1–12.

Gabler S, Ganninger M, Münnich R (2012). "Optimal Allocation of the Sample Size to Strata under Box Constraints." *Metrika*, ***75***(2), 151–161.

Hochbaum D (1995). "A Nonlinear Knapsack Problem." *Operations Research Letters*, ***17***, 103–110.

Hohnhold H (2009a). "Gerneralized Power Allocations." *Technical report*, Statistisches Bundesamt, Wiesbaden.

Hohnhold H (2009b). "Variants of Optimal Allocation in Stratified Sampling." *Technical report*, Statistisches Bundesamt, Wiesbaden.

Huddleston HF, Claypool PL, Hocking RR (1970). "Optimal Sample Allocation to Strata Using Convex Programming." *Journal of the Royal Statistical Society Series C*, ***19***(3), 273–278.

Jahn J (1986). *Mathematical Vector Optimization in Partially Ordered Linear Spaces*. Verlag Peter Lang, Frankfurt am Main.

Khan MF, Ali I, Raghav YS, Bari A (2012). "Allocation in Multivariate Stratified Surveys with Non-Linear Random Cost Function." *American Journal of Operations Research*, ***2***, 100–105.

Khan MGM, Khan EA, Ahsan MJ (2003). "An Optimal Multivariate Stratified Sampling Design Using Dynamic Programming." *Australian and New Zealand Journal of Statistics*, ***45***(1), 107–113.

Kish L (1976). "Optima and Proxima in Linear Sample Designs." *Journal of the Royal Statistics Society Series A*, ***139***(1), 80–95.

Kokan AR (1963). "Optimum Allocation in Multivariate Surveys." *Journal of the Royal Statistics Society Series A*, **126**(*4*), 557–565.

Kokan AR, Khan S (1967). "Optimum Allocation in Multivariate Surveys: An Analytical Solution." *Journal of the Royal Statistics Society Series B*, ***29***(1), 115–125.

Lin JG (2005). "On Min-Norm and Min-Max Methods of Multi-Objective Optimization." *Mathematical Programming*, ***103***, 1–33.

Lohr SL (2010). *Sampling: Design and Analysis*. 2nd edition. Cengage Learning, Boston.

Merkle H, Burgard JP, Münnich R (2016). "The AMELIA Dataset - A Synthetic Universe for Reproducible Research." In YG Berger, JP Burgard, A Byrne, A Cernat, C Giusti, P Koksel, S Lenau, S Marchetti, H Merkle, R Münnich, I Permanyer, M Pratesi, N Salvati, N Shlomo, D Smith, N Tzavidis (eds.), *Deliverable 23.1: Case studies*, volume WP23 – D23.1. URL http://inclusivegrowth.be.

Münnich R, Sachs E, Wagner M (2012a). "Calibration of Estimator-Weights via Semismooth Newton." *Journal of Global Optimization*, ***52***(3), 471–485.

Münnich R, Sachs EW, Wagner M (2012b). "Numerical Solution of Optimal Allocation Problems in Stratified Sampling under Box Constraints." *AStA Advances in Statistical Analysis*, ***96***, 435–450.

Neyman J (1934). "On the two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society*, ***97***, 558–625.

Qi L, Sun J (1993). "A Nonsmooth Version of Newton's Method." *Mathematical Programming*, ***58***(1), 353–367.

Schaich E, Münnich R (1993). "Zum Allokationsproblem bei Mehreren Untersuchungsvariablen." *Allgemeines Statistisches Archiv*, ***77***, 390–405.

Tschuprow A (1923). "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations." *Metron*, ***2***, 461–493.

Wagner M (2013). *Numerical Optimization in Survey Statistics*. Ph.D. thesis, Trier University.

Ypma J, Borchers H, Eddelbuettel D (2014). "R Package nloptr: R Interface to NLopt." http://CRAN.R-project.org/package=sampling. R package version 1.04.

A. Appendix: Optimality in multi-criteria optimization

It is in general not possible to find all Pareto-optimal points for a multi-criteria optimization problems by solving the weighted sum problem. In this section we analyze the weighted sum method for the multivariate allocation problem mathematically. While the sufficient condition of Theorem A.1 holds in a very general setting, see for example Folks and Antle (1965), this is not true for the necessary condition of Theorem A.3. We assume that the reader is familiar with the concept of optimality in multi-objective optimization and in particular with (weak) Pareto optimality. We refer to Ehrgott (2005, chapter 2) or Jahn (1986, chapter 4) for a detailed presentation.

Theorem A.1 (Sufficient Condition)**.** *Let $D \subseteq \mathbb{R}^H$ and let $f_k : D \to \mathbb{R}$, $k = 1, \ldots, K$. For every optimal solution $\bar{n}$ of $\min_{n \in D} \sum_{k=1}^{K} w_k f_k(n)$ with weights $w \in \mathbb{R}^K$, the following statements hold.*

1. *$\bar{n}$ is a weakly Pareto optimal solution for $\min_{n \in D}(f_1(n), \ldots, f_K(n))$ if $w \geq 0$.*
2. *$\bar{n}$ is a Pareto optimal solution for $\min_{n \in D}(f_1(n), \ldots, f_K(n))$ if $w > 0$.*

Proof. Proposition 3.9 in Ehrgott (2005). □

In what follows, we show that under convexity assumptions it is possible to find *all* Pareto optimal points by solving a weighted sum problem.

Lemma A.2. *Let $D \subseteq \mathbb{R}^H$ be convex and let $f_k : D \to \mathbb{R}$, $k = 1, \ldots, K$, be convex functions. Then the set $C_+(f) := \{(f_1(n), \ldots, f_K(n))^T | n \in D\} + \mathbb{R}^K_+$ is convex.*

Proof. Theorem 2.6 in Jahn (1986). □

Theorem A.3 (Necessary Condition)**.** *Let $D \subseteq \mathbb{R}^H$ be convex and let $f_k : D \to \mathbb{R}$ for $k = 1, \ldots, K$ be convex functions. Then, for each Pareto optimal solution $\bar{n}$ of the problem $\min_{n \in D}(f_1(n), \ldots, f_K(n))$ there exist weights $\bar{w} \in \mathbb{R}^K_+ \setminus \{0\}$ such that $\bar{n}$ is an optimal solution of the weighted sum problem $\min_{n \in D} \sum_{k=1}^{K} w_k f_k(n)$.*

Proof. Using the convexity of the objective function, Lemma A.2 shows that the set $C_+(f)$ mentioned in the lemma is convex. Using this property, the result follows directly from Theorem 5.4 in Jahn (1986). □

As the convexity assumption of Theorem A.3 holds for the optimal allocation problem formulated in (5), we can apply the theorem and we have proved that (up to discretization) the computations in Sections 3 and 4 describe the entire Pareto frontier of the problem.

Affiliation:

Ulf Friedrich
DFG-RTG Algorithmic Optimization
Trier University
D-54286 Trier, Germany
E-mail: friedrich@uni-trier.de
URL: https://www.optmath.de

Ralf Münnich
Economic and Social Statistics Department
Trier University
D-54286 Trier, Germany
E-mail: muennich@uni-trier.de
URL: https://www.uni-trier.de/index.php?id=58137

Martin Rupp
Economic and Social Statistics Department
Trier University
D-54286 Trier, Germany
E-mail: ruppm@uni-trier.de
URL: https://www.uni-trier.de/index.php?id=54776

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 47
February 2018

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2017-11-16
Accepted: 2017-11-16

Austrian Journal of Statistics
February 2018, Volume 47, 53–67.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i2.581

On Distribution Characteristics of a Fuzzy Random Variable

Jalal Chachi
Semnan University

Abstract

By combining two types of uncertainty randomness and vagueness the concept of fuzzy random variable was introduced in order to integrate fuzzy set theory into a branch of statistical analysis called "statistics with vague data". In this paper, a concept of fuzzy random variable will be presented. Using classical techniques in Probability Theory, some aspects and results associated to a random variable (including expectation, variance, covariance, correlation coefficient, fuzzy (empirical) cumulative distribution function) will be extended to this notion of fuzzy random variable. This notion provides a useful framework/results in order to extend statistical analysis to situations when the outcomes of random experiment are fuzzy sets.

Keywords: fuzzy random variable, fuzzy expected value, fuzzy (empirical) cumulative distribution function.

1. Introduction

Statistical data are frequently associated with an underlying imprecision due, for instance, to inexactitude in the measuring process, vagueness of the involved concepts or a certain degree of ignorance about the real values. In many cases, such an imprecision can be modeled by means of fuzzy sets in a more efficient way than considering only a single value or category (Zadeh 1965). Thus, these kinds of data are jointly affected by two sources of uncertainty: fuzziness (due to imprecision, vagueness, partial ignorance) and randomness (due to sampling or measurement errors of stochastic nature). Randomness models the stochastic variability of all possible outcomes of a situation, and fuzziness relates to the unsharp boundaries of the parameters of the model. As Zadeh (1995) states that "Probability Theory and Fuzzy Logic are complementary rather than competitive", clearly, a natural question is how fuzzy variables could interact with the type of random variables found in association with many real-life random experiments from different fields. In this way, by combining ideas, concepts and results from both theories, this article focuses on one important dimension of this issue, fuzzy random variables.

The concept of fuzzy random variable (frv) (also called "random fuzzy set" (Blanco-Fernández, Casals, Colubi, Corral, García-Bárzana, Gil, González-Rodríguez, López, Lubiano, Montenegro, Ramos-Guajardo, De La Rosa De Sá, and Sinova 2013)) was introduced in order to deal

with situations where the outcomes of a random experiment are modeled by fuzzy sets (Colubi, Domínguez-Menchero, López-Díaz, and Ralescu 2001; Colubi, Fernández-García, and Gil 2002; Colubi and Gil 2007; Colubi and González-Rodríguez 2007; Couso and Sánchez 2008; Feng 2000; Gil 2001; Gil, López-Díaz, and Ralescu 2006; González-Rodríguez, Colubi, and Gil 2006a; Krätschmer 2001; Kruse and Meyer 1987; Kwakernaak 1978, 1979; Liu and Liu 2003; Puri and Ralescu 1985, 1986; Shapiro 2009). An frv is a mapping that associates a fuzzy set of the final space to each possible result of a random experiment in a provided probability space structure. Thus, this concept generalizes the definitions of random variable and random set. Although these generalizations are not unique in the literature but they can be formalized in equivalent ways. Each definition differs from the others in the structure of the final space and the way the measurability condition is transferred to this context. For instance, Krätschmer (2001); Kruse and Meyer (1987) and Puri and Ralescu (1985, 1986) focused on the properties of the multi-valued mappings associated to the α-cuts. Kwakernaak (1978, 1979) assumes that the outcomes of the frv are fuzzy real subsets and the extreme points of their α-cuts are classical random variables. Puri and Ralescu (1985, 1986) require the α-cuts to be measurable (also different conditions for measurability of multi-valued mappings can be formulated). On the other hand, Klement, Puri, and Ralescu (1986) and Diamond and Kloeden (1994) define frv's, as classical measurable mappings. Couso and Sánchez (2008) present three different higher order possibility models that represents the imprecise information provided by an frv.

In the literature on frvs, there are only a few references to modeling the distribution of these random elements. These models are theoretically well stated, but they are not soundly supported by empirical evidence, since they correspond to restrictive random mechanisms and hence they are not realistic in practice (González-Rodríguez, Colubi, Gil, and Coppi 2006; Möller, Graf, M., and Sickert 2002). This motivated us to present in this paper another model that represents the imprecise information provided by an frv. Within this framework, we use the tools of general Probability Theory (Billingsley 1995) to define fuzzy cumulative distribution function and fuzzy empirical cumulative distribution function for an frv. We also extend the concepts of expectation, variance, covariance and correlation coefficient of an frv by reproducing classical techniques. For instance, when the images of the frv are convex fuzzy subsets of $\mathbb{R}$, we can use fuzzy arithmetic to derive a method of construction of the fuzzy expectation. On the other hand, we can make a parallel construction of the variance: let us consider a particular metric defined over the class of fuzzy subsets of the final space. In this setting, we define the variance of an frv as the mean (classical expectation of a random variable) of the squares of the distances from the images of the frv to the (fuzzy) expectation. In this context the variance of an frv is a (precise) number that quantifies the degree of dispersion of the images of the frv.

Extending these results is not just a matter of motivation, but the main issue is that the concepts of fuzzy cumulative distribution function and fuzzy empirical cumulative distribution function for an frv strongly affects the aim of the Statistics to be developed around (Hesamian and Chachi 2013). Although in the literature distributions and parameters could be defined in some senses in connection with the frv through Zadeh's extension principle (Zadeh 1965), but the objective of statistical developments refer usually to the distribution and parameters of the underlying original real-valued random variable (Wu 1999). When the distribution of an frv can be defined, the objective of statistical developments will only refer to the distribution and parameters of the frv, since either there is no underlying real-valued random variable behind the process (as happens when we deal with judgments, valuations, ratings, and so on) or the interest is just to be focused on the fuzzy perception (Blanco-Fernández *et al.* 2013). Therefore, the aim of inferential statistical developments with fuzzy data based on frvs will be to draw conclusions about the distribution of the involved frvs over populations on the basis of the information supplied by samples of (fuzzy) observations from these frvs. One of the relevant inferential problems is to estimate the parameters or measures associated with the distribution of an frv on the basis of the information provided by a sample of independent data from it. Furthermore, when Statistics are based on the concept of frv, some additional

problems arise (see also Conclusion), like

1. the lack of realistic general "parametric" families of probability distribution models for frvs (Blanco-Fernández *et al.* 2013);
2. the lack of Central Limit Theorems (CLTs) for frvs which are directly applicable for inferential purposes (Wu 2000; Krätschmer 2002a,b).

The above first item will be considered in this paper for the proposed frv by defining the concepts of fuzzy cumulative distribution function and fuzzy empirical cumulative distribution function. The second item (and also some other items in Conclusion) can be addressed in feature researches.

The paper is organized as follows. The next section provides the necessary technical background used for convenience of explaining general concepts concerned with fuzzy sets. In Section 3, we propose a new definition of frv. In Section 4, using classical techniques in Probability Theory, we extend some common characteristics of frvs including expectation, variance, covariance, correlation coefficient. In Section 5, we generalize the concept of fuzzy cumulative distribution function and fuzzy empirical cumulative distribution function for an frv. We end the paper with some general concluding remarks and open problems.

2. Preliminary concepts

In this section, first, we shall review the basic definitions and terminologies of the fuzzy set theory and uncertainty theory which are necessary for our paper (for further details, the reader is referred to Liu (2002, 2016); Peng and Liu (2004); Viertl (2011); Zimmermann (2001)). Then, a new definition of distance measure between fuzzy numbers is defined.

2.1. Fuzzy numbers

A fuzzy set $\widetilde{A}$ of the universal set $\mathbb{X}$ is defined by its membership function $\widetilde{A} : \mathbb{X} \to [0,1]$. In this paper, we consider $\mathbb{R}$ (the real line) as the universal set. We denote by $\widetilde{A}[\alpha] = \{x \in \mathbb{R} : \widetilde{A}(x) \geq \alpha\}$ the α-level set (α-cut) of the fuzzy set $\widetilde{A}$ of $\mathbb{R}$, for every $\alpha \in (0,1]$, and $\widetilde{A}[0]$ is the closure of the set $\{x \in \mathbb{R} : \widetilde{A}(x) > 0\}$. A fuzzy set $\widetilde{A}$ of $\mathbb{R}$ is called a fuzzy number if for every $\alpha \in [0,1]$, the set $\widetilde{A}[\alpha]$ is a non-empty compact interval. We denote by $\mathcal{F}(\mathbb{R})$, the set of all fuzzy numbers of $\mathbb{R}$.

A specific type of fuzzy number, which is rich and flexible enough to cover most of the applications, is the so-called LR-fuzzy number. Typically, the LR fuzzy number $\widetilde{N} = (n, l, r)_{LR}$ with central value $n \in \mathbb{R}$, left and right spreads $l \in \mathbb{R}^+$, $r \in \mathbb{R}^+$, decreasing left and right shape functions $L : \mathbb{R}^+ \to [0,1]$, $R : \mathbb{R}^+ \to [0,1]$, with $L(0) = R(0) = 1$, has the following membership function

$$\widetilde{N}(x) = \begin{cases} L(\frac{n-x}{l}) & \text{if} \quad x \leq n, \\ R(\frac{x-n}{r}) & \text{if} \quad x \geq n. \end{cases}$$

We can easily obtain the α-cut of $\widetilde{N}$ as follows

$$\widetilde{N}[\alpha] = [n - L^{-1}(\alpha)l, n + R^{-1}(\alpha)r], \quad \alpha \in [0,1].$$

For the algebraic operations of LR-fuzzy numbers, we have the following result on the basis of Zadeh's extension principle. Let $\widetilde{A} = (a, l_1, r_1)_{LR}$ and $\widetilde{B} = (b, l_2, r_2)_{LR}$ be two LR-fuzzy numbers and $\lambda \in \mathbb{R} - \{0\}$ be a real number. Then

$$\begin{aligned} \lambda \otimes \widetilde{A} &= \begin{cases} (\lambda a, \lambda l_1, \lambda r_1)_{LR} & \text{if} \quad \lambda > 0, \\ (\lambda a, |\lambda| r_1, |\lambda| l_1)_{RL} & \text{if} \quad \lambda < 0, \end{cases} \\ \widetilde{A} \oplus \widetilde{B} &= (a + b, l_1 + l_2, r_1 + r_2)_{LR}, \end{aligned}$$

2.2. Some notions from uncertainty theory

In the following, we introduce an index to compare fuzzy number $\widetilde{A} \in \mathcal{F}(\mathbb{R})$ and crisp value $x \in \mathbb{R}$. The index is used for defining a new notion of frv.

Definition 1 (Liu and Liu (2002))**.** Let $\widetilde{A} \in \mathcal{F}(\mathbb{R})$ and $x \in \mathbb{R}$. The index

$$C : \mathcal{F}(\mathbb{R}) \times \mathbb{R} \longrightarrow [0, 1],$$

which is defined by

$$C\{\widetilde{A} \leq x\} = \frac{\sup_{y \leq x} \widetilde{A}(y) + 1 - \sup_{y > x} \widetilde{A}(y)}{2},$$

shows the credibility degree that $\widetilde{A}$ is less than or equal to x. Similarly, $C\{\widetilde{A} > x\} = 1 - C\{\widetilde{A} \leq x\}$ shows the credibility degree that $\widetilde{A}$ is greater than x (see also Liu (2016)).

Definition 2 (Liu (2002))**.** Let $\widetilde{A} \in \mathcal{F}(\mathbb{R})$ and $\alpha \in [0, 1]$, then

$$\widetilde{A}_\alpha = \inf\{x \in \widetilde{A}[0] : C\{\widetilde{A} \leq x\} \geq \alpha\},$$

is called the α-pessimistic value of $\widetilde{A}$. It is clear that $\widetilde{A}_\alpha$ is a non-decreasing function of $\alpha \in (0, 1]$ (see also Peng and Liu (2004)).

Lemma 1. *Let $\widetilde{A}, \widetilde{B} \in \mathcal{F}(\mathbb{R})$ and λ be a real number. Then*

$$\begin{aligned} (\widetilde{A} \oplus \widetilde{B})_\alpha &= \widetilde{A}_\alpha + \widetilde{B}_\alpha. \\ (\lambda \otimes \widetilde{A})_\alpha &= \begin{cases} \lambda \times \widetilde{A}_\alpha & \text{if } \lambda > 0, \\ \lambda \times \widetilde{A}_{1-\alpha} & \text{if } \lambda < 0, \end{cases} \end{aligned}$$

Example 1. Suppose that $\widetilde{A} = (a, l, r)_{LR}$ is a LR-fuzzy number, and let $x \in \mathbb{R}$, then

$$C\{\widetilde{A} \leq x\} = \begin{cases} \frac{1}{2} L(\frac{a-x}{l}) & \text{if } \ x \leq a, \\ 1 - \frac{1}{2} R(\frac{x-a}{r}) & \text{if } \ x \geq a. \end{cases}$$

We can easily obtain the α-pessimistic values of $\widetilde{A}$ as follows

$$\widetilde{A}_\alpha = \begin{cases} a - lL^{-1}(2\alpha) & \text{if } \ 0.0 < \alpha \leq 0.5, \\ a + rR^{-1}(2(1-\alpha)) & \text{if } \ 0.5 \leq \alpha \leq 1.0. \end{cases}$$

As an example, consider the triangular fuzzy number $\widetilde{A} = (a, l, r)_T$, then

$$\begin{aligned} C\{\widetilde{A} \leq x\} &= \begin{cases} 0 & \text{if } \ x \in (-\infty, a - l), \\ \frac{x-a+l}{2l} & \text{if } \ x \in [a - l, a), \\ \frac{x-a+r}{2r} & \text{if } \ x \in [a, a + r), \\ 1 & \text{if } \ x \in [a + r, \infty). \end{cases} \\ \widetilde{A}_\alpha &= \begin{cases} a - l(1 - 2\alpha) & \text{if } \ 0.0 < \alpha \leq 0.5, \\ a - r(1 - 2\alpha) & \text{if } \ 0.5 \leq \alpha \leq 1.0. \end{cases} \end{aligned}$$

2.3. A new distance measure between fuzzy numbers

In the literature one can find many useful metrics between fuzzy numbers. Valuable references on this topic can be found in Blanco-Fernández *et al.* (2013); Feng and Liu (2006); Liu and Liu (2002). In the following, a new definition of metrics between fuzzy numbers is defined.

Definition 3. The distance measure is defined as the mapping $D : \mathcal{F}(\mathbb{R}) \otimes \mathcal{F}(\mathbb{R}) \to [0, \infty)$ such that it associates with two fuzzy numbers $\widetilde{A}, \widetilde{B} \in \mathcal{F}(\mathbb{R})$ the following value

$$D(\widetilde{A}, \widetilde{B}) = \int_0^1 \left(\widetilde{A}_\alpha - \widetilde{B}_\alpha\right)^2 d\alpha.$$

One can conclude that the mapping $D : \mathcal{F}(\mathbb{R}) \otimes \mathcal{F}(\mathbb{R}) \to [0, \infty)$ satisfies the following conditions:

1. For any $\widetilde{A}, \widetilde{B} \in \mathcal{F}(\mathbb{R})$, $D(\widetilde{A}, \widetilde{B}) = 0$ if and only if $\widetilde{A} = \widetilde{B}$.
2. For any $\widetilde{A}, \widetilde{B} \in \mathcal{F}(\mathbb{R})$, $D(\widetilde{A}, \widetilde{B}) = D(\widetilde{B}, \widetilde{A})$.
3. For any $\widetilde{A}, \widetilde{B}, \widetilde{C} \in \mathcal{F}(\mathbb{R})$, such that $\widetilde{A} \subseteq \widetilde{B} \subseteq \widetilde{C}$, then $D(\widetilde{A}, \widetilde{C}) \geq \max\{D(\widetilde{A}, \widetilde{B}), D(\widetilde{B}, \widetilde{C})\}$.
4. For any $\widetilde{A}, \widetilde{B}, \widetilde{C} \in \mathcal{F}(\mathbb{R})$, $D(\widetilde{A}, \widetilde{C}) \leq D(\widetilde{A}, \widetilde{B}) + D(\widetilde{B}, \widetilde{C})$.

As an example, we can easily obtain the distance between two LR-fuzzy numbers $\widetilde{A} = (a, l_1, r_1)_{LR}$ and $\widetilde{B} = (b, l_2, r_2)_{LR}$ as follows

$$\begin{aligned} D(\widetilde{A}, \widetilde{B}) \;=\; & (a-b)^2 + \frac{(l_1 - l_2)^2}{2} \int_0^1 \left(L^{-1}(\alpha)\right)^2 d\alpha + \frac{(r_1 - r_2)^2}{2} \int_0^1 \left(R^{-1}(\alpha)\right)^2 d\alpha \\ & -(a-b)(l_1 - l_2) \int_0^1 L^{-1}(\alpha)\, d\alpha + (a-b)(r_1 - r_2) \int_0^1 R^{-1}(\alpha)\, d\alpha. \end{aligned}$$

For symmetric fuzzy numbers $\widetilde{A} = (a, l, l)_L$ and $\widetilde{B} = (b, r, r)_L$, we have

$$D(\widetilde{A}, \widetilde{B}) = (a-b)^2 + (l-r)^2 \int_0^1 \left(L^{-1}(\alpha)\right)^2 d\alpha.$$

3. Fuzzy random variables

In the context of random experiments whose outcomes are not numbers (or vectors in $\mathbb{R}^p$) but they are expressed in inexact terms, the concept of frv turns out to be useful. Random fuzzy numbers (or, more generally, random fuzzy sets (Blanco-Fernández *et al.* 2013)) is a well-stated and supported model within the probabilistic setting for the random mechanisms generating fuzzy data. They integrate randomness and fuzziness, so that the first one affects the generation of experimental data, whereas the second one affects the nature of experimental data which are assumed to be intrinsically imprecise. The notion of random fuzzy set can be formalized in several equivalent ways. Thus, in this regard, different notions of frv have been introduced and investigated in the literature (Colubi *et al.* 2001; Couso and Sánchez 2008; Feng 2000; Gil *et al.* 2006; González-Rodríguez *et al.* 2006a; Hesamian and Chachi 2013; Krätschmer 2001; Kruse and Meyer 1987; Kwakernaak 1978, 1979; Liu and Liu 2003; Puri and Ralescu 1985, 1986; Shapiro 2009).

Definition 4. Suppose that a random experiment is described by a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, where Ω is a set of all possible outcomes of the experiment, $\mathcal{A}$ is a σ-algebra of subsets of Ω and $\mathbf{P}$ is a probability measure on the measurable space $(\Omega, \mathcal{A})$. The fuzzy-valued mapping $\widetilde{X} : \Omega \to \mathcal{F}(\mathbb{R})$ is called an frv if for any $\alpha \in [0, 1]$, the real-valued mapping $\widetilde{X}_\alpha : \Omega \to \mathbb{R}$ is a real-valued random variable on $(\Omega, \mathcal{A}, \mathbf{P})$. Throughout this paper, we assume that all random variables have the same probability space $(\Omega, \mathcal{A}, \mathbf{P})$

Kwakernaak (1978, 1979) introduced the notion of frvs which has been later formalized in a clear way by Kruse and Meyer (1987) as: given a probability space $(\Omega, \mathcal{A}, \mathbf{P})$, a mapping $\widetilde{X} : \Omega \to \mathcal{F}(\mathbb{R})$ is said to be an frv if for all $\alpha \in (0, 1]$ the two real-valued mappings $\widetilde{X}_\alpha^L : \Omega \to \mathbb{R}$ and $\widetilde{X}_\alpha^U : \Omega \to \mathbb{R}$ are real-valued random variables.

It can be easily investigated that the following relationships are held between the notion of frv proposed in Definition 4 and Kwakernaak and Kruse's definition of frv (see also, Example 1)

$$\begin{aligned} \widetilde{X}_\alpha &= \begin{cases} \widetilde{X}^L_{2\alpha} & \text{for} \quad 0.0 < \alpha \leq 0.5, \\ \widetilde{X}^U_{2(1-\alpha)} & \text{for} \quad 0.5 \leq \alpha \leq 1.0, \end{cases} \\ \widetilde{X}[\alpha] &= [\widetilde{X}_{\frac{\alpha}{2}}, \widetilde{X}_{1-\frac{\alpha}{2}}], \qquad \alpha \in (0,1]. \end{aligned}$$

The first relation shows that the information contained in the two-dimensional variable $(\widetilde{X}^L_\alpha, \widetilde{X}^U_\alpha)$ is summarized in the one-dimensional variable $\widetilde{X}_\alpha$ making the computational procedures in the problems more easier.

Definition 5. Two frvs $\widetilde{X}$ and $\widetilde{Y}$ are said to be independent if $\widetilde{X}_\alpha$ and $\widetilde{Y}_\alpha$ are independent, for all $\alpha \in [0,1]$. In addition, we say that two frvs $\widetilde{X}$ and $\widetilde{Y}$ are identically distributed if $\widetilde{X}_\alpha$ and $\widetilde{Y}_\alpha$ are identically distributed, for all $\alpha \in [0,1]$. Similar arguments can be used for more than two frvs. We also say that $\widetilde{X}_1, \ldots, \widetilde{X}_n$ is a fuzzy random sample if $\widetilde{X}_i$'s are independent and identically distributed frvs. We denote by $\widetilde{x}_1, \ldots, \widetilde{x}_n$ the observed values of fuzzy random sample $\widetilde{X}_1, \ldots, \widetilde{X}_n$.

4. Fuzzy expected value, variance and covariance of an frv

In analyzing fuzzy data two main types of summary measures/parameters may be distinguished:

1. fuzzy-valued summary measures, like the mean value of an frv or the median of an frv as measures for the central tendency of their distributions;
2. real-valued summary measures, like the variance of an frv as a measure for the mean error/dispersion of the distributions of the frv, or the covariance and correlation coefficient as measures of the (absolute) linear dependence/association of an frv.

Definition 6. Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a real-valued random variable. We say that X has finite mean and write $X \in L^1(\Omega, \mathcal{A}, \mathbf{P})$ if and only if $E(X) = \int_\Omega X \, d\mathbf{P} < M$, for some constant $M < \infty$.

Definition 7. Given a probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and an associated frv $\widetilde{X} : \Omega \to \mathcal{F}(\mathbb{R})$ such that for any $\alpha \in [0,1]$ the real-valued random variable $\widetilde{X}_\alpha : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{A}, \mathbf{P})$ has finite mean then the mean value of $\widetilde{X}$ is the fuzzy value $\widetilde{E}(\widetilde{X}) \in \mathcal{F}(\mathbb{R})$ such that for all $\alpha \in [0,1]$

$$\widetilde{E}(\widetilde{X})_\alpha = E(\widetilde{X}_\alpha) = \int_\Omega \widetilde{X}_\alpha \, d\mathbf{P}.$$

The mean value of an frv satisfies the usual properties of linearity and it is the Fréchet's expectation w.r.t. D, which corroborates the fact that it is a central tendency measure (Näther 2001). In this way,

Proposition 1. *$\widetilde{E}$ is additive (i.e., equivariant under the sum of frvs), that is, for frvs $\widetilde{X}$ and $\widetilde{Y}$ associated with the same probability space $(\Omega, \mathcal{A}, \mathbf{P})$ and such that $\widetilde{X}_\alpha, \widetilde{Y}_\alpha \in L^1(\Omega, \mathcal{A}, \mathbf{P})$, we have that*

1. *$\widetilde{E}(\lambda \oplus \widetilde{X}) = \lambda \oplus \widetilde{E}(\widetilde{X})$, for any constant number $\lambda \in \mathbb{R}$.*
2. *$\widetilde{E}(\lambda \otimes \widetilde{X}) = \lambda \otimes \widetilde{E}(\widetilde{X})$, for any constant number $\lambda \in \mathbb{R}$.*
3. *$\widetilde{E}(\widetilde{X} \oplus \widetilde{Y}) = \widetilde{E}(\widetilde{X}) \oplus \widetilde{E}(\widetilde{Y})$.*

Proposition 2. *$\widetilde{E}$ is the Fréchet's expectation of $\widetilde{X}$ w.r.t. D, that is,*

$$\widetilde{E}(\widetilde{X}) = argmin_{\widetilde{U}\in\mathcal{F}(\mathbb{R})}\widetilde{E}\left[D(\widetilde{X},\widetilde{U})\right],$$

so that the mean is the fuzzy value leading to the lowest mean squared D-distance (or error) with respect to the frv distribution, and this corroborates the fact that it is a central tendency measure.

Definition 8. The variance of an frv $\widetilde{X}$ is defined as

$$\begin{aligned}
\nu(\widetilde{X}) &= E\left[D(\widetilde{X},\widetilde{E}(\widetilde{X}))\right] \\
&= E\left(\int_0^1\left(\widetilde{X}_\alpha - E(\widetilde{X}_\alpha)\right)^2 d\alpha\right) \\
&= \int_\Omega\int_0^1\left(\widetilde{X}_\alpha - E(\widetilde{X}_\alpha)\right)^2 d\alpha\, d\mathbf{P} \\
&= \int_0^1\int_\Omega\left(\widetilde{X}_\alpha - E(\widetilde{X}_\alpha)\right)^2 d\mathbf{P}\, d\alpha \\
&= \int_0^1 Var(\widetilde{X}_\alpha)\, d\alpha.
\end{aligned}$$

The situation with the usual random variable is a special case of the proposed procedure. By using the indicator function $I_{\{X\}}$ as the membership function for the frv, the variance of the crisp random variable X, i.e. $Var(X)$, coincides with $\nu(X)$, therefore, we have $\nu(X) = Var(X)$.

Now, if we define the scalar multiplication between frvs $\widetilde{X}$ and $\widetilde{Y}$ as follows

$$\langle\widetilde{X},\widetilde{Y}\rangle = \int_0^1 \widetilde{X}_\alpha\widetilde{Y}_\alpha\, d\alpha,$$

then, it is easy to conclude that $\nu(\widetilde{X}) = E\langle\widetilde{X},\widetilde{X}\rangle - \langle\widetilde{E}(\widetilde{X}),\widetilde{E}(\widetilde{X})\rangle$.

Proposition 3. *Let $\widetilde{\mathbf{X}} = (\widetilde{X}_1,\widetilde{X}_2,\ldots,\widetilde{X}_n)$ be a fuzzy random sample, and*

$$S_n^2(\widetilde{\mathbf{X}}) = \frac{1}{n-1}\sum_{i=1}^n D\left(\widetilde{X}_i,\bar{\widetilde{\mathbf{X}}}\right),$$

be the crisp variance value of the fuzzy sample $\widetilde{\mathbf{X}}$, where $\bar{\widetilde{\mathbf{X}}} = \frac{1}{n}\oplus_{i=1}^n\widetilde{X}_i$ is the fuzzy sample mean value. Then the following properties are held:

1. *$E[S_n^2(\widetilde{\mathbf{X}})] = \nu(\widetilde{X})$, i.e. $S_n^2(\widetilde{\mathbf{X}})$ is an unbiased estimator of the parameter $\nu(\widetilde{X})$ (population variance).*
2. *$\lim_{n\to\infty} S_n^2(\widetilde{\mathbf{X}}) = \nu(\widetilde{X})$.*
3. *$\nu(\lambda\otimes\widetilde{X}) = \lambda^2\nu(\widetilde{X})$, for any constant number $\lambda\in\mathbb{R}$.*
4. *$\nu(\lambda\oplus\widetilde{X}) = \nu(\widetilde{X})$, for any constant number $\lambda\in\mathbb{R}$.*

Definition 9. The covariance and correlation coefficient of frvs $\widetilde{X}$ and $\widetilde{Y}$ are defined as follows, respectively,

$$\begin{aligned}
Cov(\widetilde{X},\widetilde{Y}) &= E\langle\widetilde{X},\widetilde{Y}\rangle - \langle\widetilde{E}(\widetilde{X}),\widetilde{E}(\widetilde{Y})\rangle, \\
\rho(\widetilde{X},\widetilde{Y}) &= Cov\left(\frac{\widetilde{X}\ominus\widetilde{E}(\widetilde{X})}{\sqrt{\nu(\widetilde{X})}},\frac{\widetilde{Y}\ominus\widetilde{E}(\widetilde{Y})}{\sqrt{\nu(\widetilde{Y})}}\right).
\end{aligned}$$

We can easily show that

Table 1: Data set in Example 2

$\widetilde{x}_1 = (0.23, 0.04, 0.07)_T$	$\widetilde{x}_{11} = (0.41, 0.03, 0.08)_T$	$\widetilde{x}_{21} = (0.64, 0.11, 0.07)_T$
$\widetilde{x}_2 = (0.76, 0.05, 0.02)_T$	$\widetilde{x}_{12} = (0.86, 0.08, 0.04)_T$	$\widetilde{x}_{22} = (0.94, 0.09, 0.04)_T$
$\widetilde{x}_3 = (0.98, 0.12, 0.09)_T$	$\widetilde{x}_{13} = (1.02, 0.03, 0.10)_T$	$\widetilde{x}_{23} = (1.08, 0.10, 0.06)_T$
$\widetilde{x}_4 = (1.14, 0.06, 0.09)_T$	$\widetilde{x}_{14} = (1.23, 0.03, 0.14)_T$	$\widetilde{x}_{24} = (1.37, 0.08, 0.06)_T$
$\widetilde{x}_5 = (1.46, 0.10, 0.07)_T$	$\widetilde{x}_{15} = (1.53, 0.13, 0.15)_T$	$\widetilde{x}_{25} = (1.64, 0.02, 0.08)_T$
$\widetilde{x}_6 = (1.69, 0.05, 0.12)_T$	$\widetilde{x}_{16} = (1.78, 0.04, 0.06)_T$	$\widetilde{x}_{26} = (1.83, 0.09, 0.05)_T$
$\widetilde{x}_7 = (1.95, 0.05, 0.11)_T$	$\widetilde{x}_{17} = (1.99, 0.08, 0.09)_T$	$\widetilde{x}_{27} = (2.04, 0.11, 0.06)_T$
$\widetilde{x}_8 = (2.17, 0.03, 0.05)_T$	$\widetilde{x}_{18} = (2.25, 0.04, 0.04)_T$	$\widetilde{x}_{28} = (2.36, 0.05, 0.09)_T$
$\widetilde{x}_9 = (2.40, 0.08, 0.12)_T$	$\widetilde{x}_{19} = (2.45, 0.01, 0.08)_T$	$\widetilde{x}_{29} = (2.49, 0.13, 0.05)_T$
$\widetilde{x}_{10} = (2.51, 0.10, 0.14)_T$	$\widetilde{x}_{20} = (2.57, 0.07, 0.02)_T$	$\widetilde{x}_{30} = (2.61, 0.08, 0.06)_T$

1. $Cov(\widetilde{X}, \lambda) = 0$ for any constant number $\lambda \in \mathbb{R}$.
2. $Cov(\widetilde{X}, \widetilde{X}) = \nu(\widetilde{X})$.
3. $Cov(\widetilde{X}, \widetilde{Y}) = 0$ for independent frvs $\widetilde{X}$ and $\widetilde{Y}$.
4. Let $\lambda_1, \lambda_2, \mu_1, \mu_2 \in \mathbb{R}$, then

$$\begin{aligned} Cov\left(\lambda_1 \oplus (\lambda_2 \otimes \widetilde{X}), \mu_1 \oplus (\mu_2 \otimes \widetilde{Y})\right) &= \lambda_1\mu_1 Cov(\widetilde{X}, \widetilde{Y}), \\ \rho\left(\lambda_1 \oplus (\lambda_2 \otimes \widetilde{X}), \mu_1 \oplus (\mu_2 \otimes \widetilde{Y})\right) &= \frac{\lambda_1\mu_1}{|\lambda_1\mu_1|}\rho(\widetilde{X}, \widetilde{Y}). \end{aligned}$$

5. Fuzzy cumulative distribution function

In this section, we extend the concepts of Fuzzy Cumulative Distribution Function (F.C.D.F.) and Fuzzy Empirical Cumulative Distribution Function (F.E.C.D.F.) for an frv.

Definition 10. The F.C.D.F. of frv $\widetilde{X}$ at $x \in \mathbb{R}$ is defined as fuzzy set $\widetilde{F_{\widetilde{X}}(x)}$ with the following membership function

$$\widetilde{F_{\widetilde{X}}(x)}(y) = \sup\left\{\alpha \in [0,1] : \mathbf{P}\left(\widetilde{X}_\alpha \le x\right) = y\right\}, \quad y \in [0,1],$$

Definition 11. We say that F.C.D.F. $\widetilde{F_{\widetilde{X}}(x)}$ is continuous at $x \in \mathbb{R}$, if for every $\alpha \in [0,1]$, the function $(\widetilde{F_{\widetilde{X}}(x)})_\alpha^U$ is continuous at x (or equivalently, for every $\alpha \in [0,1]$, the crisp random variable $\widetilde{X}_\alpha$ is continuous).

Definition 12. Suppose that $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ is a fuzzy random sample. The F.E.C.D.F. of fuzzy random sample $\widetilde{X}_1$, $\widetilde{X}_2$, …, $\widetilde{X}_n$, at $x \in \mathbb{R}$ is defined to be the fuzzy set $\widetilde{F_n(x)}$ with the following membership function

$$\widetilde{F_n(x)}(y) = \sup\left\{\alpha \in [0,1] : \frac{\#(\widetilde{x}_{i\alpha} \le x)}{n} = y\right\}, \qquad y \in [0,1],$$

Example 2. Suppose that, based on a fuzzy random sample of size $n = 30$, we observe the triangular fuzzy numbers given in Table 1 (Hesamian and Chachi 2013; Viertl 2011). According to Definition 12, the F.E.C.D.F. of this fuzzy random sample is obtained and the 3-dimensional curve of its membership function is shown in Fig. 1, for every $x \in [0,3]$. Moreover, in order to make the 3-dimensional curve of the membership function in Fig. 1 more clear, the α-cut of this membership function is shown in Fig. 2, for $\alpha = 0.3$.

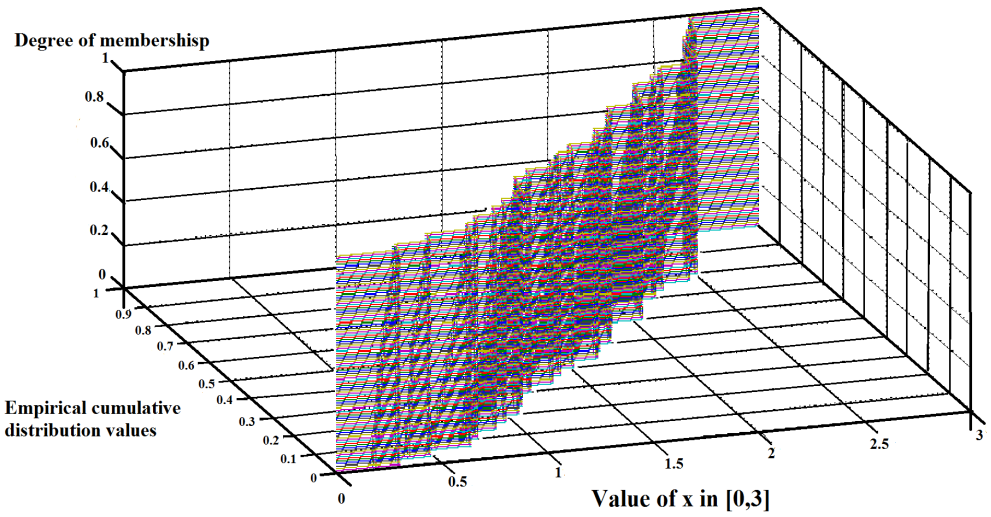

Figure 1: The plot of membership function of F.E.C.D.F. of the fuzzy observations in Table 1 for values of $x \in [0, 3]$

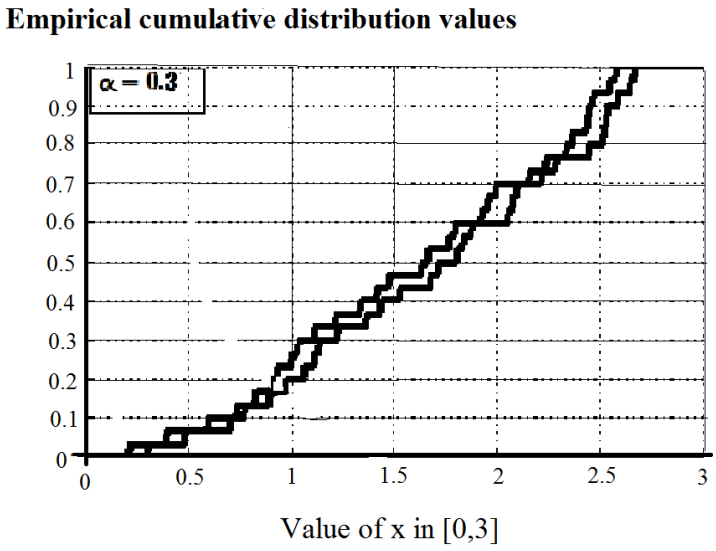

Figure 2: The α-cut of the 3-dimensional curve of the membership function shown in Figure 1 for $\alpha = 0.3$. For each value of $x \in [0, 3]$, the vertical line is the domain of the α-cut.

Example 3. Let $\widetilde{X} = \widetilde{\Theta} \oplus \Xi$, where Ξ is a (usual) normal random variable with mean 0 and variance σ^2, i.e. $\Xi \sim N(0, \sigma^2)$, and $\widetilde{\Theta}$ is a constant fuzzy set. For example, suppose $\widetilde{\Theta}$ is a LR-fuzzy number, i.e. $\widetilde{\Theta} = (\theta, l, r)_{LR}$ with known θ, l, r, and fixed functions L, and R. Therefore, $\widetilde{X} = (\Xi + \theta, l, r)_{LR}$ and for each ω, $\widetilde{X}(\omega) = (\Xi(\omega) + \theta, l, r)_{LR}$ is an observation of $\widetilde{X}$. Now, we have (see also, Example 1)

$$\widetilde{X}_\alpha = \begin{cases} \Xi + \theta - lL^{-1}(2\alpha) & \text{if} \quad \alpha \in [0, 0.5], \\ \Xi + \theta + rR^{-1}(2(1-\alpha)) & \text{if} \quad \alpha \in [0.5, 1]. \end{cases}$$

Since Ξ is a normal random variable, therefore, it is clear that $\widetilde{X}_\alpha$ is a normal random variable for each $\alpha \in [0, 1]$, i.e.

$$\widetilde{X}_\alpha \sim \begin{cases} N\left(\theta - lL^{-1}(2\alpha), \sigma^2\right) & \text{if} \quad \alpha \in [0, 0.5], \\ N\left(\theta + rR^{-1}(2(1-\alpha)), \sigma^2\right) & \text{if} \quad \alpha \in [0.5, 1]. \end{cases}$$

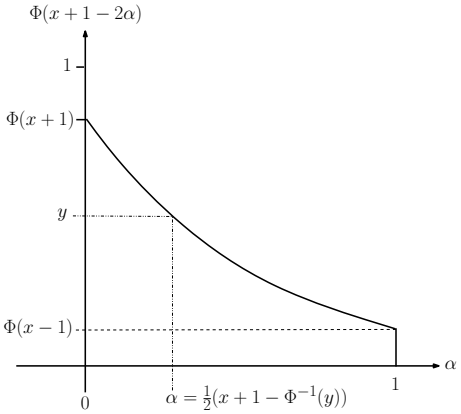

Figure 3: The graphical solution of the equation $\Phi(x+1-2\alpha) = y$ in Example 3

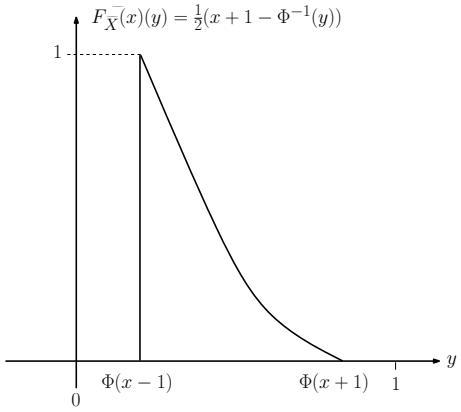

Figure 4: The membership function of the fuzzy cumulative distribution function $\widetilde{F_{\widetilde{X}}(x)}$ in Example 3

So, according to Definition 4, $\widetilde{X}$ is an frv. We can easily show that $\widetilde{E}(\widetilde{X}) = \widetilde{\Theta}$, and $\nu(\widetilde{X}) = \sigma^2$. Now, we are going to obtain the membership function of fuzzy set $\widetilde{F_{\widetilde{X}}(x)}$, i.e. the F.C.D.F. of the frv $\widetilde{X}$ at $x \in \mathbb{R}$. Its membership function is defined as

$$\widetilde{F_{\widetilde{X}}(x)}(y) = \sup\left\{\alpha \in [0,1] : \mathbf{P}\left(\widetilde{X}_\alpha \le x\right) = y\right\}, \qquad y \in [0,1],$$

in which

$$\mathbf{P}\left(\widetilde{X}_\alpha \le x\right) = \begin{cases} \Phi\left(\frac{x-\theta+lL^{-1}(2\alpha)}{\sigma}\right) & \text{if} \quad \alpha \in [0, 0.5], \\ \Phi\left(\frac{x-\theta-rR^{-1}(2(1-\alpha))}{\sigma}\right) & \text{if} \quad \alpha \in [0.5, 1], \end{cases}$$

where, Φ is the cumulative distribution function of standard normal random variable Z, i.e. if $Z \sim N(0,1)$ then $\mathbf{P}(Z \le z) = \Phi(z)$, $z \in \mathbb{R}$. We consider a simplification of the parameters $\widetilde{\Theta}$ and σ^2, therefore, we take $\widetilde{\Theta} = (0,1,1)_T$ and $\sigma = 1$ as special cases. Substituting these values in the above equations, we can easily obtain

$$\mathbf{P}\left(\widetilde{X}_\alpha \le x\right) = \Phi(x+1-2\alpha) \qquad \text{if} \quad \alpha \in [0,1].$$

Thus, the membership function of fuzzy set $\widetilde{F_{\widetilde{X}}(x)}$ is given as follows for any $y \in [0,1]$

$$\widetilde{F_{\widetilde{X}}(x)}(y) = \sup\{\alpha \in [0,1] : \Phi(x+1-2\alpha) = y\}.$$

Note that, the function $\Phi(x+1-2\alpha)$ is strictly decreasing with respect to $\alpha \in [0,1]$, for any fixed $x \in \mathbb{R}$ (see Fig. 3). Therefore, for any $y \in [0,1]$

$$\Phi(x+1-2\alpha) = y \Leftrightarrow \alpha = \frac{1}{2}(x+1-\Phi^{-1}(y)).$$

The above obtained α must be between 0 and 1, so

$$0 \leq \frac{1}{2}(x+1-\Phi^{-1}(y)) \leq 1 \Leftrightarrow 0 \leq \Phi(x-1) \leq y \leq \Phi(x+1) \leq 1.$$

Finally, according to the above equations, the membership function of $\widetilde{F_{\widetilde{X}}(x)}$ at $x \in \mathbb{R}$ is given by

$$\widetilde{F_{\widetilde{X}}(x)}(y) = \frac{1}{2}(x+1-\Phi^{-1}(y)), \quad y \in [\Phi(x-1), \Phi(x+1)] \subseteq [0,1].$$

The membership function $\widetilde{F_{\widetilde{X}}(x)}$ is depicted in Fig. 4.

This notion of frv is the definition of normality for frvs and $\widetilde{X} = \widetilde{\Theta} \oplus \Xi$, ($\Xi \sim N(0, \sigma^2)$, and $\widetilde{\Theta}$ is a constant fuzzy set) is called the normal (Gaussian) frv in the literature (Feng 2000; Puri and Ralescu 1985).

6. Conclusions

In this paper the concept of modeling fuzzy random variable is presented dealing with situations where the outcomes of a random experiment are modeled by fuzzy sets. In order to model the imprecise information of random experiments the notions of fuzzy cumulative distribution function and fuzzy empirical cumulative distribution function are considered (Möller *et al.* 2002). To achieve suitable statistical methods dealing with imprecise data and extend the usual approaches to imprecise environments several probabilistic definitions have been obtained in connection with this random element, some of them having immediate statistical implications. Fuzzy set theory seems to have suitable tools for modeling the imprecise information of random experiments and provides appropriate statistical methods based on them (see, for instance, Bandemer and Näther (1992); Chachi and Taheri (2011); Chachi, Taheri, and Viertl (2012); Colubi (2009); Colubi and Gil (2007); Colubi and González-Rodríguez (2007); Colubi, González-Rodríguez, Lubiano, and Montenegro (2006); Coppi, Gil, and Kiers (2006); Gebhardt, Gil, and Kruse (1998); González-Rodríguez, Montenegro, Colubi, and Gil (2006b); Hesamian and Chachi (2013); Kruse and Meyer (1987); Taheri and Hesamian (2011)). As a consequence, different approaches can also be provided for developing fuzzy statistical methods using the new concept of frv proposed in this paper. We end the paper with some general concluding remarks and open problems.

1- The new concept of frv proposed in this paper can be used to develop some kind of linear estimation theory. The attempt can be done to develop a certain kind of linear theory for frvs with respect to extended addition and scalar multiplication. However, the classical estimation problem in a linear regression model in view of fuzzy data can be a potential topic for further researches (see, for instance, Wünsche and Näther (2002)).

2- The new concept of frv can be studied successfully for limit theorems, and can be applied to asymptotic statistics with vague data (see, for instance, Klement *et al.* (1986)). Notice that there are lack of Central Limit Theorems (CLTs) for frvs which are directly applicable for inferential purposes (actually, there exist some CLTs for frvs according to which the normalized distance sample-population fuzzy mean converges in law to the norm of a Gaussian random element but with values often out of the cone) (Wu 2000; Krätschmer 2002a,b). Also, the essential large sample properties of the fuzzy empirical distribution function (like Cantelli-Glivenko's Lemma (Govindarajulu 2003)) can be stated and proved.

3- From a statistical point of view, fuzzy expected value and fuzzy median play important roles as central summary measures. The point estimation of these measures can be one of the first statistical analysis concerning frvs. Later, the initial hypothesis testing procedures can be studied, although they need some theoretical/practical constraints (see, for instance, Colubi (2009)).

4- The bootstrap techniques have empirically shown to be efficient and powerful in hypothesis testing. Furthermore, analogous two-sample tests and, in general, multi-sample tests for the equality of fuzzy expected values can also be obtained (see, for instance, González-Rodríguez *et al.* (2006b)).

5- As for the real/vectorial-valued case, hypotheses could either concern parameters/measures of the distribution of the frv(s) (see items 3 and 4 above) or concern the distribution itself (parametric/non-parametric). Therefore, testing hypothesis related to the distribution(s) of one-sample or multi-sample of observations can be considered. In this regard, non-parametric tests (like goodness-of-fit tests) can be developed to determine whether two underlying one dimensional distributions (or multi underlying one dimensional distributions) are the same or not. Here based on the definition of fuzzy empirical cumulative distribution functions, test statistics and test functions can be defined (see, for instance, Lin, Wu, and Watada (2010); Hesamian and Chachi (2013); Hryniewicz (2006); Taheri and Hesamian (2011))

6- It has been shown that the distribution of any real-valued random variable can be represented by means of a fuzzy set. The characterizing fuzzy sets correspond to the expected value of a certain frv based on a family of fuzzy-valued transformations of the original real-valued ones (González-Rodríguez *et al.* 2006a). They can be used for descriptive/exploratory or inferential purposes. This fact adds an extra-value to the fuzzy expected value and the preceding statistical procedures, that can be used in statistics about real distributions.

Acknowledgments

The author is very grateful to the Editor-in-Chief, Professor Matthias Templ, and the anonymous referees, for their constructive comments and suggestions that led to an improved version of this paper.

References

Bandemer H, Näther W (1992). *Fuzzy Data Analysis*. Kluwer Academic Publisher, Dordrecht.

Billingsley P (1995). *Probability and Measure*. John Wiley and Sons, New York. 3rd ed.

Blanco-Fernández A, Casals MR, Colubi A, Corral N, García-Bárzana M, Gil MA, González-Rodríguez G, López MT, Lubiano MA, Montenegro M, Ramos-Guajardo AB, De La Rosa De Sá S, Sinova B (2013). "Random Fuzzy Sets: a Mathematical Tool to Develop Statistical Fuzzy Data Analysis." *Iranian Journal of Fuzzy Systems*, **10**, 1–28.

Chachi J, Taheri SM (2011). "Fuzzy Confidence Intervals for Mean of Gaussian Fuzzy Random Variables." *Expert Systems with Applications*, **38**, 5240–5244.

Chachi J, Taheri SM, Viertl R (2012). "Testing Statistical Hypotheses Based on Fuzzy Confidence Intervals." *Austrian Journal of Statistics*, **41**, 267–286.

Colubi A (2009). "Statistical Inference About the Means of Fuzzy Random Variables: Applications to the Analysis of Fuzzy- and Real-Valued Data." *Fuzzy Sets and Systems*, **160**, 344–356.

Colubi A, Domínguez-Menchero JS, López-Díaz M, Ralescu DA (2001). "On the Formalization of Fuzzy Random Variables." *Informtion Sciences*, **133**, 3–6.

Colubi A, Fernández-García C, Gil MA (2002). "Simulation of Random Fuzzy Variables: An Empirical Approach to Statistical/Probabilistic Studies With Fuzzy Experimental Data." *IEEE Transactions on Fuzzy Systems*, **10**(3), 384–390.

Colubi A, González-Rodríguez G (2007). "Triangular Fuzzification of Random Variables and Power of Distribution Tests: Empirical Discussion." *Computational Statistics and Data Analysis*, **51**, 4742–4750.

Colubi A, González-Rodríguez G, Lubiano MA, Montenegro M (2006). "Exploratory Analysis of Random Variables Based on Fuzzification." *In: Soft Methods for Integrated Uncertainty Modelling*, **51**, 95–102.

Colubi A CRDP, Gil MA (2007). "Statistics with Fuzzy Random Variables." *METRON-International Journal of Statistics*, **LXV**, 277–303.

Coppi R, Gil MA, Kiers HAL (2006). "The Fuzzy Approach to Statistical Analysis." *Computational Statistics and Data Analysis*, **51**, 1–14.

Couso I, Sánchez L (2008). "Higher Order Models for Fuzzy Random Variables." *Fuzzy Sets and Systems*, **159**, 237–258.

Diamond P, Kloeden P (1994). *Metric Spaces of Fuzzy Sets*. World Scientific, Singapore.

Feng X, Liu YK (2006). "Measurability Criteria for Fuzzy Random Vectors." *Fuzzy Optimization and Decision Making*, **5**, 245–253.

Feng Y (2000). "Gaussian Fuzzy Random Variables." *Fuzzy Sets and Systems*, **111**, 325–330.

Gebhardt J, Gil MA, Kruse R (1998). "Concepts of Fuzzy-Valued Statistics." *in: Fuzzy Sets in Decision Analysis, Operations Research and Statistics*, **5**, 311–347.

Gil MA (2001). "Fuzzy Random Variables." *Information Sciences*, **133**, 1–2.

Gil MA, López-Díaz M, Ralescu DA (2006). "Overview on the Development of Fuzzy Random Variables." *Fuzzy Sets and Systems*, **157**, 2546–2557.

González-Rodríguez G, Colubi A, Gil MA (2006a). "A Fuzzy Representation of Random Variables: An Operational Tool in Exploratory Analysis and Hypothesis Testing." *Computational Statistics and Data Analysis*, **51**, 163–176.

González-Rodríguez G, Colubi A, Gil MA, Coppi R (2006). "A Method to Simulate Fuzzy Random Variables." *Advances in Soft Computing*, **6**, 103–110.

González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006b). "Bootstrap Techniques and Fuzzy Random Variables: Synergy in Hypothesis Testing with Fuzzy Data." *Fuzzy Sets and Systems*, **157**, 2608–2613.

Govindarajulu Z (2003). *Non-Parametric Inference*. Hackensack, World Scientific.

Hesamian G, Chachi J (2013). "Two-Sample Kolmogorov-Smirnov Fuzzy Test for Fuzzy Random Variables." *Statistical Papers*, **56**, 61–82.

Hryniewicz O (2006). "Goodman-Kruskal Measure of Dependence for Fuzzy Ordered Categorical Data." *Computational Statistics and Data Analysis*, **51**, 323–334.

Klement EP, Puri ML, Ralescu DA (1986). "Limit Theorems for Fuzzy Random Variables." *Proc. Roy. Soc. London A*, **407**, 171–182.

Krätschmer V (2001). "A Unified Approach to Fuzzy Random Variables." *Fuzzy Sets and Systems*, **123**, 1–9.

Krätschmer V (2002a). "Limit Theorems for Fuzzy-Random Variables." *Fuzzy Sets and Systems*, **126**, 253–263.

Krätschmer V (2002b). "Some Complete Metrics on Spaces of Fuzzy Subsets." *Fuzzy Sets and Systems*, **130**, 357–365.

Kruse R, Meyer KD (1987). *Statistics with Vague Data*. Reidel Publishing Company, Dordrecht, Netherlands.

Kwakernaak H (1978). "Fuzzy Random Variables, Part I: Definitions and Theorems." *Information Sciences*, **19**, 1–15.

Kwakernaak H (1979). "Fuzzy Random Variables, Part II: Algorithms and Examples for the Discrete Case." *Information Sciences*, **17**, 253–278.

Lin P, Wu B, Watada J (2010). "Kolmogorov-Smirnov Two Sample Test with Continuous Fuzzy Data." *Advances in Intelligent and Soft Computing*, **68**, 175–186.

Liu B (2002). *Theory and Practice of Uncertain Programminge*. Physica–Verlag, Heidelberg.

Liu B (2016). *Uncertainty Theory*. Springer-Verlag, Berlin. 5th ed., URL http://orsc.edu.cn/liu/ut.pdf.

Liu B, Liu YK (2002). "Expected Value of Fuzzy Variable and Fuzzy Expected Value Models." *IEEE Transactions on Fuzzy Systems*, **10**, 445–450.

Liu YK, Liu B (2003). "Fuzzy Random Variables: A Scalar Expected Value Operator." *Fuzzy Optimization and Decision Making*, **2**, 143–160.

Möller B, Graf W, M B, Sickert JU (2002). "Fuzzy Randomness - Towards a New Modeling of Uncertainty." *Fifth World Congress on Computational Mechanics, July 7-12, 2002, Vienna, Austria, Eds.: H.A. Mang, F.G. Rammerstorfer, J. Eberhardsteiner*, **WCCM**(V), 1–10.

Näther W (2001). "Random Fuzzy Variables of Second Order and Applications to Statistical Inference." *Information Sciences*, **133**, 69–88.

Peng J, Liu B (2004). "Some Properties of Optimistic and Pessimistic Values of Fuzzy." *IEEE International Conference on Fuzzy Systems*, **2**, 745–750.

Puri ML, Ralescu DA (1985). "The Concept of Normality for Fuzzy Random Variables." *The Annals of Probability*, **13**, 1373–1379.

Puri ML, Ralescu DA (1986). "Fuzzy Random Variables." *Journal of Mathematical Analysis and Applications*, **114**, 409–422.

Shapiro AF (2009). "Fuzzy Random Variables." *Insurance: Mathematics and Economics*, **44**, 307–314.

Taheri SM, Hesamian G (2011). "Goodman-Kruskal Measure of Association for Fuzzy-Categorized Variables." *Kybernetika*, **47**, 110–122.

Viertl R (2011). *Statistical Methods for Fuzzy Data*. John Wiley and Sons, Chichester.

Wu HC (1999). "Probability Density Functions of Fuzzy Random Variables." *Fuzzy Sets and Systems*, **105**, 139–158.

Wu HC (2000). "The Laws of Large Numbers for Fuzzy Random Variables." *Fuzzy Sets and Systems*, **116**, 245–262.

Wünsche A, Näther W (2002). "Least-Squares Fuzzy Regression with Fuzzy Random Variables." *Fuzzy Sets and Systems*, **130**, 43–50.

Zadeh LA (1965). "Fuzzy sets." *Informtion and Control*, **8**, 338–353.

Zadeh LA (1995). "Discussion: Probability Theory and Fuzzy Logic Are Complementary Rather Than Competitive." *Technometrics*, **37**, 271–276.

Zimmermann HJ (2001). *Fuzzy Set Theory and Its Applications*. Kluwer Nihoff, Boston. 4th ed.

Affiliation:

Jalal Chachi
Department of Mathematics, Statistics and Computer Sciences
Semnan University
Semnan 35195-363, Iran
Telephone: +98/233/336-6205
Fax: +98/233/335-4059
E-mail: jchachi@semnan.ac.ir

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 47 *Submitted:* 2016-07-16
February 2018 *Accepted:* 2017-02-27

Austrian Journal of Statistics
February 2018, Volume 47, 69–89.
http://www.ajs.or.at/
doi:10.17713/ajs.v47i2.580

The Generalized Odd Gamma-G Family of Distributions: Properties and Applications

B. Hosseini
Persian Gulf Univ.

M. Afshari*
Persian Gulf Univ.

M. Alizadeh
Persian Gulf Univ.

Abstract

Recently, new continuous distributions have been proposed to apply in statistical analysis in a way that each one solves a particular part of the classical distribution problems. In this paper, the Generalized Odd Gamma-G distribution is introduced . In particular, G has been considered as the Uniform distribution and some statistical properties such as quantile function, asymptotics, moments, entropy and order statistics have been calculated. We survey the theoretical outcomes with numerical computation by using *R* software.The fitness capability of this model has been investigated by fitting this model and others based on real data sets. The maximum likelihood estimators are assessed with simulated real data from proposed model. We present the simulation in order to test validity of maximum likelihood estimators .

Keywords: generalized odd gamma-G, maximum likelihood, moment, entropy.

1. Introduction

The classic statistical distributions which have essential limitations and problems in data modeling, has led statistical researcher to make of the new flexible distributions. The new distributions are often made through the classic distributions and give the required flexibility to the classic distributions. The most important distributions among them are Marshall-Olkin generated (MO-G) by Marshall and Olkin (1997), Kumaraswamy-G (Kw-G) by Cordeiro and de Castro (2011), McDonald-G (Mc-G) by Alexander, Cordeiro, Ortega, and Sarabia (2012), Weibull-G by Bourguignon, Silva, and Cordeiro (2014), exponentiated half-logistic by Cordeiro, Alizadeh, and Ortega (2014a), transformer (T-X) by Alzaatreh, Lee, and Famoye (2013), Logistic-X by Tahir, Cordeiro, Alzaatreh, Mansoor, and Zubair (2016) and Lomax generator by Cordeiro, Ortega, Popović, and Pescim (2014b), Kumaraswamy Marshal-Olkin family by Alizadeh, Tahir, Cordeiro, Mansoor, Zubair, and Hamedani (2015b), Beta Marshal-OLkin family by Alizadeh, Cordeiro, De Brito, and Demétrio (2015a), type I half-logistic family by Cordeiro, Alizadeh, and Diniz Marinho (2016).

Based on T-X idea by Alzaatreh *et al.* (2013), by the following definition, the Generalized

Odd Gamma-G distribution (GOGa-G) would be made

$$F(x;\alpha,\beta,\boldsymbol{\xi}) = \int_0^{\frac{G(x;\boldsymbol{\xi})^\beta}{1-G(x;\boldsymbol{\xi})^\beta}} \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}dt = \frac{\gamma\left(\alpha, \frac{G(x;\boldsymbol{\xi})^\beta}{1-G(x;\boldsymbol{\xi})^\beta}\right)}{\Gamma(\alpha)}. \tag{1}$$

where $\alpha,\beta > 0$ are two additional shape parameters, $\boldsymbol{\xi}$ is the parameter for baseline G and $\gamma(\alpha,x) = \int_0^x t^{\alpha-1}e^{-t}dt$ denote the incomplete gamma function.

In this case, the probabilty density function (pdf) of the GOGa-G distribution will be as follows:

$$f(x;\alpha,\beta,\boldsymbol{\xi}) = \frac{\beta g(x;\boldsymbol{\xi})G(x;\boldsymbol{\xi})^{\alpha\beta-1}}{\Gamma(\alpha)\left[1-G(x;\boldsymbol{\xi})^\beta\right]^{\alpha+1}} e^{\frac{-G(x;\boldsymbol{\xi})^\beta}{1-G(x;\boldsymbol{\xi})^\beta}}. \tag{2}$$

where $g(x;\boldsymbol{\xi})$ is the pdf of the $G(x;\boldsymbol{\xi})$ distribution. From now on, the random variable X with pdf (2) is shown with $X \sim$ GOGa-G$(\alpha,\beta,\boldsymbol{\xi})$. According to (1) and (2) hrt of X is as follows:

$$\tau(x;\alpha,\beta,\boldsymbol{\xi}) = \frac{\beta g(x;\boldsymbol{\xi})G(x;\boldsymbol{\xi})^{\alpha\beta-1} e^{\frac{-G(x;\boldsymbol{\xi})^\beta}{1-G(x;\boldsymbol{\xi})^\beta}}}{\left[1-G(x;\boldsymbol{\xi})^\beta\right]^{\alpha+1}\left[\Gamma(\alpha)-\gamma\left(\alpha,\frac{G(x;\boldsymbol{\xi})^\beta}{1-G(x;\boldsymbol{\xi})^\beta}\right)\right]}. \tag{3}$$

An interpretation of the GOGa-G family (1) can be given as follows:
Let T be a random variable describing a stochastic system by the cdf $G(x)^\beta$ (for $\beta > 0$). If the random variable X represents the odds ratio, the risk that the system following the lifetime T will be not working at time x is given by $\frac{G(x)^\beta}{1-G(x)^\beta}$. If we are interested in modeling the randomness of the odds ratio by the Gamma pdf $r(t) = \frac{1}{\Gamma(\alpha)}\, t^{\alpha-1}\, \mathrm{e}^{-\mathrm{t}}$ (for $t > 0$), the cdf of X is given by

$$Pr(X \le x) = R\left(\frac{G(x)^\beta}{1-G(x)^\beta}\right).$$

which is exactly the cdf (1) of the new family.

Theorem 1 provides some relations of the GOGa family with other distributions.

Theorem 1. *Let* $X \sim$*GOGa-*$G(\alpha,\beta,\boldsymbol{\xi})$ *and* $Y = \frac{G(X;\boldsymbol{\xi})^\beta}{1-G(X;\boldsymbol{\xi})^\beta}$, then $Y \sim \Gamma(\alpha,1)$.

Proof: It is clear.

The basic motivations for using the GOGa family in practice are the following:
(i) to make the kurtosis more flexible compared to the baseline model; (ii) to produce a skewness for symmetrical distributions; (iii)to construct heavy-tailed distributions that are not longer-tailed for modeling real data; (iv)to generate distributions with symmetric, left-skewed, right-skewed and reversed-J shaped; (v) to define special models with all types of the hrf; (vi) to provide consistently better fits than other generated models under the same baseline distribution.

In the following, the paper would be like this: In Section 2, a special distribution is introduced by selecting G. In Section 3, the features of the GOGa- model will be assessed using quantile function, asymptotics, functions expansion, quantile power series, moments, entropy and order statistics. In Section 4, MLE calculation method and in Section 5, estimability of the model additional parameters will be discussed using simulation. In Section 6, the proposed model is fitted based on two real data sets and compared to other famous models.

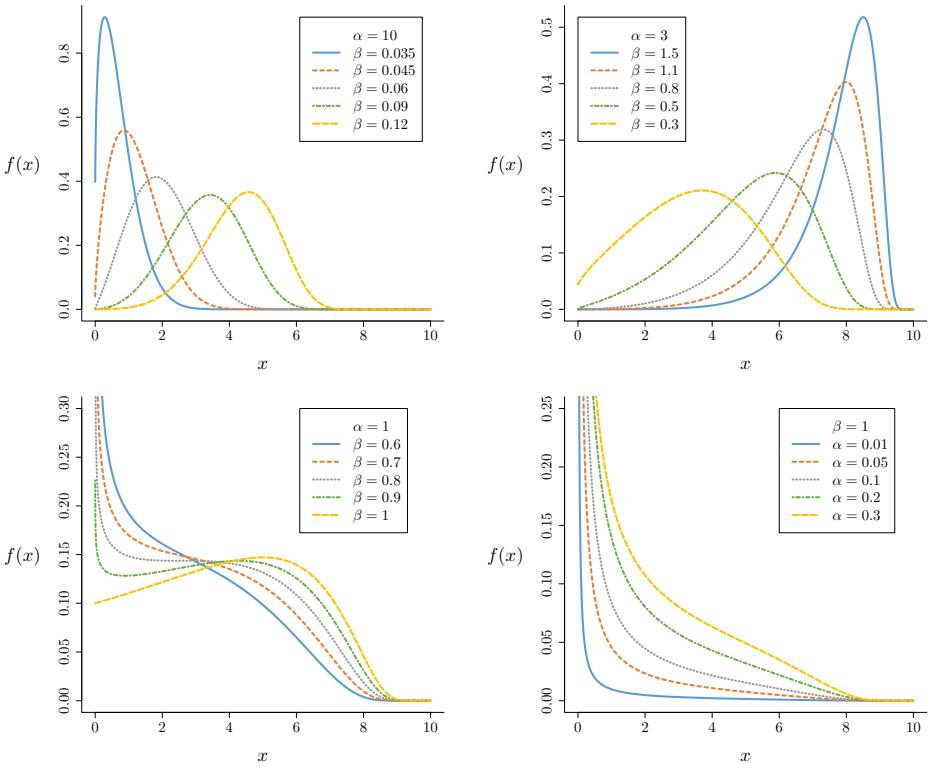

Figure 1: The sample curves of density function of GOG-U$(\alpha, \beta, 0, 10)$.

2. Special models

2.1. The generalized odd gamma-uniform (GOGa-U)

Different distributions family can be reached by selecting different Gs in equation (2). Torabi and Hedesh (2012), G has been considered as uniform distribution. In this case, by letting $\boldsymbol{\xi} = (a, b)$ equation (2) will changed as follows:

$$f(x; \alpha, \beta, a, b) = \frac{\beta(b-a)^{\beta}(x-a)^{\alpha\beta-1} e^{\frac{-(x-a)^{\beta}}{(b-a)^{\beta}-(x-a)^{\beta}}}}{\Gamma(\alpha)\left[(b-a)^{\beta}-(x-a)^{\beta}\right]^{\alpha+1}}, \quad a \leq x \leq b\,. \tag{4}$$

where $\alpha, \beta > 0, a, b \in \mathbb{R}$ and $a < b$. If X be a random variable with density function (4), then it will be displayed by GOGa-U(α, β, a, b). In Figure 1 some density and hazard functions for GOGa-U have been drawn.

One can see in the curves of Figure 1 that the different states of density function including symmetric density function (approximately), mild and high skewed (right and left) and bi-modal (in the right bottom curve, one mode is in point zero) have been produced. In Figure 2 one can see some curves of the hazard function of the GOGa-U distribution for some para-mentreters. According to Figure 2 you see that the U shape hazard functions are producible by GOGa-U.

2.2. The generalized odd gamma-Weibull (GOGa-W)

In GOGa-G, suppose G is as follows Weibull distribution function:

$$G(x; \lambda, k) = 1 - e^{-\left(\frac{x}{\lambda}\right)^{k}}, \quad x \geq 0\,.$$

In this case, by letting $\boldsymbol{\xi} = (\lambda, k)$ equation (2) will be changed as follows

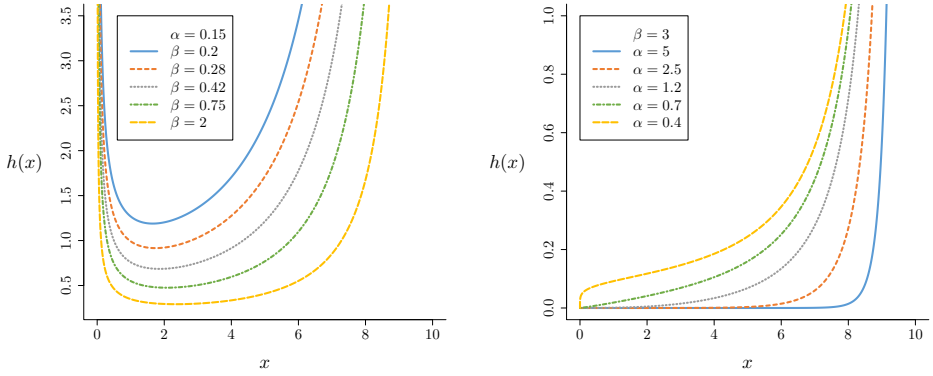

Figure 2: The sample curves of hazard function of GOGa-U$(\alpha, \beta, 0, 10)$.

$$f(x;\alpha,\beta,\lambda,k) = \frac{\beta k \left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k} \left[1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right]^{\alpha\beta-1}}{\lambda\Gamma(\alpha)\left\{1 - \left[1 - e^{-\left(\frac{x}{\lambda}\right)^k}\right]^{\beta}\right\}^{\alpha+1}} e^{\frac{-\left[1-e^{-\left(\frac{x}{\lambda}\right)^k}\right]^{\beta}}{1-\left[1-e^{-\left(\frac{x}{\lambda}\right)^k}\right]^{\beta}}}, \quad x \geq 0\,. \tag{5}$$

where $\alpha, \beta, \lambda, k > 0$. If X be a random variable with density function (5), then it will be displayed by GOGa-W$(\alpha, \beta, \lambda, k)$. In Figure 3 some pdfs for GOGa-W have been drawn.

3. Main features

3.1. Quantile function

By considering (1) quantile function (qf) X is obtained as follows: If $V \sim \Gamma(\alpha, 1)$ then the solution of nonlinear equation $x_v = Q_G\left[\left(\frac{V}{1+V}\right)^{\frac{1}{\beta}}\right]$ has cdf (1).

3.2. Asymptotics

Proposition 1. *Let $a = \inf\{x|f(x) > 0\}$, then the asymptotic of equation (1), (2) and (3) when $x \to a$ are given by*

$$F(x) \sim \frac{G(x)^{\alpha\beta}}{\alpha\Gamma(\alpha)}$$

$$f(x) \sim \frac{\beta g(x)\, G(x)^{\alpha\beta-1}}{\Gamma(\alpha)}$$

$$\tau(x) \sim \frac{\beta g(x)\, G(x)^{\alpha\beta-1}}{\Gamma(\alpha)}$$

Proposition 2. *The asymptotic of equation (1), (2) and (3) when $x \to +\infty$ are given by*

$$\overline{F}(x) \sim 1 - \frac{\gamma\left(\alpha, \frac{1}{\beta\overline{G}(x)}\right)}{\Gamma(\alpha)}$$

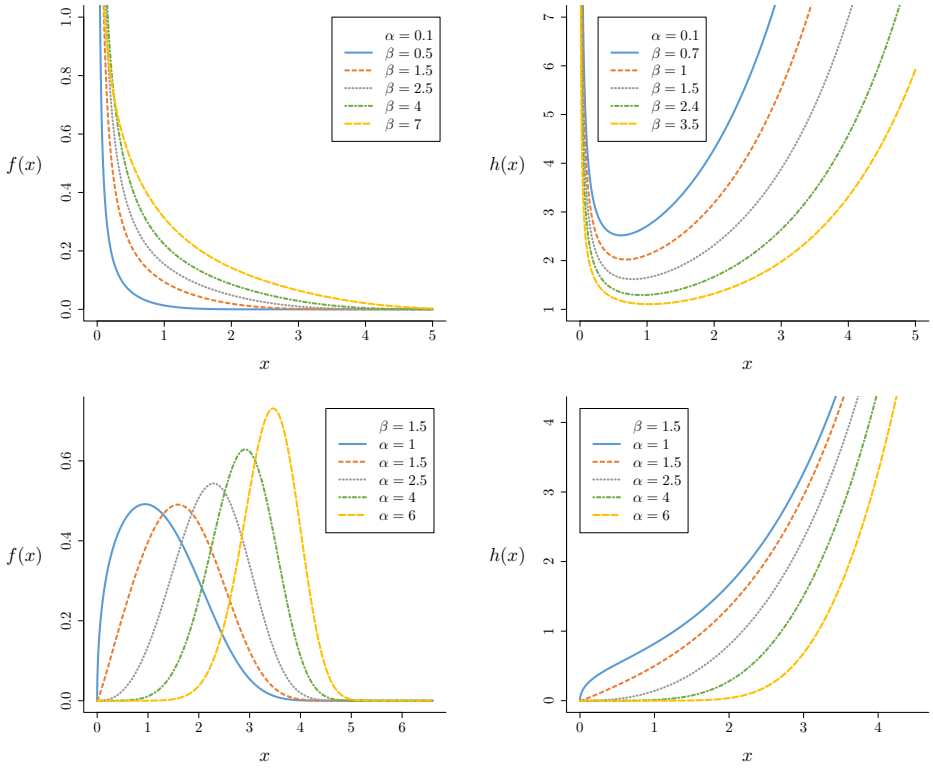

Figure 3: The sample curves of density and hazard function of GOGa-W$(\alpha, \beta, 1, 1.5)$.

$$f(x) \sim \frac{g(x)}{\beta^{\alpha}\Gamma(\alpha)\overline{G}(x)^{\alpha+1}} e^{\frac{-1}{\beta\overline{G}(x)}}$$

$$\tau(x) \sim \frac{g(x)e^{\frac{-1}{\beta\overline{G}(x)}}}{\beta^{\alpha}\left[\Gamma(\alpha) - \gamma\left(\alpha, \frac{1}{\beta\overline{G}(x)}\right)\right]\overline{G}(x)^{\alpha+1}}$$

3.3. Expansion for Pdf and Cdf and hrf

Using generalized binomial and taylor expansion one can obtain

$$\begin{aligned} f(x) &= \frac{\beta g(x) G(x)^{\alpha\beta-1}}{\Gamma(\alpha)\left[1-G(x)^{\beta}\right]^{\alpha+1}} \sum_{i=0}^{\infty} \frac{(-1)^i \left(\frac{G(x)^{\beta}}{1-G(x)^{\beta}}\right)^i}{i!} \\ &= \frac{\beta g(x)}{\Gamma(\alpha)} \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \frac{(-1)^i}{i!} \binom{-\alpha-i-1}{j} G(x)^{\beta(\alpha+i+j)-1} \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} w_{i,j} h_{\beta(\alpha+i+j)}(x)\,. \end{aligned} \tag{6}$$

where

$$w_{i,j} = \frac{(-1)^i \binom{-\alpha-i-1}{j}}{i!\,[\alpha+i+j]\,\Gamma(\alpha)}\,.$$

and $h_{\beta}(x) = \beta g(x)\, G(x)^{\beta-1}$, denote the pdf of exp-G distribution with power parameter β.

By integrating from equation (6) with respect to x, we have

$$F(x) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} w_{i,j} H_{\beta(\alpha+i+j)}(x)\,. \tag{7}$$

where $H_\beta(x) = G(x)^\beta$.
By considering $G(x) = 1 - [1 - G(x)]$ and binomial expansion we have:

$$\begin{aligned} G(x)^{\beta(\alpha+i+j)} &= \sum_{l=0}^{\infty}(-1)^l \binom{\beta(\alpha+i+j)}{l} [1-G(x)]^l \\ &= \sum_{l=0}^{\infty}\sum_{k=0}^{l}(-1)^{l+k} \binom{\beta(\alpha+i+j)}{l}\binom{l}{k} G(x)^k \\ &= \sum_{k=0}^{\infty}\sum_{l=k}^{\infty}(-1)^{l+k} \binom{\beta(\alpha+i+j)}{l}\binom{l}{k} G(x)^k \end{aligned}$$

In this case, regarding to (7) cdf extends as follows

$$F(x) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=k}^{\infty} w_{i,j}(-1)^{l+k} \binom{\beta(\alpha+i+j)}{l}\binom{l}{k} G(x)^k\,.$$

then

$$F(x) = \sum_{k=0}^{\infty} b_k G(x)^k \tag{8}$$

where

$$b_k = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=k}^{\infty} w_{i,j}(-1)^{l+k} \binom{\beta(\alpha+i+j)}{l}\binom{l}{k} \tag{9}$$

and finally regarding to (8) for cdf we also have

$$f(x) = \sum_{k=0}^{\infty} b_{k+1}\, h_{k+1}(x)$$

3.4. Moments

The rth ordinary moment of X is given by

$$\mu'_r = E(X^r) = \int_{-\infty}^{+\infty} x^r f(x)dx\,.$$

Using (1), we obtain the following:

$$\mu'_r = \sum_{k=0}^{\infty} b_{k+1} E(Y_{k+1}^r)\,. \tag{10}$$

Hereafter, Y_{k+1} denotes the Exp-G distribution with power parameter $(k+1)$. Setting $r = 1$ in (10), We have the mean f X. The last integration can be computed numericaly for most parent distributions. The skewness and kurtosis measures can be calculated from the ordinary moments using well-known relationships. The nth central moment of X, say M_n, follows as

$$M_n = E(X-\mu)^n = \sum_{h=0}^{n}(-1)^h \binom{n}{h} (\mu'_1)^n \mu'_{n-h}\,.$$

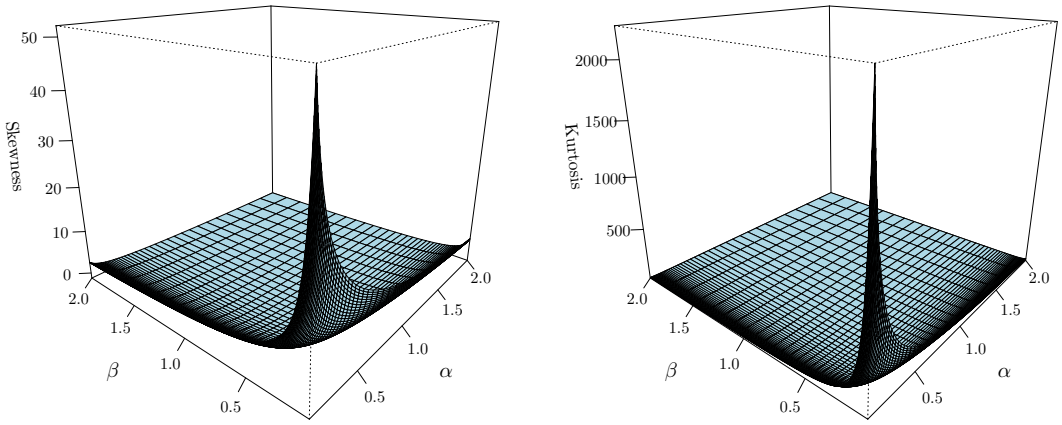

Figure 4: Skewness and kurtosis for GOGa-U.

The cumulants (κ_n) of X follow recursively from

$$\kappa_n = \mu'_n - \sum_{r=0}^{n-1} \binom{n-1}{r-1} \kappa_r \mu'_{n-r} .$$

where $\kappa_1 = \mu'_1$, $\kappa_2 = \mu'_2 - \mu'^2_1$, $\kappa_3 = \mu'_3 - 3\mu'_2\mu'_1 + \mu'^3_1$, etc. The skewness and kurtosis measures also can be calculates from the ordinary moment using well-known relationships. The moment generating function (mgf) of X, say $M_X(t) = E\left(e^{tX}\right)$, is given by

$$M_X(t) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu'_r = \sum_{k,r=0}^{\infty} \frac{t^r b_{k+1}}{r!} E(Y_{k+1}^r)$$

3.5. Incomplete moments

The main application of the first incomplete moment refers to Bonferroniand Lorenz curves. These curves are very useful in economics, reliability, demography, insurance and medicine. The answers to many important questions in economics require more than just knowing the mean of the distribution, its shape as well. This is obvious both in the study of econometrics and in areas as well. The sth incomplete moments, say $\varphi_s(t)$, is given by

$$\varphi_s(t) = \int_{-\infty}^{t} x^s f(x) dx$$

Using equation (8), we obtain

$$\varphi_s(t) = \sum_{k=0}^{\infty} b_{k+1} \int_{-\infty}^{t} x^s h_{k+1}(x) dx . \tag{11}$$

The first incomplete of the GOGa-G family, $\varphi_1(t)$, can be obtained by setting $s = 1$ in (11). Another application of the first incomplete moment is related to meanresidual life and mean waiting tie given by $m_1(t) = [1 - \varphi_1(t)]/R(t) - t$ and $M_1(t) = t - [\varphi_1(t)/F(t)]$, respectively.

3.6. Entropy

Entropy is an index for measuring variation or uncertainty of a random variable. The measure of entropy, Rrnyi (1961), is defined as follows

$$I_R(\gamma) = \frac{1}{1-\gamma} \log\left(\int_0^{\infty} f^{\gamma}(x) dx\right) .$$

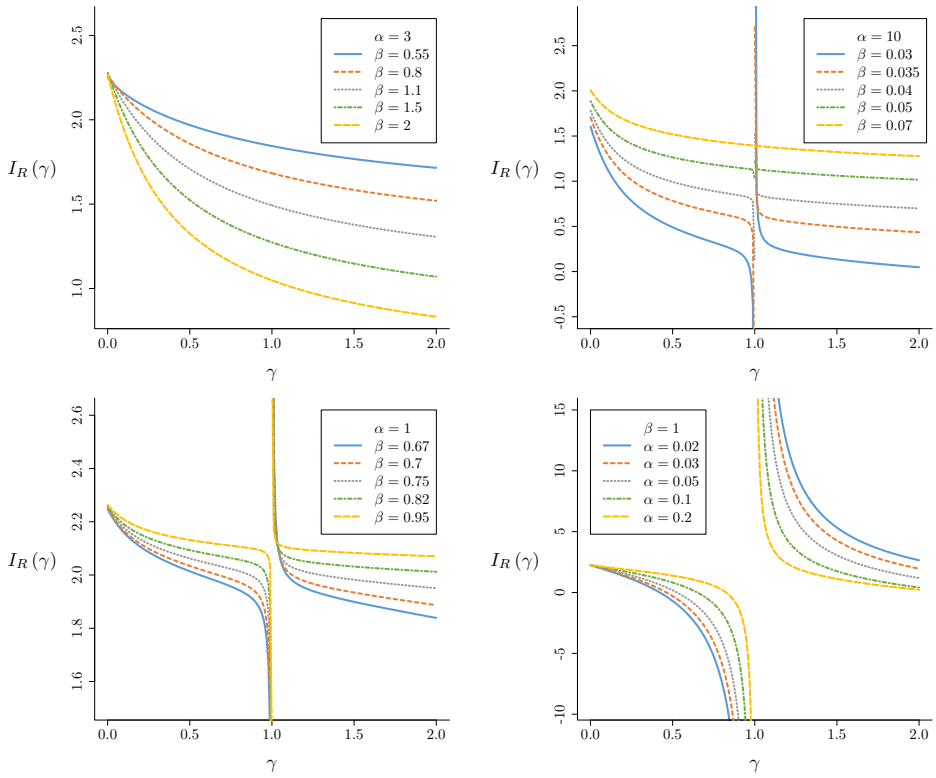

Figure 5: Curves of the GOGa-U entropy function for some parameter values.

for $\gamma > 0$ and $\gamma \neq 1$. The Shannon entropy measure is also defined by $E\{-\log[f(x)]\}$ that is a special state of the Rényi entropy when $\gamma \uparrow 1$.

$$
\begin{aligned}
f(x)^{\gamma} &= \left[\frac{\beta g G^{\alpha\beta-1} e^{\frac{-G^{\beta}}{1-G^{\beta}}}}{\Gamma(\alpha)[1-G^{\beta}]^{\alpha+1}}\right]^{\gamma} \\
&= \frac{\beta^{\gamma} g^{\gamma} G^{\gamma(\alpha\beta-1)} e^{\frac{-G^{\beta}}{1-G^{\beta}}}}{[\Gamma(\alpha)]^{\gamma}[1-G^{\beta}]^{\gamma(\alpha+1)}} \\
&= \frac{\beta^{\gamma}}{[\Gamma(\alpha)]^{\gamma}} \sum_{i=0}^{\infty} \frac{(-1)^{i}}{i!} \gamma^{i} \frac{G^{\gamma(\alpha\beta-1)+\beta i} e^{\frac{-G^{\beta}}{1-G^{\beta}}}}{[1-G^{\beta}]^{\gamma(\alpha+1)+i}} g^{\gamma} \\
&= \frac{\beta^{\gamma}}{[\Gamma(\alpha)]^{\gamma}} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^{i+j}}{i!} \binom{-\gamma(\alpha+1)-i}{j} \gamma^{i} g^{\gamma} G^{\gamma(\alpha\beta-1)+\beta(i+j)}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow I_R(\gamma) &= \frac{1}{1-\gamma} \log\left[\int_{-\infty}^{+\infty} f^{\gamma}(x) dx\right] \\
&= \frac{\gamma}{1-\gamma} \log\left[\frac{\beta}{\Gamma(\alpha)}\right] + \frac{1}{1-\gamma} \log\left[\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} v_{i,j} I(\gamma, \alpha, \beta, i, j)\right].
\end{aligned}
$$

where $v_{i,j} = \frac{(-1)^{i+j}\gamma^{i}}{i!} \binom{-\gamma(\alpha+1)-i}{j}$ and $I(\gamma, \alpha, \beta, i, j) = \int_{-\infty}^{+\infty} g(x)^{\gamma} G(x)^{\gamma(\alpha\beta-1)+\beta(i+j)} dx$.

In Figure 5 one can see some curves of the entropy function of the GOGa-U distribution for some parameters.

3.7. Order statistics

Order statistics make their appearance in many areas of statistical theory and practice. Suppose $X_1, \ldots, X_n$ is a random sample from any GOGa-G distribution. Let $X_{i:n}$ denote the ith order statistic. The pdf of $X_{i:n}$ can be expressed as

$$f_{i:n}(x) = c\, f(x)\, F^{i-1}(x)\, \{1 - F(x)\}^{n-i} = c \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} f(x)\, F(x)^{j+i-1}\,.$$

where $c = \frac{1}{B(i,n-i+1)}$.

We use the result 0.314 of Gradshteyn and Ryzhik (2000) for a power series raised to a positive integer n (for $n \geq 1$)

$$\left(\sum_{i=0}^{\infty} a_i\, u^i\right)^n = \sum_{i=0}^{\infty} c_{n,i}\, u^i\,. \tag{12}$$

where the coefficients $c_{n,i}$ (for $i = 1, 2, \ldots$) are determined from the recurrence equation (with $c_{n,0} = a_0^n$)

$$c_{n,i} = (i\, a_0)^{-1} \sum_{m=1}^{i} [m(n+1) - i]\, a_m\, c_{n,i-m}\,. \tag{13}$$

By using equations (9), (12), (13), We can demonstrate that the density function of the ith order statistic of any GOGa-G distribution can be expressed as follows:

$$f_{i:n}(x) = \sum_{r,k=0}^{\infty} m_{r,k}\, h_{r+k+1}(x)\,. \tag{14}$$

where $h_{r+k+1}(x)$ denotes the exp-G density function with parameter $r + k + 1$,

$$m_{r,k} = \frac{n!\,(r+1)\,(i-1)!\,b_{r+1}}{(r+k+1)} \sum_{j=0}^{n-i} \frac{(-1)^j\, f_{j+i-1,k}}{(n-i-j)!\,j!},$$

b_r is given by equation (9) and the quantities $f_{j+i-1,k}$ can be determined given that $f_{j+i-1,0} = b_0^{j+i-1}$ and recursively for $k \geq 1$

$$f_{j+i-1,k} = (k\, b_0)^{-1} \sum_{m=1}^{k} [m\,(j+i) - k]\, b_m\, f_{j+i-1,k-m}\,.$$

We can obtain the ordinary and incomplete moments, generating function and mean deviations of the GOGa-G order statistics from equation (14) and some properties of the exp-G model.

4. The maximum likelihood estimator

The MLE is one of the most common point estimators. This estimator is very applicable in confidence intervals and hypothesis testing. By MLE, various statistics is built for assessing the goodness-of-fit in a model, such as: the maximum log-likelihood ($\hat{\ell}_{max}$), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Anderson-Darling (A^*) and Cramér–von Mises (W^*), described by Chen and Balakrishnan (1995). The lower values of these statistics indicate that the model have better fitting. We use these statistics in section 5.

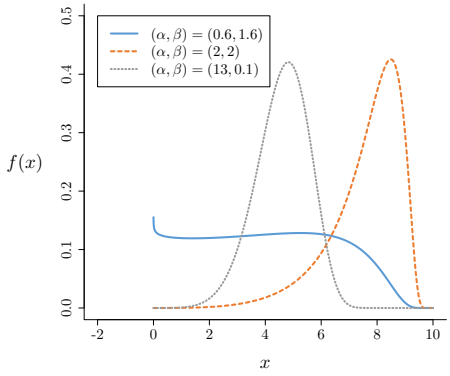

Figure 6: Three density functions for simulation study.

To calculating the MLE, let $x_1, x, ..., x_n$ are observations from pdf (2). In this case, by letting $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\xi})$ we have

$$\ell_n(\boldsymbol{\theta}) = n\ln(\beta) + \sum_{i=0}^{n} \ln(g(x_i;\boldsymbol{\xi})) + (\alpha\beta - 1)\sum_{i=0}^{n} \ln(G(x_i;\boldsymbol{\xi})) - \sum_{i=0}^{n} \frac{G(x_i;\boldsymbol{\xi})^{\beta}}{1 - G(x_i;\boldsymbol{\xi})^{\beta}} - n\ln(\Gamma(\alpha)) - (\alpha+1)\sum_{i=0}^{n} \ln(1 - G(x_i;\boldsymbol{\xi})^{\beta})$$

By numerically solving the following equations, the maximum likelihood estimators can be obtained.

$$\begin{cases} \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \alpha} = \beta \sum_{i=0}^{n} \ln G(x_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=0}^{n} \ln(1 - G(x_i)^{\beta}) = 0 \\ \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \beta} = \frac{n}{\beta} + \alpha \sum_{i=0}^{n} \ln G(x_i) - \sum_{i=0}^{n} \frac{G(x_i)^{\beta} \ln G(x_i)}{\left(1 - G(x_i)^{\beta}\right)^2} + (\alpha+1) \sum_{i=0}^{n} \frac{\ln G(x_i) G(x_i)^{\beta}}{(1 - G(x_i)^{\beta}} = 0 \\ \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}} = \sum_{i=0}^{n} \frac{g(x_i)^{(\xi)}}{g(x_i)} + (\alpha\beta - 1) \sum_{i=0}^{n} \frac{G_i^{(\xi)}}{G(x_i)} - \sum_{i=0}^{n} \frac{\beta G_i^{(\xi)} G(x_i)^{\beta-1}}{\left(1 - G(x_i)^{\beta}\right)^2} + (\alpha+1) \sum_{i=0}^{n} \frac{\beta G_i^{(\xi)} G(x_i)^{\beta-1}}{1 - G(x_i)^{\beta}} = 0 \end{cases}$$

where $g_i{}^{(\xi)} = \frac{\partial g(x_i;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$ and $G_i{}^{(\xi)} = \frac{\partial G(x_i;\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}$

5. Simulation study

In this section, the Maximum likelihood estimators for additional parameters α and β in pdf (4) for three different states, has been assessed by simulating: $(\alpha, \beta) = (0.6, 1.6)$, $(\alpha, \beta) = (2, 2)$ and $(\alpha, \beta) = (13, 0.1)$. In each three case, the uniform distribution parameters in (4) are $(a, b) = (0, 10)$. The density functions for one of the three states, has been indicated in Figure 6. One can see three different states of GOGa-U density functions, means skewed to the left, right and the symmetric .

To verify the validity of the maximum likelihood estimator, Mean Square Error of the Estimate (MSE), Coverage Probability (CP) and Coverage Lenght (CL) have been used. For example, as described in Section 3.1, for $(\alpha, \beta) = (0.6, 1.6)$, $N = 10000$ times have been simulated samples of $n = 30, 40, ..., 500$ of GOGa-U$(0.6, 1.6, 0, 10)$. To estimate the numerical value of the maximum likelihood, the *optim* function (in the *stat* package) and L-BFGS-B method in R software has been used. If $\boldsymbol{\theta} = (\alpha, \beta)$, for any simulation by n volume and $i = 1, 2, ..., N$, the maximum likelihood estimates are obtained as $\widehat{\boldsymbol{\theta}}_i = (\widehat{\alpha}_i, \widehat{\beta}_i)$. The standard deviation of estimations, which is obtained through the information matrix is shown by $s_{\widehat{\boldsymbol{\theta}}_i} = (s_{\widehat{\alpha}_i}, s_{\widehat{\beta}_i})$. In this case, the MLE, Bias, MSE, CP and CL are calculated by the following formula

$$MLE_{\widehat{\boldsymbol{\theta}}}(n) = \frac{1}{N} \sum_{i=1}^{N} \widehat{\boldsymbol{\theta}}_i$$

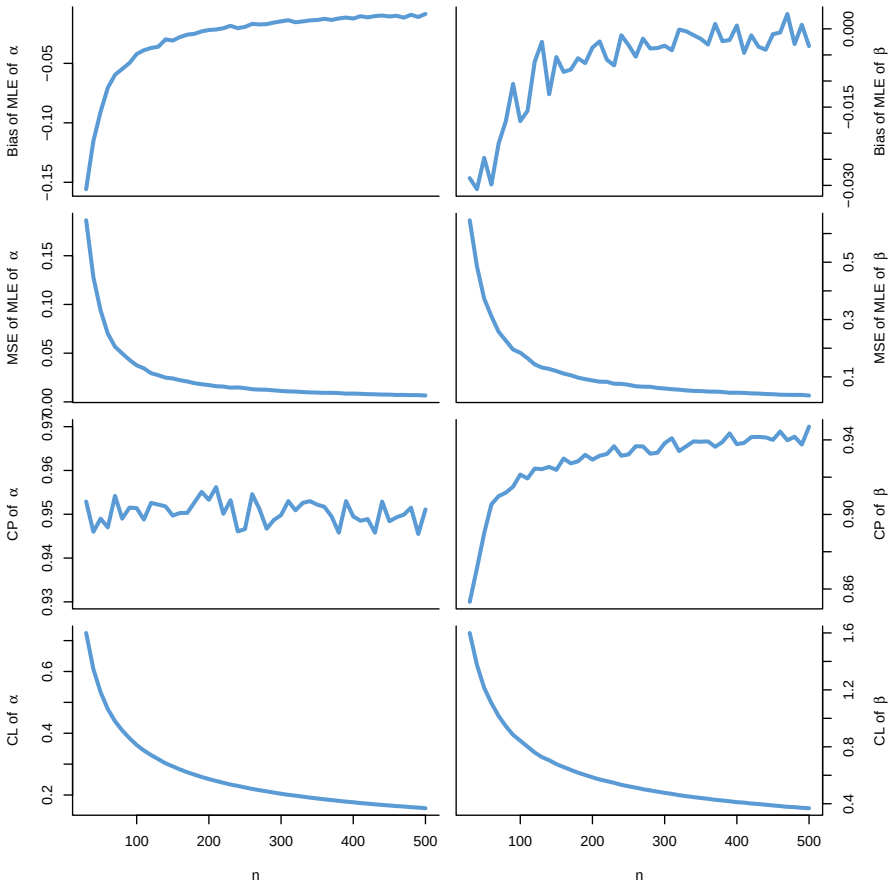

Figure 7: Biases, MSEs, CPs and CLs of $\hat{\alpha}, \hat{\beta}$ versus n when $(\alpha, \beta) = (0.6, 1.6)$.

$$Bias_{\widehat{\boldsymbol{\theta}}}(n) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)$$

$$MSE_{\widehat{\boldsymbol{\theta}}}(n) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^2$$

$$CP_{\widehat{\boldsymbol{\theta}}}(n) = \frac{1}{N}\sum_{i=1}^{N} I(\widehat{\boldsymbol{\theta}}_i - 1.96 s_{\widehat{\boldsymbol{\theta}}_i}, \widehat{\boldsymbol{\theta}}_i + 1.96 s_{\widehat{\boldsymbol{\theta}}_i})$$

$$CL_{\widehat{\boldsymbol{\theta}}}(n) = \frac{3.92}{N}\sum_{i=1}^{N} s_{\widehat{\boldsymbol{\theta}}_i}$$

In Figures 7 represent the Biases, MSEs, CPs and CLs plots for $(\alpha, \beta) = (0.6, 1.6)$. As expected, the biases and MSE of estimated parameters converges to zero while n growing. The CPs plots should converge to 0.95 and CLs plots should be descending they are correct in Figures 7. Plots of parameters vector $(\alpha, \beta) = (2, 2)$ and $(\alpha, \beta) = (13, 0.1)$ have the same position that one can see in Appendix 7.1.

6. Applications

In this section, fitting of GOGa-U and some famous models to the two real data sets has been assessed. The Akaike information criterion (AIC), Bayesian information criterion (BIC), Anderson-Darling (A^*) and Cramér-von Mises (W^*) , KolmogorovâĂŞSmirnov (K.S) and

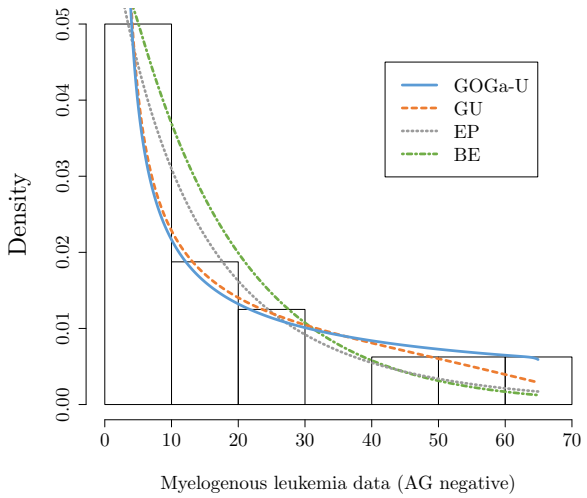

Figure 8: Histogram and estimated pdfs for the AG negative data.

the P-Value of K.S test , have been chosen to comparison of the models. The distributions: Beta Exponential (BE) (Nadarajah and Kotz (2006)), Beta Generalized Exponential (BGE) (Barreto-Souza, Santos, and Cordeiro (2010)), Beta Genearlized Half-Normal (BGHN) (Pescim, Demétrio, Cordeiro, Ortega, and Urbano (2010)), Beta Pareto (BP) (Akinsete, Famoye, and Lee (2008)), Exponentiated Pareto (EP) (Kuş (2007)), Genearlized Half-Normal (GHN) (Cooray and Ananda (2008)), Gamma-Uniform (GU) (Torabi and Hedesh (2012)), Kumaraswamy Gumbel (KwGu) (Cordeiro, Nadarajah, and Ortega (2012)) and Weibull-G$\{E\}$ (Alzaatreh, Lee, and Famoye (2015)) have been selected for comparison. The parameters of models have been estimated by the MLE method.

6.1. The myelogenous leukemia data for AG negative

This sub-section is related to study of AG data which presented by Feigl and Zelen (1965) that include 16 observations. Observed survival times (weeks) for AG negative were identified by the presence of Auer rods and significant granulative of the leukemic cells in the bone marrow at diagnosis. For the AG negative patients these factors were absent. The data set is: 56, 65, 17, 17, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43.

The Tables 1 and 2 display a summary of the fitted information criteria and MLEs for this data with different models, respectively. Models have been sorted from the lowest to the highest value of AIC. As you see, the GOGa-U is selected as the best model with all the criteria. Note that P-Value for GOGa-U is also more than all other distributions. The histogram of the AG nagative data and the plots of fitted pdf are displayed in Figure 8.

Table 1: Information criteria for the AG negative data.

Model	**AIC**	**BIC**	W^*	A^*	**K.S**	**P-Value**
GOGa-U	121.29	124.38	0.07	0.46	0.18	0.687
G-U	122.68	125.77	0.06	0.39	0.18	0.678
EP	129.09	130.64	0.1	0.65	0.21	0.475
BE	129.66	131.98	0.1	0.72	0.3	0.105
GHN	130.21	131.76	0.11	0.65	0.22	0.422
BP	131.41	134.5	0.11	0.66	0.22	0.404
Weibull-G$\{E\}$	131.55	134.64	0.11	0.68	0.22	0.441
BGHN	131.83	134.93	0.11	0.67	0.23	0.356
BGE	132.55	135.64	0.1	0.67	0.23	0.343
KwGu	134.22	137.31	0.1	0.65	0.3	0.123

Table 2: MLEs for the the AG negative data.

Model	Parameters
GOGa-U	$(\widehat{\alpha},\widehat{\beta},\widehat{a},\widehat{b}) = (0.01, 51.13, 1.99, 66.67)$ $(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{a}}, s_{\widehat{b}}) = (0.01, 62.04, 0.01, 2.63)$
G-U	$(\widehat{\alpha},\widehat{\beta},\widehat{a},\widehat{b}) = (0.40, 0.81, 1.99, 98.91)$ $(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{a}}, s_{\widehat{b}}) = (0.15, 1.20, 0.01, 53.41)$
EP	$(\widehat{\lambda},\widehat{\beta}) = 1.01, 0.04)$ $(s_{\widehat{\lambda}}, s_{\widehat{\beta}}) = (1.88, 0.02)$
BE	$(\widehat{a},\widehat{b},\widehat{\lambda}) = (8.24, 0.04, 1.54)$ $(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\lambda}}) = (40.43, 0.08, 2.85)$
GHN	$(\widehat{\alpha},\widehat{\theta}) = (0.74, 22.79)$ $(s_{\widehat{\alpha}}, s_{\widehat{\theta}}) = (0.15, 6.04)$
BP	$(\widehat{\alpha},\widehat{\beta},\widehat{\theta},\widehat{k}) = (98.66, 3.01, 0.01, 0.53)$ $(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{\theta}}, s_{\widehat{k}}) = (593.87, 17.51, 0.02, 1.80)$
Weibull-G$\{E\}$	$(\widehat{c},\widehat{\gamma},\widehat{\alpha},\widehat{\beta}) = 0.48, 3.09, 5.02, 1.40)$ $(s_{\widehat{c}}, s_{\widehat{\gamma}}, s_{\widehat{\alpha}}, s_{\widehat{\beta}}) = (0.09, 1.47, 0.50, 0.01)$
BGHN	$(\widehat{a},\widehat{b},\widehat{\alpha},\widehat{\theta}) = (0.03, 76.12, 508.34, 270.67)$ $(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\alpha}}, s_{\widehat{\theta}}) = (0.04, 4235.56, 1349.84, 471.83)$
BGE	$(\widehat{a},\widehat{b},\widehat{\lambda},\widehat{\alpha}) = (14.23, 6.84, 0.00, 0.13)$ $(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\lambda}}, s_{\widehat{\alpha}}) = (33.74, 4.52, 0.00, 0.27)$
KwGu	$(\widehat{a},\widehat{b},\widehat{\mu},\widehat{\sigma}) = (0.01, 0.11, 10.51, 1.93)$ $(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\mu}}, s_{\widehat{\sigma}}) = (0.01, 0.03, 0.01, 0.02)$

6.2. The sum of skin folds data

The second data set which contains 202 observation can be seen in Weisberg (2005) that have been used in Alzaatreh (2015) (article not yet published). These data are the sum of skin folds in 202 athletes collected at the Australian Institute of Sports and are as follows:
28.0, 98, 89.0, 68.9, 69.9, 109.0, 52.3, 52.8, 46.7, 82.7, 42.3, 109.1, 96.8, 98.3, 103.6, 110.2, 98.1, 57.0, 43.1, 71.1, 29.7, 96.3, 102.8, 80.3, 122.1, 71.3, 200.8, 80.6, 65.3, 78.0, 65.9, 38.9, 56.5, 104.6, 74.9, 90.4, 54.6, 131.9, 68.3, 52.0, 40.8, 34.3, 44.8, 105.7, 126.4, 83.0, 106.9, 88.2, 33.8, 47.6, 42.7, 41.5, 34.6, 30.9, 100.7, 80.3, 91.0, 156.6, 95.4, 43.5, 61.9, 35.2, 50.9, 31.8, 44.0, 56.8, 75.2, 76.2, 101.1, 47.5, 46.2, 38.2, 49.2, 49.6, 34.5, 37.5, 75.9, 87.2, 52.6, 126.4, 55.6, 73.9, 43.5, 61.8, 88.9, 31.0, 37.6, 52.8, 97.9, 111.1, 114.0, 62.9, 36.8, 56.8, 46.5, 48.3, 32.6, 31.7, 47.8, 75.1, 110.7, 70.0, 52.5, 67, 41.6, 34.8, 61.8, 31.5, 36.6, 76.0, 65.1, 74.7, 77.0, 62.6, 41.1, 58.9, 60.2, 43.0, 32.6, 48, 61.2, 171.1, 113.5, 148.9, 49.9, 59.4, 44.5, 48.1, 61.1, 31.0, 41.9, 75.6, 76.8, 99.8, 80.1, 57.9, 48.4, 41.8, 44.5, 43.8, 33.7, 30.9, 43.3, 117.8, 80.3, 156.6, 109.6, 50.0, 33.7, 54.0, 54.2, 30.3, 52.8, 49.5, 90.2, 109.5, 115.9, 98.5, 54.6, 50.9, 44.7, 41.8, 38.0, 43.2, 70.0, 97.2, 123.6, 181.7, 136.3, 42.3, 40.5, 64.9, 34.1, 55.7, 113.5, 75.7, 99.9, 91.2, 71.6, 103.6, 46.1, 51.2, 43.8, 30.5, 37.5, 96.9, 57.7, 125.9, 49.0, 143.5, 102.8, 46.3, 54.4, 58.3, 34.0, 112.5, 49.3, 67.2, 56.5, 47.6, 60.4, 34.9

the Tables 3 and 4 display a summary of the fitted information criteria and MLEs for this data with different models , respectively. Models have been sorted from the lowest to the highest value of AIC. As you see, the GOGa-U is selected as the best model with all the criteria. Here P-Value for GOGa-U is also more than all other distributions. The histogram of the sum of skin folds data and the plots of fitted pdf are displayed in Figure 9.

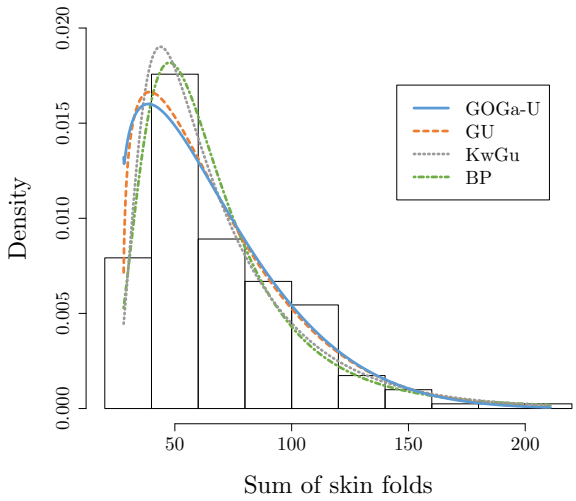

Figure 9: Histogram and estimated pdfs for the sum of skin folds data.

Table 3: Information criteria for the sum of skin folds data.

Model	**AIC**	**BIC**	W^*	A^*	**K.S**	**P-Value**
GOGa-U	1897.11	1910.34	0.09	0.55	0.05	0.731
G-U	1897.15	1910.38	0.08	0.47	0.05	0.668
KwGu	1906.25	1919.48	0.11	5.05	0.06	0.393
BP	1915.68	1928.91	0.18	1.57	0.07	0.228
Weibull-G$\{E\}$	1916.04	1929.27	0.2	2.4	0.07	0.34
BGE	1920.58	1933.81	0.26	0.76	0.08	0.179
BGHN	1925.11	1938.34	0.32	1.21	0.08	0.135
BE	1930.2	1940.13	0.4	1.25	0.09	0.063
GHN	1978.34	1984.96	0.86	2.36	0.13	0.002
EP	2119.1	2125.71	0.41	1.89	0.35	0

Table 4: MLEs for the sum of skin folds data.

Model	**Parameters**
GOGa-U	$(\widehat{\alpha},\widehat{\beta},\widehat{a},\widehat{b}) = (8.95, 0.04, 27.99, 650.04)$
	$(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{a}}, s_{\widehat{b}}) = (1.96, 0.01, 0.02, 169.55)$
G-U	$(\widehat{\alpha},\widehat{\beta},\widehat{a},\widehat{b}) = (1.27, 0.07, 27.88, 579.87)$
	$(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{a}}, s_{\widehat{b}}) = (0.17, 0.05, 0.25, 334.16)$
KwGu	$(\widehat{a},\widehat{b},\widehat{\mu},\widehat{\sigma}) = (0.01, 0.21, 68.93, 7.15)$
	$(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\mu}}, s_{\widehat{\sigma}}) = (0.01, 0.06, 2.22, 1.78)$
BP	$(\widehat{\alpha},\widehat{\beta},\widehat{\theta},\widehat{k}) = (102.94, 4.20, 3.31, 1.14)$
	$(s_{\widehat{\alpha}}, s_{\widehat{\beta}}, s_{\widehat{\theta}}, s_{\widehat{k}}) = (183.57, 4.22, 3.43, 0.63)$
Weibull-G$\{E\}$	$(\widehat{c},\widehat{\gamma},\widehat{\alpha},\widehat{\beta}) = (0.01, 0.21, 68.93, 7.15)$
	$(s_{\widehat{c}}, s_{\widehat{\gamma}}, s_{\widehat{\alpha}}, s_{\widehat{\beta}}) = (0.01, 0.06, 2.22, 1.78)$
BGE	$(\widehat{a},\widehat{b},\widehat{\lambda},\widehat{\alpha}) = (1.23, 0.73, 0.05, 8.77)$
	$(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\lambda}}, s_{\widehat{\alpha}}) = (1.98, 0.25, 0.01, 16.52)$
BGHN	$(\widehat{a},\widehat{b},\widehat{\alpha},\widehat{\theta}) = (0.27, 34.18, 43.46, 13.78)$
	$(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\alpha}}, s_{\widehat{\theta}}) = (0.10, 37.32, 31.62, 0.50)$
BE	$(\widehat{a},\widehat{b},\widehat{\lambda}) = (5.34, 5.86, 0.01)$
	$(s_{\widehat{a}}, s_{\widehat{b}}, s_{\widehat{\lambda}}) = (0.53, 1.62, 0.00)$
GHN	$(\widehat{\alpha},\widehat{\theta}) = (1.65, 86.05)$
	$(s_{\widehat{\alpha}}, s_{\widehat{\theta}}) = (0.09, 2.89)$
EP	$(\widehat{\lambda},\widehat{\beta}) = (0.01, 0.01)$
	$(s_{\widehat{\lambda}}, s_{\widehat{\beta}}) = (1.65, 86.05)$

7. Conclusions

In many applied areas there is a clear need for extended forms of the well-known distributions. Generally, the new distributions are more flexible to model real data that present a high degree of skewness and kurtosis. We propose Generalized Odd Gamma-G (GOGa-G) family of distributions. Many well-known models emerge as special cases of the GOGa-G family by using special parameter values. Some mathematical properties of the new class including explicit expansions for the ordinary and incomplete moments, quantile and generating functions, mean deviations, entropies and order statistics are provided. The model parameters are estimated by the maximum likelihood estimation method. We prove empirically by means of an application to a real data set that special cases of the proposed family can give better fits than other models generated from well-known families.

7.1. Acknowledgement

The support of Research Committee of Persian Gulf University is greatly acknowledged.

References

Akinsete A, Famoye F, Lee C (2008). "The Beta-Pareto Distribution." *Statistics*, **42**(6), 547–563.

Alexander C, Cordeiro GM, Ortega EM, Sarabia JM (2012). "Generalized Beta-Generated Distributions." *Computational Statistics & Data Analysis*, **56**(6), 1880–1897.

Alizadeh M, Cordeiro GM, De Brito E, Demétrio CGB (2015a). "The Beta Marshall-Olkin Family of Distributions." *Journal of Statistical Distributions and Applications*, **2**(1), 1.

Alizadeh M, Tahir M, Cordeiro GM, Mansoor M, Zubair M, Hamedani G (2015b). "The Kumaraswamy Marshal-Olkin Family of Distributions." *Journal of the Egyptian Mathematical Society*, **23**(3), 546–557.

Alzaatreh A, Lee C, Famoye F (2013). "A New Method for Generating Families of Continuous Distributions." *Metron*, **71**(1), 63–79.

Alzaatreh A, Lee C, Famoye F (2015). "Family of Generalized Gamma Distributions: Properties and Applications." *Hacettepe Journal of Mathematics and Statistics. Doi*, **10**.

Barreto-Souza W, Santos AH, Cordeiro GM (2010). "The Beta Generalized Exponential Distribution." *Journal of Statistical Computation and Simulation*, **80**(2), 159–172.

Bourguignon M, Silva RB, Cordeiro GM (2014). "The Weibull-G Family of Probability Distributions." *Journal of Data Science*, **12**(1), 53–68.

Chen G, Balakrishnan N (1995). "A General Purpose Approximate Goodness-of-Fit Test." *Journal of Quality Technology*, **27**(2), 154–161.

Cooray K, Ananda MM (2008). "A Generalization of the Half-Normal Distribution with Applications to Lifetime Data." *Communications in StatisticsâĂŤTheory and Methods*, **37**(9), 1323–1337.

Cordeiro GM, Alizadeh M, Diniz Marinho PR (2016). "The Type I Half-Logistic Family of Distributions." *Journal of Statistical Computation and Simulation*, **86**(4), 707–728.

Cordeiro GM, Alizadeh M, Ortega EM (2014a). "The Exponentiated Half-Logistic Family of Distributions: Properties and Applications." *Journal of Probability and Statistics*, **2014**.

Cordeiro GM, de Castro M (2011). "A New Family of Generalized Distributions." *Journal of statistical computation and simulation*, **81**(7), 883–898.

Cordeiro GM, Nadarajah S, Ortega EM (2012). "The Kumaraswamy Gumbel Distribution." *Statistical Methods & Applications*, **21**(2), 139–168.

Cordeiro GM, Ortega EM, Popović BV, Pescim RR (2014b). "The Lomax Generator of Distributions: Properties, Minification Process and Regression Model." *Applied Mathematics and Computation*, **247**, 465–486.

Feigl P, Zelen M (1965). "Estimation of Exponential Survival Probabilities with Concomitant Information." *Biometrics*, pp. 826–838.

Gradshteyn IS, Ryzhik I (2000). "Table of Integrals, Series, and Products. Translated from the Russian. Translation Edited and with a Preface by Alan Jeffrey and Daniel Zwillinger."

Kuş C (2007). "A New Lifetime Distribution." *Computational Statistics & Data Analysis*, **51**(9), 4497–4509.

Marshall AW, Olkin I (1997). "A New Method for Adding a Parameter to a Family of Distributions with Application to the Exponential and Weibull Families." *Biometrika*, **84**(3), 641–652.

Nadarajah S, Kotz S (2006). "The Beta Exponential Distribution." *Reliability engineering & system safety*, **91**(6), 689–697.

Pescim RR, Demétrio CG, Cordeiro GM, Ortega EM, Urbano MR (2010). "The Beta Generalized Half-Normal Distribution." *Computational statistics & data analysis*, **54**(4), 945–957.

Rrnyi A (1961). "On Measures of Entropy and Information." In *Fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 547–561.

Tahir M, Cordeiro GM, Alzaatreh A, Mansoor M, Zubair M (2016). "The Logistic-X Family of Distributions and Its Applications." *Communications in Statistics-Theory and Methods*, (just-accepted).

Torabi H, Hedesh NM (2012). "The Gamma-Uniform Distribution and Its Applications." *Kybernetika*, **48**(1), 16–30.

Weisberg S (2005). *Applied Linear Regression*, volume 528. John Wiley & Sons.

Appendices A

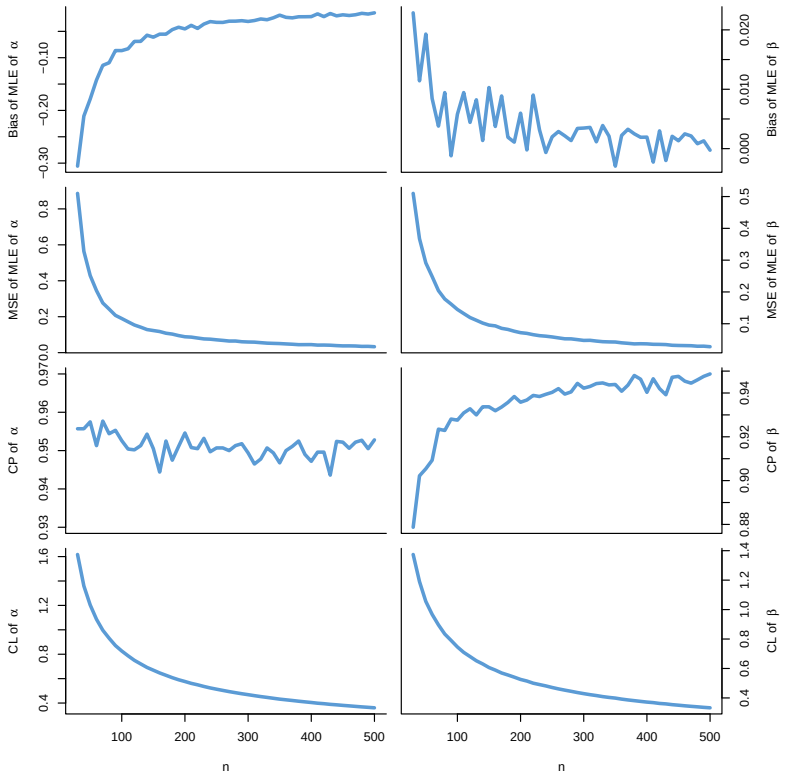

Figure 10: Biases, MSEs, CPs and CLs of $\hat{\alpha}, \hat{\theta}$ versus n when $(\alpha, \beta) = (2, 2)$.

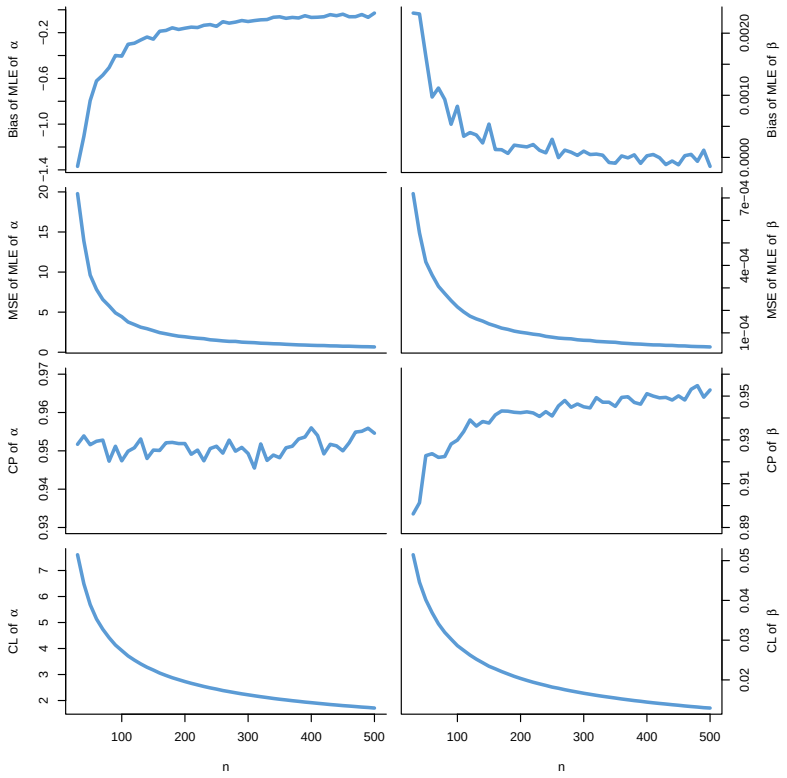

Figure 11: Biases, MSEs, CPs and CLs of $\hat{\alpha}, \hat{\theta}$ versus n when $(\alpha, \beta) = (13, 0.1)$.

Appendices B

Program developed in R to obtain the value of density (dGOGaG), distribution (pGOGaG), hazard (hGOGaG), quantile (qGOGaG) function and random generation (rGOGaG) for the GOGa-G distribution.

```
dGOGaG = function(x, par, Ge = "uniform")
        {
        if (Ge == "uniform")
            {
            G = punif(x,par[3],par[4])
            g = dunif(x,par[3],par[4])
            }
        if (Ge == "weibull")
            {
            G = pweibull(x,par[3],par[4])
            g = dweibull(x,par[3],par[4])
            }
        Gb = G^par[2]
        pdf = par[2]*g*G^(par[1]*par[2]-1)*exp(-Gb/(1-Gb))/
                        (gamma(par[1])*(1-Gb)^(par[1]+1))
        pdf[!is.finite(pdf)] = NA
        pdf
        } # end of dGOGaG

pGOGaG = function(x, par, Ge = "uniform")
        {
        if (Ge == "uniform")
            {
            G = punif(x,par[3],par[4])
            g = dunif(x,par[3],par[4])
            }
        if (Ge == "weibull")
            {
            G = pweibull(x,par[3],par[4])
            g = dweibull(x,par[3],par[4])
            }
        Gb = G^par[2]
        cdf = pgamma(Gb/(1-Gb),par[1],1)
        cdf[!is.finite(cdf)] = NA
        cdf
        } # end of pGOGaG

qGOGaG = function(p, par, Ge = "uniform")
        {
        a = qgamma(p,par[1],1)
        b = (a/(1+a))^(1/par[2])
        if (Ge == "uniform")
            {
            return(qunif(b,par[3],par[4]))
            }
        if (Ge == "weibull")
            {
            return(qweibull(b,par[3],par[4]))
```

```
        }
    } # end of qGOGaG

hGOGaG = function(x, par, Ge = "uniform")
    {
    pdf = dGOGaG(x=x, par=par, Ge = Ge)
    cdf = pGOGaG(x=x, par=par, Ge = Ge)
    hrf = pdf/(1 - cdf)
    hrf[!is.finite(hrf)] = NA
    hrf
    } # end of hGOGaG

rGOGaG = function(n, par, Ge = "uniform")
    {
    GI=rgamma(n,par[1],1)
    if (Ge == "uniform")
        {
        return(qunif((GI/(1+GI))^(1/par[2]),par[3],par[4]))
        }
    if (Ge == "weibull")
        {
        return(qweibull((GI/(1+GI))^(1/par[2]),par[3],par[4]))
        }
    } # end of rGOGaG
```

Program developed in R of claculatition for one-dimensional integral based on observations and the trapezoidal rule integration:

```
intob = function(x, y) 0.5*sum(diff(x)*(y[1:length(x)-1]+y[2:length(x)]))
```

Program developed in R of claculatition for the value of Rényi entropy:

```
REntropy = function(par, gamma)
    {
    fgamma = function(x) dGOGaG(x, par = par, Ge = "uniform")^gamma
    x = seq(par[3], par[4], le=10000)
    y = fgamma(x)
    ent = log(intob(x,y))/(1-gamma)
    ent = ent[!is.finite(ent)] = NA
    return(ent)
    } # end of REntropy
```

Program developed in R of claculatition for the value of moment, skewness and kurtosis:

```
moment = function(par, order)
    {
    x = seq(par[3], par[4], le=10000)
    y = dGOGaG(x = x, par = par, Ge = "uniform")
    return(intob(x, x^order * y))
    } # end of moment

skew = function(par)
    {
    x = seq(par[3], par[4], le=10000)
```

```
        y = dGOGaG(x = x, par = par, Ge = "uniform")
        m1 = intob(x, x*y)
        m2 = intob(x, (x-m1)^2*y)
        return(intob(x, ((x-m1)^3*y))/sqrt(m2)^3)
        } # end of skew

kurt = function(par)
        {
        x = seq(par[3], par[4], le=10000)
        y = dGOGaG(x = x, par = par, Ge = "uniform")
        m1 = intob(x, x*y)
        m2 = intob(x, (x-m1)^2*y)
        return(intob(x, (x-m1)^4*y)/sqrt(m2)^4)
        } # end of kurt
```

Program developed in R of optimization for simulations and applications. The *initpar* need to change for some observations.

```
loglikeSimulation = function(alpha,beta)
        -sum(log(dGOGaG(x, c(alpha,beta,par[3],par[4]), Ge = "uniform")))
optim(par = initpar, fn = loglikeSimulation, lower=c(0.005,0.005),
        upper=c(Inf,Inf), method="L-BFGS-B", hessian = TRUE)

loglikeApplication = function(alpha,beta,a,b)
        -sum(log(dGOGaG(x, c(alpha,beta,a,b), Ge = "uniform")))
optim(par = initpar, fn = loglikeApplication,
        lower=c(0.005,0.005, min(x)-.001,max(x)+0.001),
        upper=c(Inf,Inf,-Inf,Inf), method="L-BFGS-B", hessian = TRUE)
```

Affiliation:

Bistoon Hosseini
Department of Statistics, Faculty of Sciences
Persian Gulf University of Bushehr
Bushehr, Iran
E-mail: bistoon.hosseini@gmail.com

Mahmoud Afshari*
Department of Statistics, Faculty of Sciences
Persian Gulf University of Bushehr
Bushehr, Iran
Corresponding Author E-mail: afshar.5050@gmail.com

Morad Alizadeh
Department of Statistics, Faculty of Sciences
Persian Gulf University of Bushehr
Bushehr, Iran
E-mail: moradalizadeh78@gmail.com

Austrian Journal of Statistics http://www.ajs.or.at/
published by the Austrian Society of Statistics http://www.osg.or.at/
Volume 47 *Submitted:* 2016-07-11
February 2018 *Accepted:* 2017-01-11

ISSN 1026–597X

Contents