CONTENTS

Metodološki zvezki, Vol. 13, No. 2, 2016

Metodološki zvezki, Vol. 13, 2016

Reviewers for Volume Thirteen

Mojca Bavdaž
Rok Blagus
Gregor Dolinar
Anuška Ferligoj
Herwig Friedl
Andreja Jaklič
Borut Jurčič Zlobec
Damjana Kastelec
Tina Kogovšek
Katarina Košmelj
Vyacheslav Lyubchich
Susana Martins
Stanislav Mejza
Giovanni Millo
Irena Ograjenšek
Klemen Pavlic
Marko Robnik Sikonja
Jože Rovan
Támas Rudas
Damjam Skulj
Gregor Sočan
Janez Stare
Jordan Stoyanov
Gaj Vidmar
Blaz Zupan
Aleš Žiberna

Metodološki zvezki, Vol. 13, No. 2, 71-85

The Mixture Poisson Exponential–Inverse Gaussian Regression Model: An application in Health Services

Emilio Gómez–Déniz[1], Enrique Calderín–Ojeda[2]

Abstract

In this paper a mixed Poisson regression model for count data is introduced. This model is derived by mixing the Poisson distribution with the one–parameter continuous exponential–inverse Gaussian distribution. The obtained probability mass function is over–dispersed and unimodal with modal value located at zero. Estimation is performed by maximum likelihood. As an application, the demand for health services among people 65 and over is examined using this regression model since empirical evidence has suggested that the over–dispersion and a large portion of non–users are common features of medical care utilization data.

1 Introduction

Counting data are common in many social and biomedical studies to explain differences among cases that generate small counts of events. The Poisson distribution plays an important role in the modeling of count data. In this regard, Poisson regression models have been traditionally used to analyze data with a nonnegative integer response variable in a wide range of different applied areas, for example, biostatistics, epidemiology, accident analysis and prevention, insurance and criminology among other fields. Nevertheless, the rigidity of the Poisson mean–variance relationship makes the Poisson regression models exposed to over–dispersion (i.e. the empirical variance is larger than the empirical mean). This is a crucial modeling issue for count data since inadequate confidence interval coverage is produced when over–dispersed count data are considered. The Poisson model does not allow for heterogeneity among individuals. Often there is additional heterogeneity between individuals that is not accounted for by the predictors in the model which results in over–dispersion. To overcome this difficulty, practitioners usually use more general specifications, e.g. negative binomial regression model (Hilbe (2007) and Greene (2009)). The latter model is an example of mixed Poisson regression model. Mixed Poisson regression models are natural extensions of the Poisson regression model allowing for

[1]Department of Quantitative Methods and TiDES Institute, University of Las Palmas de Gran Canaria, Gran Canaria, Spain; emilio.gomez-deniz@ulpgc.es

[2] Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Australia; enrique.calderin@unimelb.edu.au

over–dispersion. This feature can be included in the model by assuming that the parameter of the Poisson distribution is not fixed due to the heterogeneity of the population, being likewise considered a random variable. For instance, for over–dispersed count–panel data the negative binomial and Poisson–Inverse Gaussian regression models are well–known in the statistical literature. In this regard, by using a gamma distribution for the unknown parameter θ, the former model is obtained. The latter model was proposed by Dean et al. (1989), in this case an inverse Gaussian distribution is used to describe the parameter of the Poisson distribution. These models account for over–dispersion by assuming that there will be unexplained variability among individuals who have the same predicted value. It leads to larger variance in the overall outcome distribution but has no effect on the mean.

Regrettably, other mixed Poisson regression models have not been used since they involve special functions and appropriate numerical methods are required. Nevertheless, due to the fast improvement of mathematical software these models can be handled relatively easily. In this article a new mixed Poisson regression model is proposed. As mixing distribution, a particular case of the continuous Exponential–Inverse Gaussian distribution in Bhattacharya and Kumar (1986) when one of the parameter tends to infinity is considered. Furthermore, as it arises from a mixed Poisson distribution, many of its properties can be derived from the ones of the mixing distribution. In this sense, it displays interesting features such as over–dispersion, unimodality, closed–form expressions for factorial moments of any order among other nice properties. The mixed Poisson regression model introduced in this paper does not belong to the linear exponential family of distributions. However, as Wedderburn (1974) showed, the parameter estimation and inference theory developed for the exponential family (i.e. generalized linear models), can be extended to models where a relation between the mean and variance of the response variable can be specified, even though they were not associated with a known likelihood. In this sense, the unconditional distribution obtained in the Poisson–Inverse Gaussian regression model (Dean et al. (1989)) is not part of the exponential family of distributions.

In this manuscript, the demand for health services among people 65 and over is analyzed by using this new mixed Poisson regression model. In particular, the number of hospital stays among the elderly population is considered as response variable. Moreover, as it will be shown later, the data include two important features a high proportion of zeros and over–dispersion. The use of regression model to explain the demand for health services has been studied by Gurmu and Elder (2000) where bivariate regression model for count data was used and also by Lahiri and Xing (2004) by using two–parts model based on Poisson selection model.

The remainder of the paper is structured as follows. Section 2 introduces the new Poisson distribution together with some properties; additionally parameter estimation is discussed; section 3 describes the mixed Poisson regression model derived from this distribution. Estimation is performed by maximum likelihood. Next, a numerical application to analyze factors explaining medical care of people 65 and over is examined in section 4. Finally, some conclusions are drawn in section 5.

2 The discrete model

The continuous Exponential–Inverse Gaussian distribution in Bhattacharya and Kumar (1986) can be simplified by letting one of its parameters tends to infinity. Then a more simple probability density function (pdf) is obtained. Then, the pdf of a random variable Θ following an Exponential–Inverse Gaussian distribution with a single scale parameter ϕ (henceforward $\mathcal{EIG}(\phi)$) is given by

$$f(\theta|\phi) = \sqrt{\frac{\phi}{2\,\theta}} \exp\left(-\sqrt{2\,\phi\,\theta}\right), \text{ with } \theta > 0 \text{ and } \phi > 0. \tag{2.1}$$

Let us now consider the Poisson distribution (henceforward $\mathcal{P}(\theta)$) whose probability mass function is given by

$$\Pr\{Y = y\} = e^{-\theta}\frac{\theta^y}{y!}, \;\; y = 0, 1, \ldots, \; \theta > 0. \tag{2.2}$$

Definition 1. *We say that a random variable Y has a Poisson–Exponential–Inverse Gaussian distribution if it admits the stochastic representation:*

$$\begin{aligned} Y|\theta &\sim \mathcal{P}(\theta) && (2.3)\\ \theta &\sim \mathcal{EIG}(\phi), && (2.4) \end{aligned}$$

with $\phi > 0$. We will denote this distribution by $Y \sim \mathcal{PEIG}(\phi)$.

Then, the unconditional probability mass function (pmf) of Y is given by

$$p_y = \frac{\sqrt{2\phi}\,\Gamma(2y+1)}{2^{2y+1}\,y!}\,\mathcal{U}\left(\frac{1}{2}+y, \frac{1}{2}, \frac{\phi}{2}\right), \;\; y = 0, 1, \ldots, \tag{2.5}$$

where $\mathcal{U}(a, b, z)$ represents the Tricomi confluent hypergeometric function given by ($a, z > 0$):

$$\mathcal{U}(a, b, z) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-zs}s^{a-1}(1+s)^{b-a-1}ds \tag{2.6}$$

(see Gradshteyn and Ryzhik (1994), page 1085, formula 9211–4).

The probability generating function is given by

$$G_Y(s) = \sqrt{\frac{\phi\pi}{1-s}} \exp\left\{\frac{\phi}{2(1-s)}\right\}\left[1 - \text{erf}\left(\sqrt{\frac{\phi}{2(1-s)}}\right)\right], \tag{2.7}$$

where $\text{erf}(z)$ is the error function given by

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt = \frac{2z}{\sqrt{\pi}}\,{}_1F_1(1/2, 3/2, -z^2),$$

being ${}_1F_1(\cdot,\cdot,\cdot)$ the confluent hypergeometric function.

The factorial moments of order k can be obtained from (2.5). They are provided by

$$\mu_{[k]}(Y) = E[Y(Y-1)\cdots(Y-k+1)] = \frac{2k\,\Gamma(2k)}{(2\phi)^k}, \tag{2.8}$$

with $k = 1, 2, \ldots$

From the latter expression it can be seen that (2.5) is over–dispersed, since

$$\frac{var(Y)}{E(Y)} = \frac{5}{\phi} + 1 > 1.$$

Additionally, as (2.1) has an asymptotic mode at 0, the discrete model (2.5) is unimodal with mode at 0 (see Holgate (1970)). Besides, as (2.1) is log–convex, then (2.5) is infinitely divisible and therefore, it is a compound Poisson distribution (see Propositions 8 and 9 in Karlis and Xekalaki, 2005).

Let us now suppose that $\mathbf{Y} = (Y_1, ..., Y_n)$ is a random sample of size n from the $\mathcal{PEIG}$ distribution with pmf (2.5). The log–likelihood function is proportional to

$$\ell(\phi; \mathbf{Y}) \propto \frac{n}{2} \log \phi + \sum_{i=1}^{n} \log \mathcal{U}\left(\frac{1}{2} + Y_i, \frac{1}{2}, \frac{\phi}{2}\right). \tag{2.9}$$

Having into account that

$$\frac{\partial}{\partial z}\mathcal{U}(a, b, z) = -a\, \mathcal{U}(a + 1, b + 1, z),$$

the maximum likelihood estimate of the parameter ϕ can be simply obtained by solving this normal equation

$$\frac{\partial \ell(\phi; \mathbf{Y})}{\partial \phi} = \frac{n}{\phi} - \sum_{i=1}^{n} \frac{\left(\frac{1}{2} + Y_i\right) \mathcal{U}\left(\frac{3}{2} + Y_i, \frac{3}{2}, \frac{\phi}{2}\right)}{\mathcal{U}\left(\frac{1}{2} + Y_i, \frac{1}{2}, \frac{\phi}{2}\right)} = 0. \tag{2.10}$$

The Fisher's information matrix can be approximated from

$$\frac{\partial^2 \ell(\phi; \mathbf{Y})}{\partial \phi^2} = -\frac{n}{\phi^2} - \sum_{i=1}^{n} \frac{\left(Y_i - \frac{1}{2}\right) \left\{\mathcal{M}_1(Y_i, \phi) + \left[\mathcal{M}_2(Y_i, \phi)\right]^2\right\}}{\left[\mathcal{M}_3(Y_i, \phi)\right]^2}, \tag{2.11}$$

where

$$\begin{aligned}
\mathcal{M}_1(Y_i, \phi) &= -\left(\frac{3}{2} + Y_i\right) \mathcal{U}\left(\frac{5}{2} + Y_i, \frac{5}{2}, \frac{\phi}{2}\right) \mathcal{U}\left(\frac{1}{2} + Y_i, \frac{1}{2}, \frac{\phi}{2}\right), \\
\mathcal{M}_2(Y_i, \phi) &= \left(\frac{1}{2} + Y_i\right) \left[\mathcal{U}\left(Y_i + \frac{3}{2}, \frac{3}{2}, \frac{\phi}{2}\right)\right]^2, \\
\mathcal{M}_3(Y_i, \phi) &= \mathcal{U}\left(\frac{1}{2} + Y_i, \frac{1}{2}, \frac{\phi}{2}\right).
\end{aligned}$$

This maximum likelihood estimate can also be calculated by using the EM algorithm. This method is a powerful technique that provides an iterative procedure to compute maximum likelihood estimation when data contain missing information. This methodology is suitable for distributions arising as mixtures since the mixing operation produces missing data. One of the main advantages of the EM algorithm is its numerical stability, increasing the likelihood of the observed data in each iteration. It does not guarantee convergence to the global maximum. It can be usually reached by starting the parameters at the moment estimates. The EM algorithm maximizes $\ell(\phi; \mathbf{Y})$ by iteratively

maximizing $E(\ell(\phi; \mathbf{Y}, \mathbf{Z}))$ where $\mathbf{Y} = (Y_1, ..., Y_n)$ denotes the sample observations and $\mathbf{Z} = (\theta_1, ..., \theta_n)$ denotes the missing observations and $\ell(\phi; \mathbf{Y}, \mathbf{Z})$ is the complete log–likelihood function.

The EM algorithm is based on two steps, the E–step, or expectation, fills in the missing data. Once the missing data are built–in, the parameters are estimated in the M–step (maximization step).

At the E–step of the *(j+1)*-th iteration the expected log–likelihood of the complete data model is computed by

$$E(\ell(\phi; \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}, \hat{\phi}^{(\mathbf{j})}). \tag{2.12}$$

In the M–step, the updated parameter estimate is computed from maximizing the quantity (2.12) with respect to ϕ. Then, if some terminating condition is satisfied we stop iterating, otherwise move back to E–step for more iterations.

In mixed Poisson distributions (Karlis, 2005) the unobserved quantities are the realizations of θ_i of the unobserved mixing parameter for each data point Y_i, $i = 1 \dots n$. Additionally, we assume that the distribution of $Y_i \mid \theta_i$ is Poisson with θ_i following (2.1). On the other hand, when the complete model is from the exponential family then the E–step computes the conditional expectations of its sufficient statistics. As it can be seen below, the continuous distribution given in (2.1) is a member of the exponential family of probability distributions since it can be written as

$$f(\theta|\phi) = h(\theta)\, \exp\left(A(\phi)\, T(\theta) - B(\phi)\right) \quad \text{where}$$

$h(\theta) = \dfrac{1}{\sqrt{2\theta}}$, $A(\phi) = -\sqrt{2\phi}$, $T(\theta) = \sqrt{\theta}$ and $B(\phi) = -\log\sqrt{\phi}$. Then, $T(\theta)$ is a sufficient statistic of this distribution.

The EM type algorithm for this model can be described as follows. From the current estimates $\phi^{(j)}$

- **E–step:** Calculate the pseudo–values

$$t_i = E(\sqrt{\theta_i} \mid Y_i, \hat{\phi}^{(j)})$$

  for $i = 1, \dots, n$.

- **M–step:** Find the new estimates $\hat{\phi}^{(j+1)}$

$$\hat{\phi}^{(j+1)} = \frac{1}{2}\left(\frac{n}{\sum_{i=1}^{n} t_i}\right)^2.$$

- If some convergence condition is satisfied then stop iterating, otherwise move back to the E–step for another iteration.

3 The regression model

Let us now consider a random variable Y_i denoting event counts and a vector of covariates or explanatory variables $\mathbf{x_i} = (x_{i1}, \dots, x_{ip})^t$, including an intercept, related to the *i*-th

observation that denotes a weight of observable features. In this model with fixed effects, it is assumed that

$$\begin{aligned} Y_i|\theta_i &\sim \mathcal{P}(\theta_i\mu_i) \\ \theta_i &\sim \mathcal{EIG}(\phi) \\ \mu_i &= \exp(\mathbf{x_i}^t\beta), \end{aligned} \tag{3.1}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^t$ a vector of regression coefficients.

The $\mathcal{PEIG}$ distribution has mean $\mu = 1/\phi$ and variance $1/\phi + 5/\phi^2$. If we parameterize $\mu_i = 1/\phi = \exp(\mathbf{x_i}^t\beta)$, the marginal mean and the marginal variance of the response distribution distribution are given by

$$\begin{aligned} E(Y_i|x_i) &= \exp(\mathbf{x_i}^t\beta) \text{ and} \\ var(Y_i|x_i) &= E(Y_i|x_i) + 5E(Y_i|x_i)^2, \end{aligned}$$

respectively.

Likewise the conditional mean of the response variable is related to the explanatory variables through a link function, $g(E(Y_i|x_i)) = \mathbf{x_i}^t\beta$, where $g(\cdot)$ is a monotonic function. The link function determines the function of the conditional mean that is predicted by $\mathbf{x_i}^t\beta$. As the mean of (2.5) is non-negative, the log–link is the usual choice for $\mathcal{PEIG}$ regression model since it guarantees a non-negative value for the conditional mean. Additionally, as $var(Y_i|x_i) > E(Y_i|x_i)$, this mixed Poisson regression model is over–dispersed. In addition to this, as the variance is determined by the mean, no additional variance estimate is required. Besides, this model does not nest the Poisson regression model. Maximum likelihood estimation for this fixed effect regression model involves setting the partial derivatives of the log–likelihood function with respect to regression coefficients β_j with $j = 1, \ldots, p$ equal to zero.

Let us now suppose that $(y_i, \mathbf{x_i})$, $i = 1, \ldots, n$ are n independent realizations of the regression model given in (3.1) where y_i is the response variable and $\mathbf{x_i}$ a vector of explanatory variables. Then, the log–likelihood function can be expressed as

$$\begin{aligned} \ell(\beta_1, \ldots, \beta_p) &= \sum_{i=1}^{n} \ell_i(\mu_i; \beta_1, \ldots, \beta_p) \\ &= -\frac{n}{2}\log\mu_i + \sum_{i=1}^{n}\log\Gamma(2y_i+1) - \left(2\sum_{i=1}^{n} y_i + \frac{n}{2}\right)\log 2 \\ &- \sum_{i=1}^{n}\log y_i! + \sum_{i=1}^{n}\log\mathcal{U}\left(\frac{1}{2}+y_i, \frac{1}{2}, \frac{1}{2\mu_i}\right). \end{aligned} \tag{3.2}$$

Then, the normal equations to obtain the maximum likelihood estimates are given by

$$\frac{\partial\ell}{\partial\beta_s} = \frac{n}{2}\sum_{i=1}^{n} x_{is} + \sum_{i=1}^{n}\left(\frac{1}{2}+y_i\right)\frac{x_{is}}{2}\frac{1}{\mu_i}\frac{\mathcal{U}\left(\frac{3}{2}+y_i, \frac{3}{2}, \frac{1}{2\mu_i}\right)}{\mathcal{U}\left(\frac{1}{2}+y_i, \frac{1}{2}, \frac{1}{2\mu_i}\right)},$$

with $s = 1, 2, \ldots, p$.

Furthermore, the required expressions to approximate the Fisher's information matrix associated with maximum–likelihood estimates are provided by

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta_s \partial \beta_k} &= -\sum_{t=1}^{n} \left(\frac{1}{2}+y_i\right) \frac{x_{is}x_{ik}}{2} \left(\frac{1}{2}+y_i\right) \frac{1}{\mu_i} \\ &\times \frac{\left[\mathcal{U}\left(\frac{3}{2}+y_i, \frac{3}{2}, \frac{1}{2\mu_i}\right) - \left(\frac{3}{2}+y_i\right)\mathcal{U}\left(\frac{5}{2}+y_i, \frac{5}{2}, \frac{1}{2\mu_i}\right)\right]}{\mathcal{U}\left(\frac{1}{2}+y_i, \frac{1}{2}, \frac{1}{2\mu_i}\right)} \\ &+ \left[\left(\frac{1}{2}+y_i\right) \frac{x_{is}x_{ik}}{2} \left(\frac{1}{2}+y_i\right) \frac{1}{\mu_i} \frac{\mathcal{U}\left(\frac{3}{2}+y_i, \frac{3}{2}, \frac{1}{2\mu_i}\right)}{\mathcal{U}\left(\frac{1}{2}+y_i, \frac{1}{2}, \frac{1}{2\mu_i}\right)}\right]^2,\end{aligned}$$

for $s = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, p$.

4 Application to health service data

4.1 Estimation of parameters

In the following, we are going to illustrate the performance of this mixed Poisson regression model. For that reason, let us consider now the number of hospital stays among the elderly population age 65 and over in the U.S. This amount represents a significant portion of the annual expenditures on hospital care since government insurance programs in the U.S. bear the highest financial burden for health care. Moreover, it has been forecasted that the number of elderly will continue to grow in the coming years. This set of data appears originally in Deb and Trivedi (1997) in their analysis of various measures of health–care utilization using a sample of 4406 single–person households in 1987. Data have been obtained from the Journal of Applied Econometrics 1997 Data Archive. Estimation of model and all the data analyses were done using *Mathematica 9.0* software package. All the codes used to obtain reported results and all additional information useful to make research reproducible can be found on the journal's website or it will be made available by the authors on request. Our goal is to model the number of hospital stays (HOSP) as the response variable. This measure includes two interesting features, on the one hand over–dispersion, the mean and variance of the empirical distribution are 0.30 and 0.56 respectively, and, on the other hand, a very high proportion of non–users (80.36%). Since the Poisson regression model is not able to capture the the heterogeneity among individuals found in the data, the $\mathcal{PEIG}$ regression model is used to explain the demand for health services.

Let us firstly considered the model without covariates. Parameter estimates, standard errors (in brackets) and the maximum of the log–likelihood ($\ell_{\max}$) of the distribution of the hospital stays are $\widehat{\theta} = 0.296\,(0.01)$ and $\ell_{max} = -3304.51$ for Poisson model and $\widehat{\mu} = 0.308\,(0.01)$, $\ell_{max} = -3021.92$ for $\mathcal{PEIG}$ model respectively. For the latter model, the estimate can also be obtained by using the EM algorithm after 25 iterations when the relative change of the estimate between two successive iterations is smaller than 1×10^{-10},

after taking initial starting value in the neighborhood of the moment estimate. Therefore, it can be concluded that the $\mathcal{PEIG}$ model provides a better fit to the data than Poisson distribution by considering maximum of the log–likelihood as criterion of comparison. For the standard model given in (2.5) the estimated value of ϕ is 3.24783 with a standard error of 0.138. Since the empirical distribution is over-dispersed the Poisson model seems to be inadequate for estimating these count data. Next, in Figure 1 the histogram of the empirical distribution of the number of hospital stays (Observed), together with fitted distribution, obtained from the Poisson distribution and $\mathcal{PEIG}$ distribution has been plotted. As it can be observed, there is a clear spike of extra zeros representing the non-hospitalization of the elderly population with the best fit to the data obtained with the $\mathcal{PEIG}$ model.

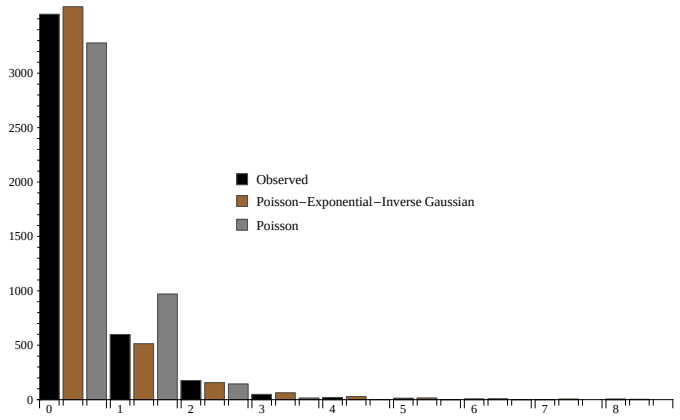

Figure 1: Observed and fitted ($\mathcal{PEIG}$ and Poisson) distribution of the number of hospital stays (HOSP)

Let us now analyze the model with covariates. The explanatory variables are as follows: (1) a dummy variable (EXCLHLTH) which takes the value 1 if self–perceived health is excellent; (2) a dummy variable (POORHLTH) which takes the value 1 if self–perceived health is poor; (3) a count variable (NUMCHRON) giving the number of chronic disease and condition (cancer, heart attack, etc.); (4) age (AGE) divided by 10; (5) a dummy variable (MALE) with value 1 if the patient is male. For the ith patient, the number of hospital stays Y_i follows a $\mathcal{PEIG}$ whose mean depends on a set of covariates trough the log–link function. The goal is to predict the number of hospital stays Y_i (response variable) using a vector of explicative variables (covariates).

At first sight, it seems logical that due to the presence of over–dispersion, a relative large long right tail, and a high proportion of zeros as compared to the proportion of other values, a simple Poisson regression model is not adequate to explain the number of hospital stays since it tends to overestimate the probability of lower values and underestimate the probability of larger values. For that reason, it is expected that a a mixed Poisson regression model will describe in a more accurate way the right tail of empirical data and the high proportion of zeros in the sample. As it can be observed in Table 1, the $\mathcal{PEIG}$ and a Poisson (in brackets) regression model have been fitted to data. From left to right parameter estimates, standard errors, t-Wald and p-values are shown for both models. After

observing the values of the estimated regressors, there exists some differences between estimated effects of both models. In this sense, the $\mathcal{PEIG}$ regression model predicts a higher use of the health service when self–perceived health is poor, the number of chronic disease and condition and age increases and the patient is male. Furthermore, when self–perceived health is excellent then the predicted change in the number of hospital stays decreases at a lower rate than in the Poisson regression model. The intercept coefficient -3.959 is the predicted logarithm of the number of hospital stays when the values of EXCLHLTH, POORHLTH, NUMCHR, AGE and MALE are equal to 0. Having said that, it can be concluded, from this numerical application, that the $\mathcal{PEIG}$ regression model predicts a higher use of the health service for this set of explanatory variables. All of parameter estimates are significant at the usual nominal level.

Table 1: Parameter estimates, standard errors, t-Wald and p-values for $\mathcal{PEIG}$ and Poisson (in brackets) regression models for the number of hospital stays.

Parameter	Estimate	S.E.	t-Wald	$\Pr > \|t\|$
INTERCEPT	–3.959(–3.220)	0.52(0.32)	–7.63(–10.19)	0.00(0.00)
EXCLHLTH	–0.688(–0.720)	0.22(0.18)	–3.15(–4.10)	0.00(0.00)
POORHLTH	0.683(0.613)	0.12(0.07)	5.60(9.18)	0.00(0.00)
NUMCHRON	0.326(0.264)	0.03(0.02)	9.72(14.48)	0.00(0.00)
AGE	0.268(0.183)	0.07(0.04)	3.93(4.39)	0.00(0.00)
MALE	0.196(0.109)	0.10(0.06)	2.17(1.94)	0.03(0.05)

Following the work of Wedderburn (1974), we have also estimated the parameters by using a quasi–likelihood model. In this case, we need only to specify the marginal response variance in terms of the marginal mean, i.e. $var(Y_i) = \mu_i + 5\mu_i^2$, $(i = 1, \ldots, n)$. Via quasi–likelihood estimation, the estimates are very close to the ones shown in Table 1. Note that they are given in the same order as in Table 1, that is, –3.92958, –0.679321, 0.605773, 0.307492, 0.262405 and 0.187604. The value of the negative of the maximum of the log–likelihood is 2896.79.

4.2 Model assessment

Several measures of model validation to compare the $\mathcal{PEIG}$ and Poisson regression model are shown in Table 2. Firstly, the value of the negative of the maximum of the log–likelihood (NLL) and Akaike Information Criterion (AIC) are given in the first two rows of this Table; as a lower value of these measures is desirable, the $\mathcal{PEIG}$ regression model is preferable. Bozdogan (1987) proposed a corrected version of AIC, the Consistent Akaike Information Criteria (CAIC), in an attempt to overcome the tendency of the AIC to overestimate the complexity of the underlying model. Bozdogan (1987) also observed that AIC does not directly depend on the sample size and, as a result, it lacks certain properties of asymptotic consistency. See also Anderson et al. (1998). When formulating the CAIC, a correction factor based on the sample size is used to compensate for the overestimating nature of AIC. The CAIC is defined as $\text{CAIC} = 2\,\text{NLL} + (1 + \log n)\, p$,

where p refers to the number of estimated parameters and n is the sample size. Again, a model that minimize the Consistent Akaike Information Criteria is preferable. As it can be observed, the $\mathcal{PEIG}$ regression model also dominates the Poisson regression model in terms of the CAIC.

Table 2: Measures of model selection for the models considered.

	Distribution	
Criterion	Poisson	$\mathcal{PEIG}$
NLL	3047.32	2895.11
AIC	6116.63	5802.22
CAIC	6150.98	5846.57
Pearson statistic, $(\epsilon_i^P)^2$	7071.90	4626.74
Deviance residual/df	–0.30183	–0.33572

Now we perform some diagnostic checks based on analysis of residuals. This is a useful method to detect outliers and check the variance assumption in a more general setting (see Cameron and Trivedi (1986), for details). Perhaps the most common choice is Pearson's residuals. They are used to identify discrepancies between models and data, and they are based upon differences between observed data points and fitted values predicted by the model. The *i*-th Pearson residual for a given model is provided by

$$\epsilon_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{var(\widehat{\mu}_i)}}, \tag{4.1}$$

where $\widehat{\mu}_i$ is the fitted marginal mean and $var(\widehat{\mu}_i)$ is the estimated marginal variance under the discussed model. Hence, if the model is correct, the variability of these residuals should appear to be fairly constant, when they are plotted against fitted values or predictors. The Pearson's residuals are often skewed for non–normal data, and this make the interpretation of the residual plots more difficult to interpret. For that reason, other quantifications of the discrepancy between observed and fitted values have been suggested in the literature. In this regard, another choice in the analysis of residual is the signed square root of the contribution to the deviance goodness–of–fit statistic (i.e. deviance residuals). This is given by $D = \sum_{i=1}^{n} d_i$, where

$$d_i = \text{sgn}(y_i - \widehat{\mu}_i)\sqrt{2(\ell(y_i) - \ell(\widehat{\mu}_i))}, \quad i = 1, 2, \ldots, n,$$

and sgn is the function that returns the sign (plus or minus) of the argument. The $\ell(y_i)$ term is the value of the log likelihood when the mean of the conditional distribution for the *i*-th individual is the individual's actual score of the response variable. The $\ell(\widehat{\mu}_i)$ is the log–likelihood when the conditional mean is plugged into the log–likelihood. Usually the deviance divided by its degree of freedom is examined by taking into account that a value much greater than one indicates a poorly fitting model. See for example De Jong and Heller (2008).

It is well–known that for the Poisson distribution with parameter θ_i the deviance residuals are given by (see Dunteman and Ho 2006))

$$d_i = \text{sgn}(y_i - \widehat{\theta}_i) \left[2 \left(y_i \log \left(\frac{y_i}{\widehat{\theta}_i} \right) - (y_i - \widehat{\theta}_i) \right) \right]^{1/2}, \quad i = 1, 2, \ldots, n. \tag{4.2}$$

For the model introduced in this manuscript the deviance residuals are easily obtained by

$$d_i = \text{sgn}(y_i - \widehat{\mu}_i) \left\{ 2 \left[\log \left(\frac{\mathcal{U}(0.5 + y_i, 0.5, (2y_i)^{-1})}{\mathcal{U}(0.5 + \widehat{\mu}_i, 0.5, (2\widehat{\mu}_i)^{-1})} \right) - \frac{1}{2} \log \left(\frac{y_i}{\widehat{\mu}_i} \right) \right] \right\}^{1/2}, \quad i = 1, 2, \ldots, n.$$

Note that the deviance does not exist whenever there are zero responses in the data. However, it is usually assumed that $d_i = 0$ when $y_i = 0$ (e.g. $y_i \log y_i$ is zero for $y_i = 0$). The Pearson's statistics together with the deviance residual divided by the degree of freedom are shown in Table 2. The $\mathcal{PEIG}$ dominates widely the Poisson distribution in terms of the Pearson's statistics and small differences appear in the value of the deviance residual. Recall that we have taken this value as zero when the observed response variable takes the value zero.

Graphical model diagnostic may also be developed using expression (4.1). In this case, for the Poisson regression model this reduces to $\epsilon_i^P = (y_i - \widehat{\theta}_i)/\sqrt{\widehat{\theta}_i}$, while for the distribution $\mathcal{PEIG}$ regression model, this expression is given by $\epsilon_i^P = (y_i - \widehat{\mu}_i)/\sqrt{\widehat{\mu}_i(1 + 5\widehat{\mu}_i)}$ as it can be easily verified. For this example, not much differences are found between these plots and those ones produced by the raw residuals, $y_i - \widehat{\theta}_i$, which are shown in Figure 2. On the other hand, the Pearson's residuals are usually standardized by dividing by $\sqrt{1 - h_i}$, where h_i are the leverages obtained from the diagonal of the hat matrix $W^{1/2}X(X'WX)^{-1}X'W^{1/2}$, being W equal to the $n \times n$ diagonal matrix with i–th entry w_i, given by $w_i = (\partial\theta_i/\partial x'\beta))^2 / var(Y_i)$. This results θ_i for the Poisson regression model and $\mu_i/(1 + 5\mu_i^2)$ for the regression based on the new distribution presented here. See Cameron and Trivedi (1986) for details about the construction of the hat matrix. The standardized Pearson's residuals have also been plotted, they are shown in Figure 3. As it can be seen, for the Poisson regression model many of the values of the Pearson's standardized residuals lie outside the range $(-2, 2)$, pointing out a poorer fit to data than the

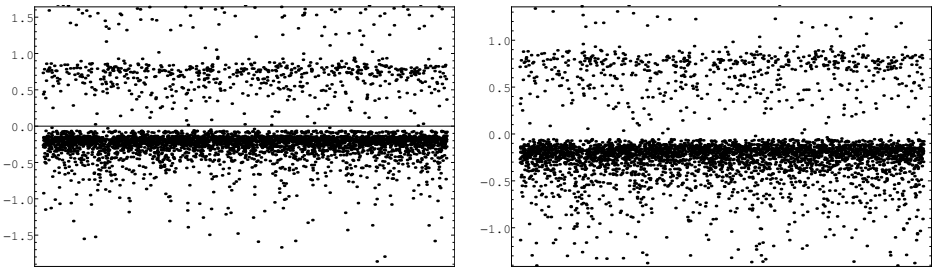

Figure 2: Plots of the raw residuals for the Poisson (left) and the $\mathcal{PEIG}$ (right) regression models.

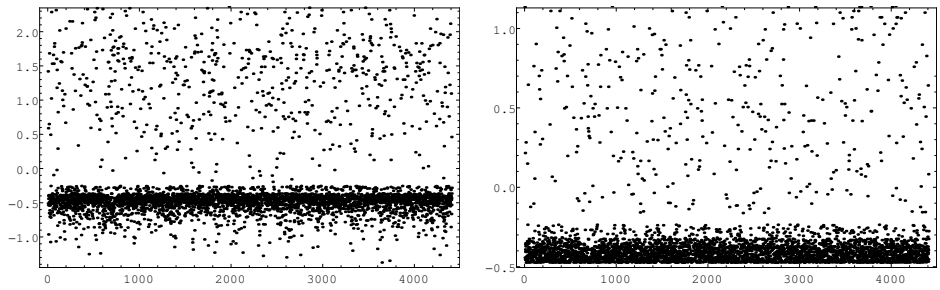

Figure 3: Standardized Pearson's residuals for the Poisson (left) and the $\mathcal{PEIG}$ (right) distributions

one obtained for the $\mathcal{PEIG}$ regression model presented in this work. See Hilbe (2007) for details.

In the following, as the regression model introduced in this paper is not nested in the Poisson regression model, the Vuong's test can be used to compare the estimates of the Poisson regression model and $\mathcal{PEIG}$ regression model. In this regard, one might be interested in testing the null hypothesis that the two models are equally close to the actual model, against the alternative one that one of the model is closer (see Vuong (1989)). The z-statistic is

$$Z = \frac{1}{\omega\sqrt{n}}\left(\ell(\widehat{\mu}) - \ell(\widehat{\theta})\right),$$

where

$$\omega^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{f(\widehat{\mu})}{g(\widehat{\theta})}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(\widehat{\mu})}{g(\widehat{\theta})}\right)\right]^2$$

and f and g represent here the $\mathcal{PEIG}$ and Poisson distributions, respectively.

Due to the asymptotic normal behaviour of the Z statistic under the null hypothesis, rejection of null hypothesis in favour of the alternative one that f occurs with significance level α, when $Z > z_{1-\alpha}$ being $z_{1-\alpha}$ the $(1-\alpha)$ quantile of the standard normal distribution. For the Vuong's test, $Z = 3.95754$, then the $\mathcal{PEIG}$ model is preferred at the usual nominal levels.

4.3 Comparisons with other models

Finally the fit obtained with the $\mathcal{PEIG}$ regression model is compared to two other mixed Posisson regression models traditionally used in the statistical literature, the negative binomial and the Poisson–Inverse Gaussian regression models (see Dean et al. (1989)). Furthermore, when the empirical data includes a high presence of zeros it is usual to consider a reparameterization of the parent distribution to capture all zeros in the sample, the zero–inflated (ZI) model. If the parent distribution is $p(x)$, a ZI distribution is built as follows (see Cohen (1966))

$$p(x) = \begin{cases} (1-\psi) + \psi\, p(0), & x = 0, \\ \psi\, p(x), & x > 0, \end{cases}$$

where $p(x)$ is the parent distribution and $0 < \psi \leq 1$ is the inflated parameter. The $\mathcal{PEIG}$, negative binomial and Poisson–Inverse Gaussian distributions have been reparameterized to obtain the maximum likelihood estimates under the ZI model and the results, together with the homogeneous models (without inflation), are displayed in Table 3.

Table 3: Maximum of the log–likelihood and Consistent Akaike Information Criteria (CAIC) for different homogeneous and ZI models.

	Homegeneous		ZI	
Distribution	NLL	CAIC	NLL	CAIC
$\mathcal{PEIG}$	2895.11	5846.57	2851.90	5769.74
NB	2857.11	5779.95	2853.37	5781.87
PIG	2877.33	5820.40	2847.69	5770.51

As it can be seen in this Table, the (ZI) $\mathcal{PEIG}$ regression model provides the best fit to data for this particular dataset when the CAIC is used as a criterion of comparison since the other two mixed Poisson regression models include an additional parameter. Since the global maximum of the log–likelihood surface is not guaranteed, different initial values of the parametric space were considered as a seed point. The calculations have been completed by using the `FindMaximum` function of Mathematica software package v.9.0 (Wolfram (2003)) (the derivative of the modified Bessel function of the third kind is available in this package). Additionally, by using other different methods such as `Newton`, `PrincipalAxis` and `QuasiNewton` the same results were obtained.

5 Conclusions

In this paper, a new mixed Poisson regression model to explain the demand for health services among people 65 and over to account for a large portion of non–users has been proposed. This model has been derived by mixing the Poisson distribution with a particular case of the continuous Exponential–Inverse Gaussian distribution when one of its parameter tends to infinity. Additionally, it is over–dispersed and unimodal with modal value located at zero. The model might be considered an alternative to Poisson regression model when the empirical data include a high proportion of zeros. In this regard, several measures of model assessment, including the Vuong's test for non-nested model selection, have been provided to support this goal. Apart from that, due to the high proportion of zeros in the empirical data, a zero–inflated version of this model has also been used to explain the demand for health services of elderly people.

Acknowledgment

The authors would like to thank the editor and two anonymous referees for their relevant and useful comments. Research partially funded by grant by grant ECO2013-47092 (Ministerio de Economía y Competitividad, Spain).

References

[1] Anderson, D.R., Burnham, K.P. and White, G.C. (1998). Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture–recapture studies. *Journal of Applied Statistics*, **25**, 2, 263–282.

[2] Bhattacharya, S.K., Kumar, S. (1986): E–IG model in life testing. *Calcutta Statistical Association Bulletin*, **35**, 85–90.

[3] Bozdogan, H. (1987): Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**, 3, 345–370.

[4] Cameron, A.C. and Trivedi, P.K. (1986): Econometric models based on counts data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, **1**, 1, 29–53.

[5] Cohen, A. C. (1966): A note on certain discrete mixed distributions. Biometrics, **22**, 3, 566–572.

[6] De Jong, P. and Heller, G.H. (2008): *Generalized Linear Models for Insurance Data*. Cambridge University Press.

[7] Dean, C.B., Lawless, J., and Willmot, G.E. (1989): A mixed Poisson–inverse Gaussian regression model. *Canadian Journal of Statistics*, **17**, 2, 171–181.

[8] Deb, P. and Trivedi, P.K. (1997): Demand for Medical Care by the Elderly: A Finite Mixture Approach. *Journal of Applied Econometrics*, **12**, 3, 313–336.

[9] Dunteman, G.H. and Ho, M–H.R. (2006): *An Introduction to Generalized Linear Models*. SAGE Publications.

[10] Gradshteyn, I.S., Ryzhik, I.M. (1994): Table of Integrals, Series, and Products. Alan Jeffrey, Editor. Fifth Edition. Academic Press, Boston.

[11] Greene, W. (2009): Models for count data with endogenous participation. *Empirical Economics*, **36**, 133–173.

[12] Gurmu, S. and Elder, J. (2000): Generalized bivariate count data regression models. *Economics Letters*, **68**, 31–36.

[13] Hilbe, J.M. (2007): *Negative Binomial Regression.* New York: Cambridge University Press.

[14] Holgate, P. (1970): The modality of some compound Poisson distribution. *Biometrika*, **57**, 666–667.

[15] Karlis, D. (2005): EM algorithm for mixed Poisson and other discrete distributions. *Astin Bulletin*, **35**, 3–24.

[16] Karlis, D. and Xekalaki, E. (2005): Mixed Poisson distributions. *International Statistical Review*, **73**, 35–58.

[17] Lahiri, K. and Xing, G. (2004): An econometric analysis of veterans' health care utilization using two–part models. *Empirical Economics*, **29**, 431–449.

[18] Vuong, Q. (1989): Likelihood ratio tests for model selection and non–nested hypotheses. *Econometrica*, **57**, 307–333.

[19] Wedderburn, R.W.M. (1974): Quasi–likelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika*, **61**, 439–447.

[20] Wolfram, S. (2003): *The Mathematica Book*. Wolfram Media, Inc.

Metodološki zvezki, Vol. 13, No. 2, 87-100

X bar control chart for non-normal symmetric distributions

Kristina Veljkovic [1]

Abstract

In statistical quality control, X bar control chart is extensively used to monitor a change in the process mean. In this paper, X bar control chart for non-normal symmetric distributions is proposed. For chosen Student, Laplace, logistic and uniform distributions of quality characteristic, we calculated theoretical distribution of standardized sample mean and fitted Pearson type II or type VII distributions. Width of control limits and power of the X bar control chart were established, giving evidence of the goodness of fit of the corresponding Pearson distribution to the theoretical distribution of standardized sample mean. For implementation of X bar control chart in practice, numerical example of construction of a proposed chart is given.

1 Introduction

The X bar chart is extensively used in practice to monitor a change in the process mean. It is usually assumed that measured quality characteristic has normal or approximately normal distribution. On the other hand, occurrence of non-normal data in industry is quite common (see Alloway and Raghavachari, 1991; Janacek and Meikle, 1997). Violation of normality assumption results in incorrect control limits of control charts (Alwan, 1995). Misplaced control limits lead to inappropriate charts that will either fail to detect real changes in the process or which will generate spurious warnings when the process has not changed.

In the case of non-normal symmetric distribution of quality characteristics, no recommendations, except the use of the normal distribution, are given in the quality control literature. Approximation of the distribution of sample mean with normal distribution is based on the central limit theorem, but in practice small sample sizes are usually used.

We will consider four types of non-normal symmetric distributions of quality characteristic: Student, Laplace, logistic and uniform distributions. These distributions are chosen because of their applications in various disciplines (economics, finance, engineering, hydrology, etc., see for instance Ahsanullah, et al., 2014; Balakrishnan, 1992; Kotz et al., 2001). For each of these distributions, we calculated theoretical distribution of the standardized sample mean (or its best approximation) and approximated it with Pearson type II or type VII distributions. Pearson system of distributions is known to provide approximations to a wide variety of observed distributions (Johnson et al., 1994).

[1] Department of Probability and Statistics, Faculty of Mathematics, University of Belgrade, Serbia; kristina@matf.bg.ac.rs

It is presumed that a process begins in in-control state with mean μ_0 and that single assignable cause of magnitude δ results in a shift in the process mean from μ_0 to either $\mu_0 - \delta\sigma$ or $\mu_0 + \delta\sigma$, where σ is the process standard deviation (Montgomery, 2005). It is also assumed that the standard deviation remains stable. Center line of the X bar chart is set at μ_0 and upper and lower control limits, respectively, $\mu_0 + k\sigma/\sqrt{n}$ and $\mu_0 - k\sigma/\sqrt{n}$, where n represents the sample size and k width of control limits. Samples of size n are taken from the process and the sample mean is plotted on the X bar chart. If a sample mean exceeds control limits, it is assumed that some shift in the process mean has occurred and a search for the assignable cause is initiated.

The rest of the paper is organized as follows. In Sections 2, 3 and 4, respectively, descriptions of chosen distributions of quality characteristic, distributions of standardized sample mean and Pearson types II and VII distributions are given. Construction of the X bar control chart and its power are examined in Section 5, along with the comparisons of theoretical distribution of sample mean with the corresponding Pearson distribution. In Section 6, implementation of proposed X bar chart is considered. Finally, conclusions are drawn in Section 7.

2 Distribution of quality characteristic

We considered four types of non-normal symmetric distributions of quality characteristic X: Student distribution $t(10)$, standard Laplace $L(1)$ distribution and logistic distribution $LGS(1)$ (see Johnson et al. 1994; Johnson et al. 1995) as representatives of symmetric distributions with heavier tails than normal distribution (Figure 1) and uniform $U(0,1)$ distribution as a representative of symmetric distributions with lighter tails than normal distribution. For simplicity, we have chosen standard forms of all four distributions.

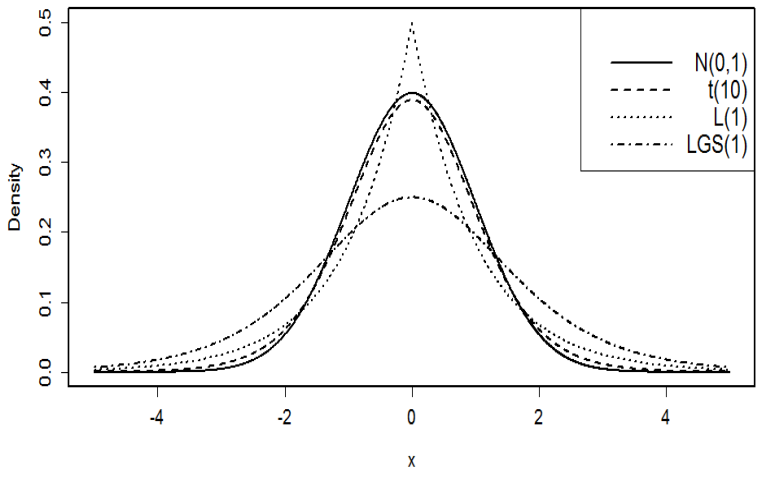

Figure 1: Probability density functions of Student $t(10)$, Laplace $L(1)$, logistic $LGS(1)$ and standard normal $N(0,1)$ distributions

Distribution	f_X	μ	σ^2	α_4
$t(10)$	$\frac{315}{256\sqrt{10}}\left(1+\frac{x^2}{10}\right)^{-5.5},\ x\in\mathbb{R}$	0	1.25	4
$L(1)$	$\frac{1}{2}e^{-\lvert x\rvert}\ x\in\mathbb{R}$	0	2	6
$LGS(1)$	$\frac{e^{-x}}{(1+e^{-x})^2}\ x\in\mathbb{R}$	0	$\frac{\pi^2}{3}$	4.2
$U(0,1)$	$x,\ x\in[0,1]$	0.5	$\frac{1}{12}$	1.8

Table 1: Chosen distributions of quality characteristics

Distributions are given in Table 1 by their probability density function f_X, mean μ, variance $\sigma^2 = Var(X)$ and kurtosis $\alpha_4 = \frac{E(X-E(X))^4}{\sigma^4}$. As all chosen distributions are symmetric around the zero, skewness $\alpha_3 = \frac{E(X-E(X))^3}{\sigma^{\frac{3}{2}}} = 0$.

3 Distribution of standardized sample mean

For chosen distributions of quality characteristic, we will derive the distribution of standardized sample mean $T_n = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$. As all chosen distributions are symmetric, skewness of standardized sample mean will also be equal to 0.

3.1 Sample from Student's distribution

Witkowský (2001, 2004) proposed a method for numerical evaluation of the distribution function of a linear combination of independent Student variables. The method is based on the inversion formula which leads to the one-dimensional numerical integration.

Let $(X_1, X_2, \ldots, X_n)$ be a sample from Student $t(\nu)$ distribution. Further, let $Y = \sum_{k=1}^{n} X_k$ be sum of these variables and $\phi_{X_k}(t)$ denote the characteristic function of X_k. The characteristic function of Y is

$$\phi_Y(t) = \prod_{k=1}^{n} \phi_{X_k}(t) = \prod_{k=1}^{n} \frac{1}{2^{\frac{\nu}{2}-1}\Gamma\left(\frac{\nu}{2}\right)} \left(\nu^{\frac{1}{2}}|t|\right)^{\frac{\nu}{2}} K_{\nu/2}\left(\nu^{\frac{1}{2}}|t|\right),$$

where $K_\alpha(z)$ denotes modified Bessel function of the second kind.

The cumulative distribution function $F_Y(y)$ of random variable Y is, according to the inversion formula due to Gil-Pelaez (1951), given by

$$F_Y(y) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{\sin(ty)\phi_Y(t)}{t}\,\mathrm{d}t \tag{3.1}$$

For any chosen y algorithm $tdist$ in R package $tdist$ (Witkowský and Savin, 2005) evaluates the integral in (3.1) by multiple p-points Gaussian quadrature over the real interval $t \in (0, 10\pi)$. The whole interval is divided in m subintervals and the integration over each subinterval is done with p-points Gaussian quadrature which involves base points

b_{ij}, and weight factors w_{ij}, $i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, m$. So,

$$F_Y(y) \approx \frac{1}{2} + \frac{1}{\pi} \sum_{j=1}^{m} \sum_{i=1}^{p} \frac{\sin(b_{ij} y)}{b_{ij}} w_{ij} \phi_Y(b_{ij}).$$

Then, cumulative distribution function of standardized sample mean is equal to

$$F_{T_n}(t) = F_Y\left(\frac{\sqrt{5n}}{2} t\right),\ t \in \mathbb{R}.$$

Kurtosis of T_n is equal to $\alpha_{4,T_n} = 3 + \frac{1}{n}$

3.2 Sample from Laplace distribution

Let $(X_1, X_2, \ldots, X_n)$ be a sample from standard Laplace $L(1)$ distribution. Difference of two independent random variables with standard exponential $\varepsilon(1)$ distribution has standard Laplace distribution. Further, standard exponential distribution is gamma distribution, $\Gamma(1, 1)$. Sum of n independent variables with $\Gamma(1, 1)$ distribution is gamma distribution $\Gamma(n, 1)$. In that way, we conclude that sum Y of n independent random variables $X_1, X_2, \ldots, X_n$ with standard Laplace distribution can be written as the difference of two random variables with gamma distribution $\Gamma(n, 1)$ which is called bilateral gamma distribution.

Bilateral gamma distribution is symmetric around 0 (Küchler and Tappe, 2008), with cumulative distribution function for $y > 0$

$$F_Y(y) = \frac{1}{2} + \frac{1}{2^n} \cdot \frac{1}{(n-1)!} \sum_{k=0}^{n} a_k \gamma(k+1, y)$$

where the coefficients $(a_k)_{k=0,\ldots,n-1}$ are given by

$$a_k = \binom{n-1}{k} \frac{1}{2^{n-1-k}} \prod_{l=0}^{n-2-k} (n+l),\ a_{n-1} = 1.$$

and $\gamma(n, y)$ is incomplete gamma function.

Then, cumulative distribution function of standardized sample mean is equal to

$$F_{T_n}(t) = F_Y\left(\sqrt{2n} t\right),\ t \in \mathbb{R}.$$

Kurtosis of standardized sample mean is equal $\alpha_{4,T_n} = 3 + \frac{3}{n}$.

3.3 Sample from logistic distribution

Let $(X_1, X_2, \ldots, X_n)$ be a random sample from logistic $LGS(1)$ distribution. Insofar, the best approximation of the distribution of standardized sample mean T_n is given by Gupta and Han (1992). They considered the Edgeworth series expansions up to order n^{-3} for

the distribution of the standardized sample mean. Cumulative distribution function of T_n is given by

$$F_{T_n}(t) \approx \Phi(t) - \varphi(t)\left(\frac{1}{n}\left(\frac{1}{4!}\frac{6}{5}H_3(t)\right) + \frac{1}{n^2}\left(\frac{1}{6!}\frac{48}{7}H_5(t) + \right.\right.$$
$$+ \left.\left.\frac{35}{8!}\left(\frac{6}{5}\right)^2 H_7(t)\right)\right) + \frac{1}{n^3}\left(\frac{1}{8!}\frac{432}{5}H_7(t) + \frac{210}{10!}\frac{48}{7}\frac{6}{5}H_9(t) + \right.$$
$$+ \left.\frac{5775}{12!}\left(\frac{6}{5}\right)^3 H_{11}(t)\right)\right),\ t \in \mathbb{R},$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are standard normal pdf and cdf and $H_j(x)$ is the Hermite polynomial.

Kurtosis of standardized sample mean is $\alpha_{4,T_n} = 3 + \frac{1.2}{n}$.

3.4 Sample from uniform distribution

Let $(X_1, X_2, \ldots, X_n)$ be a random sample from uniform $U(0,1)$ distribution. The sum $Y = \sum_{k=1}^{n} X_k$ has Irwin-Hall distribution (Johnson et al., 1995) with cumulative distribution function

$$F_Y(y) = \frac{1}{2} + \frac{1}{2n!}\sum_{k=0}^{n}(-1)^k\binom{n}{k}\operatorname{sgn}(y-k)(y-k)^n,\ x \in \mathbb{R}.$$

Then, standardized sample mean has cumulative distribution function equal to

$$F_{T_n}(t) = F_Y\left(\left(\frac{t}{\sqrt{12n}} + \frac{1}{2}\right)n\right),\ t \in \mathbb{R}.$$

Kurtosis of standardized sample mean is $\alpha_{4,T_n} = 3 - \frac{1.2}{n}$.

4 Symmetric Pearson distributions

4.1 Pearson type II distribution

Pearson type II distribution can be used for approximation of the distribution of random variable with skewness $\alpha_3 = 0$ and kurtosis $\alpha_4 < 3$ (Johnson et al., 1994). Cumulative distribution function of Pearson type II distribution is equal to

$$F(t) = I_{\frac{t-\lambda}{s}}(a,a),\ 0 < \frac{t-\lambda}{s} < 1,$$

where

$$\lambda = -\sqrt{\frac{2\alpha_4}{3-\alpha_4}},\quad s = 2\sqrt{\frac{2\alpha_4}{3-\alpha_4}},\quad a = \frac{5\alpha_4 - 9}{2(3-\alpha_4)} + 1, \tag{4.1}$$

$I_t(a,b) = \frac{B_t(a,b)}{B(a,b)}$, $B(a,b)$ is beta function and $B_t(a,b)$ is incomplete beta function.

In other words, random variable $\frac{T-\lambda}{s}$ has beta distribution $\mathcal{B}(a,a)$.

4.2 Pearson type VII distribution

Pearson type VII distribution can be used for approximation of the distribution of random variable with skewness $\alpha_3 = 0$ and kurtosis $\alpha_4 > 3$ (Johnson et al., 1994). Cumulative distribution function of Pearson type VII distribution is equal to

$$F(t) = \frac{1}{2} I_{a^2/(a^2+t^2)} \left(m - \frac{1}{2}, \frac{1}{2} \right), \quad t < 0$$

and

$$F(t) = 1 - \frac{1}{2} I_{a^2/(a^2+t^2)} \left(m - \frac{1}{2}, \frac{1}{2} \right), \quad t > 0,$$

where

$$m = \frac{5\alpha_4 - 9}{2(\alpha_4 - 3)}, \quad a = \sqrt{\frac{2\alpha_4}{\alpha_4 - 3}}. \tag{4.2}$$

5 Design of X bar control chart

For sample sizes $n = 3, 4, \ldots, 10$, we calculated theoretical distribution of the standardized sample mean of considered distributions, using results from Section 3 and then we approximated it with Pearson type II distribution in the case of uniform distribution and with Pearson type VII distribution in the case of Student, Laplace and logistic distributions. Parameters of the fitted Pearson types II and VII distributions are calculated using formulas (4.1) and (4.2). Code for all calculations was written, by the author, in statistical software R and is available as supplementary code on the web site of the Journal. Width of control limits of the X bar control chart is calculated from

$$\alpha = 1 - P\{\mu_0 - k\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu_0 + k\frac{\sigma}{\sqrt{n}} | \mu = \mu_0\} = 2(1 - F_{T_n}(k)), \tag{5.1}$$

where F_{T_n} is cumulative distribution function of standardized sample mean, using Brent's root-finding method (Brent, 1973). Same procedure was followed for both the theoretical distribution of standardized sample mean and corresponding Pearson distribution.

Control limits of the X bar control chart for non-normal symmetric distributions are calculated for specified probability 0.0027 of type I error, in analogy with X bar control chart for normal distribution. When quality characteristics is normally distributed, the probability that sample mean falls outside three standard deviations from the center line is 0.0027, for in-control process. These are so called three-sigma control limits (here sigma refers to the standard deviation of sample mean) and they are frequently used in construction of X bar control chart (Montgomery, 2005).

Calculated widths of control limits, for considered distributions of quality characteristic, sample sizes $n = 3, 4, \ldots, 10$, probability of false alarm $\alpha = 0.0027$, for theoretical distribution of the standardized sample mean and Pearson types II and VII distributions, are given in Table 2.

As it can be seen in the Table 2, the values of the width of the control limits calculated from theoretical distribution and corresponding Pearson distribution are very close, i.e. corresponding Pearson distribution fits very well to the theoretical distribution of the

Sample size	Width of control limits							
	Student $t(10)$		*Laplace* $L(1)$		*Logistic* $LGS(1)$		*Uniform* $U(0,1)$	
	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson
$n=3$	3.21966	3.22227	3.54221	3.53915	3.25580	3.26074	2.59834	2.65308
$n=4$	3.16998	3.17156	3.43224	3.43628	3.20035	3.20234	2.72926	2.74902
$n=5$	3.13867	3.13966	3.36034	3.36606	3.16405	3.16527	2.79650	2.80355
$n=6$	3.11712	3.11775	3.30939	3.31520	3.13877	3.13966	2.83511	2.83866
$n=7$	3.10136	3.10178	3.27130	3.27668	3.12021	3.12091	2.86060	2.86314
$n=8$	3.08934	3.08962	3.24168	3.24652	3.10602	3.10660	2.87932	2.88118
$n=9$	3.07987	3.08005	3.21796	3.22227	3.09482	3.09531	2.89366	2.89502
$n=10$	3.07221	3.07233	3.19852	3.20234	3.08577	3.08619	2.90489	2.90597

Table 2: Width of control limits of X bar control chart

standardized sample mean. On the other hand, normal approximation would give value of $k = 2.99998$, for all n and all distributions of quality characteristics.

Now, we are interested to see what is the power of X bar control charts for detecting shifts $\delta = 0.5, 1.0, \ldots, 3.0$, for calculated width of control limits. Power of X bar control chart for detecting shifts from mean μ_0 to $\mu_1 = \mu_0 \pm \delta\sigma$ can be calculated from

$$\begin{aligned} 1-\beta &= 1 - P\{\mu_0 - k\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu_0 + k\frac{\sigma}{\sqrt{n}} | \mu = \mu_1\} = \\ &= F_{T_n}(-k - \delta\sqrt{n}) + F_{T_n}(-k + \delta\sqrt{n}). \end{aligned}$$

We should note that power of proposed X bar control chart for detecting shift $\delta = 0$ is 0.0027 for all considered distributions and sample sizes, i.e. it maintains probability of type I error.

Mainly, we want to investigate what is the minimum shift that X bar control chart can detect with a power of at least 90%.

Calculated power of X bar control chart, for considered distributions of quality characteristic, sample sizes $n = 3, 4, \ldots, 10$, shifts $\delta = 0.5, 1.0, \ldots, 3.0$ for both theoretical distribution of standardized sample mean and corresponding Pearson distribution, are given in Table 3.

¿From the Table 3, we see that X bar control chart can detect shifts of $\delta = 1.5$ with power of at least 90% for sample sizes of $n = 9$ and greater for all considered distributions. In order for the X bar chart to detect shifts of $\delta = 2.0$ with power of 90% and greater, it is necessary to take samples of size at least $n = 4$ for Student, Laplace and logistic distributions and sample sizes of $n = 5$ and greater for uniform distribution of quality characteristic. Also, we can once more notice that the corresponding Pearson distribution approximates the distribution of standardized sample mean rather well. In general, it can be concluded that X bar control chart can detect shifts of at least $\delta = 1.5$ with power of 90% and greater for non-normal symmetric distribution of quality characteristic.

6 Implementation of proposed X bar control chart

Now we are interested to see how proposed X bar control chart can be implemented in practice, in case when the distribution function of the quality characteristic is non-normal,

Distribution	**Power**											
	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson	Theor.	Pearson
$t(10)$	$\delta = 0.5$		$\delta = 1.0$		$\delta = 1.5$		$\delta = 2.0$		$\delta = 2.5$		$\delta = 3.0$	
$n = 3$	0.0110	0.0110	0.0665	0.06620	0.2609	0.2597	0.5998	0.5989	0.8715	0.8711	0.9750	0.9748
$n = 4$	0.0162	0.0162	0.1175	0.1171	0.4307	0.4300	0.8016	0.8014	0.9661	0.9659	0.9967	0.9967
$n = 5$	0.0224	0.0224	0.1795	0.1791	0.5871	0.5868	0.9108	0.9107	0.9919	0.9918	0.9996	0.9996
$n = 6$	0.0296	0.0296	0.2488	0.2484	0.7145	0.7144	0.9625	0.9625	0.9982	0.9982	1	1
$n = 7$	0.0377	0.0377	0.3218	0.3216	0.8100	0.8100	0.9850	0.9850	0.9996	0.9996	1	1
$n = 8$	0.0468	0.0468	0.3957	0.3955	0.8775	0.8775	0.9942	0.9942	0.9999	0.9999	1	1
$n = 9$	0.0566	0.0566	0.4678	0.4677	0.9231	0.9231	0.9979	0.9978	1	1	1	1
$n = 10$	0.0674	0.0673	0.5363	0.5363	0.9528	0.9528	0.9992	0.9992	1	1	1	1
$L(1)$												
$n = 3$	0.0077	0.0071	0.0370	0.0355	0.1541	0.1588	0.4642	0.4674	0.8061	0.8014	0.9514	0.9532
$n = 4$	0.0111	0.0104	0.0718	0.0708	0.3179	0.3207	0.7317	0.7258	0.9433	0.9438	0.9916	0.9921
$n = 5$	0.0156	0.0148	0.1215	0.1212	0.4973	0.4949	0.8761	0.8740	0.9841	0.9846	0.9986	0.9986
$n = 6$	0.0209	0.0201	0.1842	0.1842	0.6511	0.6469	0.9452	0.9451	0.9957	0.9958	0.9998	0.9997
$n = 7$	0.0273	0.0265	0.2561	0.2559	0.7670	0.7638	0.9765	0.9767	0.9989	0.9988	1	0.9999
$n = 8$	0.0347	0.0339	0.3327	0.3321	0.8487	0.8469	0.9901	0.9903	0.9997	0.9997	1	1
$n = 9$	0.0430	0.0423	0.4101	0.4089	0.9039	0.90305	0.9960	0.9960	0.9999	0.9999	1	1
$n = 10$	0.0523	0.0517	0.4850	0.4835	0.9400	0.9397	0.9984	0.9983	1	1	1	1
$LGS(1)$												
$n = 3$	0.0106	0.0103	0.0619	0.0613	0.2469	0.2460	0.5864	0.5840	0.8653	0.8638	0.9726	0.9726
$n = 4$	0.0155	0.0153	0.1109	0.1106	0.4178	0.4172	0.7945	0.7935	0.9637	0.9637	0.9962	0.9963
$n = 5$	0.0215	0.0213	0.1718	0.1718	0.5775	0.5769	0.9072	0.9070	0.9911	0.9911	0.9995	0.9995
$n = 6$	0.0285	0.0283	0.2409	0.2408	0.7079	0.7074	0.9607	0.9607	0.9979	0.9979	0.9999	0.9999
$n = 7$	0.0364	0.0363	0.3143	0.3142	0.8056	0.8053	0.9841	0.9842	0.9995	0.9995	1	1
$n = 8$	0.0452	0.0451	0.3887	0.3886	0.8746	0.8744	0.9938	0.9938	0.9999	0.9999	1	1
$n = 9$	0.0550	0.0549	0.4616	0.4614	0.9211	0.9211	0.9977	0.9977	1	1	1	1
$n = 10$	0.0655	0.0655	0.5309	0.5308	0.9515	0.9515	0.9991	0.9991	1	1	1	1
$U(0, 1)$												
$n = 3$	0.0424	0.0357	0.2022	0.1890	0.4999	0.4794	0.7976	0.7803	0.9575	0.9531	0.9986	0.9976
$n = 4$	0.0419	0.0397	0.2403	0.2348	0.6030	0.5951	0.8937	0.8898	0.9906	0.9903	1	1
$n = 5$	0.0470	0.0461	0.2929	0.2909	0.7062	0.7034	0.9527	0.9521	0.9986	0.9985	1	1
$n = 6$	0.0544	0.0538	0.3533	0.3523	0.7946	0.7934	0.9816	0.9816	0.9999	0.9998	1	1
$n = 7$	0.0630	0.0626	0.4167	0.4159	0.8629	0.8622	0.9936	0.9936	1	1	1	1
$n = 8$	0.0726	0.0723	0.4801	0.4794	0.9120	0.9117	0.9980	0.9980	1	1	1	1
$n = 9$	0.0830	0.0828	0.5416	0.5411	0.9455	0.9453	0.9994	0.9994	1	1	1	1
$n = 10$	0.0942	0.0941	0.6001	0.5996	0.9673	0.9672	0.9999	0.9998	1	1	1	1

Table 3: Power of X bar control chart

symmetric but unknown. For fitting Pearson type II or type VII distributions to data, we need an estimate of kurtosis based on sample of means.

6.1 Measures of sample kurtosis

We have three measures of sample kurtosis

$$g_2^* = \frac{m_4}{m_2^2}, \quad G_2^* = \frac{N-1}{(N-2)(N-3)}\left((N+1)g_2 + 6\right) + 3, \quad b_2^* = \frac{m_4}{s^4},$$

where m_k are sample central moments.

Joanes and Gill (1998) investigated three measures $g_2 = g_2^* - 3$, $G_2 = G_2^* - 3$ and $b_2 = b_2^* - 3$ of sample excess kurtosis. They showed that, generating 100000 samples of different sizes from Student t_5 distribution, g_2 generally has the smallest mean-squared error. We followed the same procedure for measures g_2^*, G_2^* and b_2^* and generated 100000 samples of different sizes from distributions of standardized sample mean of Student $t(10)$, Laplace $L(1)$, logistic $LGS(1)$ and uniform $U(0,1)$ distributions. We confirm Joanes and Gill's findings. So, we will use, for calculation of the parameters of Pearson types II and VII distributions, measure g_2^* as an estimate of sample kurtosis.

6.2 Empirical power of X bar control chart

In this section, we will calculate the empirical power of proposed X bar control chart in order to investigate its performance in practice. We will take, by Monte Carlo simulations, $m = 25, 50, 100$ samples of sizes 3 to 10 from Student t(10), Laplace $L(1)$, logistic $LGS(1)$ and uniform $U(0,1)$ distributions. Sample means, as well as estimates of mean and standard deviation, are calculated. Further, we estimated kurtosis of the distribution of sample mean with g_2^*. Then, corresponding Pearson type II or type VII distribution is fitted to m sample means and control limits and power of the X bar control chart are calculated. This procedure is repeated 100000 times. The average power of the X bar control chart, for considered distributions, is presented in Table 4 (rounded to four decimal places). It is expected that sample size and number of groups will affect sample estimates, i.e. values of parameters of fitted Pearson distribution and therefore power of proposed X bar control chart.

We compared the values of empirical power for a number of groups $m = 25, 50, 100$ with theoretical power from Table 3, giving accent on the values of theoretical power of 90% and greater. We made the following conclusions for shift sizes of 1.5 and greater. Zero difference is present at sample sizes of at least 7 and $\delta = 3$. Absolute difference between theoretical and empirical power gets smaller as a number of groups and shift sizes rise. In most of the cases, the difference exists on third to the fourth decimal place. In other words, proposed X bar control chart has quite satisfactory performance. General advice for its use in practice would be to choose preferably more than 25 groups of sample size of 9 and greater, in order to detect shift $\delta = 1.5$ with the power of at least 90%.

Distribution	**Power**																	
	$m = 25$						$m = 50$						$m = 100$					
$t(10)$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$	$\delta = 3.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$	$\delta = 3.0$	$\delta = 0.5$	$\delta = 1.0$	$\delta = 1.5$	$\delta = 2.0$	$\delta = 2.5$	$\delta = 3.0$
$n = 3$	0.0522	0.1857	0.4231	0.7150	0.9163	0.9842	0.0302	0.1369	0.3643	0.6701	0.8986	0.9808	0.0201	0.1073	0.3252	0.6423	0.8873	0.9786
$n = 4$	0.0688	0.2541	0.5670	0.8627	0.9776	0.9972	0.0424	0.2014	0.5161	0.8378	0.9734	0.9968	0.0292	0.1672	0.4822	0.8219	0.9705	0.9967
$n = 5$	0.0852	0.3225	0.6914	0.9394	0.9937	0.9993	0.0551	0.2688	0.6495	0.9281	0.9928	0.9993	0.0391	0.2328	0.6233	0.9206	0.9924	0.9993
$n = 6$	0.1019	0.3909	0.7909	0.9742	0.9980	0.9998	0.0682	0.3376	0.7588	0.9699	0.9979	0.9998	0.0499	0.3014	0.7385	0.9667	0.9978	0.9998
$n = 7$	0.1180	0.4565	0.8629	0.9888	0.9993	0.9999	0.0821	0.4060	0.8405	0.9872	0.9992	0.9999	0.0615	0.3717	0.8264	0.9863	0.9993	0.9999
$n = 8$	0.1341	0.5199	0.9128	0.9949	0.9997	1.0000	0.0963	0.4728	0.8984	0.9944	0.9997	1.0000	0.0736	0.4405	0.8886	0.9942	0.9997	1.0000
$n = 9$	0.1517	0.5820	0.9460	0.9976	0.9999	1.0000	0.1105	0.5360	0.9366	0.9974	0.9999	1.0000	0.0865	0.5075	0.9307	0.9975	0.9999	1.0000
$n = 10$	0.1679	0.6380	0.9664	0.9988	0.9999	1.0000	0.1253	0.5958	0.9609	0.9987	0.9999	1.0000	0.0996	0.5702	0.9575	0.9988	0.9999	1.0000
$L(1)$																		
$n = 3$	0.0364	0.1420	0.3542	0.6481	0.8828	0.9751	0.0188	0.0930	0.2814	0.5841	0.8531	0.9677	0.0119	0.0655	0.2325	0.5401	0.8326	0.9622
$n = 4$	0.0528	0.2117	0.5102	0.8263	0.9681	0.9956	0.0292	0.1536	0.4427	0.7886	0.9596	0.9944	0.0187	0.1172	0.3953	0.7628	0.9533	0.9936
$n = 5$	0.0691	0.2828	0.6481	0.9227	0.9911	0.9990	0.0408	0.2205	0.5913	0.9040	0.9889	0.9988	0.0271	0.1795	0.5532	0.8913	0.9873	0.9986
$n = 6$	0.0861	0.3539	0.7587	0.9667	0.9972	0.9997	0.0535	0.2912	0.7158	0.9589	0.9966	0.9996	0.0366	0.2483	0.6870	0.9532	0.9962	0.9996
$n = 7$	0.1030	0.4235	0.8412	0.9856	0.9990	0.9999	0.0665	0.3617	0.8101	0.9824	0.9988	0.9999	0.0471	0.3203	0.7903	0.9802	0.9988	0.9999
$n = 8$	0.1197	0.4902	0.8986	0.9935	0.9996	0.9999	0.0804	0.4316	0.8776	0.9922	0.9995	0.9999	0.0585	0.3927	0.8645	0.9915	0.9995	1.0000
$n = 9$	0.1364	0.5542	0.9367	0.9969	0.9998	1.0000	0.0947	0.4993	0.9232	0.9964	0.9998	1.0000	0.0708	0.4638	0.9148	0.9962	0.9998	1.0000
$n = 10$	0.1536	0.6136	0.9605	0.9984	0.9999	1.0000	0.1100	0.5643	0.9527	0.9982	0.9999	1.0000	0.0833	0.5302	0.9470	0.9982	0.9999	1.0000
$LGS(1)$																		
$n = 3$	0.0495	0.1787	0.4130	0.7058	0.9120	0.9831	0.0284	0.1301	0.3526	0.6590	0.8932	0.9793	0.0185	0.0998	0.3105	0.6281	0.8803	0.9766
$n = 4$	0.0664	0.2483	0.5599	0.8584	0.9766	0.9970	0.0402	0.1943	0.5060	0.8314	0.9717	0.9966	0.0275	0.1600	0.4711	0.8151	0.9686	0.9964
$n = 5$	0.0831	0.3177	0.6864	0.9376	0.9935	0.9993	0.0531	0.2625	0.6425	0.9253	0.9924	0.9992	0.0373	0.2255	0.6147	0.9172	0.9918	0.9993
$n = 6$	0.0996	0.3862	0.7870	0.9734	0.9979	0.9998	0.0662	0.3315	0.7533	0.9685	0.9977	0.9998	0.0481	0.2953	0.7332	0.9654	0.9977	0.9998
$n = 7$	0.1160	0.4524	0.8605	0.9884	0.9992	0.9999	0.0801	0.4005	0.8370	0.9867	0.9992	0.9999	0.0594	0.3650	0.8222	0.9857	0.9993	0.9999
$n = 8$	0.1331	0.5175	0.9117	0.9948	0.9997	1.0000	0.0940	0.4675	0.8959	0.9942	0.9997	1.0000	0.0716	0.4348	0.8859	0.9940	0.9997	1.0000
$n = 9$	0.1500	0.5792	0.9450	0.9975	0.9999	1.0000	0.1085	0.5314	0.9349	0.9973	0.9999	1.0000	0.0847	0.5028	0.9290	0.9973	0.9999	1.0000
$n = 10$	0.1669	0.6359	0.9658	0.9987	0.9999	1.0000	0.1239	0.5930	0.9602	0.9987	0.9999	1.0000	0.0976	0.5662	0.9565	0.9988	0.9999	1.0000
$U(0, 1)$																		
$n = 3$	0.0788	0.2527	0.5190	0.7998	0.9553	0.9938	0.0573	0.2227	0.4978	0.7874	0.9543	0.9949	0.0463	0.2061	0.4876	0.7826	0.9538	0.9959
$n = 4$	0.0918	0.3107	0.6377	0.9049	0.9879	0.9988	0.0669	0.2772	0.6157	0.8975	0.9885	0.9991	0.0536	0.2578	0.6054	0.8936	0.9893	0.9994
$n = 5$	0.1068	0.3731	0.7436	0.9584	0.9965	0.9997	0.0784	0.3372	0.7222	0.9557	0.9969	0.9998	0.0627	0.3157	0.7122	0.9542	0.9975	0.9999
$n = 6$	0.1219	0.4356	0.8276	0.9823	0.9989	0.9999	0.0914	0.4002	0.8103	0.9819	0.9991	0.9999	0.0733	0.3779	0.8012	0.9818	0.9993	1.0000
$n = 7$	0.1377	0.4976	0.8886	0.9924	0.9996	1.0000	0.1044	0.4623	0.8759	0.9925	0.9997	1.0000	0.0844	0.4400	0.8685	0.9928	0.9998	1.0000
$n = 8$	0.1525	0.5554	0.9290	0.9965	0.9998	1.0000	0.1181	0.5226	0.9215	0.9967	0.9999	1.0000	0.0968	0.5023	0.9167	0.9971	0.9999	1.0000
$n = 9$	0.1690	0.6132	0.9562	0.9983	0.9999	1.0000	0.1318	0.5801	0.9514	0.9985	0.9999	1.0000	0.1094	0.5614	0.9487	0.9987	1.0000	1.0000
$n = 10$	0.1849	0.6659	0.9729	0.9991	1.0000	1.0000	0.1469	0.6359	0.9706	0.9993	1.0000	1.0000	0.1226	0.6179	0.9691	0.9994	1.0000	1.0000

Table 4: Empirical power of X bar control chart

6.3 Example

Montgomery (2005) gave data set on thickness of a printed circuit board (in inches), for 25 samples of three boards each.

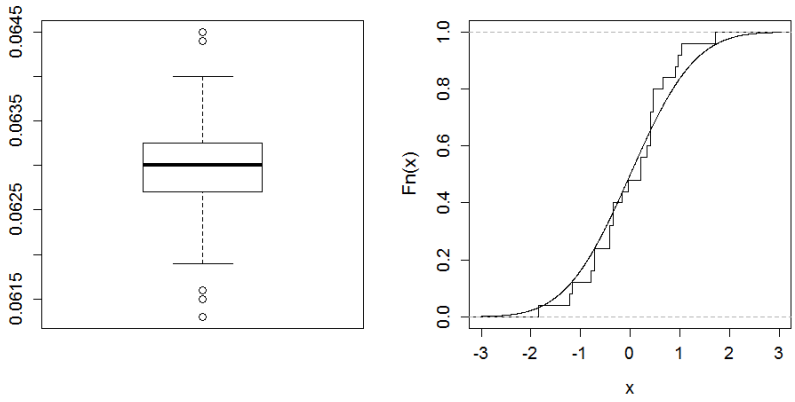

Figure 2: Boxplot of the thickness data (left graph) and empirical cumulative distribution function of standardized sample means with fitted Pearson type II distribution (right graph)

As we can see on boxplot (Figure 2, left graph), sample distribution seems symmetric. We tested symmetry of data distribution using Mira test (Mira, 1999), the Cabilio-Masaro test (Cabilio and Masaro, 1996) and Miao-Gel-Gastwirth (MGG) test (Miao et al., 2006). Based on results of all three tests, we can conclude that data distribution is symmetric (Mira test: Test Statistic = 0.9029, p-value = 0.3666; Cabilio-Masaro test: Test Statistic = 0.8846, p-value = 0.3764; MGG test: Test Statistic = 1.0162, p-value = 0.3095). R function *symmetry.test* for these tests can be found in R package *lawstat* (Gastwirth et al., 2015).

Now we will test the normality of the sample distribution using Shapiro-Wilk, Anderson-Darling and Lilliefors normality tests (Razali and Wah, 2011). Based on results of all three tests, we conclude that data distribution is not normal (Shapiro-Wilk test: W = 0.9589, p-value = 0.01584; Anderson-Darling test: A = 1.4759, p-value = 0.00076; Lilliefors test D = 0.1467, p-value = 0.00039). We used R function *shapiro.test* (package *stats*) for Shapiro-Wilk test and *ad.test*, *lillie.test* from R package *nortest* (Gross and Ligges, 2015) for Anderson-Darling and Lilliefors normality tests, respectively.

For each of 25 samples, we calculated sample mean. Mean of all sample means is equal to $\overline{\overline{X}} = 0.06295$ and this is the estimate of unknown process mean and center line of X bar control chart. Further, we estimated process standard deviation with mean range, $\hat{\sigma} = \overline{R} = 0.00092$. Now, we can calculate standardized sample means and kurtosis of standardized sample means. We got $\hat{\alpha}_4 = g_2^* = 2.83154$ (measures of sample excess kurtosis can be found in R package $e1071$ (Meyer et al., 2014)). So, as the distribution of standardized sample means is symmetric with kurtosis smaller than 3, we will approximate its distribution with Pearson type II distribution. We calculated parameters of distribution using equation (4.1). Empirical distribution function along with fitted Pearson

type II distribution of standardized sample means is given on Figure 2, right graph.

For probability of false alarm $\alpha = 0.0027$, we get, using equation (5.1), that width of control limits is equal to $k = 2.83665$. Now we may calculate lower and upper control limits of X bar control chart, $LCL = \overline{\overline{X}} - k\frac{\overline{R}}{\sqrt{n}} = 0.06143$, $UCL = \overline{\overline{X}} + k\frac{\overline{R}}{\sqrt{n}} = 0.06448$ and construct X bar chart (Figure 3). As we can see on Figure 3, all sample means are within the control limits and we can conclude that process is in-control and keep the estimates of unknown process mean, standard deviation, as well as the width of control limits.

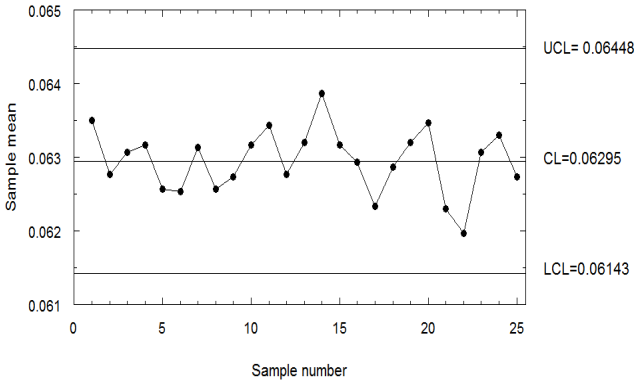

Figure 3: X bar control chart for the thickness data

7 Conclusions

We considered design of the X bar control chart when quality characteristic has one of the following non-normal symmetric distributions: Student distribution with 10 degrees of freedom, standard Laplace, standard logistic and standard uniform distributions. We calculated theoretical distribution of the standardized sample mean (or its best approximation) and approximated it with Pearson type II or type VII distributions. Then we calculated width of control limits of the X bar chart, which gave evidence of the goodness of fit of the corresponding Pearson distribution to the theoretical distribution of the standardized sample mean. Further, we examined the power of X bar control chart in detecting the shifts. Results suggest that the X bar chart can detect shifts of at least $\delta = 1.5$ with power of 90% and greater. Then we undertook Monte Carlo study in order to calculate empirical power of proposed X bar control chart, confirming its quite satisfactory performance. Finally, we constructed X bar chart for a given data set, when data distribution is non-normal and symmetric, but unknown.

References

[1] Ahsanullah, M., Golam Kibria, B.M. and Shakil, M. (2014): *Normal and Student's t Distributions and Their Applications*. Atlantis Press, Paris.

[2] Alloway, J.A. and Raghavachari, M. (1991): Control charts based on Hodges-Lehmann estimator. *Journal of Quality Technology*, **23**, 336-347.

[3] Alwan, L.C. (1995): The Problem of Misplaced Control Limits, *Journal of the Royal Statistical Society, Series C*, **44**, 269-278.

[4] Balakrishnan, N. (1992): *Handbook of the Logistic Distribution*. Marcel Dekker, New York.

[5] Brent, R.P. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, New Jersey.

[6] Cabilio, P. and Masaro, J. (1996): A simple test of symmetry about an unknown median.*The Canadian Journal of Statistics*, **24**, 349-361.

[7] Gastwirth, J.L., Gel, Y.R., Wallace Hui, W.L., Miao, W. and Noguchi, K. (2015): *lawstat: Tools for Biostatistics, Public Policy, and Law*. R package version 2.5.

[8] Gil-Pelaez, J. (1951): Note on the inversion theorem. *Biometrika*, **38**, 481-482.

[9] Gross, J. and Ligges, U. (2015): *nortest: Tests for Normality*. R package version 1.0-3.

[10] Gupta, S.S. and Han, S. (1992): Selection and ranking procedures for logistic populations, In: *Order Statistics and Nonparametrics: Theory and Applications* (Edited by P.K. Sen and I.A. Salama). Elsevier, Amsterdam, 377-404.

[11] Janacek, G.J. and Meikle, S.E. (1997): Control charts based on medians. *The Statistician*, **46**, 19-31.

[12] Joanes, D.N. and Gill, C.A. (1998): Comparing measures of sample skewness and kurtosis. *The Statistician*, **47**, 183-189.

[13] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994): *Continuous Univariate Distributions Volume 1*. Wiley, New York.

[14] Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995): *Continuous Univariate Distributions Volume 2*. Wiley, New York.

[15] Kotz, S., Kozubowski, T.J. and Podgórski, K. (2001): *The Laplace Distribution and Generalizations : a Revisit with Applications to Communications, Economics, Engineering, and Finance*. Springer, New York.

[16] Küchler, U. and Tappe, S. (2008): On the shapes of bilateral Gamma densities. *Statistics & Probability Letters*, **78**, 2478-2484.

[17] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch, F. (2014): *e1071: Misc Functions of the Department of Statistics, TU Wien*. R package version 1.6-4.

[18] Miao, W., Gel, Y.R. and Gastwirth, J.L. (2006): A New Test of Symmetry about an Unknown Median, In: *Random Walk, Sequential Analysis and Related Topics - A Festschrift in Honor of Yuan-Shih Chow* (Edited by A. Hsiung, C.-H. Zhang and Z. Ying). World Scientific Publisher, Singapore.

[19] Mira, A. (1999): Distribution-free test for symmetry based on Bonferroni's measure. *Journal of Applied Statistics*, **26**, 959-972.

[20] Montgomery, D.C. (2005): *Introduction to Statistical Quality Control*. Wiley, New York.

[21] Razali, N. and Wah, Y.B. (2011): Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, **2**, 21-33.

[22] Witkovský, V. (2001): On the exact computation of the density and of the quantiles of linear combinations of t and F random variables. *Journal of Statistical Planning and Inference*, **94**, 1-13.

[23] Witkovský, V. (2004): Matlab algorithm tdist: the distribution of a linear combination of Student's t random variables. *COMPSTAT 2004 Symposium*, Prague.

[24] Witkovský, V. and Savin, A. (2005): *tdist: Distribution of a linear combination of independent Student's t-variables*. R package version 0.1.1.

Metodološki zvezki, Vol. 13, No. 2, 2016, 101-116

Estimating the Coefficient of Asymptotic Tail Independence: a Comparison of Methods

Marta Ferreira[1]

Abstract

Many multivariate analyses require the account of extreme events. Correlation is an insufficient measure to quantify tail dependence. The most common tail dependence coefficients are based on the probability of simultaneous exceedances. The coefficient of asymptotic tail independence introduced in Ledford and Tawn ([18] 1996) is a bivariate measure often used in the tail modeling of data in finance, environment, insurance, among other fields of applications. It can be estimated as the tail index of the minimum component of a random pair with transformed unit Pareto marginals. The literature regarding the estimation of the tail index is extensive. Semi-parametric inference requires the choice of the number k of the largest order statistics that lead to the best estimate, where there is a tricky trade-off between variance and bias. Many methodologies have been developed to undertake this choice, most of them applied to the Hill estimator (Hill, [16] 1975). We are going to analyze, through simulation, some of these methods within the estimation of the coefficient of asymptotic tail independence. We also compare with a minimum-variance reduced-bias Hill estimator presented in Caeiro *et al.* ([3] 2005). A pure heuristic procedure adapted from Frahm *et al.* ([13] 2005), used in a different context but with a resembling framework, will also be implemented. We will see that some of these simple tools should not be discarded in this context. Our study will be complemented by applications to real datasets.

1 Introduction

It is undeniable that extreme events have been occurring in areas like environment (e.g. climate changes due to pollution and global heating), finance (e.g., market crashes due to less regulation and globalization), telecommunications (e.g., growing traffic due to a high expanding technological development), among others. Extreme values are therefore the subject of concern of many analysts and researchers, who have come to realize that they should be dealt with some care, requiring their own treatment. For instance, the classical linear correlation is not a suitable dependence measure if the dependence characteristics in the tail differ from the remaining realizations in the sample. An illustration is addressed in Embrechts *et al.* ([9] 2002). To this end, the tail dependence coefficient (TDC) defined in

[1] Center of Mathematics of University of Minho, Center for Computational and Stochastic Mathematics of University of Lisbon and Center of Statistics and Applications of University of Lisbon, Portugal; msferreira@math.uminho.pt

Joe ([17] 1997), usually denoted by λ, is more appropriate. More precisely, for a random pair (X, Y) with respective marginal distribution functions (dfs) F_X and F_Y, the TDC is given by

$$\lambda = \lim_{t \downarrow 0} P(F_Y(Y) > 1 - t | F_X(X) > 1 - t), \tag{1.1}$$

whenever the limit exists. Roughly speaking, the TDC evaluates the probability of one variable exceeding a large value given that the other also exceeds it. A positive TDC means that X and Y are tail dependent and whenever null we conclude the random pair is tail independent. In this latter case, the rate of convergence towards zero is a kind of residual tail dependence that, once ignored, may lead to an under-estimation of the risk underlying the simultaneous exceedance of a large value. On the other hand, by considering that the random variables (rv's) X and Y are tail dependent when they are actually asymptotically independent, it will result in an over-estimation of such risk. The degree of misspecification depends on the degree of asymptotic independence given by the mentioned rate of convergence, denoted η in Ledford and Tawn ([18] 1996). More precisely, it is assumed that

$$P(F_X(X) > 1 - t, F_Y(Y) > 1 - t) = t^{1/\eta} L(t),\ \eta \in (0, 1], \tag{1.2}$$

where $L(t)$ is a slowly varying function at zero, i.e., $L(tx)/L(t) \to 1$ as $t \downarrow 0$ for all $x > 0$. We call the parameter η the coefficient of asymptotic tail independence. Whenever $\eta < 1$, X and Y are asymptotically independent and, if $\eta = 1$, asymptotic dependence holds if $L(t) \to c > 0$, as $t \downarrow 0$. In case X and Y are exactly independent then $\eta = 1/2$ and we can also discern between asymptotically vanishing negative dependence and asymptotically vanishing positive dependence if, respectively, $\eta \in (0, 1/2)$ and $\eta \in (1/2, 1)$. Observe that we can state (1.2) as

$$P\left(\min\left(\frac{1}{1 - F_X(X)}, \frac{1}{1 - F_Y(Y)}\right) > t\right) = t^{-1/\eta} L(1/t), \tag{1.3}$$

and thus η corresponds to the tail index of the minimum of the two marginals standardized as unit Pareto. The tail index, also denoted extreme value index, quantifies the "weight" of the tail of a univariate distribution: whenever negative, null or positive it means that the tail of the underlying model is, respectively, "light", "exponential" or "heavy". In what concerns univariate extreme values, it is the primary parameter as it is implicated in all other extremal parameters, such as, extremal quantiles, right end-point of distributions, probability of exceedance of large levels, as well as return periods, among others. Therefore, the estimation of the tail index is a crucial issue, with numerous contributions in the literature. A survey on this topic can be seen, for instance, in Beirlant *et al.* ([2] 2004).

Under a semi-parametric framework in the domain of heavy tails, the Hill estimator, introduced in Hill ([16] 1975), have proved to possess good properties, being an essential tool in any application on this topic. For a random sample $(T_1, \ldots, T_n)$, the Hill estimator corresponds to the sample mean of the log-excesses of the $k + 1$ larger order statistics $T_{n:n} \geq \ldots \geq T_{n-k:n}$, i.e.,

$$H_n(k) \equiv H(k) := \frac{1}{k} \sum_{i=1}^{k} \log \frac{T_{n-i+1:n}}{T_{n-k:n}},\ 1 \leq k < n, \tag{1.4}$$

Consistency requires that k must be intermediate, that is, a sequence of integers $k \equiv k_n$, $1 \leq k < n$, such that

$$k_n \to \infty \text{ and } k_n/n \to 0, \text{ as } n \to \infty.$$

There is no definite formula to obtain k and it must be chosen not too small to avoid high variance but also not to large to prevent high bias. Figure 1 illustrates this issue, particularly the dashed line corresponding to a unit Frchet model where the tail index is 1. Observe also that there is a kind of stable area of the sample path around the true value of the tail index, where the variance is no longer high and the bias haven't started to increase. This disadvantage is transversal to the semi-parametric tools concerning extreme values inference. In the particular case of the Hill estimator, many efforts have been made to minimize the problem, ranging from bias-corrected versions to the implementation of procedures to compute k. The minimum-variance reduced-bias (MVRB) Hill estimator presented in Caeiro *et al.* ([3] 2005; see also Neves *et al.* [21] 2015) was developed for the Hall-Welsh class (within Generalized Pareto distributions), with reciprocal quantile function

$$F^{-1}(1-1/x) = Cx^{\gamma}\left(1+\gamma\beta x^{\rho}/\rho + o(x^{\rho})\right),\ x \to \infty, \tag{1.5}$$

where $\gamma > 0$ is the tail index of model F, $C > 0$, and $\beta \neq 0$ and $\rho < 0$ are second order parameters. The MVRB Hill estimator is given by

$$CH_n(k) \equiv CH(k) := H(k)\left(1 - \frac{\widehat{\beta}(n/k)^{\widehat{\rho}}}{1-\widehat{\rho}}\right),\ 1 \leq k < n, \tag{1.6}$$

where $\widehat{\beta}$ and $\widehat{\rho}$ are suitable estimators of β and ρ, respectively. Details about these latter are addressed in Caeiro *et al.* ([4] 2009) and references therein. We will denote it "corrected Hill" (CH). Our aim is to compare, through simulation, several methods regarding the Hill and corrected Hill estimators applied to the estimation of η. We also consider the graphical and pure heuristic procedure presented in Frahm *et al.* ([13] 2005) in the context of estimating the TDC λ in (1.1), also relying on the choice of k upper order statistics with the same bias/variance controversy. All the estimation procedures are described in Section 2. The simulation study is conducted in Section 3 and applications to real datasets appear in Section 4. A small discussion ends this work in Section 5.

2 Estimation methods

In this section we describe the procedures that we are going to consider in the estimation of the coefficient of asymptotic tail independence η given in (1.3) and therefore corresponding to the tail index of

$$T = \min((1-F_X(X))^{-1}, (1-F_Y(Y))^{-1}). \tag{2.1}$$

Coefficient η is positive and we can use positive tail index estimators such as Hill. Observe that T is the minimum between two unit Pareto r.v.'s Alternatively, we can also undertake

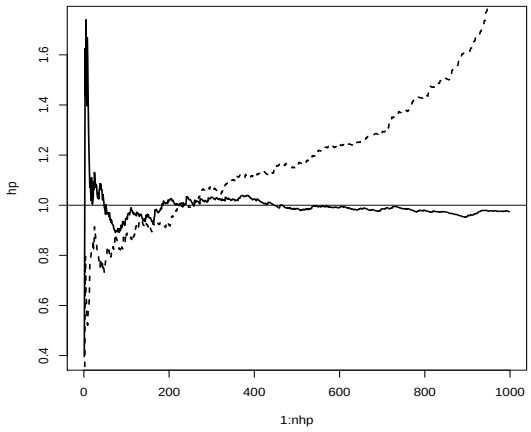

Figure 1: Hill plots of 1000 realizations of a unit Pareto (full line) and a unit Fréchet (dashed line), both with tail index equal to 1 (horizontal line).

a unit Frchet marginal transformation since $1 - F_X(X) \sim -\log F_X(X)$. However, in the sequel, we prosecute with unit Pareto marginals, since the Hill estimator has smaller bias in the Pareto models than in the Frchet ones (see Figure 1; see also Draisma *et al.* [6] 2004 and references therein). In order to estimate the unknown marginal df's F_X and F_Y we consider their empirical counterparts (ranks of the components), more precisely,

$$T_i^{(n)} := \min((n+1)/(n+1-R_i^X), (n+1)/(n+1-R_i^Y)),\ i = 1, \ldots, n$$

where R_i^X denotes the rank of X_i among $(X_1, \ldots, X_n)$ and R_i^Y denotes the rank of Y_i among $(Y_1, \ldots, Y_n)$.

The estimation of η through the tail index estimators Hill and maximum likelihood (Smith, [24] 1987) was addressed in Draisma *et al.* ([6] 2004). Other estimators were also considered in Poon *et al.* ([23] 2003; see also references therein) and more recently in Goegebeur and Guillou ([14] 2013) and Dutang *et al.* ([8] 2014). However, no method was analyzed in order to attain the best choice of k in estimation.

In the domain of positive tail indexes, the Hill estimator is the most widely studied and many developments have been appearing around it. The main topics concern methods to obtain the value of k related to the number of tail observations to use in estimation and procedures to control the bias without increasing the variance. The corrected Hill version in (1.6), for instance, removes from Hill its dominant bias component estimated by $H(k)(\widehat{\beta}(n/k)^{\widehat{\rho}})/(1-\widehat{\rho})$.

In the following, we describe the methods developed in literature for the Hill estimator to compute the value of k, that will be used to estimate η (the tail index of rv T in (2.1)) in our simulation study.

Based on Beirlant *et al.* ([1] 2002) and little restrictive conditions on the underlying

model, we have

$$Y_i := (i+1)\log\frac{T^{(n)}_{n-i:n}\,H(i)}{T^{(n)}_{n-(i+1):n}\,H(i+1)} = \eta + b(n/k)\left(\frac{i}{k}\right)^{-\rho} + \epsilon_i,\ i = 1, ..., k, \quad (2.2)$$

where the error term ϵ_i is zero-centered and b is a positive function such that $b(x) \to 0$, as $x \to \infty$. Extensive simulation studies conclude that the results tend to be better when ρ is considered fixed, even if misspecified. Matthys and Beirlant ([19] 2000) suggest $\rho = -1$. From model (2.2), the resulting least squares estimators of η and $b(n/k)$ are given by

$$\widetilde{\eta}^{LS}_{k,n} = \overline{Y}_k - \widetilde{b}^{LS}_{k,n}/(1-\rho) \quad \text{and} \quad \widetilde{b}^{LS}_{k,n} = \frac{(1-\rho)^2(1-2\rho)}{\rho^2}\frac{1}{k}\sum_{i=1}^{k}\left(\left(\frac{i}{k}\right)^{-\rho} - \frac{1}{1-\rho}\right)Y_i. \quad (2.3)$$

Thus, by replacing these estimates in the Hill's asymptotic mean squared error (AMSE)

$$\text{AMSE}(H(k)) = \frac{\eta^2}{k} + \left(\frac{b(n/k)}{1-\rho}\right)^2,$$

we are able to compute $\widehat{k}^1_{opt}$ as the value of k that minimizes the obtained estimates of the AMSE and estimate η as $H(\widehat{k}^1_{opt})$.

On the other hand, we can compute the approximate value of k that minimizes the AMSE, given by

$$k_{opt} \sim b(n/k)^{-2/(1-2\rho)}k^{-2\rho/(1-2\rho)}\left(\frac{\eta^2(1-\rho)^2}{-2\rho}\right)^{1/(1-2\rho)}. \quad (2.4)$$

See, e.g., Beirlant *et al.* ([1] 2002). Replacing again η and $b(n/k)$ by the respective least squares estimates in (2.3) with fixed $\rho = -1$, we derive $\widehat{k}_{opt,k}$, for $k = 3, ..., n$, using (2.4). Then compute $\widehat{k}^2_{opt} = median\{\widehat{k}_{opt,k},\ k = 3, ..., \lfloor\frac{n}{2}\rfloor\}$, where $\lfloor x\rfloor$ denotes the largest integer not exceeding x and consider η estimated by $H(\widehat{k}^2_{opt})$.

Further reading of the methods is referred to Beirlant *et al.* ([1] 2002), Matthys and Beirlant ([19] 2000) and references therein. In the sequel, they are shortly denoted, respectively, AMSE and KOPT.

The adaptive procedure of Drees and Kaufmann ([6] 1998) looks for the optimum k under which the bias starts to dominate the variance. The method is developed for the Hall-Welsh class of models defined in (1.5), for which it is proved that the maximum random fluctuation of $\sqrt{i}(H(i) - \eta)$, $i = 1, ..., k-1$, with $k \equiv k_n$ an intermediate sequence, is of order $\sqrt{\log\log n}$. More precisely, for ρ fixed at -1, we have:

1. Consider $r_n = 2.5 \times \widetilde{\eta} \times n^{0.25}$, with $\widetilde{\eta} = \widehat{\eta}_{2\sqrt{n},n}$.

2. Calculate $\widetilde{k}(r_n) := \min\{k = 1, ..., n-1 :\ \max_{i<k}\sqrt{i}|H(i) - H(k)| > r_n\}$. If $\sqrt{i}|H(i) - H(k)| > r_n$ doesn't hold for any k, consider $0.9 \times r_n$ to r_n and repeat step 2, otherwise move to step 3.

3. For $\varepsilon \in (0, 1)$, usually $\varepsilon = 0.7$, obtain

$$\widehat{k}_{\text{DK}} = \left\lfloor \frac{1}{3}(2\widetilde{\eta}^2)^{1/3}\left(\frac{\widetilde{k}(r_n^{\varepsilon})}{(\widetilde{k}(r_n))^{\varepsilon}}\right)^{1/(1-\varepsilon)}\right\rfloor$$

This method will be shortly referred DK.

Sousa and Michailidis (2004) method is based on the Hill sum plot, (k, S_k), $k = 1, ..., n-1$, where $S_k = kH(k)$. We have $E(S_k) = k\eta$, an thus the sumplot must be approximately linear for the values of k where $H(k) \approx \eta$, with the respective slope being an estimator of η. The method essentially seeks the breakdown of linearity. Their approach is based on a sequential testing procedure implemented in McGee and Carleton ([20] 1970), leaning over approximately Pareto tail models. More precisely, consider the regression model $y = X\eta + \delta$, with $y = (S_1, ..., S_k)'$, $X = [1\ i]_{i=1}^{k}$ and δ the error term. It is checked the null hypothesis that a new point y_0 is adjacent to the left or to the right of the set of points $y = (y_1, ..., y_k)$, through the statistics

$$TS = s^{-2}\left((y_0 - \widehat{y}_0^*)^2 + \sum_{i=1}^{k}(\widehat{y}_i - \widehat{y}_i^*)^2\right),$$

where $*$ denotes the predictions based on $k+1$ and $s^2 = (k-2)^{-1}(y'y - \widehat{\eta}X'y)$. Since TS is approximately distributed by $F_{1,k-2}$, the null hypothesis is rejected if TS is larger than the $(1-\alpha)$-quantile, $F_{1,k-2;1-\alpha}$. The method, shortly denoted SP from now on, is described in the following algorithm:

1. Fit a least-squares regression line to the initial $k = \nu n$ upper observations, $y = [y_i]_{i=1}^{k}$ (usually $\nu = 0.02$).

2. Using the test statistic TS, determine if a new point $y_0 = y_j$ for $j > k$, belongs to the original set of points y. Go adding points until the null hypothesis is rejected.

3. Consider $k_{\texttt{new}} = \max(0, \{j : TS < F_{1,k-2;1-\alpha}\})$. If $k_{\texttt{new}} = 0$, no new points are added to y and thus move forward to step 4. Return to step 1. if $k_{\texttt{new}} > 0$ by considering $k = k_{\texttt{new}}$.

4. Estimate η by considering the obtained k.

The heuristic procedure introduced in Gomes *et al.* ([15] 2013), searches for the supposed stable region encompassing the best k to be estimated. More precisely, we need first to obtain the minimum value j_0, such that the rounded values to j decimal places of $H(k)$, $1 \leq k < n$, denoted $H(k; j)$ are not all equal. Identify the set of values of k associated to equal consecutive values of $H(k; j_0)$. Consider the set with largest range $\ell := k_{\texttt{max}} - k_{\texttt{min}}$. Take all the estimates $H(k; j_0 + 2)$ with $k_{\texttt{max}} \leq k \leq k_{\texttt{min}}$, i.e., the estimates with two additional decimal points and calculate the mode. Consider $\mathcal{K}$ the set of k-values corresponding to the mode. Take $H(\widehat{k})$, with $\widehat{k}$ being the maximum of $\mathcal{K}$. Since it was specially designed for reduced-bias estimators, we shortly referred it as RB method hereinafter.

Frahm *et al.* ([13] 2005) also presented a heuristic procedure that can be applied to all estimators depending on a number k of rv's whose choice bears the mentioned trade-off between bias and variance. Indeed is was developed within the estimation of the TDC λ defined in (1.1). It was adapted to the Hill estimator in Ferreira ([11, 12] 2014, 2015) as follows:

1. Smooth the Hill plot $(k, H(k))$ by taking the means of $2b+1$ successive points, $\overline{H}(1), ..., \overline{H}(n-2b)$, with bandwidth $b = \lfloor w \times n \rfloor$.

2. Define the regions $p_k = (\overline{H}(k), ..., \overline{H}(k+m-1))$, $k = 1, ..., n-2b-m+1$, with length $m = \lfloor \sqrt{n-2b} \rfloor$. The algorithm stops at the first region satisfying

$$\sum_{i=k+1}^{k+m-1} \left| \overline{H}(i) - \overline{H}(k) \right| \leq 2s,$$

   where s is the empirical standard-deviation of $\overline{H}(1), ..., \overline{H}(n-2b)$.

3. Consider the chosen plateau region p_k* and estimate η as the mean of the values of p_k* (consider the estimate zero if no plane region fulfills the stopping condition).

The estimation of η through the plateau method was analyzed in Ferreira and Silva ([10] 2014) with respect to the sensibility of the bandwidth. The value $w = 0.005$ seems a reasonable choice (thus each moving average in step 1. consists in 1% of the data), also suggested in Frahm *et al.* ([13] 2005). In the sequel it will be referred as plateau method (in short PLAT).

Both RB and PLAT are simultaneously graphical and free-assumption methods since they are based on the search of a plane region of the estimator's plot that presumably contains the best sample fraction k to be estimated through a totally "ad-hoc" procedure. The sumplot is also a graphical method and the remaining procedures are neither graphical nor free-assumption.

3 Simulation study

In this section we compare through simulation the performance of the methods described above within the estimation of η through the under study estimators Hill in (1.4) and corrected Hill in (1.6).

We have generated 100 runs of samples of sizes $n = 100, 1000, 5000$ from the following models:

- Bivariate Normal distribution ($\eta = (1+\rho)/2$; see, e.g., Draisma *et al.* [6] 2004); we consider correlation $\rho = -0.2$ ($\eta = 0.4$), $\rho = 0.2$ ($\eta = 0.6$) and $\rho = 0.8$ ($\eta = 0.9$); we use notation, respectively, $N(-0.2)$, $N(0.2)$ and $N(0.8)$.

- Bivariate t-Student distribution t_ν with correlation coefficient given by $\rho \neq -1$ ($\lambda = 2F_{t_{\nu+1}}\left(-\sqrt{(\nu+1)(1-\rho)/(1+\rho)}\right)$, see Embrechts *et al.* [9] 2002; we have $\lambda > 0$ and thus $\eta = 1$); we consider $\nu = 4$ and $\rho = 0.25$ ($\lambda = 0.1438$) and $\nu = 1$ and $\rho = 0.75$ ($\lambda = 0.6464$); we use notation, respectively, t_4 and t_1.

- Bivariate extreme value distribution with a asymmetric-logistic dependence function $\ell(x, y) = (1-a_1)x + (1-a_2)y + ((a_1 x)^{1/\alpha} + (a_2 y)^{1/\alpha})^\alpha$, with $x, y \geq 0$,

dependence parameter $\alpha \in (0, 1]$ and asymmetric parameters $a_1, a_2 \in (0, 1]$ ($\lambda = 2 - l(1, 1)$, see Beirlant *et al.* [1] 2004; we have $\lambda > 0$ and thus $\eta = 1$); we consider $\alpha = 0.7$ and $a_1 = 0.4$, $a_2 = 0.2$ ($\lambda = 0.1010$) and $\alpha = 0.3$ and $a_1 = 0.6$, $a_2 = 0.8$ ($\lambda = 0.5182$); we use notation, respectively, $AL(0.7)$ and $AL(0.3)$.

- Farlie-Gumbel-Morgenstern distribution with dependence 0.5 ($\eta = 0.5$, see Dutang *et al.* [8] 2014); we use notation $FGM(0.5)$.

- Frank distribution with dependence 2 ($\eta = 0.5$, see Dutang *et al.* [8] 2014); we use notation $Fr(2)$.

Observe that the case $N(0.8)$ is an asymptotic tail independent model close to tail dependence since $\eta = 0.9 \approx 1$. On the other hand, the cases t_4 and $AL(0.7)$ are tail dependent cases ($\eta = 1$) near asymptotic tail independence since $\lambda = 0.1438 \approx 0$ and $\lambda = 0.1010 \approx 0$, respectively. We consider these examples in order to assess the robustness of the methods in border cases.

In Figures 2 and 3 are plotted, respectively, the results of the simulated values of the absolute bias and root mean squared error (rmse), for the Hill and corrected Hill estimators, in the case $n = 1000$. All the results are presented in Table 1 concerning the Hill estimator and Table 2 with respect to the corrected Hill. Observe that this latter case requires the estimation of additional second order parameters (β and ρ). To this end, we have followed the indications in Caeiro *et al.* ([4] 2009). For the ρ estimation, there was an overall best performance whenever it was taken fixed at value -1, leading to the reported results.

The largest differences between Hill and corrected Hill can be noticed in the above mentioned border cases, with the corrected one presenting lower absolute bias and rmse. The other models also show this difference but in a small amount. We remark that we are working with the minimum of Pareto rv's and the Hill estimator is unbiased in the Pareto case. The FGM and Frank models behave otherwise with a little lower absolute bias and rmse within the Hill estimator, for either estimated or several fixed values tried for ρ.

The failure cases in the DK method (column "NF" of Tables 1 and 2) correspond to an estimate of k out of the range $\{1, \ldots, n - 1\}$, which were ignored in the results. It sets up the worst performance, which may be justified by the fact that the class of models underlying the scope of application of this method excludes the simple Pareto law.

The corrected Hill exhibits better results in general, particularly for methods KOPT, PLAT and AMSE, followed by SP and RB, in large sample sizes (n¿=1000). The PLAT procedure also performs well with the Hill estimator unlike the SP.

For $n = 100$, we have good results within RB and SP based on corrected Hill. Once again, the PLAT method behaves well in both estimators.

The border cases of weak tail dependence (t_4 and $AL(0.7)$) are critical throughout all evaluated procedures and estimators. On the other hand, the methods are robust in the border case of tail independence near dependence expressed in model $N(0.8)$.

4 Applications

In this section we illustrate the methods with three datasets analyzed in literature:

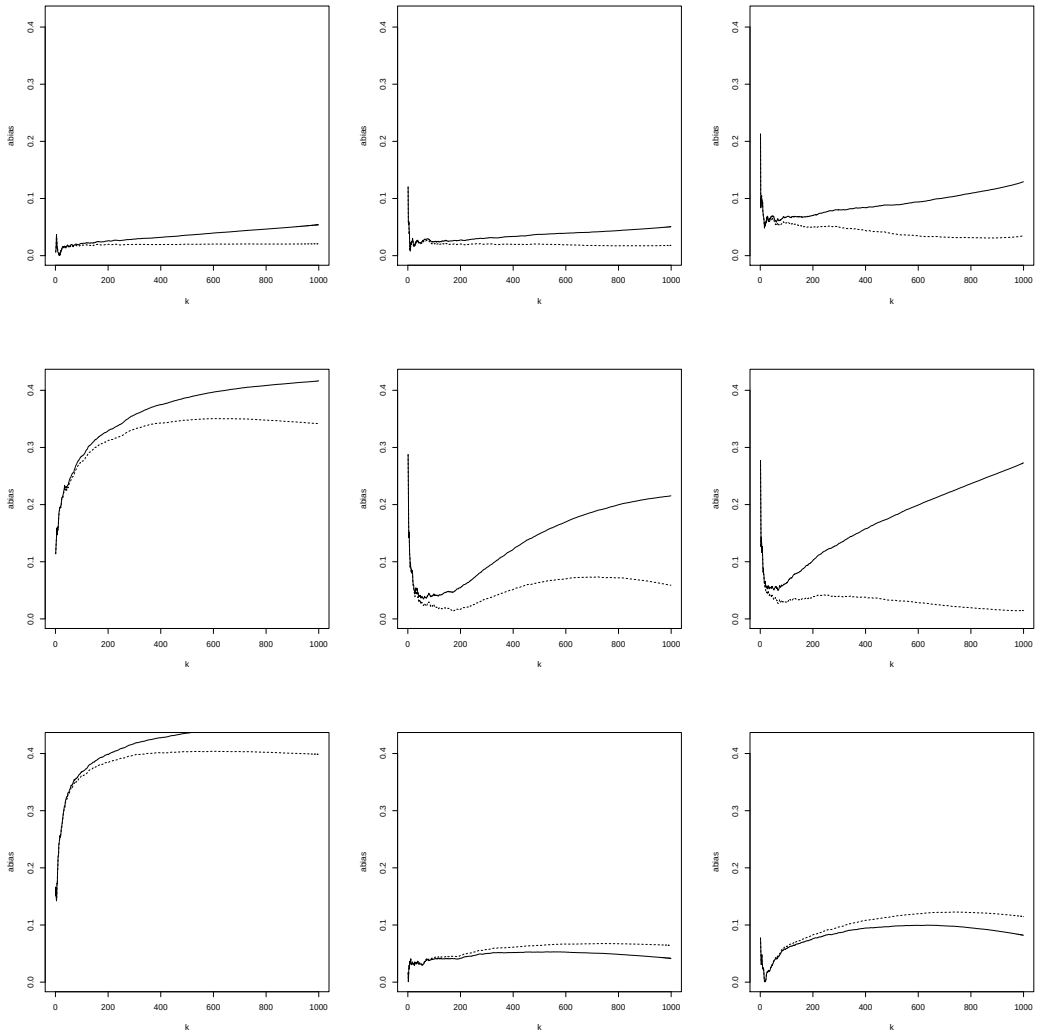

Figure 2: Simulated results of the absolute bias of Hill (full) and corrected Hill (dashed), for $n = 1000$, of the models (left-to-right and top-to-down): $N(-0.2)$, $N(0.2)$, $N(0.8)$, t_4, t_1, $AL(0.3)$, $AL(0.7)$, $FGM(0.5)$ and $Fr(2)$.

- I: The data consists of closing stock index levels of S&P 500 from the US and FTSE 100 from the UK, over the period 11 December 1989 to 31 May 2000, totalizing 2733 observed pairs (see, e.g., Poon *et al.* ([23] 2003)).

- II: The wave-surge data corresponding to 2894 paired observations collected during 1971-77 in Cornwall (England); it was analyzed in Coles and Tawn ([5] 1994) and later also in Ramos and Ledford ([22] 2009) under a parametric view.

- III: The Loss-ALAE data analyzed in Beirlant *et al.* ([2] 2004; see also references therein) consisting of 1500 pairs of registered claims (in USD) corresponding to an

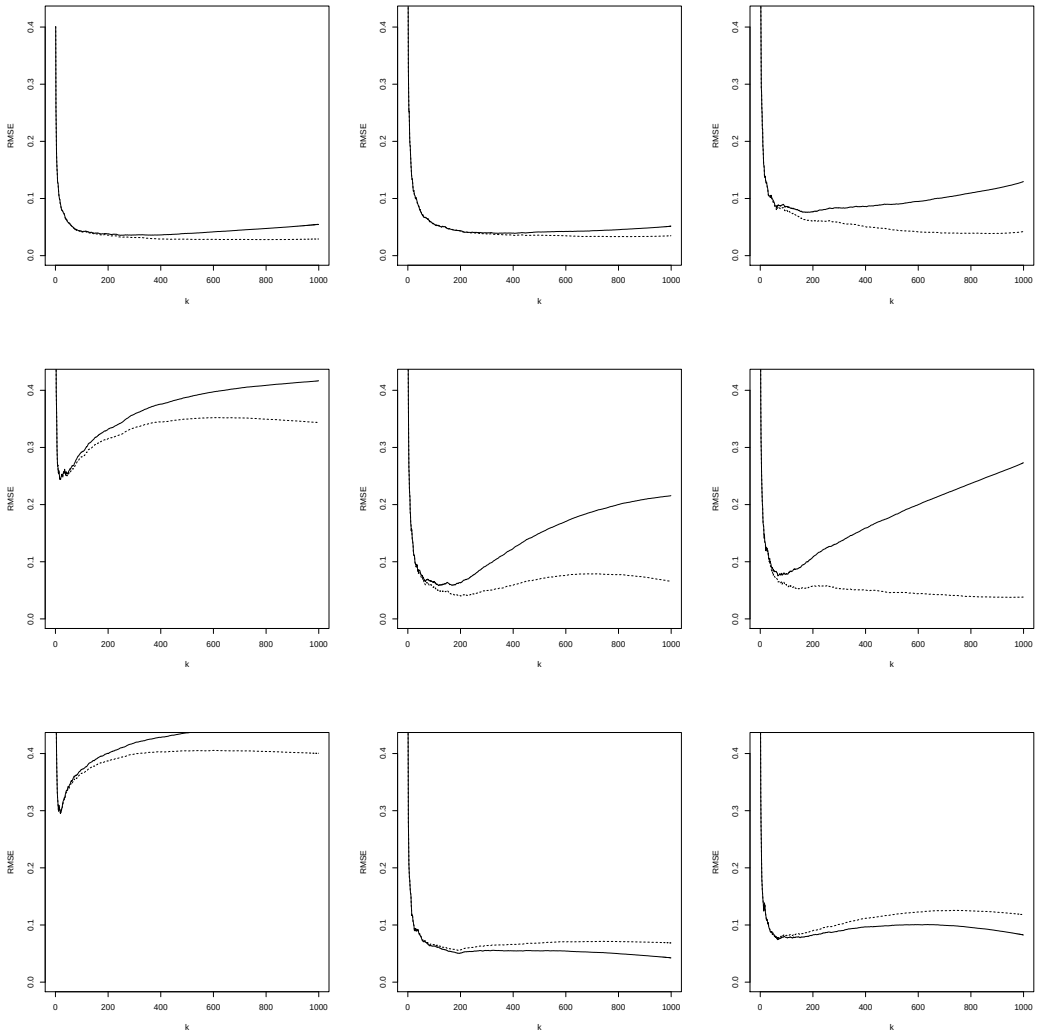

Figure 3: Simulated results of the rmse of Hill (full) and corrected Hill (dashed), for $n = 1000$, of the models (left-to-right and top-to-down): $N(-0.2)$, $N(0.2)$, $N(0.8)$, t_4, t_1, $AL(0.3)$, $AL(0.7)$, $FGM(0.5)$ and $Fr(2)$.

indemnity payment (loss) and an allocated loss adjustment expense (ALAE).

The respective scatter-plots are placed in Figure 4. For the US and UK stock market returns, the largest values in each tail for one variable correspond to reasonably large values of the same sign for the other variable, hinting an asymptotic independence but not exactly independence. In the wave-surge data, the dependence seems a bit more persistent within large values, as well as in Loss-ALAE data. The Hill and corrected Hill sample paths of η estimates are pictured in Figure 5. Table 3 reproduces the estimates obtained with each method and estimators under study. The estimation results found in literature for the financial (I), environmental (II) and insurance datasets (III) are respec-

tively approximated by 0.731, 0.85 and 0.9. The results seem to be in accordance with the simulation study.

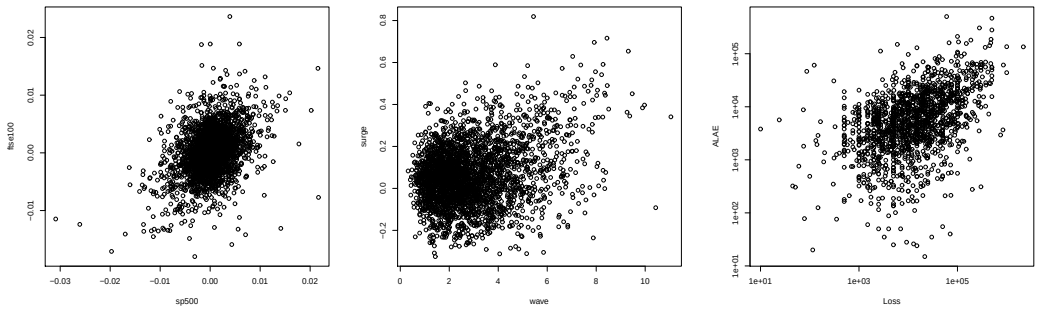

Figure 4: From left to right: scatter-plots of datasets I, II and III.

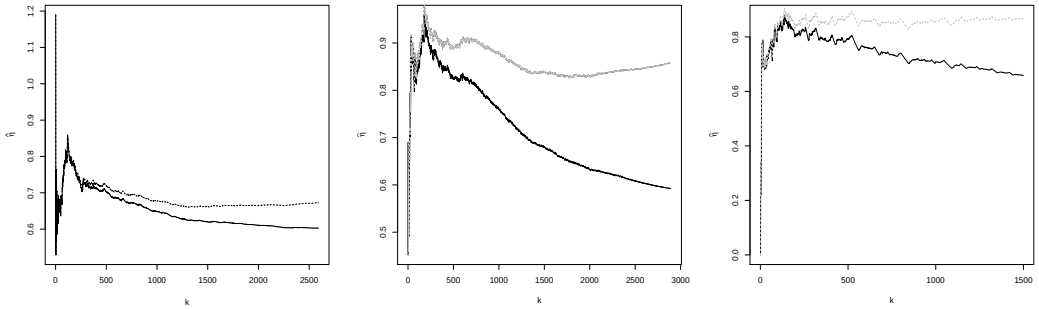

Figure 5: From left to right: sample paths of Hill (full;black) corrected Hill (dashed;grey) of datasets I, II and III.

5 Discussion

In this paper we have analyzed some simple estimation methods for the coefficient of asymptotic tail independence, with some of them revealing promising results. However, the choice of the estimator is not completely straightforward. It can be seen from simulation results that the ordinary Hill estimator may be still preferred over the corrected one in some situations. Also in boundary cases of tail dependence near independence, there are still some worrying errors to correct. These will be topics of a future research.

Acknowledgment

The author wishes to thank the reviewers for their constructive and valuable comments that have improved this work. This research was financed by Portuguese Funds through FCT - Fundação para a Ciência e a Tecnologia, within the Project UID/MAT/00013/2013 and by the research centre CEMAT (Instituto Superior Técnico, Universidade de Lisboa) through the Project UID/Multi/04621/2013.

References

[1] Beirlant, J., Dierckx, G., Guillou, A. and Stărică, C. (2002): On Exponential Representation of Log-Spacings of Extreme Order Statistics. *Extremes*, **5**, 157-180.

[2] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J.L. (2004): *Statistics of Extremes: Theory and Applications*. J. Wiley & Sons.

[3] Caeiro, F., Gomes, M.I. and Pestana, D.D. (2005): Direct reduction of bias of the classical Hill estimator. *Revstat*, **3(2)**, 111-136.

[4] Caeiro, F., Gomes, M.I. and Henriques-Rodrigues, L. (2009): Reduced-Bias Tail Index Estimators Under a Third-Order Framework. *Communications in Statistics - Theory and Methods*, **38(7)**, 1019-1040.

[5] Coles, S.G. and Tawn, J.A. (1994): Statistical methods for multivariate extremes: an application to structural design (with discussion). *Appl. Statist.*, **43**, 1-48.

[6] Draisma, G., Drees, H., Ferreira, A. and de Haan, L. (2004): Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, **10(2)**, 251-280.

[7] Drees, H. and Kaufmann, E. (1998): Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process Appl.*, **75**, 149-172.

[8] Dutang, C., Goegebeur, Y. and Guillou, A. (2014): Robust and bias-corrected estimation of the coefficient of tail dependence. *Insurance: Mathematics and Economics*, **57**, 46-57.

[9] Embrechts, P., McNeil, A. and Straumann, D. (2002): Correlation and dependency in risk management: properties and pitfalls. In: *Risk Management: Value at Risk and Beyond*, M.A.H. Dempster, Ed. Cambridge University Press, 176–223.

[10] Ferreira, M. and Silva, S. (2014): An Analysis of a Heuristic Procedure to Evaluate Tail (in)dependence. *Journal of Probability and Statistics*, Vol. 2014, Article ID 913621, 15 pages.

[11] Ferreira, M. (2014): A Heuristic Procedure to Estimate the Tail Index. Proceedings of the 14th International Conference in Computational Science and Its Applications - ICCSA 2014, June 30 - July 3 (2014), Guimares, Portugal, *IEEE-Computer Society*, 4264a241, 241-245.

[12] Ferreira, M. (2015): Estimating the tail index: Another algorithmic method. *Prob-Stat Forum*, **08**, 45-53.

[13] Frahm, G., Junker, M. and Schmidt R. (2005): Estimating the tail-dependence coefficient: properties and pitfalls. *Insurance: Mathematics & Economics*, **37(1)**, 80-100.

[14] Goegebeur, Y. and Guillou, A. (2013): Asymptotically unbiased estimation of the coefficient of tail dependence. *Scand. J. Stat.*, **40**, 174-189

[15] Gomes, M.I., Henriques-Rodrigues, L., Fraga Alves, M.I. and Manjunath, B.G. (2013): Adaptive PORT-MVRB estimation: an empirical comparison of two heuristic algorithms. *Journal of Statistical Computation and Simulation*, **83(6)**, 1129-1144.

[16] Hill, B.M. (1975): A Simple General Approach to Inference About the Tail of a Distribution. *Ann. Stat.*, **3**, 1163-1174.

[17] Joe, H. (1997): *Multivariate Models and Dependence Concepts*. Harry Joe, Chapman & Hall.

[18] Ledford, A. and Tawn, J. (1996): Statistics for near independence in multivariate extreme values. *Biometrika*, **83(1)**, 169-187.

[19] Matthys, G. and Beirlant, J. (2000): Adaptive Threshold Selection in Tail Index Estimation. In: *Extremes and Integrated Risk Management*, (Edited by P. Embrechts), 37-49. Risk Books, London.

[20] McGee, V.E. and Carleton, W.T. (1970): Piecewise Regression. *Journal of the American Statistical Association*, **65**, 1109-1124.

[21] Neves, M., Gomes, M.I., Figueiredo, F. and Prata-Gomes, D. (2015): Modeling Extreme Events: Sample Fraction Adaptive Choice in Parameter Estimation. *Journal of Statistical Theory and Practice*, **9(1)**, 184-199.

[22] Ramos, A. and Ledford, A. (2009): A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society, Series B*, **71**, 219-241.

[23] Poon, S.-H., Rockinger, M. and Tawn, J. (2003): Modelling extreme-value dependence in international stock markets. *Statistica Sinica*, **13**, 929-953.

[24] Smith, R.L. (1987): Estimating tails of probability distributions. *Ann. Statist.*, **15**, 1174-1207.

[25] Sousa, B. and Michailidis, G. (2004): A Diagnostic Plot for Estimating the Tail Index of a Distribution. *Journal of Computational and Graphical Statistics*, **13(4)**, 1–22.

	SP			KOPT			AMSE			RB			DK				PLAT	
$n = 100$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0449	0.0590	90	0.0387	0.1232	12	0.0258	0.0579	68	0.0286	0.0470	69	0.0350	0.2883	3	4	0.0111	0.0780
$N(0.2)$	0.0574	0.0698	89	0.1202	0.2002	15	0.0878	0.1224	64	0.0532	0.0714	75	0.0388	0.4878	4	2	0.0384	0.1042
$N(0.8)$	0.1372	0.1460	93	0.1881	0.2726	16	0.1935	0.2402	77	0.1323	0.1397	75	0.1320	0.4158	8	7	0.1133	0.1440
t_4	0.4187	0.4223	96	0.4121	0.4458	20	0.4309	0.4362	79	0.4155	0.4188	76	0.3007	0.5849	3	5	0.3539	0.3734
t_1	0.2266	0.2323	96	0.1605	0.2297	14	0.2318	0.2344	95	0.2144	0.2199	76	0.1923	0.3481	12	5	0.1300	0.1507
$AL(0.7)$	0.4642	0.4658	94	0.4625	0.4895	18	0.4784	0.4863	92	0.4572	0.4594	78	0.3447	0.6026	4	3	0.4199	0.4342
$AL(0.3)$	0.2825	0.2855	98	0.1686	0.2364	17	0.2877	0.3024	73	0.2498	0.2556	74	0.1991	0.3459	14	6	0.1585	0.1864
$FGM(0.5)$	0.0383	0.0578	90	0.0507	0.1683	12	0.0163	0.1117	56	0.0362	0.0585	75	0.0508	0.3649	6	8	0.0302	0.1052
$Fr(2)$	0.0805	0.0954	88	0.2065	0.1762	13	0.0320	0.1265	61	0.0839	0.0960	77	0.0041	0.3391	5	5	0.0764	0.1293
$n = 1000$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0425	0.0546	819	0.0059	0.0515	121	0.0378	0.0474	652	0.0437	0.0455	755	0.0242	0.3225	48	2	0.0247	0.0399
$N(0.2)$	0.0462	0.0642	826	0.0370	0.0687	171	0.0519	0.0690	777	0.0394	0.0432	754	0.0223	0.3651	39	0	0.0297	0.0452
$N(0.8)$	0.1178	0.1266	866	0.0832	0.0907	277	0.1231	0.1239	920	0.0926	0.0940	625	0.0991	0.3588	84	1	0.0716	0.0784
t_4	0.3921	0.4013	893	0.3303	0.3339	220	0.3703	0.3737	460	0.4056	0.4061	822	0.0431	0.6092	29	1	0.3114	0.3172
t_1	0.1975	0.2095	933	0.0777	0.0896	238	0.1530	0.1562	509	0.1886	0.1906	779	0.0479	0.1042	78	0	0.0554	0.0664
$AL(0.7)$	0.4518	0.4544	941	0.3906	0.3931	197	0.4245	0.4270	592	0.4392	0.4398	643	0.1613	0.6207	45	4	0.3827	0.3864
$AL(0.3)$	0.2369	0.2597	885	0.1282	0.1356	303	0.1821	0.1859	496	0.1940	0.1945	580	0.0800	0.1506	108	1	0.0868	0.0961
$FGM(0.5)$	0.0358	0.0430	846	0.0303	0.0525	178	0.0429	0.0600	630	0.0487	0.0516	762	0.0216	0.3347	50	0	0.0415	0.0532
$Fr(2)$	0.0630	0.0859	696	0.0305	0.0791	132	0.0409	0.1136	405	0.0952	0.0963	786	0.0380	0.3451	50	3	0.0691	0.0795
$n = 5000$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0485	0.0515	4369	0.0217	0.0280	629	0.0424	0.0445	3353	0.0399	0.0406	3135	0.0920	0.3383	572	1	0.0214	0.0271
$N(0.2)$	0.0486	0.0490	4804	0.0288	0.0346	847	0.0410	0.0422	3684	0.0384	0.0391	3590	0.0601	0.4406	402	1	0.0261	0.0330
$N(0.8)$	0.1253	0.1261	4902	0.0725	0.0745	1343	0.1021	0.1043	3357	0.0907	0.0915	3052	0.0696	0.2242	737	0	0.0585	0.0625
t_4	0.4103	0.4117	4853	0.2709	0.2745	548	0.2746	0.2829	648	0.4106	0.4107	4418	0.0636	0.4472	34	1	0.2653	0.2688
t_1	0.2075	0.2090	4902	0.0499	0.0543	1062	0.0804	0.0843	1442	0.2039	0.2043	4573	0.0209	0.0393	235	0	0.0201	0.0328
$AL(0.7)$	0.4594	0.4595	4999	0.3428	0.3448	457	0.3558	0.3633	1178	0.4411	0.4413	3222	0.1898	0.5659	20	2	0.3511	0.3534
$AL(0.3)$	0.2694	0.2712	4950	0.0956	0.0989	969	0.1100	0.1137	1101	0.1989	0.1998	3024	0.0499	0.0641	298	0	0.0529	0.0642
$FGM(0.5)$	0.0391	0.0422	4562	0.0277	0.0387	705	0.0415	0.0460	2053	0.0487	0.0494	3655	0.0421	0.3120	190	0	0.0313	0.0379
$Fr(2)$	0.0831	0.0842	4854	0.0620	0.0684	617	0.0862	0.0926	1590	0.1027	0.1030	3650	0.0035	0.2501	286	0	0.0692	0.0738

Table 1: Simulation results from Hill estimator, where *abias* denotes the absolute bias, NF the number of fails and $\overline{k}$ correspond to the mean of the k values obtained in the 100 runs.

	SP			KOPT			AMSE			RB			DK				PLAT	
$n = 100$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0186	0.0653	91	0.0427	0.1287	12	0.0032	0.0738	57	0.0137	0.0603	74	0.0416	0.2880	2	5	0.0076	0.0795
$N(0.2)$	0.0164	0.0977	90	0.1085	0.2044	15	0.0458	0.1295	58	0.0202	0.0961	74	0.0514	0.4604	3	1	0.0241	0.1130
$N(0.8)$	0.0594	0.1066	93	0.1717	0.2675	17	0.1014	0.1860	66	0.0658	0.1050	77	0.1025	0.4099	7	6	0.0959	0.1436
t_4	0.3446	0.3618	96	0.3846	0.4268	20	0.3649	0.3810	66	0.3566	0.3704	71	0.2871	0.6015	3	2	0.3361	0.3610
t_1	0.0952	0.1261	96	0.1369	0.2112	15	0.1104	0.1337	78	0.1118	0.1387	75	0.1437	0.3297	5	0	0.0850	0.1215
$AL(0.7)$	0.3995	0.4123	93	0.4528	0.4846	18	0.4245	0.4410	60	0.4122	0.4227	76	0.3313	0.5980	4	4	0.4046	0.4237
$AL(0.3)$	0.0437	0.1355	96	0.1187	0.2105	21	0.0781	0.1698	66	0.0609	0.1418	71	0.1537	0.3491	7	3	0.0865	0.1519
$FGM(0.5)$	0.0659	0.1121	89	0.0439	0.1749	13	0.0199	0.1345	55	0.0565	0.1036	72	0.0468	0.3775	3	8	0.0393	0.1170
$Fr(2)$	0.1237	0.1549	88	0.0199	0.1794	13	0.0733	0.1718	58	0.1210	0.1482	73	0.0048	0.3401	4	7	0.0912	0.1499
$n = 1000$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0165	0.0357	819	0.0008	0.0514	120	0.0119	0.0495	463	0.0204	0.0286	948	0.0103	0.3473	9	2	0.0206	0.0367
$N(0.2)$	0.0200	0.0539	808	0.0305	0.0662	169	0.0273	0.0608	515	0.0179	0.0342	913	0.0432	0.3660	18	1	0.0222	0.0442
$N(0.8)$	0.0353	0.0552	848	0.0545	0.0674	253	0.0450	0.0505	527	0.0359	0.0438	837	0.1318	0.4158	23	7	0.0514	0.0622
t_4	0.3255	0.3343	893	0.3061	0.3109	197	0.3275	0.3317	296	0.3471	0.3489	838	0.0806	0.6097	23	0	0.3042	0.3100
t_1	0.0514	0.0680	924	0.0278	0.0474	238	0.0525	0.0617	331	0.0667	0.0731	827	0.1303	0.2741	54	2	0.0276	0.0439
$AL(0.7)$	0.3937	0.3969	941	0.3751	0.3786	183	0.3920	0.3948	365	0.4009	0.4023	935	0.1170	0.6324	14	5	0.3781	0.3817
$AL(0.3)$	0.0063	0.0538	857	0.0371	0.0572	210	0.0610	0.1185	241	0.0239	0.0409	797	0.1388	0.2870	42	2	0.0413	0.0559
$FGM(0.5)$	0.0547	0.0649	846	0.0356	0.0572	180	0.0617	0.0671	600	0.0657	0.0698	904	0.0288	0.3346	42	0	0.0446	0.0585
$Fr(2)$	0.0854	0.1104	668	0.0371	0.0841	140	0.0845	0.1253	516	0.1172	0.1200	814	0.0442	0.3355	62	1	0.0729	0.0843
$n = 5000$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	abias	rmse	$\overline{k}$	NF	abias	rmse
$N(-0.2)$	0.0199	0.0240	4368	0.0156	0.0248	584	0.0206	0.0254	1823	0.0208	0.0225	4686	0.0520	0.3914	59	0	0.0200	0.0248
$N(0.2)$	0.0173	0.0223	4804	0.0223	0.0304	865	0.0193	0.0232	2720	0.0159	0.0205	4766	0.0292	0.4502	71	1	0.0212	0.0289
$N(0.8)$	0.0324	0.0346	4902	0.0458	0.0495	1110	0.0481	0.0502	1291	0.0308	0.0337	4466	0.0992	0.4592	78	3	0.0475	0.0521
t_4	0.3349	0.3360	4853	0.2495	0.2535	454	0.2549	0.2620	487	0.3447	0.3451	3940	0.1093	0.5322	39	2	0.2666	0.2696
t_1	0.0446	0.0473	4901	0.0008	0.0191	747	0.0127	0.0216	979	0.0535	0.0566	3782	0.0849	0.2497	311	2	0.0032	0.0214
$AL(0.7)$	0.3967	0.3971	4999	0.3303	0.3325	397	0.3210	0.3279	465	0.4003	0.4007	4203	0.1226	0.6323	26	3	0.3509	0.3529
$AL(0.3)$	0.0157	0.0260	4902	0.0315	0.0401	550	0.0414	0.0485	633	0.0194	0.0292	4520	0.1619	0.3958	210	1	0.0355	0.0427
$FGM(0.5)$	0.0561	0.0602	4562	0.0316	0.0431	727	0.0511	0.0567	2440	0.0619	0.0634	4196	0.0484	0.2880	218	0	0.0328	0.0399
$Fr(2)$	0.1186	0.1208	4854	0.0693	0.0758	691	0.1225	0.1162	2431	0.1264	0.1270	4100	0.0333	0.2429	241	1	0.0703	0.0757

Table 2: Simulation results from corrected Hill estimator, where *abias* denotes the absolute bias, NF the number of fails and $\overline{k}$ correspond to the mean of the k values obtained in the 100 runs.

$H(k)$	I	k	II	k	III	k
DK	0.6510	21	0.8255	83	0.7827	78
SP	0.6025	2592	0.5922	2893	0.6584	1499
KOPT	0.6733	744	0.9137	738	0.8444	135
AMSE	0.6494	955	0.7076	1244	0.6850	1172
RB	0.6041	2477	0.5967	2772	0.7428	708
PLAT	0.7148	–	0.8755	–	0.8110	–
$CH(k)$	I	k	II	k	III	k
DK	0.7654	5	0.4521	1	0.7044	27
SP	0.6725	2592	0.8581	2893	0.8671	1499
KOPT	0.7070	585	0.8991	412	0.8661	176
AMSE	0.6925	726	0.8997	596	0.8386	678
RB	0.6652	2264	0.8300	2040	0.8671	1499
PLAT	0.7261	–	0.8908	–	0.8524	–

Table 3: Estimates of η and respective values k, of datasets I, II and III.

Metodološki zvezki, Vol. 13, No. 2, 2016, 117-130

Approximate Confidence Interval for the Reciprocal of a Normal Mean with a Known Coefficient of Variation

Wararit Panichkitkosolkul[1]

Abstract

An approximate confidence interval for the reciprocal of a normal population mean with a known coefficient of variation is proposed. This has applications in the area of nuclear physics, agriculture and economic when the researcher knows the coefficient of variation. The proposed confidence interval is based on the approximate expectation and variance of the estimator by Taylor series expansion. A Monte Carlo simulation study was conducted to compare the performance of the proposed confidence interval with the existing confidence interval. Simulation results show that the proposed confidence interval performs as well as the existing one in terms of coverage probability. However, the approximate confidence interval is very easy to calculate compared with the exact confidence interval.

1 Introduction

The reciprocal of a normal mean is applied in the area of nuclear physics, agriculture and economic. For example, Lamanna et al. (1981) studied a charged particle momentum, $p = \mu^{-1}$ where μ is the track curvature of a particle. The reciprocal of a normal mean is defined by $\theta = \mu^{-1}$, where μ is the population mean. Many researchers studied the reciprocal of a normal mean. For instance, Zaman (1981) discussed the estimators without moments in the case of the reciprocal of a normal mean. The maximum likelihood estimate of the reciprocal of a normal mean with a class of zero-one loss functions was proposed by Zaman (1985). Withers and

[1] Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand; wararit@mathstat.sci.tu.ac.th

Nadarajah (2013) presented the theorem to construct the point estimators for the inverse powers of a normal mean.

Wongkhao et al. (2013) proposed two confidence intervals for the reciprocal of a normal mean with a known coefficient of variation. Their confidence intervals can be applied when the coefficient of variation of a control group is known. One of their confidence intervals is developed based on an exact method in which this confidence interval is constructed from the pivotal statistics Z, where Z follows the standard normal distribution. The other confidence interval is constructed based on the generalized confidence interval (Weerahandi, 1993). Simulation results show that the coverage probabilities of the two confidence intervals are not significantly different. However, the confidence interval based on the exact method is shorter than the generalized confidence interval. The exact method uses Taylor series expansion to find the expectation and variance of the estimator of θ and uses these results for constructing the confidence interval for θ. The lower and upper limits of the confidence interval based on the exact method are difficult to compute since they depend on an infinite summation. Therefore, our main aim in this paper is to propose an approximate confidence interval for the reciprocal of a normal mean with a known coefficient of variation. The computation of the new proposed confidence interval is easier than the exact confidence interval proposed by Wongkhao et al. (2013). In addition, we also compare the estimated coverage probabilities of the new proposed confidence interval and existing confidence interval using a Monte Carlo simulation.

The paper is organized as follows. In Section 2, the theoretical background of the existing confidence interval for θ is discussed. We provide the theorem for constructing the approximate confidence interval for θ in Section 3. In Section 4, the performance of the confidence intervals for θ is investigated through a Monte Carlo simulation study. Conclusions are provided in the final section.

2 Existing Confidence Interval

In this section, we review the theorem and corollary proposed by Wongkhao et al. (2013) and use these to construct the exact confidence interval for θ.

Theorem 1. *(Wongkhao et al., 2013) Let* $Y_1,...,Y_n$ *be a random sample of size* n *from a normal distribution with mean* μ *and variance* σ^2. *The estimator of* θ *is* $\hat{\theta} = \bar{Y}^{-1}$ *where* $\bar{Y} = n^{-1}\sum_{j=1}^{n} Y_j$. *The expectation of* $\hat{\theta}$ *and* $\hat{\theta}^2$ *when a coefficient of variation,* $\tau = \frac{\sigma}{\mu}$ *is known, are respectively*

$$E(\hat{\theta}) = \theta\left[1+\sum_{k=1}^{\infty}\frac{(2k)!}{2^k k!}\left(\frac{\tau^2}{n}\right)^k\right] \quad (1)$$

and

$$E(\hat{\theta}^2) = \theta^2\sum_{k=1}^{\infty}\frac{(2k+1)!}{2^k k!}\left(\frac{\tau^2}{n}\right)^k.$$

Proof of Theorem 1 See Wongkhao et al. (2013) ■

From (1), $\lim_{n\to\infty} E(\hat{\theta}) = \theta$ and $E\left(\frac{\hat{\theta}}{w}\right) = \theta$, where $w = 1+\sum_{k=1}^{\infty}\frac{(2k)!}{2^k k!}\left(\frac{\tau^2}{n}\right)^k$. Thus, the unbiased estimator of θ is $\frac{\hat{\theta}}{w} = \frac{1}{w\bar{Y}}$.

Corollary 1. *From Theorem 1,* $\text{var}(\hat{\theta}) \approx \frac{\theta^2}{n}\tau^2$.

Proof of Corollary 1 See Wongkhao et al. (2013) ■

Now we will use the fact that, from the central limit theorem,

$$Z = \frac{\hat{\theta}-\theta}{\sqrt{\text{var}(\hat{\theta})}} \sim N(0,1).$$

Based on Theorem 1 and Corollary 1, we get

$$Z = \frac{\frac{\hat{\theta}}{w}-\theta}{\sqrt{\frac{\theta^2}{n}\tau^2}} \sim N(0,1). \quad (2)$$

Therefore, the $100(1-\alpha)\%$ exact confidence interval for θ based on Equation (2) is

$$CI_{exact} = \frac{\hat{\theta}}{w} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}^2}{n}\tau^2},$$

where $w = 1+\sum_{k=1}^{\infty}\frac{(2k)!}{2^k k!}\left(\frac{\tau^2}{n}\right)^k$ and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution.

3 Proposed Confidence Interval

To find a simple approximate expression for the expectation of $\hat{\theta}$, we use a Taylor series expansion of $\frac{1}{y}$ around μ:

$$\frac{1}{y} \approx \frac{1}{y}\bigg|_{\mu} + (y-\mu)\frac{\partial}{\partial y}\left(\frac{1}{y}\right)\bigg|_{\mu} + \frac{1}{2}(y-\mu)^2\frac{\partial^2}{\partial y^2}\left(\frac{1}{y}\right)\bigg|_{\mu} + O\left(\left((y-\mu_y)\frac{\partial}{\partial y}\right)^3\left(\frac{1}{y}\right)\right). \tag{3}$$

Theorem 2. *Let* $Y_1,...,Y_n$ *be a random sample of size* n *from a normal distribution with mean* μ *and variance* σ^2. *The estimator of* θ *is* $\hat{\theta} = \bar{Y}^{-1}$ *where* $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$. *The approximate expectation and variance of* $\hat{\theta}$ *when a coefficient of variation,* $\tau = \frac{\sigma}{\mu}$ *is known, are respectively*

$$E(\hat{\theta}) \approx \frac{1}{\mu}\left(1+\frac{\tau^2}{n}\right) \tag{3}$$

and

$$\text{var}(\hat{\theta}) \approx \frac{\theta^2}{n}\tau^2. \tag{4}$$

Proof of Theorem 2.

Consider random variable $\bar{Y}$ where $\bar{Y}$ has support $(0,\infty)$. Let $\hat{\theta} = \bar{Y}^{-1}$. Find approximations for $E(\hat{\theta})$ and $\text{var}(\hat{\theta})$ using Taylor series expansion of $\hat{\theta}$ around μ as in Equation (3). The mean of $\hat{\theta}$ can be found by applying the expectation operator to the individual terms (ignoring all terms higher than two),

$$\begin{aligned}
E(\hat{\theta}) &= E\left(\frac{1}{\bar{Y}}\right) \\
&\approx E\left(\frac{1}{\bar{Y}}\right)\bigg|_{\mu} + E\left[\frac{\partial}{\partial \bar{Y}}\left(\frac{1}{\bar{Y}}\right)(\bar{Y}-E(\bar{Y}))\right]\bigg|_{\mu} + \frac{1}{2}E\left[\frac{\partial^2}{\partial \bar{Y}^2}\left(\frac{1}{\bar{Y}}\right)(\bar{Y}-E(\bar{Y}))^2\right]\bigg|_{\mu} + O(n^{-1}) \\
&\approx \frac{1}{\mu} + 0 + \frac{1}{2}\left(\frac{2}{(E(\bar{Y}))^3}\text{var}(\bar{Y})\right) \\
&= \frac{1}{\mu} + \frac{\text{var}(\bar{Y})}{\mu^3} \\
&= \frac{1}{\mu}\left(1+\frac{\sigma^2}{n\mu^2}\right) \\
&= \frac{1}{\mu}\left(1+\frac{\tau^2}{n}\right).
\end{aligned} \tag{4}$$

An approximation of the variance of $\hat{\theta}$ is obtained by using the first-order terms of the Taylor series expansion:

$$\begin{aligned}
\text{var}(\hat{\theta}) &= \text{var}\left(\frac{1}{\bar{Y}}\right) \\
&= E\left[\left(\frac{1}{\bar{Y}} - E\left(\frac{1}{\bar{Y}}\right)\right)^2\right] \\
&\approx E\left[\left(\frac{1}{\bar{Y}} - \frac{1}{\mu}\right)^2\right] \\
&\approx E\left[\left(\frac{1}{\mu} + \frac{\partial}{\partial \bar{Y}}\left(\frac{1}{\bar{Y}}\right)(\bar{Y} - E(\bar{Y})) - \frac{1}{\mu}\right)^2\right]\Bigg|_{\mu} \\
&= \left(\frac{\partial}{\partial \bar{Y}}\left(\frac{1}{\bar{Y}}\right)\right)^2 \text{var}(\bar{Y})\Bigg|_{\mu} \\
&\approx \frac{\text{var}(\bar{Y})}{\mu^4} \\
&= \frac{\sigma^2}{n\mu^4} \\
&= \frac{\theta^2}{n}\tau^2. \qquad (5)
\end{aligned}$$

■

It is clear from Equation (4) that $\hat{\theta}$ is asymptotically unbiased $\left(\lim_{n\to\infty} E(\hat{\theta}) = \theta\right)$ and $E\left(\frac{\hat{\theta}}{v}\right) = \theta$, where $v = 1 + \frac{\tau^2}{n}$. Therefore, the unbiased estimator of θ is $\frac{\hat{\theta}}{v} = \frac{1}{v\bar{Y}}$. From Equation (5), $\hat{\theta}$ is consistent $\left(\lim_{n\to\infty} \text{var}(\hat{\theta}) = 0\right)$.

We then apply the central limit theorem and Theorem 2,

$$Z = \frac{\frac{\hat{\theta}}{v} - \theta}{\sqrt{\frac{\theta^2}{n}\tau^2}} \sim N(0,1).$$

Therefore, it is easily seen that the $(1-\alpha)100\%$ approximate confidence interval for θ is

$$CI_{approx} = \frac{\hat{\theta}}{v} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\theta}^2}{n}\tau^2},$$

where $v = 1 + \frac{\tau^2}{n}$ and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution.

4 Simulation Study

A Monte Carlo simulation was conducted using the R statistical software [16] version 3.2.1 to compare the estimated coverage probabilities of the new proposed confidence interval and the exact confidence interval. Source code is available in Appendix. The estimated coverage probability (based on M replicates) are given by $\widehat{1-\alpha} = \#(L \le \theta \le U)/M$, where $\#(L \le \theta \le U)$ denotes the number of simulation runs for which the true reciprocal of a normal mean θ lies within the confidence interval. From two previous sections, we found that the lengths of both confidence intervals are equal to $2z_{1-\alpha/2}\sqrt{\hat{\theta}^2\tau^2/n}$ which the expected lengths are not considered in simulation study. The sets of normal data were generated with θ = 0.1, 0.2, 0.5, 1, 5 and 10, and the coefficient of variation τ = 0.05, 0.1, 0.2, 0.33, 0.5 and 0.67. The sample sizes were set at n = 10, 20, 30, 50, 100 and 500. The number of simulation runs was 10,000 and the nominal confidence level $1-\alpha$ was fixed at 0.95.

The results are demonstrated in Figure 1 and Tables 1-4. Both confidence intervals have estimated coverage probabilities close to the nominal confidence level for almost situations. However, the estimated coverage probabilities of the exact confidence interval are very poor when the coefficient of variation τ is close to 1 and small sample sizes. Additionally, the estimated coverage probabilities of the confidence intervals do not increase or decrease according to the values of τ and n. The estimated coverage probabilities of the proposed confidence interval are not significantly different from these of the exact confidence interval in any situation. However, the approximate confidence interval is very easy to calculate compared with the exact confidence interval because the exact confidence interval is based on an infinite summation.

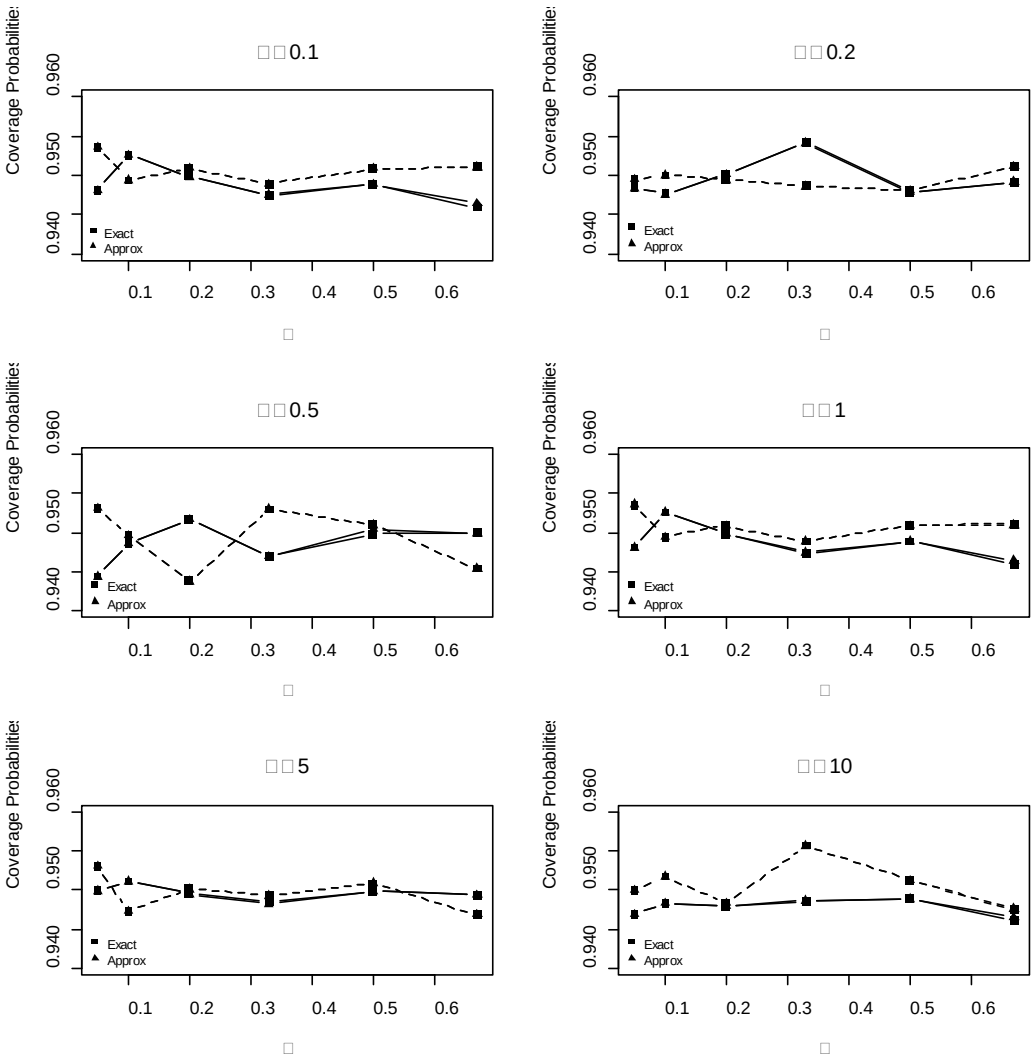

Figure 1: Estimated coverage probabilities of confidence intervals for the reciprocal of a normal mean with a known coefficient of variation when $n = 30$ (solid line) and $n = 100$ (dash line)

Table 1: Estimated coverage probabilities of confidence intervals for the reciprocal of a normal mean with a known coefficient of variation when $\theta = 0.1$ and 0.2.

n	τ	$\theta = 0.1$		$\theta = 0.2$	
		Exact	Approx.	Exact	Approx.
10	0.05	0.9475	0.9475	0.9489	0.9489
	0.10	0.9471	0.9471	0.9493	0.9493
	0.20	0.9498	0.9499	0.9500	0.9500
	0.33	0.9482	0.9483	0.9480	0.9486
	0.50	0.9325	0.9469	0.9334	0.9502
	0.67	0.0019	0.9456	0.0030	0.9455
20	0.05	0.9543	0.9543	0.9489	0.9489
	0.10	0.9489	0.9489	0.9529	0.9529
	0.20	0.9519	0.9519	0.9480	0.9479
	0.33	0.9514	0.9514	0.9492	0.9491
	0.50	0.9500	0.9505	0.9447	0.9452
	0.67	0.9475	0.9480	0.9457	0.9459
30	0.05	0.9481	0.9481	0.9484	0.9484
	0.10	0.9526	0.9526	0.9476	0.9476
	0.20	0.9498	0.9498	0.9501	0.9501
	0.33	0.9474	0.9475	0.9542	0.9541
	0.50	0.9489	0.9489	0.9479	0.9479
	0.67	0.9459	0.9464	0.9490	0.9492
50	0.05	0.9474	0.9474	0.9492	0.9492
	0.10	0.9485	0.9485	0.9500	0.9500
	0.20	0.9494	0.9494	0.9496	0.9496
	0.33	0.9499	0.9499	0.9476	0.9475
	0.50	0.9514	0.9517	0.9496	0.9497
	0.67	0.9485	0.9486	0.9476	0.9475
100	0.05	0.9536	0.9536	0.9495	0.9495
	0.10	0.9494	0.9494	0.9500	0.9500
	0.20	0.9509	0.9509	0.9494	0.9494
	0.33	0.9489	0.9489	0.9486	0.9486
	0.50	0.9509	0.9509	0.9481	0.9481
	0.67	0.9511	0.9510	0.9511	0.9511
500	0.05	0.9479	0.9479	0.9467	0.9467
	0.10	0.9488	0.9488	0.9511	0.9511
	0.20	0.9517	0.9517	0.9511	0.9511
	0.33	0.9519	0.9519	0.9481	0.9481
	0.50	0.9469	0.9469	0.9476	0.9476
	0.67	0.9484	0.9484	0.9480	0.9479

Table 2: Estimated coverage probabilities of confidence intervals for the reciprocal of a normal mean with a known coefficient of variation when $\theta = 0.5$ and 1.

n	τ	$\theta = 0.5$		$\theta = 1$	
		Exact	Approx.	Exact	Approx.
10	0.05	0.9489	0.9489	0.9475	0.9475
	0.10	0.9482	0.9482	0.9471	0.9471
	0.20	0.9491	0.9491	0.9498	0.9499
	0.33	0.9462	0.9463	0.9482	0.9483
	0.50	0.9357	0.9501	0.9325	0.9469
	0.67	0.0032	0.9471	0.0019	0.9456
20	0.05	0.9515	0.9515	0.9543	0.9543
	0.10	0.9482	0.9481	0.9489	0.9489
	0.20	0.9502	0.9502	0.9519	0.9519
	0.33	0.9518	0.9520	0.9514	0.9514
	0.50	0.9515	0.9518	0.9500	0.9505
	0.67	0.9445	0.9453	0.9475	0.9480
30	0.05	0.9444	0.9444	0.9481	0.9481
	0.10	0.9486	0.9486	0.9526	0.9526
	0.20	0.9517	0.9517	0.9498	0.9498
	0.33	0.9469	0.9470	0.9474	0.9475
	0.50	0.9499	0.9505	0.9489	0.9489
	0.67	0.9500	0.9498	0.9459	0.9464
50	0.05	0.9474	0.9474	0.9474	0.9474
	0.10	0.9520	0.9520	0.9485	0.9485
	0.20	0.9490	0.9490	0.9494	0.9494
	0.33	0.9485	0.9485	0.9499	0.9499
	0.50	0.9475	0.9475	0.9514	0.9517
	0.67	0.9503	0.9502	0.9485	0.9486
100	0.05	0.9531	0.9531	0.9536	0.9536
	0.10	0.9496	0.9496	0.9494	0.9494
	0.20	0.9438	0.9438	0.9509	0.9509
	0.33	0.9530	0.9530	0.9489	0.9489
	0.50	0.9510	0.9510	0.9509	0.9509
	0.67	0.9454	0.9454	0.9511	0.9510
500	0.05	0.9527	0.9527	0.9515	0.9515
	0.10	0.9469	0.9469	0.9507	0.9507
	0.20	0.9520	0.9520	0.9442	0.9442
	0.33	0.9500	0.9500	0.9495	0.9495
	0.50	0.9507	0.9507	0.9500	0.9500
	0.67	0.9507	0.9507	0.9519	0.9519

Table 3: Estimated coverage probabilities of confidence intervals for the reciprocal of a normal mean with a known coefficient of variation when $\theta = 5$ and 10.

n	τ	$\theta = 5$		$\theta = 10$	
		Exact	Approx.	Exact	Approx.
10	0.05	0.9489	0.9489	0.9508	0.9508
	0.10	0.9476	0.9476	0.9473	0.9473
	0.20	0.9516	0.9515	0.9482	0.9482
	0.33	0.9500	0.9501	0.9497	0.9497
	0.50	0.9326	0.9475	0.9335	0.9481
	0.67	0.0020	0.9457	0.0028	0.9505
20	0.05	0.9490	0.9490	0.9514	0.9514
	0.10	0.9490	0.9490	0.9478	0.9478
	0.20	0.9522	0.9521	0.9440	0.9440
	0.33	0.9497	0.9497	0.9504	0.9504
	0.50	0.9474	0.9479	0.9475	0.9479
	0.67	0.9454	0.9462	0.9472	0.9478
30	0.05	0.9499	0.9499	0.9469	0.9469
	0.10	0.9511	0.9511	0.9483	0.9483
	0.20	0.9495	0.9495	0.9479	0.9479
	0.33	0.9485	0.9482	0.9486	0.9487
	0.50	0.9498	0.9498	0.9489	0.9488
	0.67	0.9494	0.9494	0.9461	0.9465
50	0.05	0.9516	0.9516	0.9512	0.9512
	0.10	0.9521	0.9521	0.9496	0.9496
	0.20	0.9510	0.9510	0.9480	0.9480
	0.33	0.9496	0.9496	0.9481	0.9481
	0.50	0.9498	0.9497	0.9506	0.9505
	0.67	0.9513	0.9512	0.9471	0.9471
100	0.05	0.9531	0.9531	0.9500	0.9500
	0.10	0.9473	0.9473	0.9517	0.9517
	0.20	0.9501	0.9501	0.9483	0.9483
	0.33	0.9493	0.9493	0.9556	0.9556
	0.50	0.9509	0.9509	0.9512	0.9512
	0.67	0.9469	0.9469	0.9475	0.9476
500	0.05	0.9497	0.9497	0.9516	0.9516
	0.10	0.9510	0.9510	0.9505	0.9505
	0.20	0.9502	0.9502	0.9528	0.9528
	0.33	0.9486	0.9486	0.9521	0.9521
	0.50	0.9484	0.9484	0.9525	0.9525
	0.67	0.9518	0.9518	0.9493	0.9493

5 An Illustrative Example

To illustrate an example of two confidence interval for the reciprocal of a normal mean proposed in the previous section, we used the weights (in kilograms) of 61 one-month old infants listed as follows:

4.960 5.130 4.260 5.160 4.050 5.240 4.350 4.360 3.930 4.410

4.610 4.550 4.460 2.940 4.160 4.110 4.410 4.800 5.130 3.670
4.550 4.290 4.950 5.210 3.210 4.030 3.580 4.360 4.360 3.920
4.050 4.630 3.756 4.586 5.336 2.828 4.172 4.256 4.594 4.866
4.784 4.520 5.238 4.320 5.330 3.836 5.916 5.010 4.344 3.496
4.148 4.044 5.192 4.368 4.180 4.102 5.210 4.382 5.070 5.044
3.530

The data were taken from the study by Ziegler et al. (2007) (cited in Ledolter and Hogg, 2010, p.287). From past experience, we assume that the coefficient of variation of the weights of 61 one-month old infants is about 0.14. The histogram, density plot, Box-and-Whisker plot and normal quantile-quantile plot are displayed in Figure 2. Algorithm 1 shows the result of the Shapiro-Wilk normality test.

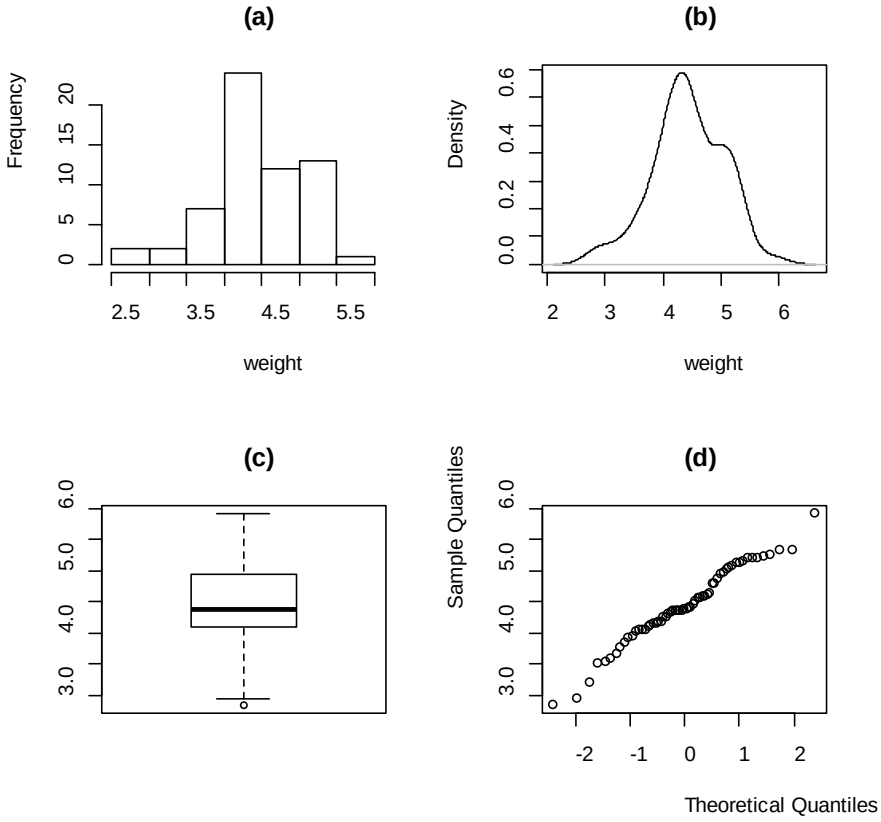

Figure 2: (a) Histogram, (b) density plot, (c) Box-and-Whisker plot and (d) normal quantile-quantile plot of the weight of a one-month old infant

```
Shapiro-Wilk normality test

data:  weight
W = 0.978, p-value = 0.3383
```

Algorithm 1: Shapiro-Wilk test for normality of the weight of a one-month old infant

The 95% exact and approximate confidence intervals for the reciprocal of a normal mean are calculated and reported in Table 4. The lower and upper limits of the both confidence intervals are not different.

Table 4: The 95% confidence intervals for the reciprocal of a normal mean of the weight of a one-month old infant.

Methods	Confidence Intervals		Lengths
	Lower Limit	Upper Limit	
Exact	0.2176837	0.2335416	0.0158579
Approximate	0.2176838	0.2335416	0.0158578

6 Conclusions

In this paper, we proposed an approximate confidence interval for the reciprocal of a normal population mean with a known coefficient of variation. Normally, this arises when the coefficient of variation of the control group is known. The approximate confidence interval proposed uses the approximation of the expectation and variance of the estimator. The proposed new confidence interval is compared with the exact confidence interval constructed by Wongkhao et al. (2013) through a Monte Carlo simulation study. The approximate confidence interval performs as efficiently as the exact confidence interval in terms of coverage probability. Moreover, approximate confidence interval also is easy to compute compared with the exact confidence interval.

Appendix: Source R code for all confidence intervals

```
ci.exact <- function(y,tao,alpha) {
        n <- length(y)
        ybar <- mean(y)
        zeta.hat <- 1/ybar
        w <- cal.w(tao,n)
```

```
        z <- qnorm(1-alpha/2)
        T1 <- (tao^2)/(n*(ybar^2))
        lower <- (zeta.hat/w)-z*sqrt(T1)
        upper <- (zeta.hat/w)+z*sqrt(T1)
        out <- cbind(lower,upper)
        return(out)
}

ci.approx <- function(y,tao,alpha) {
        n <- length(y)
        ybar <- mean(y)
        zeta.hat <- 1/ybar
        v <- 1+(tao^2)/n
        z <- qnorm(1-alpha/2)
        T1 <- ((zeta.hat^2)*(tao^2))/n
        lower <- (zeta.hat/v)-z*sqrt(T1)
        upper <- (zeta.hat/v)+z*sqrt(T1)
        out <- cbind(lower,upper)
        return(out)

}

cal.w <- function(tao,n) {
        temp <- rep(0,50)
        for (k in 1:50) {
        temp[k] <- (factorial(2*k)/((2^k)*factorial(k)))*(((tao^2)/n)^k)
        }
        w <- 1+sum(temp)
        return(w)
}
```

Acknowledgements

The author is grateful to two anonymous referees for their valuable comments and comments, which have significantly enhanced the quality and presentation of this paper. The author is also thankful for the support in the form of the research funds awarded by Thammasat University.

References

[1] Ihaka, R. and Gentleman, R. (1996): R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics,* **5**, 299-314.

[2] Lamanna, E., Romano, G. and Sgrbi, C. (1981): Curvature measurements in nuclear emulsions. *Nuclear Instruments and Methods*, **187**, 387-391.

[3] Ledolter, L., Hogg, R.V. (2010): *Applied Statistics for Engineers and Physical Scientists*, Pearson, New Jersey.

[4] Weerahandi, S. (1993): Generalized confidence intervals, *Journal of the American Statistical Association*, **88**, 899-905.

[5] Withers, C.S. and Nadarajah, S. (2013): Estimators for the inverse powers of a normal mean, *Journal of Statistical Planning and Inference*, **143**, 441-455.

[6] Wongkhao, A., Niwitpong, S. and Niwitpong, S. (2013): Confidence interval for the inverse of a normal mean with a known coefficient of variation. *International Journal of Mathematical, Computational, Statistical, Natural and Physical Engineer*, **7**, 877-880.

[7] Zaman, A. (1981): Estimators without moments: the case of the reciprocal of a normal mean. *Journal of Econometrics*, **15**, 289-298.

[8] Zaman, A. (1985): Admissibility of the maximum likelihood estimate of the reciprocal of a normal mean with a class of zero-one loss functions. *Sankhyā: The Indian Journal of Statistics, Series A*, **47**, 239-246.

[9] Ziegler, E., Nelson, S.E., Jeter, J.M. (2007): *Early iron supplementation of breastfed infants*, Department of Pediatrics, University of Iowa, Iowa City, USA.

Metodološki zvezki, Vol. 13, No. 2, 2016, 131-150

Exploring the Relationship between two Compositions using Canonical Correlation Analysis

Glòria Mateu-Figueras[1], Josep Daunis-i-Estadella[2], Germà Coenders[3], Berta Ferrer-Rosell[4], Ricard Serlavós[5], Joan Manuel Batista-Foguet[6]

Abstract

The aim of this article is to describe a method for relating two compositions which combines compositional data analysis and canonical correlation analysis (CCA), and to examine its main statistical properties. We use additive log-ratio (alr) transformation on both compositions and apply standard CCA to the transformed data. We show that canonical variates are themselves log-ratios and log-contrasts. The first pair of canonical variates can be interpreted as the log-contrast of a composition that has the maximum correlation with a log-contrast of the other composition. The second pair can be interpreted as the log-contrast of a composition that has the maximum correlation with a log-contrast of the other composition, under the restriction that they are uncorrelated with the first pair, and so on.

Using properties from changes of basis, we prove that both canonical correlations and canonical variates are invariant to the choice of divisors in alr transformation. We show how to implement the analysis and interpret the results by means of an illustration from the social sciences field using data from Kolb's Learning Style Inventory and Boyatzis' Philosophical Orientation Questionnaire, which distribute a fixed total score among several learning modes and philosophical orientations.

[1] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, 17003 Girona, Spain; gloria.mateu@udg.edu

[2] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, 17003 Girona, Spain; pepus.daunis@udg.edu

[3] Department of Economics, University of Girona, Campus Montilivi, 17003 Girona, Spain; germa.coenders@udg.edu

[4] Department of Economics, University of Girona, Edifici Sant Domènec, Plaça Ferrater Mora 1, 17004 Girona, Spain; berta.ferrer@udg.edu

[5] Department of People Management and Organization, ESADE, University Ramon Llull, Av. Pedralbes, 60-62, 08034 Barcelona, Spain; ricard.serlavos@esade.edu

[6] Department of People Management and Organization, ESADE, University Ramon Llull, Av. Pedralbes, 60-62, 08034 Barcelona, Spain; joanm.batista@esade.edu

1 Introduction

Compositional data lie in a constrained positive space with a fixed sum and convey information on the relative importance of components. Typical examples are chemical and geological compositions (adding to 100% in weight or volume), genotype frequencies (adding to 1), time use (adding to 24 hours), voting (adding to 100% of votes), or household budget allocation (adding to 100% of the budget). The fixed sum is typically normalized to one, and a D-term composition ($x_1, x_2,..., x_D$) is thus constrained as follows:

$$0<x_d<1 \text{ and } \sum_{d=1}^{D} x_d = 1 \tag{1.1}$$

Serious problems arise when using standard statistical analysis tools on compositional data (Aitchison, 1986, 2001; Pawlowsky-Glahn & Buccianti, 2011):

1. Compositional data have a bounded distribution. This implies at least non-normality and heteroscedasticity (lower variance close to the boundary).

2. One component can only increase if some others decrease. This results in negative spurious correlations among the components (Pearson, 1897) and prevents interpreting effects of linear models in the usual way "keeping everything else constant".

3. The true dimensionality of a set of compositional variables is D–1. Analysis of all D dimensions leads to perfect collinearity.

4. Compositional data lie in a (D–1)-dimensional Euclidean space called the simplex, with different operations and distance from real space (Billheimer et al., 2001; Pawlowsky-Glahn & Egozcue, 2001).

The compositional data analysis (CoDa) tradition started with Aitchison's seminal work (1986) on treating chemical and biological compositions. Nowadays, however, it spans almost all of the hard sciences and has started to be used in the social sciences, which often face similar problems (Batista-Foguet et al., 2015; Coenders et al., 2011; van Eijnatten et al., 2015; Ferrer-Rosell & Coenders, 2016; Ferrer-Rosell et al., 2015, 2016a, 2016b; Fry, 2011; Hlebec et al., 2012; Kogovšek et al., 2013; Vives-Mestres et at., 2016).

The literature on CoDa has extensively dealt with relating one composition to non-compositional data (Egozcue et al., 2012; Hron et al., 2012; Martín-Fernández et al., 2015) and with analyzing one single composition. As far as the exploratory data analysis of one

single composition is concerned (Egozcue & Pawlowsky-Glahn, 2011), available methods include the variation array (Aitchison, 1986), principal component analysis (Aitchison, 1983; Aitchison & Greenacre, 2012), the CoDa-dendrogram (Pawlowsky-Glahn & Egozcue, 2011), and the CoDa-biplot (Aitchison & Greenacre, 2012). As regards exploratory tools to relate two compositions, the natural choice is canonical correlation analysis – CCA (Aitchison, 1986). Typical problems relating two compositions include the relationship between the composition of species and the chemical composition of the environment (ter Braak, 1996); between the composition of foods and the composition of their energy and nutrients; or between the composition of materials and the composition of spectral curves in image processing. The use of CCA for compositional data was foreshadowed in Aitchison (1986), without much mention of its properties or interpretation. At a later date, van den Boogaart and Tolosana-Delgado (2013) devised an advanced procedure for compositional CCA requiring software designed for this purpose.

Drawing from Aitchison (1986), in this article we develop and illustrate a simple procedure for carrying out CCA of two compositional vectors and examine its interpretation and main statistical properties. Even if specialized techniques for compositional data have appeared (van den Boogaart & Tolosana-Delgado, 2013; Pawlowsky-Glahn & Buccianti, 2011; Pawlowsky Glahn et al., 2015; Thió-Henestrosa & Martín-Fernández, 2005), compositional data can also be transformed so that they can be subject to standard and well-understood statistical techniques carried out using standard software. This is the approach we take in this article.

Given the fact that only information on the relative size of components is available in a compositional data context, logarithms of ratios between component values are a meaningful way of expressing the data and guaranteeing the principles of CoDa (Aitchison, 2001). A logarithm of a ratio is scale invariant, meaning that it does not change if the values involved are multiplied by an arbitrary constant. Adding or dropping components from a composition does not modify the log-ratios computed from the remaining components. This is related to the principles of scale invariance and subcompositional coherence. For full details on CoDa principles, see Pawlowsky Glahn et al. (2015).

Several log-ratio transformations have been suggested in the literature (Egozcue et al., 2003). Additive log-ratio transformation (alr) is the easiest to compute since it is simply the log-ratio between each component and the last:

$$y_d=\ln(x_d/x_D)=\ln(x_d)-\ln(x_D) \text{ with } d=1,2,...,D-1 \quad (1.2)$$

Alr-transformed y_d variables recover the full unconstrained real space. It must be noted that one dimension is lost. Although alr transformation is used in this article due to its simplicity, there are alternatives (see Egozcue et al., 2003 for a general background on the transformations and Section 3.3. for a discussion of their applicability to CCA).

Since the decision on which component to leave in last place and serve as a reference in the alr transformation is often arbitrary, there is concern regarding whether the results of a statistical analysis are invariant to this arbitrary choice. Of course, different log-ratios constitute different variables and the raw results will never be invariant. However, it is considered desirable that overall goodness of fit measures be invariant to this choice. Once results are reexpressed as a function of the log components $\ln(x_d)$, they should ideally also be invariant.

The structure of the article is as follows. First, we review the basics of CCA. We then come to the particular case in which CCA is applied to compositions that have been subjected to alr transformation, showing how to interpret the key results, proving that they are invariant to the choice of reference component, and discussing alternative transformations. Following this, we present an illustration from the field of education using data from Kolb's Learning Style Inventory (Batista-Foguet et al., 2015; Kolb, 1984, 1999) and Boyatzis' Philosophical Orientation Questionnaire (Boyatzis et al., 2000). The final section discusses the strengths and weaknesses of the method.

2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate analysis technique which studies the relationships between two sets of variables $\mathbf{Y}_a=(y_{a1}, y_{a2}, ..., y_{ap})$ and $\mathbf{Y}_b=(y_{b1}, y_{b2}, ..., y_{bq})$ usually defined in the real space. The method was first introduced in Hotelling (1936) and a non-technical description can be found in Hair et al. (2009).

CCA builds pairs of linear combinations of each set of variables called canonical variates. The first canonical variate cv_{a1} for set $\mathbf{Y}_a$ is derived so that it is maximally correlated with the first canonical variate cv_{b1} for set $\mathbf{Y}_b$. The second canonical variate cv_{a2} for set $\mathbf{Y}_a$ is derived so that it is maximally correlated with the second canonical variate cv_{b2} for set $\mathbf{Y}_b$ under the restriction that both new canonical variates are uncorrelated with cv_{a1} and cv_{b1}. The following

pairs are extracted in a similar manner and have the maximum mutual correlation, while being uncorrelated with the previous pairs. The process may be continued up to $\min\{p,q\}$ times.

The raw canonical coefficients w_{aij} and w_{bij} are the weights used to compute the i-th pair of canonical variates from the j-th original variables:

$$\begin{aligned} cv_{a1} &= w_{a11}y_{a1} + w_{a12}y_{a2} + \cdots + w_{a1p}y_{ap} \\ cv_{b1} &= w_{b11}y_{b1} + w_{b12}y_{b2} + \cdots + w_{b1q}y_{bq} \\ cv_{a2} &= w_{a21}y_{a1} + w_{a22}y_{a2} + \cdots + w_{a2p}y_{ap} \\ cv_{b2} &= w_{b21}y_{b1} + w_{b22}y_{b2} + \cdots + w_{b2q}y_{bq} \\ &\cdots\cdots \end{aligned} \tag{2.1}$$

In practice, the canonical coefficients are computed from three covariance matrices: the square matrix $\mathbf{S}_{aa}$ contains covariances in the first variable set, the square matrix $\mathbf{S}_{bb}$ covariances in the second set, and the rectangular matrix $\mathbf{S}_{ab}$ covariances between variables of one set and the other. Canonical variates are obtained from an eigenvalue analysis of matrix:

$$\mathbf{S}_{aa}^{-1}\mathbf{S}_{ab}\mathbf{S}_{bb}^{-1}\mathbf{S}_{ba} \tag{2.2}$$

The correlation between cv_{a1} and cv_{b1} is the first canonical correlation $\hat{\rho}_1$, the correlation between cv_{a2} and cv_{b2} is the second canonical correlation $\hat{\rho}_2$, and so on. These canonical correlations are obtained as the square root of the eigenvalues of the matrix in Equation (2.2).

The maximum number of canonical variates that can be extracted is the smallest dimension of the two sets of variables. For instance, if p=5 and q=8, then a maximum of 5 pairs of variates can be obtained. As with many other multivariate analysis techniques, the researcher is interested in a parsimonious solution and in interpreting only the most relevant variates. The relevance of a pair of canonical variates can be assessed by the sheer size of the canonical correlation, the interpretability of the canonical variates from the canonical weights, or the statistical significance of the canonical correlations according to Wilks' Λ tests, which are also a function of the eigenvalues. Since, $\hat{\rho}_1 > \hat{\rho}_2 > \ldots > \hat{\rho}_{\min\{p,q\}}$, a common strategy is to sequentially test the following hypotheses:

$$H_{01}: \rho_1=\rho_2=\rho_3=\cdots=\rho_{\min\{p,q\}}=0$$

$$H_{02}: \rho_2=\rho_3=\cdots=\rho_{\min\{p,q\}}=0$$

$$\ldots\ldots \tag{2.3}$$

$$H_{0\min\{p,q\}-1}: \rho_{\min\{p,q\}-1}=\rho_{\min\{p,q\}}=0$$

$$H_{0\min\{p,q\}}: \rho_{\min\{p,q\}}=0$$

The rejection of H_{01} to H_{0i} and the failure to reject H_{0i+1} to $H_{0\min\{p,q\}}$ shows the first i canonical correlations to be statistically significant.

Other common results of a CCA which provide a useful aid to interpreting the canonical variates require standardization in some form or other (Hair et al., 2009) and are:

1. Standardized canonical coefficients (coefficients used to compute canonical variates from standardized *y* variables).

2. Canonical loadings (correlations between the canonical variates and the *y* variables they are computed from).

3. Canonical cross-loadings (correlations between canonical variates and the other set of *y* variables).

4. Redundancy analysis (percentages of variance for the *y* variables explained by their own canonical variates and from the canonical variates computed from the other set of *y* variables).

3 Canonical Correlation Analysis of Compositional Data Transformed by Means of alr

3.1 Interpretation

Given two compositions with D_a and D_b components, $X_a=(x_{a1},x_{a2},\ldots,x_{aDa})$ and $X_b=(x_{b1},x_{b2},\ldots,x_{bDb})$, following Aitchison (1986) we first apply alr transformation with the last component in the denominator. The results are the following two real vectors with $p=D_a-1$ and $q=D_b-1$ elements:

$$\mathbf{Y}_a = \left(\ln\left(\frac{x_{a1}}{x_{aDa}}\right), \ln\left(\frac{x_{a2}}{x_{aDa}}\right), \ldots, \ln\left(\frac{x_{ap}}{x_{aDa}}\right) \right)$$

$$\mathbf{Y}_b = \left(\ln\left(\frac{x_{b1}}{x_{bDb}}\right), \ln\left(\frac{x_{b2}}{x_{bDb}}\right), \ldots, \ln\left(\frac{x_{bq}}{x_{bDb}}\right) \right) \tag{3.1}$$

We can rewrite Equation (3.1) as:

$$\mathbf{Y}_a = \left(\ln(x_{a1}) - \ln(x_{aDa}), \ln(x_{a2}) - \ln(x_{aDa}), \ldots, \ln(x_{ap}) - \ln(x_{aDa}) \right)$$

$$\mathbf{Y}_b = \left(\ln(x_{b1}) - \ln(x_{bDb}), \ln(x_{b2}) - \ln(x_{bDb}), \ldots, \ln(x_{bq}) - \ln(x_{bDb}) \right) \tag{3.2}$$

$\mathbf{Y}_a$ and $\mathbf{Y}_b$ are two sets of real variables to which we can apply the standard CCA procedure from the covariance matrices of each set of transformed variables and the covariance matrix between the transformed variables of one set and the other in Equation (2.2).

The first pair of canonical variates in Equation (2.1), when expressed in terms of logarithms of components, becomes:

$$cv_{a1} = w_{a11}\ln(x_{a1}) + w_{a12}\ln(x_{a2}) + \cdots + w_{a1p}\ln(x_{ap}) - (w_{a11} + w_{a12} + \cdots + w_{a1p})\ln(x_{aDa})$$

$$cv_{b1} = w_{b11}\ln(x_{b1}) + w_{b12}\ln(x_{b2}) + \cdots + w_{b1q}\ln(x_{bq}) - (w_{b11} + w_{b12} + \cdots + w_{b1q})\ln(x_{bDb}) \tag{3.3}$$

Since the raw canonical coefficients are applied from $\ln(x_{a1})$ to $\ln(x_{ap})$ and again to $\ln(x_{aDa})$ with reversed signs, the weights of all D_a logarithms add up to zero, and the same occurs with the weights of the D_b logarithms of the x_b variables. This would also hold for the remaining canonical variates.

This is the same as saying that the canonical variates are log ratios of the product of components with a positive weight raised to a power equal to that weight, over the product of components with a negative weight raised to a power equal to the absolute weight. Let us show an example of the former for a canonical variate of a 5-dimensional composition with:

$$cv_{a1} = 1y_{a1} + 1.5y_{a2} + 0.5y_{a3} - 0.5\,y_{a4} \tag{3.4}$$

The reexpression of this canonical variate as a log-ratio is:

$$cv_{a1} = 1\ln(x_{a1}) + 1.5\ln(x_{a2}) + 0.5\ln(x_{a3}) - 0.5\ln(x_{a4}) - 2.5\ln(x_{a5}) = \ln\left(\frac{x_{a1}x_{a2}^{1.5}x_{a3}^{0.5}}{x_{a4}^{0.5}x_{a5}^{2.5}}\right) \tag{3.5}$$

The cv_{a1} log-ratio in this example is high mainly when x_{a1} and x_{a2} are high and x_{a5} is low. Since the sum of positive exponents equals the sum of negative exponents, the log-ratio is also a log-contrast, that is, a log-linear combination where the sum of the coefficients is 0 (Aitchison, 1986: 84).

The first pair of canonical variates can thus be interpreted as the log-contrast of one of the compositions that has the maximum correlation with a log-contrast of the other composition. The second pair can be interpreted as the log-contrast of one of the compositions that has the maximum correlation with a log-contrast of the other composition, under the restriction that they are uncorrelated with the first pair of canonical variates. A similar interpretation would hold for the third pair, subject to zero correlation with the first two pairs, and so on.

3.2 Invariance of the Results to the Choice of Reference Component in alr

Although the last component in each composition was chosen as the common divisor in our alr transformation, this could equally have been any other component. Consequently, for any analysis involving alr vectors, it is important to check the invariance of the key results to component permutations, or in other words, their invariance with respect to the choice of divisor in alr transformation. In this section we show specifically that Wilks' Λ tests, canonical correlations, and canonical variates as functions of log components –Equation (3.3)– are invariant to this choice.

It is easy to see how two alr-transformed vectors using different components as a divisor are related using a change-of-basis matrix. Following Mateu-Figueras et al. (2011), the elements of an alr vector are the coefficients of the original composition with respect to a particular non-orthonormal basis on the simplex, the sample space of compositional data. The effect of changing the common divisor is to obtain the coefficients with respect to another particular basis, which is analogous to performing an oblique rotation of the data.

Let $\mathbf{Y}_a$ and $\mathbf{Y}_b$ be the alr transformed vectors using the last components as common divisors and let $\mathbf{Y}_a^*$ and $\mathbf{Y}_b^*$ be the alr-transformed vectors using other components as denominators. Then, $\mathbf{Y}_a^*=\mathbf{Q}\mathbf{Y}_a$ and $\mathbf{Y}_b^*=\mathbf{P}\mathbf{Y}_b$. We can obtain the exact expression of matrices **Q** and **P** (see Aitchison, 1986: 94), but the important point here is that matrices **Q** and **P** are change-of-basis matrices. From the usual properties of covariance matrices we know that:

$$\mathbf{S}^*_{aa} = \mathbf{Q}\mathbf{S}_{aa}\mathbf{Q}' \tag{3.6}$$

$$\mathbf{S}^*_{bb} = \mathbf{P}\mathbf{S}_{bb}\mathbf{P}' \tag{3.7}$$

$$\mathbf{S}^*_{ab} = \mathbf{Q}\mathbf{S}_{ab}\mathbf{P}' \text{ and } \mathbf{S}^*_{ba} = \mathbf{P}\mathbf{S}_{ba}\mathbf{Q}' \tag{3.8}$$

When using different common divisors in alr transformation, the analyzed matrix in Equation (2.2) becomes:

$$\left(\mathbf{S}^*_{aa}\right)^{-1}\mathbf{S}^*_{ab}\left(\mathbf{S}^*_{bb}\right)^{-1}\mathbf{S}^*_{ba} \tag{3.9}$$

By using the relationships in Equations (3.6)–(3.8), Equation (3.9) becomes:

$$\begin{aligned}&\left(\mathbf{S}^*_{aa}\right)^{-1}\mathbf{S}^*_{ab}\left(\mathbf{S}^*_{bb}\right)^{-1}\mathbf{S}^*_{ba} = \left(\mathbf{Q}\mathbf{S}_{aa}\mathbf{Q}'\right)^{-1}\left(\mathbf{Q}\mathbf{S}_{ab}\mathbf{P}'\right)\left(\mathbf{P}\mathbf{S}_{bb}\mathbf{P}'\right)^{-1}\left(\mathbf{P}\mathbf{S}_{ba}\mathbf{Q}'\right) = \\ &\left(\mathbf{Q}'\right)^{-1}\mathbf{S}^{-1}_{aa}\mathbf{S}_{ab}\mathbf{S}^{-1}_{bb}\mathbf{S}_{ba}\mathbf{Q}'\end{aligned} \tag{3.10}$$

From linear algebra properties, we know that the eigenvalues of a matrix are invariant under changes of basis. Consequently, both the canonical correlations and Wilks' Λ tests are invariant under change of common divisor in alr transformation.

It is easy to see how the normalized eigenvectors of matrices in Equations (3.9) and (2.2), denoted as $\mathbf{w}^*_{ai}$ and $\mathbf{w}_{ai}$ respectively, must be related by $\mathbf{Q}'\mathbf{w}^*_{ai} = \mathbf{w}_{ai}$ or $\mathbf{w}^*_{ai} = \left(\mathbf{Q}'\right)^{-1}\mathbf{w}_{ai}$. Then we obtain the invariance of the corresponding canonical variates as:

$$cv^*_{ai} = \left(\mathbf{w}^*_{ai}\right)'\mathbf{Y}^*_a = \left(\left(\mathbf{Q}'\right)^{-1}\mathbf{w}_{ai}\right)'\mathbf{Q}\mathbf{Y}_a = \mathbf{w}'_{ai}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Y}_a = \mathbf{w}'_{ai}\mathbf{Y}_a = cv_{ai} \tag{3.11}$$

Conversely, all results that imply standardization, like standardized canonical coefficients, canonical loadings/cross-loadings and redundancy analysis, are not invariant to the choice of reference component in alr transformation. In the case of CoDa, however, given the facts that canonical variates can be readily interpreted as log-ratios and log-contrasts on their own, and that standardization is extremely uncommon for log-contrasts, standardized information is not needed to enhance interpretation and is not considered in this article.

3.3 Appropriateness of Alternative Log-ratio Transformations for Canonical Correlation Analysis

One key issue when working with CoDa is the choice of the log-ratio transformation, since different possibilities are available. Additive log-ratio (alr) and centered log-ratio (clr) transformations were introduced in Aitchison (1986), while isometric log-ratio transformation (ilr) was introduced in Egozcue et al. (2003).

Aitchison's (1986) proposal for compositional CCA involved alr transformation. Although alr transformation is simple and easy to interpret, it is asymmetric in its parts. By changing the part in the denominator, a different alr-transformed vector is obtained. For this reason, when alr transformation is used, it is important to check the invariance of the results with respect to the choice of common denominator, as we have done in Section 3.2. However, as Egozcue et al. (2003) noted, the main drawback of alr transformation is that it is not an isometric transformation from the simplex to the real space. It was later shown that an alr vector can be viewed as the coefficients of a composition with respect to a non-orthonormal basis on the simplex (Mateu-Figueras et al., 2011). Consequently, it is not suitable for statistical techniques that use distances or angles between alr vectors, such as cluster analysis. Note that these problems do not occur when using CCA because eigenvalues and eigenvectors of a product of covariance matrices are involved. Due to the non-orthonormality of the basis, the equality $\mathbf{Q}'\mathbf{w}_{ai}^{*} = \mathbf{w}_{ai}$ is only true if the vector product $\mathbf{Q}'\mathbf{w}_{ai}^{*}$ is normalized, although this does not affect the analyses considered in this article.

Clr transformation is defined as the logarithm of the ratio of each part over the geometric mean. It is a symmetric transformation with respect to the compositional parts and also an isometric transformation. Nevertheless, clr transformation has the disadvantage that the clr covariance matrix is singular. In our case, clr transformation would not be a good choice because CCA uses covariance matrices and their inverses. Conversely, it would be a good choice for cluster analysis or other statistical techniques in which distances are crucial and covariances do not need to be inverted.

Ilr transformation is isometric and consequently makes it possible to associate distances in the simplex with distances in the transformed space. Additionally, an ilr vector can be viewed as the coordinates of a composition with respect to an orthonormal basis on the simplex. Finally, covariance matrices can be inverted. It can thus be used in virtually all statistical analyses. The expression of the ilr using a particular orthonormal basis is given in Egozcue et

al. (2003). Nevertheless, in inner product spaces, an orthonormal basis is not uniquely determined and in some cases it is not straightforward to determine which basis is the most appropriate to solve a specific problem and how it can be interpreted. Faced with the problem of interpreting CCA on ilr transformation, van den Boogaart and Tolosana-Delgado (2013) devise a graphical back-transformation of the canonical coefficients. In any case, the invariance of the ilr results with respect to the choice of the orthonormal basis also holds.

In this article, although alr transformation was used due to its simplicity, ilr transformation could also have been used, and we actually did rerun the illustration analysis with ilr transformation. The final canonical variates expressed in terms of the log components and as log-contrasts are invariant, because alr and ilr vectors are also related through a change-of-basis matrix.

4 Illustration

4.1 Background

In this illustration of compositional CCA, our aim is to relate students' learning styles to their philosophical orientations. Philosophical orientation is a good means of understanding the relationship between people's values and beliefs, and their behavior and approach to learning (Boyatzis et al., 2000). Since a person's behavior is related to his or her values and beliefs, philosophy is important for comprehending and predicting behavior, with the added advantage that a person's philosophy goes beyond social context. Philosophical orientation is useful for answering questions such as how individuals 'act across various social settings' or 'think about establishing the value of things, activities and others' (Boyatzis et al., 2000: 50). Three major clusters of philosophical systems have traditionally been proposed. These clusters define the extent to which a person is pragmatic (PR), intellectual (IN) or humanistic (HU).

A person with a predominantly PR philosophical orientation will make decisions based on the benefits of the action, measured in terms of utility or comparing input and output. If the objectives to be achieved are not clear or measuring utility is difficult, then an activity will be less valuable to a person with this orientation.

Someone with a predominantly IN philosophical orientation will be rational, logical and focus on comprehending everything. The central concern underlying this philosophical orientation is analytical.

Someone with a predominantly HU orientation is thought to be committed to human values. This kind of person will tend to determine whether an activity is worthy in terms of its impact on other people and the quality of the relationship with these people. The central issue underlying HU orientation is a concern for close and personal relationships.

According to Experiential Learning Theory, learning is a process whereby knowledge is created through the transformation of experience (Kolb, 1984). Learning requires abilities to grasp and transform knowledge that are polar opposites. In grasping knowledge, some people perceive new information through experiencing the concrete, tangible, and felt qualities of the world, which is referred to as concrete experience (CE), while others tend to take hold of new information through symbolic representation or abstract conceptualization (AC). In transforming knowledge, some people tend to carefully watch others who are involved in the experience and reflect on what happens (reflective observation – RO), while others choose to start doing things (active experimentation - AE). Learning can also be conceived as a four-stage cycle, where each stage is represented by a learning mode.

At the CE stage, one tends to rely more on intuition than on a systematic focus. Moreover, in this stage, a learner relies on the ability to be open, receptive and adaptive to changes. At the RO stage, one comprehends situations by taking into account different perspectives. In this stage, a learner relies on patience and objectivity, as well as thoughts and feelings. At the AC stage, logic and ideas are needed to understand a problem, rather than feelings. A learner in this stage relies on systematic planning and the theoretical development of ideas. Finally, at the AE stage, one learns by experimenting with changing situations. In this stage, a learner will find it more useful to put ideas into practice and see what really works than to simply observe.

4.2 Data and Measures

Multidimensional forced-choice questionnaires to measure philosophical orientations and learning modes were designed in Boyatzis et al. (2000) and Kolb (1999). In these questionnaires, each question consists of a set of *D* statements, and each statement is an indicator of a different dimension, in our case, of a philosophical orientation (D=3) or a

learning mode (D=4). Respondents are instructed to rank these statements. In this article, we assume that ranks are coded as $D-1$ for the most preferred statement, $D-2$ to the second most preferred, down to 0 for the least preferred. The Philosophical Orientation Questionnaire consists of k=20 questions designed as in this example:

"I think of my value, or worth, in terms of:

(a) My relationships (e.g. family, friends).

(b) My ideas or ability to invent new concepts or ability to analyse things.

(c) My financial net worth or income."

Statement (a) reflects the HU orientation, (b) the IN orientation, and (c) the PR orientation.

The Learning Style Inventory includes k=12 questions designed as in this example:

"*When I am learning:*

(a) I like to experience sensations.

(b) I like to observe and listen.

(c) I like to think about ideas.

(d) I like to do things."

Statement (a) reflects the CE mode, (b) the RO mode, (c) the AC mode, and (d) the AE mode.

The ranks of each dimension are summed across the k questions to produce D global scores, one for each dimension. These D scores have a fixed sum for all respondents, equal to $kD(D-1)/2$. Once the global scores have been computed, forced-choice instruments can be understood as compositions, in which the $kD(D-1)/2$ total is allocated to the D dimensions (components), so that data only convey information about the relative importance of dimensions (learning modes and philosophical orientations) for a given individual. Under this coding scheme, the dimension score is the number of times the dimension has been preferred over other dimensions in all possible pair-wise comparisons over the k questions. For instance, if a component is always ranked as the lowest, it has never been preferred to any other mode and receives a 0 score. If a component is always ranked as the highest, it is preferred k times to the other $D-1$ modes and receives a $k(D-1)$ score. Scores can thus be understood as having ratio scale properties: a component with a score of 6 has been preferred

to other components twice as many times across the k items than a mode with score of 3 (Batista-Foguet et al., 2015). Alternative ways of coding these questionnaires are discussed in de Vries and van der Ark (2008).

In this illustration, we use the same data as those used by Batista-Foguet et al. (2015), which cover 7 consecutive years (2006-2013) of candidates on an international MBA program at a leading European business school. The sample size was 1,194 full time participants from 86 countries, of which the most common were Spain (15.9%), the US (13.7%), India (9.6%), and Germany (5.6%). 69.7% were male and 30.3% female. Average age was 31.4 years (SD 2.8 years). Previous student background was heterogeneous, including not only economics (11%) and management studies (32%), but also engineering (36.4%), social sciences (9.3%), arts (5.7%) and hard sciences (5.5%).

The philosophical orientation components were labeled x_{p1}=pragmatic (PR), x_{p2}=intellectual (IN), and x_{p3}=humanistic (HU); while the learning mode components were labeled x_{l1}=abstract conceptualization (AC), x_{l2}=concrete experience (CE), x_{l3}=active experimentation (AE), and x_{l4}=reflective observation (RO). The final two components, HU and RO, were used as a reference for the alr transformation:

$$\begin{aligned}
\text{log - ratio of PR over HU } y_{p1} &= \ln\left(\frac{x_{p1}}{x_{p3}}\right) = \ln\left(x_{p1}\right) - \ln\left(x_{p3}\right) \\
\text{log - ratio of IN over HU } y_{p2} &= \ln\left(\frac{x_{p2}}{x_{p3}}\right) = \ln\left(x_{p2}\right) - \ln\left(x_{p3}\right) \\
\text{log - ratio of AC over RO } y_{l1} &= \ln\left(\frac{x_{l1}}{x_{l4}}\right) = \ln\left(x_{l1}\right) - \ln\left(x_{l4}\right) \\
\text{log - ratio of CE over RO } y_{l2} &= \ln\left(\frac{x_{l2}}{x_{l4}}\right) = \ln\left(x_{l2}\right) - \ln\left(x_{l4}\right) \\
\text{log - ratio of AE over RO } y_{l3} &= \ln\left(\frac{x_{l3}}{x_{l4}}\right) = \ln\left(x_{l3}\right) - \ln\left(x_{l4}\right)
\end{aligned} \tag{4.1}$$

4.3 Results

After submitting the sets (y_{p1}, y_{p2}) and (y_{l1}, y_{l2}, y_{l3}) to a CCA using SPSS v.23, the resulting canonical correlations are $\hat{\rho}_1$=0.246 and $\hat{\rho}_2$=0.163. Their significance tests are in Table 1. The raw (unstandardized) canonical coefficients are in Table 2.

Table 1: Significance Tests for the Canonical Correlations

H_0	Wilk's Λ	χ^2	DF	*p*-value
$\rho_1=\rho_2=0$	0.914	93.854	6	0.000
$\rho_2=0$	0.973	28.295	2	0.000

Table 2: Raw Canonical Coefficients as a Function of the Log-ratios

	Variate 1	Variate 2
Philosophical orientations		
y_{p1} (log-ratio of PR over HU)	-0.524	1.730
y_{p2} (log-ratio of IN over HU)	2.085	-0.274
Learning modes		
y_{l1} (log-ratio of AC over RO)	1.720	-0.177
y_{l2} (log-ratio of CE over RO)	-0.447	-1.347
y_{l3} (log-ratio of AE over RO)	-1.032	1.311

The original canonical variates are functions of the log ratios and are easily re-expressed by hand as a function of the log-components as in Equation (3.3). For instance, in the philosophical orientation composition the first canonical variate is:

$$\begin{aligned} cv_{p1} &= -0.524\, y_{p1} + 2.085\, y_{p2} = \\ &-0.524 \ln(x_{p1}) + 0.524 \ln(x_{p3}) + 2.085 \ln(x_{p2}) - 2.085 \ln(x_{p3}) = \\ &-0.524 \ln(x_{p1}) + 2.085 \ln(x_{p2}) - 1.561 \ln(x_{p3}) \end{aligned} \tag{4.2}$$

Table 3: Raw Canonical Coefficients as a Function of the Log-components

		Variate 1	Variate 2
Philosophical orientations			
	$ln(x_{p1})$ (PR)	-0.524	1.730
	$ln(x_{p2})$ (IN)	2.085	-0.274
	$ln(x_{p3})$ (HU)	-1.561	-1.456
Learning modes			
	$ln(x_{l1})$ (AC)	1.720	-0.177
	$ln(x_{l2})$ (CE)	-0.447	-1.347
	$ln(x_{l3})$ (AE)	-1.032	1.311
	$ln(x_{l4})$ (RO)	-0.241	0.213

Canonical variates as a function of log components are shown in Table 3. As in Equation (4.2), the coefficients in Table 2 apply to all rows in Table 3 but the last one of each composition, which receives their sum with reversed sign.

The canonical variates in Table 3 correspond to the following log-contrasts:

$$cv_{p1} = \ln\left(\frac{x_{p2}^{2.085}}{x_{p1}^{0.524}x_{p3}^{1.561}}\right) \quad cv_{p2} = \ln\left(\frac{x_{p1}^{1.730}}{x_{p2}^{0.274}x_{p3}^{1.456}}\right)$$

$$cv_{l1} = \ln\left(\frac{x_{l1}^{1.720}}{x_{l2}^{0.447}x_{l3}^{1.032}x_{l4}^{0.241}}\right) \quad cv_{l2} = \ln\left(\frac{x_{l3}^{1.311}x_{l4}^{0.213}}{x_{l1}^{0.177}x_{l2}^{1.347}}\right) \tag{4.3}$$

The first pair of canonical variates can therefore be interpreted as follows: when the IN (x_{p2}) orientation is high and the HU (x_{p3}) orientation is low, then the AC (x_{l1}) mode is high and the AE (x_{l3}) mode is low. The second pair of canonical variates can be interpreted as follows: when the PR (x_{p1}) orientation is high and the HU (x_{p3}) orientation is low, then the AE (x_{l3}) mode is high and the CE (x_{l2}) is low. Our results are similar to those of Boyatzis et al. (2000), who reported the PR orientation as correlating positively with AE and negatively with CE; and the IN orientation as correlating positively with AC and negatively with AE.

5 Discussion

The increasing awareness of CoDa leads to an increasing interest in problems involving more than one composition. Standard statistical analysis includes many tools for relating two sets of variables, and one of the most popular in multivariate exploratory analysis is CCA. Within CoDa, tools for relating several compositions are still underdeveloped. In this article we have shown how to adapt CCA to compositional data in order to explore the relationship between two compositions. In our illustration we have found learning styles to be related to philosophical orientations in an interpretable manner in accordance with the literature, which supports the practical usefulness of the method as an exploratory tool.

The appeal of the CoDa log-ratio approach for applied researchers lies in the fact that once the data have been transformed using appropriate log-ratios, standard and well-understood statistical techniques such as CCA can be used. Once log-ratios have been computed, a compositional CCA is no more complicated than a standard CCA and standard statistical software dealing with CCA can be used. In order to be used with compositional

data, software must be able to derive the canonical variates from the covariance product in Equation (2.2) and include raw canonical coefficients as a part of the output. We recommend either SPSS, the *cca* function in the *yacca* R library (setting *xscale=FALSE*, *yscale=FALSE*), or the *cc* function in the *CCA* R library. It must be taken into account that some software for CCA either analyzes correlation matrices rather than covariance matrices (like the *canocor* function in the R library of the same name) or reports only standardized coefficients (like the *CCorA* function in the *vegan* R library). For the computation of canonical correlations and their significance tests, standardization or the use of correlations are irrelevant.

In some cases, the interpretation of the results of a statistical method on compositional data differs to some extent from its interpretation on unconstrained data. In the case of CCA, standardized results are neither usable nor needed, because unstandardized canonical variates can be interpreted as log-contrasts in a straightforward manner. This way of interpreting the results as log-contrasts fits well with the CoDa way of thinking and increases the attractiveness of the approach within an exploratory CoDa. CCA can also be applied to relate one composition to a set of numeric variables defined in the real space. In this case, the canonical variates are log-contrasts in the composition and linear combinations of the set of numeric variables with maximum mutual correlation.

The CoDa approach focuses on relative rather than absolute differences in the data. Treating compositional data directly without the log-ratio transformation implies assuming that the difference between scores 1 and 2 is the same as the difference between scores 10 and 11, while in the former case they differ by 100% and in the second by only 10%. A commonly mentioned limitation of the CoDa approach is the presence of zeros in the x_d variables, which prevents the analyst from computing log-ratios. Details on methods available for treating zeros prior to analysis, which perform well if the percentage of cases with zeros is not large, can be found in Martín-Fernández et al. (2011).

Further research could include adapting other multivariate techniques that relate sets of variables to compositional data, such as redundancy analysis, in order to derive a specified number of new latent variables from a composition that explains as much variance as possible from the other compositions. Related methods in the statistical modeling arena include simultaneous regression systems in which both explanatory and dependent variables are compositional (Tolosana-Delgado and van den Boogaart, 2013) and compositional partial least squares (Kalivodová et al., 2015).

Acknowledgements

The authors would like to acknowledge the support provided by Spanish Health Ministry Grant CB06/02/1002 funding the research group "Epidemiology and Public Health (CIBERESP)"; Catalan Autonomous Government Consolidated Research Group Grants 2014SGR551 and 2014SGR582 funding the research groups "Compositional and Spatial Data Analysis (COSDA)" and "Leadership Development Research Centre (GLEAD)"; Spanish Economy and Competitiveness Ministry grants MINECO/FEDER-EU MTM2015-65016-C2-1-R and EDU2015-68610-R funding the projects "COmpositional Data Analysis and RElated meThOdS (CODA-RETOS)" and "Assessing Individual and Team Entrepreneurial Potential"; and University of Girona grants MPCUdG2016/069 and MPCUdG2016/098.

References

[1] Aitchison, J. (1983): Principal component analysis of compositional data. *Biometrika*, **70**, 57–65.

[2] Aitchison, J. (1986): *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

[3] Aitchison, J. (2001): Simplicial inference. In M.A. Viana and D.S. Richards (Eds): *Algebraic Methods in Statistics and Probability*. *Contemporary Mathematics Series of the American Mathematical Society, vol. 287,* 1-22. Providence, RI: American Mathematical Society.

[4] Aitchison, J. and Greenacre, M. (2002): Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 375-392.

[5] Batista-Foguet, J.M., Ferrer-Rosell, B., Serlavós, R., Coenders, G. and Boyatzis, R.E. (2015): An alternative approach to analyze ipsative data. Revisiting Experiential Learning Theory. *Frontiers in Psychology*, **6**, 1742.

[6] Billheimer, D., Guttorp, P. and Fagan, W. (2001): Statistical interpretation of species composition. *Journal of the American Statistical Association,* **96**, 1205-1214.

[7] Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013): *Analyzing Compositional Data with R*. Berlin: Springer.

[8] Boyatzis, R.E., Murphy, A.J. and Wheeler, J.V. (2000): Philosophy as a missing link between values and behaviour. *Psychological Reports,* **86**, 47-64.

[9] ter Braak, C.J.F. (1996). *Unimodal Models to Relate Species to Environment*. Wageningen, NL: DLO-Agricultural Mathematics Group.

[10] Coenders, G., Hlebec, V. and Kogovšek, T. (2011): Measurement quality in indicators of compositions. A compositional multitrait-multimethod approach. *Survey Research Methods,* **5**, 63-74.

[11] Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K. and Filzmoser, P. (2012): Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, **6**, 87–108.

[12] Egozcue, J.J., and Pawlowsky-Glahn, V. (2011): Basic concepts and procedures. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 12-28. New York: Wiley.

[13] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003): Isometric logratio transformations for compositional data analysis. *Mathematical Geology,* **35**, 279-300.

[14] van Eijnatten, F.M., van der Ark, L.A. and Holloway, S.S. (2015): Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis. *Quality & Quantity*, **49**, 559-579.

[15] Ferrer-Rosell, B. and Coenders, G. (2016): Destinations and crisis. Profiling tourists' budget share from 2006 to 2012. *Journal of Destination Marketing & Management*. doi: 10.1016/j.jdmm.2016.07.002

[16] Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2015): Determinants in tourist expenditure composition- the role of airline types. *Tourism Economics*, **21**, 9-32.

[17] Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2016a): Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. *Tourism Analysis,* **21***,* 589-602.

[18] Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2016b): Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics*, **22**, 269–291.

[19] Fry, T. (2011): Applications in economics. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 318-326. New York: Wiley.

[20] Hair, J.F, Black, W.C., Babin, B.J. and Anderson, R.E. (2009): *Multivariate Data Analysis. A Global Perspective (7th ed.).* Upper Saddle River, NJ: Pearson/Prentice Hall.

[21] Hlebec, V., Kogovšek, T. and Coenders, G. (2012): The measurement quality of social support survey measurement instruments. *Metodološki Zvezki*, **9**, 1-24.

[22] Hotelling, H. (1936): Relations between two sets of variates. *Biometrika,* **28**, 321-377.

[23] Hron, K., Filzmoser, P. and Thompson, K. (2012): Linear regression with compositional explanatory variables. *Journal of Applied Statistics,* **39,**. 1115-1128.

[24] Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H. and Adam, T. (2015): PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, **29**, 21-28.

[25] Kogovšek, T., Coenders, G. and Hlebec, V. (2013): Predictors and outcomes of social network compositions. A compositional structural equation modeling approach. *Social Networks,* **35**, 1-10.

[26] Kolb, D.A. (1984): *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice Hall.

[27] Kolb, D.A. (1999): *Learning Style Inventory, Version 3.* Boston, MA: Hay Resources Direct.

[28] Martín-Fernández, J.A., Daunis-i-Estadella, J. and Mateu-Figueras, G. (2015): On the interpretation of differences between groups for compositional data. *SORT- Statistics and Operations Research Transactions*, **39**, 231–252.

[29] Martín-Fernández, J.A., Palarea-Albaladejo, J. and Olea, R.A. (2011): Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 47-62. New York: Wiley.

[30] Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J.J. (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 31-42. New York: Wiley.

[31] Pawlowsky-Glahn, V. and Buccianti, A. (2011): *Compositional Data Analysis. Theory and Applications*. New York: Wiley.

[32] Pawlowsky-Glahn, V. and Egozcue, J.J. (2001): Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment,* **15**, 384-398.

[33] Pawlowsky-Glahn, V. and Egozcue, J.J. (2011): Exploring compositional data with the CoDa-dendrogram*. Austrian Journal of Statistics,* **40***,* 103–113.

[34] Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015): *Modeling and Analysis of Compositional Data*. Chichester: Wiley.

[35] Pearson, K. (1897): Mathematical contributions to the theory of evolution. On a form of spurious correlations which may arise when indices are used in the measurements of organs. *Proceedings of the Royal Society*, **60**, 489-498.

[36] Thió-Henestrosa, S. and Martín-Fernández, J.A. (2005): Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology,* **37**, 773-793.

[37] Tolosana-Delgado, R. and van den Boogaart, K.G. (2013): Regression between compositional data sets. In K. Hron, P. Filzmoser and M. Templ (Eds): *Proceedings of the 5th International Workshop on Compositional Data Analysis (CoDaWork 2013)*, 163-176. Vienna: Vienna University of Technology.

[38] Vives-Mestres, M., Martín-Fernández, J.A. and Kenett, R. (2016): Compositional data methods in customer survey analysis*. Quality and Reliability Engineering International,* **32**, 2115-2125.

[39] de Vries, A.L.M. and van der Ark, L.A. (2008): Scoring methods for ordinal multidimensional forced-choice items*.* In J. Daunis-i-Estadella and J.A. Martín-Fernández (Eds): *Proceedings of the 3rd International Workshop on Compositional Data Analysis (CoDaWork 2008)*, 1-18. Girona: University of Girona.

INSTRUCTIONS TO AUTHORS

Language: *Metodološki zvezki – Advances in Methodology and Statistics* is published in English.

Submission of papers: Authors are requested to submit their articles (complete in all respects) to the Editor by e-mail (MZ@stat-d.si). Contributions are accepted on the understanding that the authors have obtained the necessary authority for publication. Submission of a paper will be held to imply that it contains original unpublished work and is not being submitted for publication elsewhere. Articles must be prepared in LaTeX or Word. Appropriate styles and example files can be downloaded from the Journal's web page (http://www.stat-d.si/mz/).

Review procedure: Manuscripts are reviewed by two referees. The editor reserves the right to reject any unsuitable manuscript without requesting an external review.

Preparation of manuscripts

Tables and figures: Tables and figures must appear in the text (not at the end of the text). They are numbered in the following way: Table 1, Table 2,…, Figure 1, Figure 2,…

References within the text: The basic reference format is (Smith, 1999). To cite a specific page or pages use (Smith, 1999: 10-12). Use "et al." when citing a work by more than three authors (Smith et al., 1999). The letters a, b, c etc. should be used to distinguish different citations by the same author(s) in the same year (Smith, 1999a; Smith, 1999b).

Notes: Essential notes, or citations of unusual sources, should be indicated by superscript number in the text and corresponding text under line at the bottom of the same page.

Equations: Equations should be centered and labeled with two numbers separated by a dot enclosed by parentheses. The first number is the current section number and the second a sequential equation number within the section, e.g., (2.1)

Author notes and acknowledgements: Author notes identify authors by complete name, affiliation and his/her e-mail address. Acknowledgements may include information about financial support and other assistance in preparing the manuscript.

Reference list: All references cited in the text should be listed alphabetically and in full after the notes at the end of the article.

References to books, part of books or proceedings:

[1] Smith, J.B. (1999): *Title of the Book.* Place: Publisher.

[2] Smith, J.B. and White A.B. (2000): *Title of the Book.* Place: Publisher.

[3] Smith, J. (2001): Title of the chapter. In A.B. White (Ed): *Title of the Proceedings*, 14-39. Place: Publisher.

Reference to journals:

[4] Smith, J.B. (2002): Title of the article**.** *Name of Journal***, 2**, 46-76.

Metodološki zvezki
Advances in Methodology and Statistics

Published by		Izdajatelj
Faculty of Social Sciences University of Ljubljana, for Statistical Society of Slovenia		**Fakulteta za družbene vede Univerze v Ljubljani za Statistično društvo Slovenije**
Editors	**Valentina Hlebec Lara Lusa**	Urednika
Founding Editors	**Anuška Ferligoj Andrej Mrvar**	Prva urednika
Cover Design	**Bojan Senjur Gregor Petrič**	Oblikovanje naslovnice
Typesetting	**Lara Lusa**	Računalniški prelom
Printing	**DEMAT, d.o.o. Ljubljana, Slovenia**	Tisk
	MZ	
is indexed and abstracted in	**SCOPUS EBSCO ECONIS STMA-Z ProQuest**	je indeksirana in abstrahirana v
Home page URL	http://www.stat-d.si/mz/	Spletna stran

ISSN 1854 - 0023