CHILEAN JOURNAL OF **STATISTICS** Volume 10 Number 2 - December 2019

www.soche.cl/chjs

Volume 10 Number 2 - December 2019

CHILEAN JOURNAL OF **STATISTICS**

CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva

Volume 10 Number 2
December 2019
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

Aims

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

The ChJS is an international scientific forum strongly committed to gender equality, open access of publications and data, and the new era of information. The ChJS covers a broad range of topics in statistics, data science, data mining, artificial intelligence, and big data, including research, survey and teaching articles, reviews, and material for statistical discussion. In particular, the ChJS considers timely articles organized into the following sections: Theory and methods, computation, simulation, applications and case studies, education and teaching, development, evaluation, review, and validation of statistical software and algorithms, review articles, letters to the editor.

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

Chilean Journal of Statistics

Volume 10, Number 2
December 2019

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

http://www.soche.cl/chjs

Contents

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 95–98

Tenth Volume – Second Issue
Editorial Paper

"Chilean Journal of Statistics": An international scientific forum committed to gender equality, open access, and the new era of information

We introduce the second issue of the tenth volume of the Chilean Journal of Statistics (ChJS). In this opportunity, and before presenting the interesting papers to be published in the current issue, I would like to make some reflections about the international character of our journal, but also about gender equality and the important challenges for statistics in the era of information, as well as for the open access to publications and data.

Regarding gender equality, it allows us to accelerate progress and opportunities for everyone. However, our journal is in debt in relation a such an equality, so that this is big challenge which we are assuming from now. Indeed, our editorial board is increasing the number of women starting from this issue, because it is necessary and we must do justice to the talent and empowerment of women in science, and particularly in statistics. Welcome on board Dr. Alejandra Tapia from Chile, Dr. Viviana Giampaoli from Argentina, Dr. Michelli Barros from Brazil, Dr. Teresa Oliveira from Portugal, and Dr. Ana B. Nieto from Spain. We are sure that with your enthusiasm, dynamism and talent, our journal will benefit greatly. We are honored with your acceptance to be part of the ChJS.

Related to our international character, we must recall that the ChJS is published by the Chilean Statistical Society (www.soche.cl) and belongs to the Chilean statistical community, but our prestigious Editorial Board, presented at http://chjs.mat.utfsm.cl/board.html, is composed of researchers from Argentina, Australia, Austria, Bulgaria, Brazil, Canada, Chile, China, Colombia, Greece, India, Italy, Mexico, Netherlands, Peru, Portugal, Romania, Saudi Arabia, Spain, Switzerland, UK, and US, which are distributed according to the graphical plot displayed in Figure 1 according to data visualization of data science, such as in Figure 2 below. Currently, a 20% of this Board are women and we hope to increase this number promptly. In addition to our Editorial Board, we have received papers to be evaluated from different countries and from the five continents as shown in Figure 2. The ChJS is publishing about 20% of the papers submitted to our journal. We are an open access journal, which publishes original papers free of charges for publication, allowing the international community to disseminate the statistical knowledge at no cost. This is a contribution from the Chilean Statistical Society to the knowledge economy, defined as the use of knowledge to create goods and services. Furthermore, our journal will promote the concept of open data in our next published papers.

Considering the relevant challenges for statistics in the new era of information, which we are living during these days, the ChJS is open to publish papers related to artificial intelligence, big data, data mining, data science, and text mining. We have taken into account this challenge incorporating expert researchers on these topics as part of our Editorial Board to evaluate and process possible papers regarding such topics. We hope to publish an adequate number of papers on this interesting thematic during 2020.

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

http://www.soche.cl/chjs

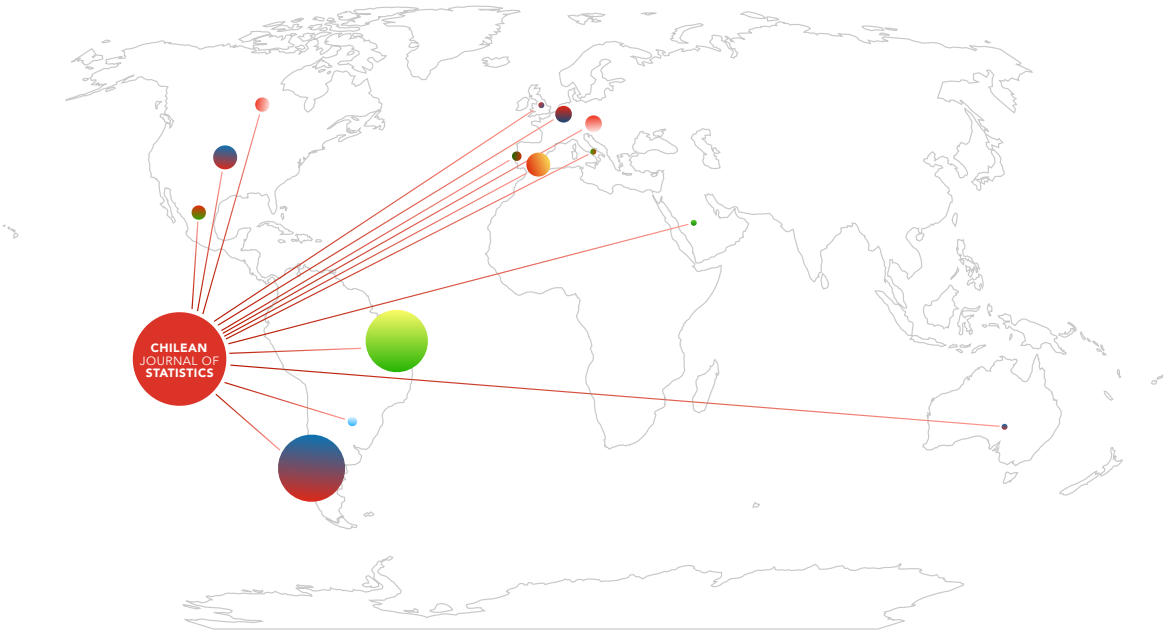

Figure 1. Distribution of the Editorial Board's members in relation to their country.

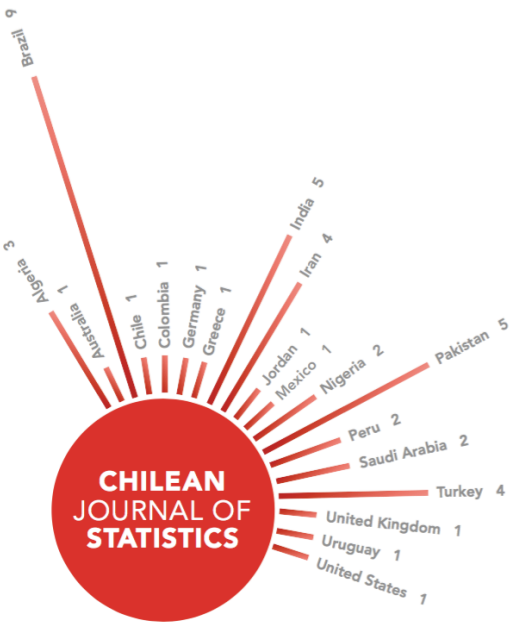

Figure 2. Dstribution of the authors who submit papers to the ChJS in relation to their country.

About the Directory of the Chilean Statistical Society (https://soche.cl/quienes-somos), firstly, I would like to thank its President, Dr. Mauricio Castro, for the trust placed in me to be appointed as the Editor-in-Chief of the ChJS from March, 2019. Also, I want to congratulate Dr. Castro for his excellent work in this high position of the SOCHE. He assumed as President with great leadership during a complicated period for our beloved scientific society. Secondly, I wish to congratulate as well to Dr. Jorge Figueroa for being recently elected as the new President of the SOCHE from 2020, to whom I thank also his trust to continue in my editorial position. They and all the community can rest assured that I will make my best effort to bring the ChJS to the highest standards of professionalism, impartiality and quality that all scientific journal must strive for.

In addition to this presentation note, the second issue of the tenth volume of the ChJS comprises five papers, which correspond to valuable contributions of renowned international researchers who have honored us by publishing their interesting works in our journal; all of these papers are available for free at http://chjs.mat.utfsm.cl/issues.html. We also thank our Associated Editors and the anonymous reviewers who have contributed to keeping the top quality standards of the ChJS. Our first paper is authored by Paulo H. Ferreira, Taciana K.O. Shimizu, Adriano K. Suzuki, and Francisco Louzada. The authors introduced an asymmetric extension to the tobit model by assuming that the error term follows a tilted-normal distribution. The model parameters were estimated by a standard method, whose performance was evaluated with Monte Carlo simulations for different sample sizes and parameter settings. Also, adequacy of the tobit formulation was assessed by using model selection criteria. An illustration with real data was also given including a diagnostic analysis. The second paper is authored by Eduardo Horta and Flavio Ziegelmann. The authors provided sufficient conditions ensuring that mixing properties hold for the sequence of empirical cumulative distribution functions associated with a conjugate process. Also, numerical examples were provided to illustrate the results obtained in this work. The third paper is authored by Guilherme Parreira da Silva, Cesar Augusto Taconeli, Walmes Marques Zeviani, and Isadora Aparecida Sprengoski do Nascimento. The authors evaluated the performance of Shewhart control charts based on neoteric ranked set sampling to monitor the mean of normal and non-normal processes. They used mean, median and standard deviation of run lengths to make this evaluation based on Monte Carlo simulations. The impact of imperfect ranking and non-normality were also assessed and an illustration with real data was provided to show the potential applications. The fourth paper is authored by Lucas Pereira Lopes, Vicente Garibay Cancho, and Francisco Louzada, who provided a GARCH methodology to describe a more realistic pricing option using stocks from two Brazilian companies. The methodology was confronted with a type Black-Scholes model, obtaining good results. Concepts of copulas, marginal models and asymmetry were considered in this study making the joint modeling more flexible and realistic. The empirical aspects of the obtained results were relevant in financial emerging markets, where non-normality seems to be evident. This second issue closes with a fifth paper authored by Boubaker Mechab, Nesrine Hamidi, and Samir Benaissa. The authors investigated nonparametric estimation of the relative error in functional regression and censored data. The almost complete consistency and the asymptotic normality of the estimator of the regression operator in the case of a censored response given a functional explanatory variable were studied. The finite sample performance based on the mean square error between standard and relative error regressions was assessed by simulations. A real data illustration was carried out to apply the results obtained.

Finally, I would like the statistical and data science communities, our prestigious Editorial Board, and authors to champion the ChJS as an international scientific forum committed to gender equality, open access, and the new era of information to encourage others to submitting new investigations to the ChJS. We are indexed by serious and rigorous international systems, including the ISI Web of Science. The ChJS continues facing big challenges for the future and we need of all the community in meeting them.

Víctor Leiva
Editor-in-Chief
Chilean Journal of Statistics
http://www.victorleiva.cl

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 99–122

Statistical Modeling
Research Paper

On an asymmetric extension of the tobit model based on the tilted-normal distribution

Paulo H. Ferreira[1,*], Taciana K.O. Shimizu[2], Adriano K. Suzuki[2], and Francisco Louzada[2]

[1]Department of Statistics, Federal University of Bahia, Salvador, Brazil
[2]Department of Applied Mathematics and Statistics, University of São Paulo, São Carlos, Brazil

(Received: 25 February 2019 · Accepted in final form: 24 April 2019)

Abstract

In this paper, we introduce an asymmetric extension to the tobit model by assuming that the error term follows a tilted-normal distribution. The new model, namely tilted-normal tobit model, can be an useful alternative to other skewed tobit models, such as the skew-normal and power-normal tobit models. The method of maximum likelihood is used for estimating the model parameters. We provide some simulation studies for different sample sizes and parameter settings. In addition, we perform residual and influence diagnostic analysis. Finally, we use American food consumption data to illustrate the better performance of the model introduced.

Keywords: Censored regression model · Influence · Maximum likelihood estimation · Residual and influence diagnostic analysis · Tilted-normal distribution.

Mathematics Subject Classification: Primary 62J05 · Secondary 62N01.

1. Introduction

Tobit models are regression models whose range of the dependent variable is somehow constrained. They were first suggested in a pioneering work by Tobin (1958), to describe the relationship between a non-negative dependent variable (the ratio of total durable goods expenditure to total disposable income, per household) and a vector of independent variables (the age of the household head, and the ratio of liquid asset holdings to total disposable income). Tobin called his model the limited dependent variable model, however it and its various generalizations are popularly known among economists as tobit models, a phrase coined by Goldberger (1964) due to similarities with probit models (the term tobit aims to synthesize in one word Tobin's probit concept). Tobit models are also known as censored regression models. For discussion on properties, parameter estimation and asymptotic properties of estimators, see, e.g., Amemiya (1973, 1984, 1985) and Fair (1977).

*Corresponding author. Email: paulohenri@ufba.br

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

http://www.soche.cl/chjs

The tobit specification is adequate for the situation where the sample proportion of zero observations is roughly equivalent to the left tail area of the assumed parametric distribution. The Cragg model (Cragg, 1971), which in the classical literature is known as the two-part model, is an alternative to tobit when the rate of zero observations is quite different from the probability of the left tail obtained with the assumed parametric model.

An interesting way of extending the tobit model is supposing that the probability distribution of the perturbations is no longer normal. For instance, Arellano-Valle et al. (2012) proposed an extension of the tobit model using the Student-t distribution, which is useful for statistical modeling of censored data sets involving observed variables with heavier tails than the normal distribution. Martínez-Flórez et al. (2013) assumed the power-normal distribution (Gupta and Gupta, 2008), thus providing an asymmetric alternative to tobit model. However, such a probability distribution is problematic, that is, of limited use, since it only accommodates low to moderate left-skewness. Moreover, Castro et al. (2014) extended the tobit model to the class of scale mixtures of normal distributions (Andrews and Mallows, 1974) from the Bayesian viewpoint. Other important contributions extending the tobit model by using asymmetric and/or heavy-tailed distributions are Garay et al. (2016, 2017), Mattos et al. (2018), Barros et al. (2018) and Desousa et al. (2018) among many others.

The main purpose of this paper is to focus on the study of the censored regression model, under the assumption that the error term follows the tilted-normal distribution (Maiti and Dey, 2012). Such probability distribution has received some attention in the recent literature, e.g. Louzada et al. (2018) applied the tilted-normal model to compositional data on percentages of players' points in the Brazilian men's volleyball super league 2014/2015. Parameter estimation is performed by using the maximum likelihood (ML) approach and its large sample properties. Application is implemented to American food consumption data set (USDA, 2000), where it is demonstrated that the proposed model can be very useful in fitting real data sets.

The paper is organized as follows. In Section 2, we define the tilted-normal distribution and discuss some of its properties. We present the tilted-normal tobit model and implement inference using the ML approach in Section 3. In Section 4, results of simulation studies reveal the good performance of the estimation approach and the appropriateness of some information criteria in distinguishing among candidate models. Section 5 presents an application to real data on consumption of tomato in the United States in 1994-1996 (USDA, 2000). Model fitting evaluation indicates that the data set in question is much better fitted by the tilted-normal tobit model than by the classic (standard or Type I) tobit model (Tobin, 1958), as well as by other asymmetric models, like the skew-normal tobit model (Hutton and Stanghellini, 2011) and the power-normal tobit model (Martínez-Flórez et al., 2013). Finally, some concluding remarks and directions for future work are given in Section 6. In the work of Hutton and Stanghellini (2011), the skew-normal tobit model was used to address the skewness and right-censoring problems in bounded health scores.

2. The Tilted-Normal Distribution

In this section, we present some basic properties of the tilted-normal distribution, including the probability density function (PDF) and the cumulative distribution function –CDF– (Subsection 2.1), the moments (Subsection 2.2), as well as other relevant issues (Subsection 2.3).

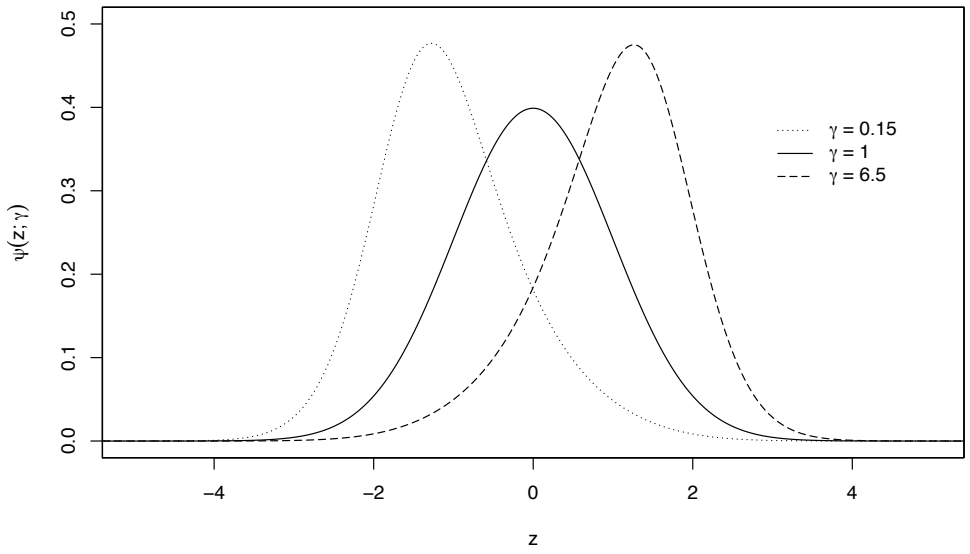

Figure 1. Tilted-normal PDF $\psi(z;\gamma)$ for some values of γ.

2.1 Probabilistic functions

Following the proposition of García et al. (2010) and Maiti and Dey (2012), the tilted-normal distribution is defined as follows. Let Z be a standard normal random variable, that is, $Z \sim \mathrm{N}(0,1)$. Following Marshall and Olkin (1997), the standard tilted-normal distribution, denoted by $\mathrm{TN}(0,1,\gamma)$, has PDF given by

$$\psi(z;\gamma) = \frac{\gamma\phi(z)}{\left[1-(1-\gamma)\left\{1-\Phi(z)\right\}\right]^2}, \quad z \in \mathrm{R},$$

where $\gamma > 0$ is a shape/skewness parameter, ϕ is the PDF of the standard normal distribution and Φ is the CDF of the standard normal distribution. The standard tilted-normal PDF is a unimodal function, which is skewed to the left if $\gamma > 1$ and to the right if $0 < \gamma < 1$, while $\gamma = 1$ indicates a standard normal PDF (Maiti and Dey, 2012). Figure 1 displays a few PDF graphs for different values of γ.

If Z is a random variable from a $\mathrm{TN}(0,1,\gamma)$ distribution, then the location-scale extension of Z, $Y = \mu + \sigma Z$, has PDF given by

$$\psi(y;\mu,\sigma,\gamma) = \frac{\frac{\gamma}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{\left[1-(1-\gamma)\left\{1-\Phi\left(\frac{y-\mu}{\sigma}\right)\right\}\right]^2}, \tag{1}$$

as well as its CDF given by

$$\Psi(y;\mu,\sigma,\gamma) = \frac{\Phi\left(\frac{y-\mu}{\sigma}\right)}{1-(1-\gamma)\left\{1-\Phi\left(\frac{y-\mu}{\sigma}\right)\right\}}, \tag{2}$$

where $\mu \in \mathrm{R}$ and $\sigma > 0$. We will denote this extension by using the notation $Y \sim \mathrm{TN}(\mu,\sigma,\gamma)$.

2.2 Moments

For the model (1), García et al. (2010) showed that the k-th moment about the origin of the random variable Y is given by

$$\begin{aligned}\mu_k^{'} = \mathrm{E}\left[Y^k\right] &= \int_{-\infty}^{\infty} y^k \psi(y;\mu,\sigma,\gamma)\mathrm{d}y \\ &= \int_0^1 \left[\mu + \sigma\sqrt{2}\,\mathrm{erf}^{-1}\left(\frac{-u+\gamma-u\gamma}{u+\gamma-u\gamma}\right)\right]^k \mathrm{d}u,\end{aligned} \tag{3}$$

where $\mathrm{erf}^{-1}(w) = w\sqrt{\pi}/2 + \mathrm{O}\left(w^3\right) \simeq w\sqrt{\pi}/2$ is the inverse error function.

Although the expression (3) seems to be not available in compact form, the authors verified the following approximations:

$$\begin{aligned}\mu_1^{'} = \mathrm{E}\left[Y\right] &\simeq \frac{2(1-\gamma)^2\mu - \sigma\sqrt{2\pi}\left(1-\gamma^2+2\gamma\log(\gamma)\right)}{2\gamma(1-\gamma)}, \\ \mu_2^{'} = \mathrm{E}\left[Y^2\right] &\simeq \frac{\gamma}{2(1-\gamma)^3}\Big\{2(1-\gamma)^2\mu^2 + 2\left[\left(\gamma^2-1\right)\mu\sigma\sqrt{2\pi} + \left(1+6\gamma\left(1+\gamma\right)\pi\sigma^2\right)\right] \\ &\quad + 4\gamma\sigma\left[(1-\gamma)\mu\sqrt{2\pi} - (1+\gamma)\pi\sigma\right]\log(\gamma)\Big\}, \\ \mu_3^{'} = \mathrm{E}\left[Y^3\right] &\simeq \frac{-1}{4(1-\gamma)\gamma}\Big\{-4(1-\gamma)^3\mu^2\left[-1+\gamma(1+\gamma)\right] + 6(1-\gamma)^2\mu\pi\sigma^2\left[1+\gamma\left(2+\gamma^2\right)\right] \\ &\quad - \sqrt{2\pi}\pi\sigma^3\left[1+2\gamma-5\gamma^2+11\gamma^3+4\gamma^4-\gamma^5\right] - 6(1-\gamma)\gamma\sigma\sqrt{\pi}\Big[2\sqrt{2}(1-\gamma)\mu^2 \\ &\quad + 4\mu\sigma\sqrt{\pi}\left(1-\gamma^2\right) + \sqrt{2}(1+\gamma)^2\pi\sigma^2\Big]\log(\gamma)\Big\}.\end{aligned} \tag{4}$$

These quantities can be used to compute the approximate mean ($\mathrm{E}[Y] = \mu_1^{'}$), variance ($\mathrm{Var}[Y] = \mu_2^{'} - (\mu_1^{'})^2$) and skewness index ($\beta_1 = \mu_3^{'}/(\mu_2^{'})^{3/2}$) of the random variable Y, and are particularly useful for estimating the parameters by the method of moments.

2.3 Others

The model (1) can be extended by considering $\mu_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is an unknown vector of regression coefficients and $\boldsymbol{x}_i$ is a vector of known regressors correlated with the response vector, for $i = 1, \ldots, n$.

Regarding the other skewed distributions that could be used instead of the tilted-normal distribution, Gupta and Gupta (2008) observed that the estimation of the shape parameter of the skew-normal distribution (Azzalini, 1985) is problematic, among others, in the cases where the sample size is not large enough. Monti (2003) noticed that the estimate of the shape parameter is $\hat{\gamma} = \pm\infty$, even when the data are generated by a model with finite γ. Moreover, Pewsey et al. (2012) showed that the Fisher information matrix for the skew-normal distribution is singular under the symmetry hypothesis and, therefore, regularity conditions are not satisfied for the likelihood approach. The same authors also derived the Fisher information matrix for the location-scale version of the power-normal model (Gupta and Gupta, 2008) and have shown that, in addition to its several nice properties, it is not singular for the shape parameter $\gamma = 1$. However, as pointed out by Maiti and

Dey (2012), left-skewness is not so clear and modeling of left-skewed data will be misfit. This is due to the fact that such a distribution can only accommodate low to moderate left-skewness of the data distribution. Hence, the power-normal model is not appropriate for the cases where the data distribution exhibits strong left-skewness. This limitation also applies to the tilted-normal distribution, which can not capture high or moderate levels of skewness (when measured in an appropriate manner). In fact, Rubio and Steel (2012) and Jones (2015) discuss the restrictions of using the Marshall-Olkin transformation for inducing skewness in many symmetric models (including the normal one). Despite such limitation, we demonstrate here that the proposed tobit model based on the tilted-normal distribution can still be very useful in fitting real data sets as in Section 5.

3. The Tilted-Normal Tobit Model

In this section, we introduce the proposed extension of the tobit model using the tilted-normal distribution (Subsection 3.1) and discuss statistical inference based on the ML method (Subsection 3.2).

3.1 Formulation

Let $D_i = I(Y_i > 0)$, where $I(\cdot)$ is the indicator function. The tilted-normal tobit model can be defined by relating the observed dependent variable Y_i^{o} to the original (that is, of theoretical interest), but censored, dependent variable Y_i, as follows:

$$Y_i^{\mathrm{o}} = D_i Y_i \quad \text{and} \quad Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{5}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameter vector, $\boldsymbol{x}_i$ is a $p \times 1$ vector of known independent variables, and the errors $\epsilon_i \sim \mathrm{TN}(0, \sigma, \gamma)$.

The value of the location parameter, 0, of ϵ_i implies, from the first expression of (4), that $\mathrm{E}[\epsilon_i] \simeq -\sigma\sqrt{2\pi}\left(1 - \gamma^2 + 2\gamma\log(\gamma)\right)/\left(2\gamma(1-\gamma)\right) < 0$, $\forall \sigma, \gamma > 0$ and $\gamma \neq 1$. Also, for $\sigma > 0$ fixed, $\mathrm{E}[\epsilon_i] \to -\infty$ when $\gamma \to 0^+$ and $\mathrm{E}[\epsilon_i] \to 0$ as $\gamma \to 1^-$. This location parameter choice follows from the work of Martínez-Flórez et al. (2013). However, it could also have been chosen in order to obtain $\mathrm{E}[\epsilon_i] = 0$, as in the normal model, and similarly as in the work of Mattos et al. (2018). Although, even in this case, the expected value of the observed dependent variable Y_i^{o} differs from the location parameter $\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, that is, $\mathrm{E}[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i] = \mathrm{E}[Y_i \mid Y_i > 0, \boldsymbol{x}_i]\, P(Y_i > 0 \mid \boldsymbol{x}_i)$, which, after some steps and considering $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, results in $\mathrm{E}[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i] = \Phi\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}/\sigma\right)\left[\boldsymbol{x}_i^\top\boldsymbol{\beta} + \sigma\phi\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}/\sigma\right)/\Phi\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}/\sigma\right)\right] \neq \boldsymbol{x}_i^\top\boldsymbol{\beta}$ (see, e.g., Greene, 2012, Chapter 19).

Note, however, that for the case where $\epsilon_i \sim \mathrm{TN}(0, \sigma, \gamma)$, the main difficulty in obtaining $\mathrm{E}[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i]$, which would further allow us to analyze the effects of the inequality $\mathrm{E}[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i] \neq \boldsymbol{x}_i^\top\boldsymbol{\beta}$ on the intercept β_0 of the tilted-normal tobit model, is that there seems to be no explicit known expression for the conditional expectation $\mathrm{E}[Y_i \mid Y_i > 0, \boldsymbol{x}_i]$. Nevertheless, such expected value can be obtained numerically (as shown in Figure 4) or via approximations, e.g., by using some general results of the Marshall and Olkin (1997) family of distributions shown in Cordeiro et al. (2014), among others. We will leave this part of research for our future work.

The tilted-normal tobit model is basically a censored tilted-normal regression model with the tilted-normal distribution replacing the normal distribution for the error term. Thus, parameter estimation for the proposed model is related to parameter estimation for the censored tilted-normal distribution.

For the more general case, where the (known) left-censoring point is $c_i \in \mathrm{R}$, or even for the right-censoring case, we can obtain the estimation results by using the previous model (5), in the same way as stated in Martínez-Flórez et al. (2013).

The next subsection is devoted to implementation of parameter estimation by ML approach and discusses its properties in large samples.

3.2 Estimation

The ML estimators are the most commonly used in the literature. These estimators enjoy desirable properties and can be used for constructing confidence intervals for the model parameters. The normal approximation for the ML estimators in large sample distribution theory is easily handled either analytically or numerically.

In this work, we consider the ML estimation of the unknown parameters of the tilted-normal tobit model. The approach is described as follows.

Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \sigma, \gamma\right)^\top$ be the vector of parameters of interest. Also suppose that the data consist of $n = n_0 + n_1$ observations $(\boldsymbol{x}_1, d_1 y_1), \ldots, (\boldsymbol{x}_n, d_n y_n)$, where n_0 and n_1 are the number of observations on the sets $N_0 = \{i : d_i = 0\} = \{i : y_i = 0\}$ and $N_1 = \{i : d_i = 1\} = \{i : y_i > 0\}$, respectively. Since the unobserved random variables $Y_1, \ldots, Y_n$ are independent, with $Y_i \sim \mathrm{TN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma, \gamma)$, we have $P\left(Y_i^{\mathrm{o}} = 0\right) = P\left(Y_i \leq 0\right) = \Phi\left(-\boldsymbol{x}_i^\top \boldsymbol{\beta}/\sigma\right) / \left(1 - (1-\gamma)\left\{1 - \Phi\left(-\boldsymbol{x}_i^\top \boldsymbol{\beta}/\sigma\right)\right\}\right)$, for $i \in N_0$, while for the non-nulls Y_i^{o}s we have that they are distributed as their respective Y_is, that is, $Y_i^{\mathrm{o}} \sim \mathrm{TN}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma, \gamma)$, for $i \in N_1$. Thus, from the relations mentioned above, the likelihood function for the tilted-normal tobit model is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{\Phi\left(\frac{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)}{1 - (1-\gamma)\left\{1 - \Phi\left(\frac{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right\}}\right]^{1-d_i} \left[\frac{\frac{\gamma}{\sigma}\phi\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)}{\left(1 - (1-\gamma)\left\{1 - \Phi\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right\}\right)^2}\right]^{d_i}.$$

Then, the corresponding log-likelihood function is expressed as

$$\begin{aligned}\ell(\boldsymbol{\theta}) = &\sum_{i=1}^{n}(1-d_i)\,\log\left(\Phi\left(\frac{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right) - \sum_{i=1}^{n}(1-d_i)\log\left(1 - (1-\gamma)\left\{1 - \Phi\left(\frac{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right\}\right) \\ &+ \log(\gamma)\sum_{i=1}^{n} d_i - \log(\sigma)\sum_{i=1}^{n} d_i + \sum_{i=1}^{n} d_i\,\log\left(\phi\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right) \\ &- 2\sum_{i=1}^{n} d_i\,\log\left(1 - (1-\gamma)\left\{1 - \Phi\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right\}\right). \end{aligned} \tag{6}$$

The ML estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained by directly solving the nonlinear equations: $U(\boldsymbol{\beta}) = \mathbf{0}$, $U(\sigma) = 0$ and $U(\gamma) = 0$, where $U(\cdot)$ denotes the score function (see Appendix for analytic description). Note that these equations can not be solved analytically, but we can use, for instance, the `optim` routine (method = "L-BFGS-B") of the R software to solve them numerically. Since regularity conditions are satisfied using the large sample distribution, the distribution of $\hat{\boldsymbol{\theta}}$ can be approximated by a multivariate normal distribution, that is, $\hat{\boldsymbol{\theta}} \sim \mathrm{N}_{p+2}(\boldsymbol{\theta}, [J_{p+2}(\hat{\boldsymbol{\theta}})]^{-1})$, to obtain confidence intervals and hypothesis testing for the parameters of the tilted-normal tobit model, where $J_{p+2}\hat{\boldsymbol{\theta}})$ is the $(p+2)\times(p+2)$ observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$. The elements of the diagonal of $[J_{p+2}(\hat{\boldsymbol{\theta}})]^{-1}$ can be used to approximate the corresponding standard errors.

4. Simulation Studies

In this section, we present the main results obtained from Monte Carlo simulation studies aimed at verifying properties of the ML estimators of the tilted-normal tobit model parameters, with different sample sizes and censoring percentages (Subsection 4.1), as well as investigating the appropriateness of the chosen model selection criteria (Subsection 4.2).

4.1 Parameter recovery study

The first simulation study was based on $M = 2,000$ generated samples of sizes $n = 50$, 100, 300 and 500.

Without loss of generality, we took $\sigma = 1$ and $\beta_1 = 3.5$. It was considered a linear model with a single covariate X whose values were generated according to a $\mathrm{N}(0, 1)$ distribution. We assumed errors $\epsilon_i \sim \mathrm{TN}(0, \sigma, \gamma)$. To ensure a censoring percentage (that is, of zero y_i observations) of approximately 5%, 25%, 50% and 75%, we set the following true values for β_0, respectively (and also for different values of γ):

- For $\gamma = 0.5$: $\beta_0 = 6.4$, 2.8, 0.4 and -2.1;
- For $\gamma = 1$: $\beta_0 = 6$, 2.4, 0.05 and -2.5;
- For $\gamma = 2$: $\beta_0 = 5.5$, 2.1, -0.4 and -2.9;
- For $\gamma = 5$: $\beta_0 = 5$, 1.5, -0.9 and -3.4.

Observed data y_i were taken as $y_i = \max\{\beta_0 + \beta_1 x_i + \epsilon_i, 0\}$. In order to evaluate estimators performance for point estimates, the following quantities were considered: means, biases and mean squared errors (MSEs) of the parameter estimates, and estimated coverage lengths (CLs). We also assessed the performance of the proposed model through the coverage probabilities (CPs) of the 95% normal confidence intervals. ML estimates were computed by using the optim routine (method = "L-BFGS-B") of the R software.

Let $\hat{\boldsymbol{\theta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}, \widehat{\gamma})^\top$ be the ML estimators of the tilted-normal tobit model parameters and $(s_{\widehat{\beta}_0}, s_{\widehat{\beta}_1}, s_{\widehat{\sigma}}, s_{\widehat{\gamma}})$ be their standard errors, which were computed by inverting the observed information matrix. The means, biases, MSEs, CLs and CPs can be estimated by the following equations:

$$\mathrm{Mean}\left(\hat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_j^{(m)}, \quad \mathrm{Bias}\left(\hat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_j^{(m)} - \theta_j\right),$$

$$\mathrm{MSE}\left(\hat{\theta}_j\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\widehat{\theta}_j^{(m)} - \theta_j\right)^2, \quad \mathrm{CL}\left(\hat{\theta}_j\right) = \frac{3.919928}{M}\sum_{m=1}^{M} s_{\widehat{\theta}_j^{(m)}}$$

and

$$\mathrm{CP}\left(\theta_j\right) = \frac{1}{M}\sum_{m=1}^{M} I\left(\widehat{\theta}_j^{(m)} - 1.959964\, s_{\widehat{\theta}_j^{(m)}} < \theta_j < \widehat{\theta}_j^{(m)} + 1.959964\, s_{\widehat{\theta}_j^{(m)}}\right),$$

for $j = 1, 2, 3, 4$, where $\widehat{\theta}_j^{(m)}$ is the ML estimate of θ_j obtained from the m^{th} replicated sample.

From Tables 1-4, it can be seen that the ML estimates of β_0 and β_1 are unstable, because these parameters are affected by the skewness parameter γ and the proportion of zero observations in the sample. However, the ML estimates become more stable as the

Table 1. Estimation results for the tilted-normal tobit model ($\gamma = 0.5$).

Sample size	Censoring percentage	Parameter	True value	Mean	Bias	MSE	CP	CL
50	5	β_0	6.4	6.4519	0.0519	0.5945	0.9800	3.0781
		β_1	3.5	3.5128	0.0128	0.0242	0.9415	0.5863
		σ	1	1.0012	0.0012	0.0165	0.9340	0.5361
		γ	0.5	0.8487	0.3487	0.9849	0.8565	4.2149
	25	β_0	2.8	2.8342	0.0342	0.6226	0.9840	3.3211
		β_1	3.5	3.5184	0.0184	0.0440	0.9220	0.7732
		σ	1	0.9980	-0.0020	0.0183	0.9400	0.5851
		γ	0.5	0.9119	0.4119	1.2976	0.8615	5.0279
	50	β_0	0.4	0.4397	0.0397	0.7172	0.9870	3.8817
		β_1	3.5	3.5226	0.0226	0.0935	0.9150	1.0962
		σ	1	0.9887	-0.0113	0.0246	0.9345	0.6950
		γ	0.5	0.9361	0.4361	1.3936	0.8435	6.3242
	75	β_0	-2.1	-2.0622	0.0378	1.0908	0.9850	5.1825
		β_1	3.5	3.5664	0.0664	0.3317	0.8945	1.9142
		σ	1	0.9630	-0.0370	0.0449	0.9180	0.9563
		γ	0.5	0.9565	0.4565	1.6456	0.8140	8.5444
100	5	β_0	6.4	6.4122	0.0122	0.3380	0.9770	2.1728
		β_1	3.5	3.5070	0.0070	0.0120	0.9365	0.4178
		σ	1	1.0034	0.0034	0.0083	0.9510	0.3611
		γ	0.5	0.7327	0.2327	0.5402	0.8945	2.5704
	25	β_0	2.8	2.8077	0.0077	0.3922	0.9845	2.4075
		β_1	3.5	3.5094	0.0094	0.0228	0.9325	0.5490
		σ	1	1.0060	0.0060	0.0096	0.9470	0.4015
		γ	0.5	0.7782	0.2782	0.6988	0.8825	3.1039
	50	β_0	0.4	0.4195	0.0195	0.5152	0.9870	2.8807
		β_1	3.5	3.5090	0.0090	0.0447	0.9300	0.7719
		σ	1	1.0072	0.0072	0.0138	0.9500	0.4878
		γ	0.5	0.8287	0.3287	0.8875	0.8645	4.0961
	75	β_0	-2.1	-2.0563	0.0437	0.7604	0.9820	4.0274
		β_1	3.5	3.5217	0.0217	0.1345	0.9295	1.3138
		σ	1	0.9945	-0.0055	0.0232	0.9475	0.6972
		γ	0.5	0.8941	0.3941	1.3086	0.8340	6.2404
300	5	β_0	6.4	6.3988	-0.0012	0.1036	0.9610	1.1943
		β_1	3.5	3.5026	0.0026	0.0039	0.9460	0.2414
		σ	1	1.0002	0.0002	0.0025	0.9555	0.1908
		γ	0.5	0.5756	0.0756	0.1010	0.9210	1.1327
	25	β_0	2.8	2.7921	-0.0079	0.1209	0.9655	1.3199
		β_1	3.5	3.5052	0.0052	0.0068	0.9430	0.3166
		σ	1	1.0012	0.0012	0.0028	0.9560	0.2090
		γ	0.5	0.5978	0.0978	0.1455	0.9180	1.3146
	50	β_0	0.4	0.3847	-0.0153	0.1704	0.9830	1.5937
		β_1	3.5	3.5063	0.0063	0.0135	0.9410	0.4446
		σ	1	1.0023	0.0023	0.0040	0.9530	0.2515
		γ	0.5	0.6397	0.1397	0.2584	0.9150	1.7361
	75	β_0	-2.1	-2.1111	-0.0111	0.3248	0.9865	2.3749
		β_1	3.5	3.5081	0.0081	0.0394	0.9335	0.7416
		σ	1	1.0041	0.0041	0.0083	0.9540	0.3754
		γ	0.5	0.7279	0.2279	0.5313	0.8890	2.9640
500	5	β_0	6.4	6.3938	-0.0062	0.0546	0.9630	0.9079
		β_1	3.5	3.5020	0.0020	0.0023	0.9470	0.1873
		σ	1	0.9999	-0.0001	0.0013	0.9465	0.1436
		γ	0.5	0.5463	0.0463	0.0498	0.9435	0.8300
	25	β_0	2.8	2.7915	-0.0085	0.0667	0.9670	1.0014
		β_1	3.5	3.5040	0.0040	0.0040	0.9445	0.2453
		σ	1	1.0004	0.0004	0.0016	0.9545	0.1566
		γ	0.5	0.5572	0.0572	0.0678	0.9305	0.9389
	50	β_0	0.4	0.3887	-0.0113	0.0900	0.9765	1.2022
		β_1	3.5	3.5031	0.0031	0.0082	0.9405	0.3438
		σ	1	1.0006	0.0006	0.0022	0.9540	0.1865
		γ	0.5	0.5770	0.0770	0.1084	0.9320	1.1725
	75	β_0	-2.1	-2.1190	-0.0190	0.1888	0.9845	1.8003
		β_1	3.5	3.5043	0.0043	0.0228	0.9335	0.5727
		σ	1	1.0035	0.0035	0.0047	0.9530	0.2768
		γ	0.5	0.6499	0.1499	0.2955	0.9050	2.0154

sample size increases. It can also be noted that the MSEs of the ML estimates of β_0, β_1, σ and γ decrease as the sample size increases, which is expected by us since ML estimators are consistent. As pointed out by Martínez-Flórez et al. (2013), bias correction methods, such as bootstrap or jackknife (Efron, 1982; Efron and Tibshirani, 1993), could be tried to improve small sample performance. The main conclusion here is that we are quite safe to work with the ML estimation method if sample sizes are large (that is, greater than 100).

Table 2. Estimation results for the tilted-normal tobit model ($\gamma = 1$).

Sample size	Censoring percentage	Parameter	True value	Mean	Bias	MSE	CP	CL
50	5	β_0	6	6.1417	0.1417	0.5485	0.9840	2.9170
		β_1	3.5	3.5138	0.0138	0.0248	0.9410	0.5952
		σ	1	0.9993	-0.0007	0.0143	0.9420	0.4983
		γ	1	1.3193	0.3193	1.6982	0.8365	6.4760
	25	β_0	2.4	2.5572	0.1572	0.5809	0.9840	3.2330
		β_1	3.5	3.5164	0.0164	0.0434	0.9300	0.7780
		σ	1	0.9934	-0.0066	0.0167	0.9435	0.5595
		γ	1	1.3130	0.3130	1.7376	0.8215	7.2424
	50	β_0	0.05	0.2379	0.1879	0.6702	0.9830	3.8073
		β_1	3.5	3.5213	0.0213	0.0895	0.9220	1.0849
		σ	1	0.9800	-0.0200	0.0232	0.9325	0.6784
		γ	1	1.3061	0.3061	1.8358	0.8025	8.6561
	75	β_0	-2.5	-2.2420	0.2580	1.0763	0.9775	5.1703
		β_1	3.5	3.5570	0.0570	0.3211	0.8985	1.9001
		σ	1	0.9463	-0.0537	0.0459	0.8975	0.9457
		γ	1	1.2107	0.2107	1.9302	0.7510	10.8050
100	5	β_0	6	6.0401	0.0401	0.2902	0.9905	2.0510
		β_1	3.5	3.5072	0.0072	0.0124	0.9390	0.4236
		σ	1	1.0026	0.0026	0.0068	0.9600	0.3323
		γ	1	1.2999	0.2999	1.2101	0.8715	4.5938
	25	β_0	2.4	2.4526	0.0526	0.3278	0.9920	2.3338
		β_1	3.5	3.5107	0.0107	0.0227	0.9350	0.5513
		σ	1	1.0035	0.0035	0.0082	0.9550	0.3836
		γ	1	1.3147	0.3147	1.3749	0.8605	5.3230
	50	β_0	0.05	0.1408	0.0908	0.4298	0.9890	2.8498
		β_1	3.5	3.5107	0.0107	0.0431	0.9285	0.7646
		σ	1	1.0006	0.0006	0.0119	0.9545	0.4777
		γ	1	1.3014	0.3014	1.4665	0.8425	6.4662
	75	β_0	-2.5	-2.3055	0.1945	0.7475	0.9720	4.0618
		β_1	3.5	3.5254	0.0254	0.1306	0.9290	1.3025
		σ	1	0.9843	-0.0157	0.0223	0.9380	0.7052
		γ	1	1.2642	0.2642	1.8509	0.7850	8.9251
300	5	β_0	6	5.9950	-0.0050	0.0855	0.9710	1.1322
		β_1	3.5	3.5029	0.0029	0.0040	0.9475	0.2447
		σ	1	1.0001	0.0001	0.0019	0.9535	0.1745
		γ	1	1.1421	0.1421	0.3988	0.9165	2.2529
	25	β_0	2.4	2.3905	-0.0095	0.1100	0.9760	1.2999
		β_1	3.5	3.5050	0.0050	0.0069	0.9445	0.3177
		σ	1	1.0016	0.0016	0.0025	0.9540	0.2016
		γ	1	1.1843	0.1843	0.5630	0.9055	2.6950
	50	β_0	0.05	0.0414	-0.0086	0.1468	0.9845	1.6296
		β_1	3.5	3.5061	0.0061	0.0132	0.9420	0.4390
		σ	1	1.0021	0.0021	0.0037	0.9625	0.2564
		γ	1	1.2282	0.2282	0.7680	0.9010	3.5501
	75	β_0	-2.5	-2.4638	0.0362	0.2994	0.9780	2.5066
		β_1	3.5	3.5097	0.0097	0.0377	0.9395	0.7342
		σ	1	1.0002	0.0002	0.0081	0.9505	0.4031
		γ	1	1.2854	0.2854	1.2776	0.8595	5.7356
500	5	β_0	6	5.9930	-0.0070	0.0487	0.9665	0.8641
		β_1	3.5	3.5022	0.0022	0.0024	0.9455	0.1899
		σ	1	1.0003	0.0003	0.0011	0.9520	0.1317
		γ	1	1.0855	0.0855	0.1869	0.9355	1.6243
	25	β_0	2.4	2.3894	-0.0106	0.0649	0.9695	0.9893
		β_1	3.5	3.5041	0.0041	0.0041	0.9450	0.2461
		σ	1	1.0012	0.0012	0.0015	0.9565	0.1515
		γ	1	1.1152	0.1152	0.2923	0.9235	1.9189
	50	β_0	0.05	0.0383	-0.0117	0.0913	0.9760	1.2347
		β_1	3.5	3.5035	0.0035	0.0079	0.9360	0.3394
		σ	1	1.0016	0.0016	0.0023	0.9520	0.1912
		γ	1	1.1501	0.1501	0.4136	0.9180	2.4745
	75	β_0	-2.5	-2.5073	-0.0073	0.2004	0.9780	1.9543
		β_1	3.5	3.5052	0.0052	0.0220	0.9375	0.5661
		σ	1	1.0031	0.0031	0.0053	0.9570	0.3070
		γ	1	1.2613	0.2613	0.9612	0.8865	4.4039

4.2 Misspecification study

The second simulation study was based on 3,000 generated samples of size $n = 500$. The main goal was to verify if we could distinguish between the proposed model and the candidate ones, in the light of the data set, based on the adopted model selection criteria: Akaike information criterion (AIC) (Akaike, 1977), corrected AIC (AICc) (Sugiura, 1978; Hurvich and Tsai, 1989), consistent AIC (CAIC) (Bozdogan, 1987; Anderson et al., 1998), Bayesian information criterion (BIC) (Schwarz, 1978), and Hannan-Quinn information criterion (HQIC) (Hannan and Quinn, 1979).

Table 3. Estimation results for the tilted-normal tobit model ($\gamma = 2$).

Sample size	Censoring percentage	Parameter	True value	Mean	Bias	MSE	CP	CL
50	5	β_0	5.5	5.7563	0.2563	0.5080	0.9745	2.8682
		β_1	3.5	3.5112	0.1548	0.0241	0.9395	0.5863
		σ	1	0.9814	0.1141	0.0134	0.9280	0.4887
		γ	2	1.9730	1.6222	2.6308	0.7890	9.9064
	25	β_0	2.1	2.3960	0.2960	0.5682	0.9750	3.1662
		β_1	3.5	3.5123	0.0123	0.0399	0.9335	0.7515
		σ	1	0.9745	-0.0255	0.0161	0.9355	0.5509
		γ	2	1.9013	-0.0987	2.6386	0.7630	10.5477
	50	β_0	-0.4	-0.0271	0.3729	0.7246	0.9690	3.7854
		β_1	3.5	3.5204	0.0204	0.0869	0.9210	1.0689
		σ	1	0.9564	-0.0436	0.0237	0.9165	0.6806
		γ	2	1.7980	-0.2020	2.8972	0.7410	12.1170
	75	β_0	-2.9	-2.4003	0.4997	1.2122	0.9485	5.1567
		β_1	3.5	3.5541	0.0541	0.3009	0.8955	1.8605
		σ	1	0.9143	-0.0857	0.0494	0.8740	0.9397
		γ	2	1.5492	-0.4508	3.0574	0.6790	14.1499
100	5	β_0	5.5	5.6038	0.1038	0.2478	0.9710	2.0787
		β_1	3.5	3.5049	0.0049	0.0118	0.9445	0.4156
		σ	1	0.9935	-0.0065	0.0064	0.9520	0.3417
		γ	2	2.1771	0.1771	2.3208	0.8435	8.0449
	25	β_0	2.1	2.2390	0.1390	0.2746	0.9730	2.3670
		β_1	3.5	3.5076	0.0076	0.0202	0.9405	0.5298
		σ	1	0.9887	-0.0113	0.0077	0.9485	0.3940
		γ	2	2.1057	0.1057	2.4177	0.8260	8.8352
	50	β_0	-0.4	-0.1877	0.2123	0.4311	0.9595	2.9385
		β_1	3.5	3.5110	0.0110	0.0396	0.9325	0.7483
		σ	1	0.9825	-0.0175	0.0121	0.9420	0.5042
		γ	2	2.0225	0.0225	2.6717	0.8000	10.6324
	75	β_0	-2.9	-2.5234	0.3766	0.7913	0.9530	4.0722
		β_1	3.5	3.5261	0.0261	0.1224	0.9285	1.2666
		σ	1	0.9539	-0.0461	0.0237	0.9125	0.7083
		γ	2	1.7361	-0.2639	2.8601	0.7240	12.4367
300	5	β_0	5.5	5.5060	0.0060	0.0872	0.9630	1.2106
		β_1	3.5	3.5026	0.0026	0.0038	0.9455	0.2393
		σ	1	0.9989	-0.0011	0.0023	0.9490	0.1945
		γ	2	2.2276	0.2276	1.3580	0.9025	4.7415
	25	β_0	2.1	2.1138	0.0138	0.1114	0.9630	1.4184
		β_1	3.5	3.5040	0.0040	0.0061	0.9445	0.3043
		σ	1	0.9991	-0.0009	0.0030	0.9475	0.2310
		γ	2	2.2532	0.2532	1.7165	0.8855	5.6553
	50	β_0	-0.4	-0.3504	0.0496	0.1487	0.9525	1.8175
		β_1	3.5	3.5068	0.0068	0.0124	0.9450	0.4281
		σ	1	0.9951	-0.0049	0.0046	0.9470	0.3008
		γ	1	2.1821	0.1821	1.9111	0.8760	7.1213
	75	β_0	-2.9	-2.7586	0.1414	0.2937	0.9545	2.6569
		β_1	3.5	3.5126	0.0126	0.0349	0.9390	0.7116
		σ	1	0.9831	-0.0169	0.0094	0.9260	0.4415
		γ	2	2.0438	0.0438	2.3756	0.8085	9.5972
500	5	β_0	5.5	5.4934	-0.0066	0.0549	0.9615	0.9335
		β_1	3.5	3.5023	0.0023	0.0023	0.9455	0.1857
		σ	1	1.0006	0.0006	0.0014	0.9530	0.1492
		γ	2	2.1799	0.1799	0.8502	0.9230	3.5385
	25	β_0	2.1	2.0939	-0.0061	0.0728	0.9660	1.0957
		β_1	3.5	3.5040	0.0040	0.0037	0.9460	0.2358
		σ	1	1.0012	0.0012	0.0019	0.9525	0.1777
		γ	2	2.2187	0.2187	1.1618	0.9085	4.2419
	50	β_0	-0.4	-0.3887	0.0113	0.1047	0.9570	1.4216
		β_1	3.5	3.5032	0.0032	0.0075	0.9405	0.3305
		σ	1	0.9994	-0.0006	0.0031	0.9510	0.2342
		γ	2	2.2231	0.2231	1.5191	0.8950	5.5508
	75	β_0	-2.9	-2.8224	0.0776	0.2014	0.9575	2.1558
		β_1	3.5	3.5070	0.0070	0.0203	0.9375	0.5472
		σ	1	0.9914	-0.0086	0.0063	0.9375	0.3557
		γ	2	2.1280	0.1280	2.0353	0.8575	8.0622

As in the simulation study presented in the previous subsection, we considered a linear model with a single covariate $X \sim \mathrm{N}(0, 1)$ and set $\beta_1 = 3.5$. We also assumed the following distributions for the errors:

- Normal: that is, $\epsilon_i \sim \mathrm{N}(0, 1)$. To ensure a censoring percentage of about 5%, 25%, 50% and 75%, we took the following true values for β_0, respectively: 6, 2.4, 0.1 and -2.4;
- Skew-normal: that is, $\epsilon_i \sim \mathrm{SN}(0, 1, \gamma)$ (for details on the skew-normal distribution, see Azzalini, 1985). To consider the two kinds of skewness this distribution has (left-skewed if $\gamma < 0$ and right-skewed if $\gamma > 0$, while for $\gamma = 0$ the distribution reduces to the

Table 4. Estimation results for the tilted-normal tobit model ($\gamma = 5$).

Sample size	Censoring percentage	Parameter	True value	Mean	Bias	MSE	CP	CL
50	5	β_0	5	5.3540	0.3540	0.3342	0.9460	2.8536
		β_1	3.5	3.5111	0.0111	0.0201	0.9475	0.5493
		σ	1	0.9412	-0.0588	0.0162	0.8680	0.5013
		γ	5	3.5575	-1.4425	8.1214	0.7440	19.0512
	25	β_0	1.5	1.8797	0.3797	0.3676	0.9520	3.2018
		β_1	3.5	3.5134	0.0134	0.0358	0.9400	0.7148
		σ	1	0.9292	-0.0708	0.0202	0.8700	0.5740
		γ	5	3.4673	-1.5327	8.3702	0.7460	21.0448
	50	β_0	-0.9	-0.4860	0.4140	0.4478	0.9570	3.8100
		β_1	3.5	3.5140	0.0140	0.0727	0.9260	1.0123
		σ	1	0.9083	-0.0917	0.0307	0.8520	0.7005
		γ	5	3.3800	-1.6200	8.8377	0.7575	24.6978
	75	β_0	-3.4	-2.9325	0.4675	0.8327	0.9380	5.1829
		β_1	3.5	3.5171	0.0171	0.2475	0.9065	1.7592
		σ	1	0.8662	-0.1338	0.0603	0.8165	0.9525
		γ	5	3.2562	-1.7438	9.4535	0.7985	31.3246
100	5	β_0	5	5.2705	0.2705	0.2364	0.9070	2.2265
		β_1	3.5	3.5057	0.0057	0.0098	0.9495	0.3863
		σ	1	0.9601	-0.0399	0.0088	0.8905	0.3841
		γ	5	3.8660	-1.1340	6.6349	0.7725	15.6729
	25	β_0	1.5	1.8164	0.3164	0.2865	0.9030	2.5228
		β_1	3.5	3.5080	0.0080	0.0169	0.9440	0.4980
		σ	1	0.9500	-0.0500	0.0111	0.8795	0.4426
		γ	5	3.6951	-1.3049	7.4738	0.7565	17.3047
	50	β_0	-0.9	-0.5265	0.3735	0.3687	0.9255	2.9950
		β_1	3.5	3.5063	0.0063	0.0335	0.9410	0.7008
		σ	1	0.9354	-0.0646	0.0160	0.8755	0.5327
		γ	5	3.4990	-1.5010	8.3924	0.7285	19.4424
	75	β_0	-3.4	-2.9793	0.4207	0.5580	0.9445	4.1795
		β_1	3.5	3.5194	0.0194	0.1060	0.9355	1.2020
		σ	1	0.9051	0.1512	0.0319	0.8505	0.7477
		γ	5	3.3094	-1.6906	9.1418	0.7580	25.4297
300	5	β_0	5	5.1046	0.1046	0.0970	0.9320	1.4321
		β_1	3.5	3.5023	0.0023	0.0031	0.9510	0.2211
		σ	1	0.9848	-0.0152	0.0035	0.9125	0.2501
		γ	5	4.6549	-0.3451	4.4836	0.8570	11.5500
	25	β_0	1.5	1.6473	0.1473	0.1269	0.9185	1.6924
		β_1	3.5	3.5036	0.0036	0.0052	0.9505	0.2844
		σ	1	0.9781	-0.0219	0.0045	0.8970	0.2970
		γ	5	4.4234	-0.5766	5.0965	0.8290	13.1186
	50	β_0	-0.9	-0.6767	0.2233	0.1929	0.8930	2.0705
		β_1	3.5	3.5073	0.0073	0.0105	0.9445	0.3966
		σ	1	0.9638	-0.0362	0.0071	0.8735	0.3650
		γ	5	4.0483	-0.9517	6.0402	0.7815	14.9564
	75	β_0	-3.4	-3.0605	0.3395	0.3464	0.9010	2.8726
		β_1	3.5	3.5092	0.0092	0.0295	0.9345	0.6632
		σ	1	0.9402	-0.0598	0.0142	0.8540	0.5025
		γ	5	3.5831	-1.4169	7.9578	0.7305	18.6356
500	5	β_0	5	5.0518	0.0518	0.0607	0.9445	1.1470
		β_1	3.5	3.5014	0.0014	0.0019	0.9450	0.1711
		σ	1	0.9928	-0.0072	0.0021	0.9460	0.2015
		γ	5	4.9160	-0.0840	3.4303	0.8955	9.5602
	25	β_0	1.5	1.5847	0.0847	0.0826	0.9315	1.3724
		β_1	3.5	3.5032	0.0032	0.0032	0.9470	0.2198
		σ	1	0.9880	-0.0120	0.0029	0.9295	0.2425
		γ	5	4.7424	-0.2576	4.1078	0.8560	11.1365
	50	β_0	-0.9	-0.7555	0.1445	0.1282	0.9100	1.7211
		β_1	3.5	3.5030	0.0030	0.0062	0.9480	0.3060
		σ	1	0.9774	-0.0226	0.0047	0.9030	0.3051
		γ	5	4.4380	-0.5620	5.0809	0.8275	13.2592
	75	β_0	-3.4	-3.1236	0.2764	0.2595	0.8970	2.4194
		β_1	3.5	3.5058	0.0058	0.0173	0.9395	0.5098
		σ	1	0.9553	-0.0447	0.0094	0.8605	0.4235
		γ	5	3.8121	-1.1879	6.8930	0.7495	16.4207

normal one), and ensure a censoring percentage of approximately 5%, 25%, 50% and 75%, we set the following true values for β_0, respectively (and also for different values of shape/skewness parameter γ):

- For $\gamma = -2.2$: $\beta_0 = 6.6$, 3.1, 0.9 and -1.7;
- For $\gamma = -1.2$: $\beta_0 = 6.5$, 3, 0.2 and -1.8;
- For $\gamma = 1.2$: $\beta_0 = 5.4$, 1.8, -0.7 and -3.1;
- For $\gamma = 2.2$: $\beta_0 = 5$, 3, -0.7 and -3.3.

- Power-normal: that is, $\epsilon_i \sim \text{PN}(0, 1, \gamma)$ (for details on the power-normal distribution, see Gupta and Gupta, 2008). To consider the two kinds of skewness this distribution has (left-skewed if $0 < \gamma < 1$ and right-skewed if $\gamma > 1$, while for $\gamma = 1$ the distribution reduces to the normal one), and ensure a censoring percentage of approximately 5%, 25%, 50% and 75%, we assumed the following true values for β_0, respectively (and also for different values of shape/skewness parameter γ):
  - For $\gamma = 0.35$: $\beta_0 = 7.2$, 3.6, 1.1 and -1.5;
  - For $\gamma = 2.8$: $\beta_0 = 5$, 1.5, -1 and -3.2;
  - For $\gamma = 10$: $\beta_0 = 4.2$, 0.7, -1.5 and -4.3.
- Tilted-normal: that is, $\epsilon_i \sim \text{TN}(0, 1, \gamma)$. In order to consider the two kinds of skewness this distribution has, and ensure a censoring percentage of approximately 5%, 25%, 50% and 75%, we set the following true values for β_0, respectively (and also for different values of γ):
  - For $\gamma = 6.5$: $\beta_0 = 5$, 1.5, -1 and -3.5;
  - For $\gamma = 2$: $\beta_0 = 5.5$, 2, -0.5 and -2.8;
  - For $\gamma = 0.5$: $\beta_0 = 6.5$, 2.7, 0.3 and -2.1;
  - For $\gamma = 0.15$: $\beta_0 = 7$, 3.5, 1 and -1.4.

It is important to note that the shape/skewness parameter values presented above, were chosen in order to ensure a skewness measure of approximately -0.5, -0.2, 0.2 and 0.5, respectively (in the order that such values appear), for each error distribution (with the exception of the power-normal distribution for the first case, since -0.5 is less than ≈ -0.48, which is the lowest skewness measure that can be accommodated by such a model). The observed data y_i were taken as $y_i = \max\{\beta_0 + \beta_1 x_i + \epsilon_i, 0\}$, for $i = 1, \ldots, n$.

For each obtained sample and for each situation described above, we applied the following procedures: all four models (tobit-N, tobit-SN, tobit-PN and tobit-TN, where tobit-N stands for the normal tobit model, tobit-SN is the skew-normal tobit model, tobit-PN is the power-normal tobit model, and tobit-TN is the tilted-normal tobit model) were fitted to the data set and then the best one was selected according to the AIC, AICc, CAIC, BIC and HQIC criteria. The proportion of times each model was chosen is shown in Tables 5-9. The results in these tables indicate that the true model from which the sample was generated shows a higher proportion, except for the cases where the degree of asymmetry is weak.

5. Application

In this section, we illustrate the applicability of our proposed tobit-TN model (Subsection 5.2) and its diagnostics (Subsection 5.3) using an American food consumption data set (Subsection 5.1) extracted from the 1994-1996 Continuing Survey of Food Intakes by Individuals (CSFII) (USDA, 2000).

5.1 Data

In the CSFII, two nonconsecutive days of dietary data for individuals of all ages residing in the United States were collected via in-person interviews using 24 hours recall. Each sample person reported the amount of each food item consumed. Where two days were reported, there is also a third record regarding daily averages. Socioeconomic and demographic data for the sample households and their members were also collected in the survey. Here, the size of the extracted sample is $n = 304$ adults aged 20 or older (we only consider one member per household). In our application, presented in detail in this section, we select the amount of tomatoes consumed (in 400 grams) by them as the response variable.

Table 5. The proportion of times each tobit model is selected as the best one according to the AIC criterion.

True model	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
tobit-N 5%	0.8000	0.0130	0.0950	0.0920
tobit-N 25%	0.8037	0.0147	0.0933	0.0883
tobit-N 50%	0.7863	0.0130	0.1047	0.0960
tobit-N 75%	0.7817	0.0280	0.1010	0.0893
tobit-SN 5% ($\gamma = -2.2$)	0.0007	0.5393	0.2753	0.1847
tobit-SN 25%	0.0033	0.4817	0.3367	0.1783
tobit-SN 50%	0.0260	0.4233	0.3490	0.2017
tobit-SN 75%	0.1693	0.2967	0.3347	0.1993
tobit-SN 5% ($\gamma = -1.2$)	0.3130	0.1893	0.2683	0.2293
tobit-SN 25%	0.3860	0.1657	0.2447	0.2037
tobit-SN 50%	0.5240	0.1067	0.2010	0.1683
tobit-SN 75%	0.6037	0.0773	0.1683	0.1507
tobit-SN 5% ($\gamma = 1.2$)	0.3410	0.1273	0.2773	0.2543
tobit-SN 25%	0.4187	0.1143	0.2287	0.2383
tobit-SN 50%	0.5077	0.1027	0.1863	0.2033
tobit-SN 75%	0.6460	0.1057	0.1033	0.1450
tobit-SN 5% ($\gamma = 2.2$)	0.0017	0.5117	0.3120	0.1747
tobit-SN 25%	0.0033	0.5120	0.2967	0.1880
tobit-SN 50%	0.0297	0.4517	0.2837	0.2350
tobit-SN 75%	0.2013	0.3620	0.1860	0.2507
tobit-PN 5% ($\gamma = 0.35$)	0.2917	0.0000	0.5177	0.1907
tobit-PN 25%	0.3367	0.0000	0.4820	0.1813
tobit-PN 50%	0.4227	0.0010	0.4180	0.1583
tobit-PN 75%	0.5593	0.0030	0.3080	0.1297
tobit-PN 5% ($\gamma = 2.8$)	0.3530	0.1263	0.2850	0.2357
tobit-PN 25%	0.4067	0.1150	0.2477	0.2307
tobit-PN 50%	0.5173	0.1073	0.1760	0.1993
tobit-PN 75%	0.6273	0.1013	0.1257	0.1457
tobit-PN 5% ($\gamma = 10$)	0.0070	0.3173	0.4350	0.2407
tobit-PN 25%	0.0587	0.2870	0.3987	0.2557
tobit-PN 50%	0.0967	0.2727	0.3657	0.2650
tobit-PN 75%	0.3153	0.2593	0.1913	0.2340
tobit-TN 5% ($\gamma = 6.5$)	0.0030	0.0000	0.1137	0.8833
tobit-TN 25%	0.0147	0.0027	0.1460	0.8367
tobit-TN 50%	0.0607	0.0013	0.1970	0.7410
tobit-TN 75%	0.2023	0.0007	0.2613	0.5357
tobit-TN 5% ($\gamma = 2$)	0.3023	0.0003	0.2733	0.4240
tobit-TN 25%	0.3733	0.0000	0.2660	0.3607
tobit-TN 50%	0.4643	0.0007	0.2420	0.2930
tobit-TN 75%	0.5910	0.0057	0.2103	0.1930
tobit-TN 5% ($\gamma = 0.5$)	0.3283	0.1277	0.1737	0.3703
tobit-TN 25%	0.4010	0.1273	0.1477	0.3240
tobit-TN 50%	0.4893	0.1123	0.1257	0.2727
tobit-TN 75%	0.6457	0.1200	0.0810	0.1533
tobit-TN 5% ($\gamma = 0.15$)	0.0033	0.1537	0.1577	0.6853
tobit-TN 25%	0.0047	0.1690	0.1887	0.6377
tobit-TN 50%	0.0237	0.1720	0.2080	0.5963
tobit-TN 75%	0.1053	0.2527	0.1787	0.4633

Table 10 presents the definitions and sample statistics for all considered variables, where we see that the proportion of tomato-consuming individuals in the data set is around 70%. Among those consuming, an individual on average consumes 66.12 grams of tomatoes per day. The histogram and boxplots of tomato consumption are presented in Figures 2 and 3, respectively. Proposed by Hubert and Vandervieren (2008) and used when the data are skewed distributed, the adjusted boxplot (see Figure 3 right panel) indicates that some potential outliers identified by the usual boxplot (see Figure 3 left panel) are not outliers.

Table 11 shows asymmetry and kurtosis coefficients for complete data and also for positive ys. Note that values for the asymmetry and kurtosis coefficients justify using the skewed alternatives to the tobit-N model, e.g. the proposed tobit-TN model.

5.2 Model results

Following Martínez-Flórez et al. (2013), a more emphatic indication that an asymmetric model should be considered comes from testing the hypothesis of a tobit-N model against

Table 6. The proportion of times each tobit model is selected as the best one according to the AICc criterion.

True model	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
tobit-N 5%	0.8050	0.0123	0.0933	0.0893
tobit-N 25%	0.8093	0.0147	0.0900	0.0860
tobit-N 50%	0.7923	0.0123	0.1013	0.0940
tobit-N 75%	0.7887	0.0273	0.0977	0.0863
tobit-SN 5% ($\gamma = -2.2$)	0.0007	0.5393	0.2753	0.1847
tobit-SN 25%	0.0033	0.4817	0.3367	0.1783
tobit-SN 50%	0.0273	0.4233	0.3480	0.2013
tobit-SN 75%	0.1720	0.2967	0.3323	0.1990
tobit-SN 5% ($\gamma = -1.2$)	0.3177	0.1887	0.2660	0.2277
tobit-SN 25%	0.3907	0.1633	0.2433	0.2027
tobit-SN 50%	0.5317	0.1057	0.1977	0.1650
tobit-SN 75%	0.6087	0.0773	0.1660	0.1480
tobit-SN 5% ($\gamma = 1.2$)	0.3493	0.1267	0.2737	0.2503
tobit-SN 25%	0.4240	0.1137	0.2263	0.2360
tobit-SN 50%	0.5147	0.1013	0.1833	0.2007
tobit-SN 75%	0.6517	0.1050	0.1010	0.1423
tobit-SN 5% ($\gamma = 2.2$)	0.0020	0.5113	0.3120	0.1747
tobit-SN 25%	0.0033	0.5120	0.2967	0.1880
tobit-SN 50%	0.0303	0.4513	0.2833	0.2350
tobit-SN 75%	0.2037	0.3617	0.1860	0.2487
tobit-PN 5% ($\gamma = 0.35$)	0.2963	0.0000	0.5150	0.1887
tobit-PN 25%	0.3440	0.0000	0.4767	0.1793
tobit-PN 50%	0.4310	0.0010	0.4133	0.1547
tobit-PN 75%	0.5660	0.0030	0.3030	0.1280
tobit-PN 5% ($\gamma = 2.8$)	0.3597	0.1257	0.2817	0.2330
tobit-PN 25%	0.4123	0.1140	0.2453	0.2283
tobit-PN 50%	0.5233	0.1070	0.1737	0.1960
tobit-PN 75%	0.6340	0.1007	0.1233	0.1420
tobit-PN 5% ($\gamma = 10$)	0.0077	0.3173	0.4347	0.2403
tobit-PN 25%	0.0603	0.2867	0.3977	0.2553
tobit-PN 50%	0.0983	0.2723	0.3653	0.2640
tobit-PN 75%	0.3210	0.2583	0.1883	0.2323
tobit-TN 5% ($\gamma = 6.5$)	0.0033	0.0000	0.1137	0.8830
tobit-TN 25%	0.0153	0.0027	0.1457	0.8363
tobit-TN 50%	0.0627	0.0013	0.1963	0.7397
tobit-TN 75%	0.2070	0.0007	0.2600	0.5323
tobit-TN 5% ($\gamma = 2$)	0.3073	0.0003	0.2717	0.4207
tobit-TN 25%	0.3783	0.0000	0.2643	0.3573
tobit-TN 50%	0.4707	0.0007	0.2393	0.2893
tobit-TN 75%	0.5957	0.0057	0.2073	0.1913
tobit-TN 5% ($\gamma = 0.5$)	0.3333	0.1273	0.1720	0.3673
tobit-TN 25%	0.4053	0.1273	0.1460	0.3213
tobit-TN 50%	0.4957	0.1110	0.1237	0.2697
tobit-TN 75%	0.6537	0.1183	0.0787	0.1493
tobit-TN 5% ($\gamma = 0.15$)	0.0037	0.1537	0.1577	0.6850
tobit-TN 25%	0.0050	0.1690	0.1887	0.6373
tobit-TN 50%	0.0243	0.1717	0.2080	0.5960
tobit-TN 75%	0.1083	0.2517	0.1777	0.4623

an asymmetric tobit model (e.g. the tobit-TN model), that is,

$$\mathrm{H}_0 : \gamma = 1 \quad \text{versus} \quad \mathrm{H}_1 : \gamma \neq 1,$$

using the likelihood ratio statistic:

$$\Lambda = \frac{L_{\text{tobit-N}}(\boldsymbol{\theta})}{L_{\text{tobit-TN}}(\boldsymbol{\theta})}.$$

This leads to the observed value: $-2\log(\Lambda) = 50.5177$, which is greater than the 5% critical value of the Chi-square distribution with one degree of freedom, given by $\chi^2_{1;0.95} = 3.8415$. Therefore, we can conclude that the tobit-TN model fits the American food consumption data set (tomato consumption) better than the standard tobit model (that is, the tobit-N model).

Table 12 presents the parameter estimates for the tobit-N and tobit-TN models, as well as for the other asymmetric alternatives, such as the tobit-SN and tobit-PN models. Notice that all the information criteria choose the tobit-TN model as the best one.

Table 7. The proportion of times each tobit model is selected as the best one according to the CAIC criterion.

True model	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
tobit-N 5%	0.9897	0.0017	0.0063	0.0023
tobit-N 25%	0.9900	0.0013	0.0037	0.0050
tobit-N 50%	0.9873	0.0020	0.0057	0.0050
tobit-N 75%	0.9787	0.0050	0.0087	0.0077
tobit-SN 5% ($\gamma = -2.2$)	0.0220	0.5363	0.2640	0.1777
tobit-SN 25%	0.0670	0.4707	0.2997	0.1627
tobit-SN 50%	0.2410	0.3657	0.2390	0.1543
tobit-SN 75%	0.6440	0.1720	0.1020	0.0820
tobit-SN 5% ($\gamma = -1.2$)	0.7730	0.0703	0.0833	0.0733
tobit-SN 25%	0.8320	0.0507	0.0633	0.0540
tobit-SN 50%	0.9110	0.0287	0.0313	0.0290
tobit-SN 75%	0.9373	0.0177	0.0200	0.0250
tobit-SN 5% ($\gamma = 1.2$)	0.7913	0.0487	0.0840	0.0760
tobit-SN 25%	0.8450	0.0377	0.0563	0.0610
tobit-SN 50%	0.8960	0.0280	0.0340	0.0420
tobit-SN 75%	0.9457	0.0220	0.0107	0.0217
tobit-SN 5% ($\gamma = 2.2$)	0.0373	0.4980	0.3020	0.1627
tobit-SN 25%	0.0667	0.4823	0.2793	0.1717
tobit-SN 50%	0.2513	0.3603	0.2193	0.1690
tobit-SN 75%	0.6440	0.1723	0.0793	0.1043
tobit-PN 5% ($\gamma = 0.35$)	0.7610	0.0000	0.1753	0.0637
tobit-PN 25%	0.8137	0.0000	0.1330	0.0533
tobit-PN 50%	0.8637	0.0003	0.0983	0.0377
tobit-PN 75%	0.9437	0.0010	0.0430	0.0123
tobit-PN 5% ($\gamma = 2.8$)	0.7900	0.0453	0.0907	0.0740
tobit-PN 25%	0.8363	0.0380	0.0677	0.0580
tobit-PN 50%	0.8940	0.0333	0.0367	0.0360
tobit-PN 75%	0.9440	0.0203	0.0147	0.0210
tobit-PN 5% ($\gamma = 10$)	0.0947	0.2923	0.3963	0.2167
tobit-PN 25%	0.3253	0.2150	0.2793	0.1803
tobit-PN 50%	0.4680	0.1657	0.2073	0.1590
tobit-PN 75%	0.7557	0.0950	0.0627	0.0867
tobit-TN 5% ($\gamma = 6.5$)	0.0570	0.0000	0.1033	0.8397
tobit-TN 25%	0.1447	0.0027	0.1180	0.7347
tobit-TN 50%	0.3437	0.0013	0.1237	0.5313
tobit-TN 75%	0.6633	0.0003	0.0950	0.2413
tobit-TN 5% ($\gamma = 2$)	0.7707	0.0000	0.0977	0.1317
tobit-TN 25%	0.8223	0.0000	0.0757	0.1020
tobit-TN 50%	0.8717	0.0000	0.0593	0.0690
tobit-TN 75%	0.9400	0.0007	0.0310	0.0283
tobit-TN 5% ($\gamma = 0.5$)	0.7747	0.0567	0.0530	0.1157
tobit-TN 25%	0.8467	0.0467	0.0307	0.0760
tobit-TN 50%	0.8917	0.0330	0.0277	0.0477
tobit-TN 75%	0.9503	0.0270	0.0070	0.0157
tobit-TN 5% ($\gamma = 0.15$)	0.0503	0.1487	0.1463	0.6547
tobit-TN 25%	0.0860	0.1570	0.1680	0.5890
tobit-TN 50%	0.2110	0.1473	0.1557	0.4860
tobit-TN 75%	0.4707	0.1560	0.0993	0.2740

In Figure 4, we show a scatter plot of $\hat{\mathrm{E}}\left[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i\right]$ (calculated numerically using adaptive quadrature implemented by the integrate function in R) versus $\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, for $i = 1, 2, \ldots, 304$. Besides the fact that $\hat{\mathrm{E}}\left[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i\right] \neq \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, there seems to be a slightly quadratic relationship between these two quantities.

5.3 Residual and influence diagnostic analysis

Next, we perform a residual analysis to detect atypical observations and/or model misspecification. We can generate envelopes as suggested by Atkinson (1981), based on the generalized Cox-Snell (GCS) residuals, which for the case of tilted-normal distribution are defined as $r_i^{\mathrm{GCS}} = -\log\left(1 - \hat{\Psi}(y_i; \hat{\mu}_i, \hat{\sigma}, \hat{\gamma})\right)$, $i = 1, \ldots, n$, where $\hat{\Psi}$ denotes the CDF (2) fitted to the data. The results (half-normal plots with simulated envelopes) are shown in Figure 5, from which we can see that the tobit-TN model fits better the American food consumption data set.

In order to identify influential observations, we can generate graphs of the generalized Cook distance (Cook, 1977, 1986), where a high value of this measure indicates that the

Table 8. The proportion of times each tobit model is selected as the best one according to the BIC criterion.

True model	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
tobit-N 5%	0.9857	0.0020	0.0090	0.0033
tobit-N 25%	0.9827	0.0017	0.0080	0.0077
tobit-N 50%	0.9807	0.0023	0.0093	0.0077
tobit-N 75%	0.9703	0.0057	0.0120	0.0120
tobit-SN 5% ($\gamma = -2.2$)	0.0120	0.5380	0.2697	0.1803
tobit-SN 25%	0.0460	0.4760	0.3110	0.1670
tobit-SN 50%	0.1800	0.3850	0.2680	0.1670
tobit-SN 75%	0.5503	0.2033	0.1380	0.1083
tobit-SN 5% ($\gamma = -1.2$)	0.7137	0.0830	0.1097	0.0937
tobit-SN 25%	0.7873	0.0637	0.0820	0.0670
tobit-SN 50%	0.8737	0.0343	0.0477	0.0443
tobit-SN 75%	0.9067	0.0247	0.0317	0.0370
tobit-SN 5% ($\gamma = 1.2$)	0.7330	0.0613	0.1060	0.0997
tobit-SN 25%	0.8017	0.0487	0.0750	0.0747
tobit-SN 50%	0.8623	0.0370	0.0457	0.0550
tobit-SN 75%	0.9270	0.0310	0.0147	0.0273
tobit-SN 5% ($\gamma = 2.2$)	0.0250	0.5037	0.3057	0.1657
tobit-SN 25%	0.0470	0.4937	0.2840	0.1753
tobit-SN 50%	0.2010	0.3850	0.2330	0.1810
tobit-SN 75%	0.5657	0.2083	0.0987	0.1273
tobit-PN 5% ($\gamma = 0.35$)	0.6983	0.0000	0.2197	0.0820
tobit-PN 25%	0.7567	0.0000	0.1760	0.0673
tobit-PN 50%	0.8130	0.0003	0.1373	0.0493
tobit-PN 75%	0.9107	0.0010	0.0660	0.0223
tobit-PN 5% ($\gamma = 2.8$)	0.7363	0.0577	0.1117	0.0943
tobit-PN 25%	0.7937	0.0453	0.0850	0.0760
tobit-PN 50%	0.8547	0.0437	0.0527	0.0490
tobit-PN 75%	0.9147	0.0330	0.0227	0.0297
tobit-PN 5% ($\gamma = 10$)	0.0657	0.3030	0.4070	0.2243
tobit-PN 25%	0.2693	0.2340	0.3020	0.1947
tobit-PN 50%	0.3963	0.1863	0.2377	0.1797
tobit-PN 75%	0.6933	0.1153	0.0810	0.1103
tobit-TN 5% ($\gamma = 6.5$)	0.0417	0.0000	0.1053	0.8530
tobit-TN 25%	0.1083	0.0027	0.1240	0.7650
tobit-TN 50%	0.2810	0.0013	0.1377	0.5800
tobit-TN 75%	0.5893	0.0003	0.1200	0.2903
tobit-TN 5% ($\gamma = 2$)	0.7033	0.0000	0.1223	0.1743
tobit-TN 25%	0.7773	0.0000	0.0973	0.1253
tobit-TN 50%	0.8260	0.0000	0.0777	0.0963
tobit-TN 75%	0.9147	0.0007	0.0407	0.0440
tobit-TN 5% ($\gamma = 0.5$)	0.7193	0.0670	0.0670	0.1467
tobit-TN 25%	0.7950	0.0567	0.0457	0.1027
tobit-TN 50%	0.8497	0.0443	0.0350	0.0710
tobit-TN 75%	0.9313	0.0340	0.0100	0.0247
tobit-TN 5% ($\gamma = 0.15$)	0.0350	0.1513	0.1497	0.6640
tobit-TN 25%	0.0627	0.1597	0.1743	0.6033
tobit-TN 50%	0.1647	0.1553	0.1673	0.5127
tobit-TN 75%	0.4023	0.1803	0.1110	0.3063

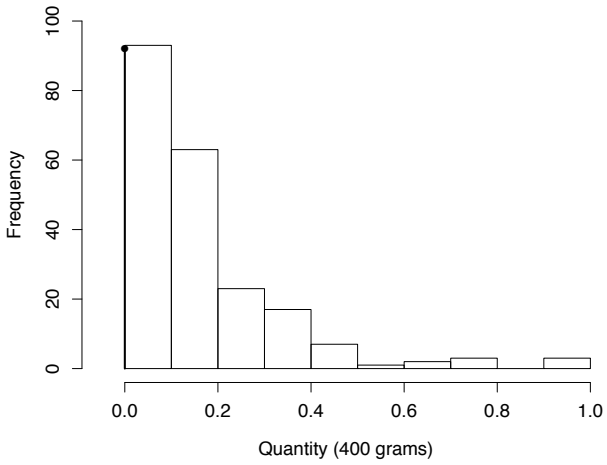

Figure 2. Distribution of the tomato consumption. The vertical line at zero on x axis represents individuals that did not consume tomatoes during the survey period.

Table 9. The proportion of times each tobit model is selected as the best one according to the HQIC criterion.

True model	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
tobit-N 5%	0.9300	0.0047	0.0327	0.0327
tobit-N 25%	0.9330	0.0070	0.0337	0.0263
tobit-N 50%	0.9260	0.0053	0.0380	0.0307
tobit-N 75%	0.9037	0.0140	0.0433	0.0390
tobit-SN 5% ($\gamma = -2.2$)	0.0033	0.5390	0.2743	0.1833
tobit-SN 25%	0.0117	0.4813	0.3313	0.1757
tobit-SN 50%	0.0653	0.4167	0.3250	0.1930
tobit-SN 75%	0.3270	0.2653	0.2460	0.1617
tobit-SN 5% ($\gamma = -1.2$)	0.4983	0.1393	0.1913	0.1710
tobit-SN 25%	0.5780	0.1143	0.1730	0.1347
tobit-SN 50%	0.7083	0.0653	0.1263	0.1000
tobit-SN 75%	0.7707	0.0520	0.0893	0.0880
tobit-SN 5% ($\gamma = 1.2$)	0.5360	0.1010	0.1877	0.1753
tobit-SN 25%	0.6053	0.0833	0.1557	0.1557
tobit-SN 50%	0.6940	0.0680	0.1147	0.1233
tobit-SN 75%	0.8113	0.0643	0.0510	0.0733
tobit-SN 5% ($\gamma = 2.2$)	0.0053	0.5110	0.3110	0.1727
tobit-SN 25%	0.0143	0.5080	0.2933	0.1843
tobit-SN 50%	0.0820	0.4347	0.2687	0.2147
tobit-SN 75%	0.3580	0.3020	0.1463	0.1937
tobit-PN 5% ($\gamma = 0.35$)	0.4790	0.0000	0.3800	0.1410
tobit-PN 25%	0.5470	0.0000	0.3297	0.1233
tobit-PN 50%	0.6297	0.0010	0.2707	0.0987
tobit-PN 75%	0.7597	0.0023	0.1723	0.0657
tobit-PN 5% ($\gamma = 2.8$)	0.5430	0.0960	0.1943	0.1667
tobit-PN 25%	0.6030	0.0817	0.1657	0.1497
tobit-PN 50%	0.7023	0.0790	0.1057	0.1130
tobit-PN 75%	0.8033	0.0663	0.0593	0.0710
tobit-PN 5% ($\gamma = 10$)	0.0193	0.3153	0.4290	0.2363
tobit-PN 25%	0.1337	0.2703	0.3667	0.2293
tobit-PN 50%	0.2130	0.2457	0.3123	0.2290
tobit-PN 75%	0.4873	0.1940	0.1417	0.1770
tobit-TN 5% ($\gamma = 6.5$)	0.0097	0.0000	0.1120	0.8783
tobit-TN 25%	0.0387	0.0027	0.1400	0.8187
tobit-TN 50%	0.1307	0.0013	0.1770	0.6910
tobit-TN 75%	0.3650	0.0007	0.2043	0.4300
tobit-TN 5% ($\gamma = 2$)	0.4937	0.0000	0.2083	0.2980
tobit-TN 25%	0.5740	0.0000	0.1870	0.2390
tobit-TN 50%	0.6540	0.0000	0.1583	0.1877
tobit-TN 75%	0.7703	0.0027	0.1127	0.1143
tobit-TN 5% ($\gamma = 0.5$)	0.5143	0.1037	0.1197	0.2623
tobit-TN 25%	0.6013	0.0977	0.0900	0.2110
tobit-TN 50%	0.6953	0.0793	0.0710	0.1543
tobit-TN 75%	0.8103	0.0773	0.0330	0.0793
tobit-TN 5% ($\gamma = 0.15$)	0.0073	0.1533	0.1573	0.6820
tobit-TN 25%	0.0177	0.1663	0.1863	0.6297
tobit-TN 50%	0.0637	0.1673	0.1970	0.5720
tobit-TN 75%	0.0637	0.1673	0.1970	0.5720

Table 10. Variable definitions and sample statistics ($n = 304$).

Variable	Definition	Mean	Standard Deviation
Dependent variable: amount consumed			
Tomato (in 400 grams)	Quantity of tomatoes consumed	0.1153	0.1598
	Among the consuming ($n = 212$; 69.74%)	0.1653	0.1684
Continuous explanatory variable			
Income	Household income as the proportion of poverty threshold	2.3730	0.8489
Binary explanatory variables (yes = 1; no = 0)			
Age 20-30	Age is 20-30	0.1480	
Age 31-40	Age is 31-40	0.1776	
Age 41-50	Age is 41-50	0.1974	
Age 51-60	Age is 51-60	0.1743	
Age > 60	Age > 60 (reference)	0.3026	
Northeast	Resides in the Northeastern states	0.1579	
Midwest	Resides in the Midwestern states	0.2336	
West	Resides in the Western states	0.2204	
South	Resides in the Southern states (reference)	0.3882	

Source: Compiled from the CSFII, USDA, 1994-1996.

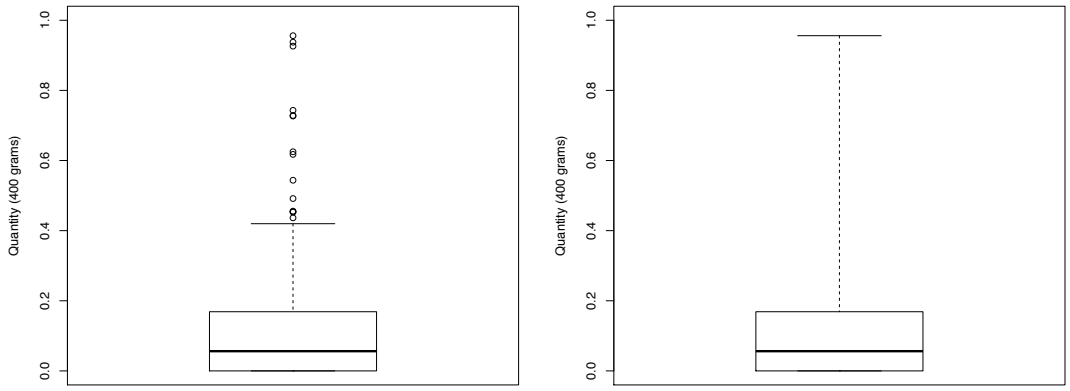

Figure 3. Usual boxplot (left panel) and adjusted boxplot (right panel) for the tomato consumption data.

Table 11. Descriptive statistics for GCS residuals of the tobit-N model.

n	Mean	Standard deviation	Skewness	Kurtosis
304	1.0970	1.3159	4.4515	26.8052
212	1.3603	1.4983	3.8351	20.0321

Table 12. Parameter estimates (standard errors in parenthesis) for tobit-N, tobit-SN, tobit-PN and tobit-TN models.

Parameter	Fitted model			
	tobit-N	tobit-SN	tobit-PN	tobit-TN
β_0 (Intercept)	-0.0025	-0.1541	-0.9440	0.5662
	(0.0446)	(0.0378)	(0.1849)	(0.0852)
β_{11} (Age 20-30)	-0.0419	-0.0164	-0.0228	-0.0222
	(0.0397)	(0.0328)	(0.0323)	(0.0284)
β_{12} (Age 31-40)	-0.0744	-0.0439	-0.0503	-0.0503
	(0.0371)	(0.0317)	(0.0306)	(0.0274)
β_{13} (Age 41-50)	-0.0142	0.0053	-0.0032	-0.0081
	(0.0353)	(0.0277)	(0.0283)	(0.0254)
β_{14} (Age 51-60)	-0.0152	0.0094	0.0017	-0.0029
	(0.0369)	(0.0293)	(0.0296)	(0.0264)
β_{21} (Northeast)	0.0845	0.0511	0.0516	0.0344
	(0.0368)	(0.0281)	(0.0296)	(0.0268)
β_{22} (Midwest)	0.0499	0.0240	0.0261	0.0173
	(0.0326)	(0.0259)	(0.0262)	(0.0234)
β_{23} (West)	0.0253	0.0166	0.0191	0.0160
	(0.0328)	(0.0269)	(0.0267)	(0.0235)
β_3 (Income)	0.0292	0.0151	0.0171	0.0133
	(0.0147)	(0.0120)	(0.0118)	(0.0103)
σ	0.2024	0.2675	0.3930	0.2325
	(0.0104)	(0.0147)	(0.0354)	(0.0185)
γ	-	4.3851	99.9963	0.0114
		(2.1513)	(75.3681)	(0.0059)
Log-likelihood	-37.3939	-22.7356	-20.2673	-12.1351
AIC	94.7879	67.4711	62.5347	46.2702
AICc	95.6920	68.5433	63.6068	47.3424
CAIC	141.9582	119.3584	114.4220	98.1576
BIC	131.9582	108.3584	103.4220	87.1576
HQIC	109.6569	83.8270	78.8905	62.6261

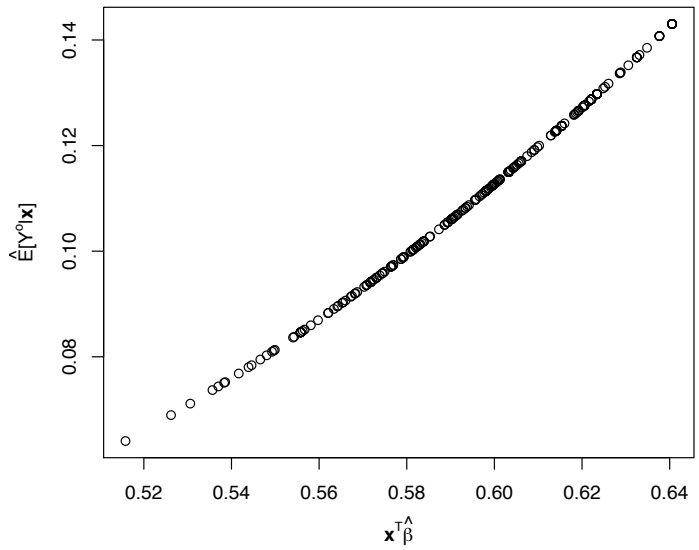

Figure 4. Scatter plot of $\hat{\mathrm{E}}\left[Y_i^{\mathrm{o}} \mid \boldsymbol{x}_i\right]$ versus $\boldsymbol{x}_i\hat{\boldsymbol{\beta}}$, $i = 1, 2, \ldots, 304$, for tobit-TN model.

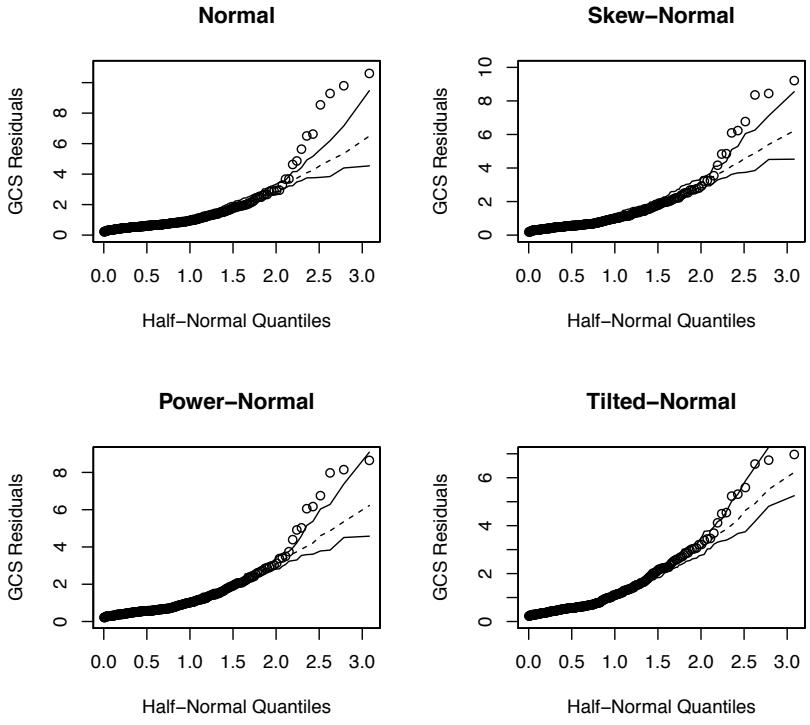

Figure 5. Half-normal plots with simulated envelopes for the GCS residuals.

corresponding observation has a high impact on the ML estimates of the parameters. We can use 1.0 as the cut-off value, as employed by some authors, like Imon (2005). From Figure 6, we note that, under the tobit-N model fitting, the observations 187 and 237 are influential on the ML estimates. However, with the tobit-SN, tobit-PN and tobit-TN models fitted, the scenario has changed: no observation is considered influential on the

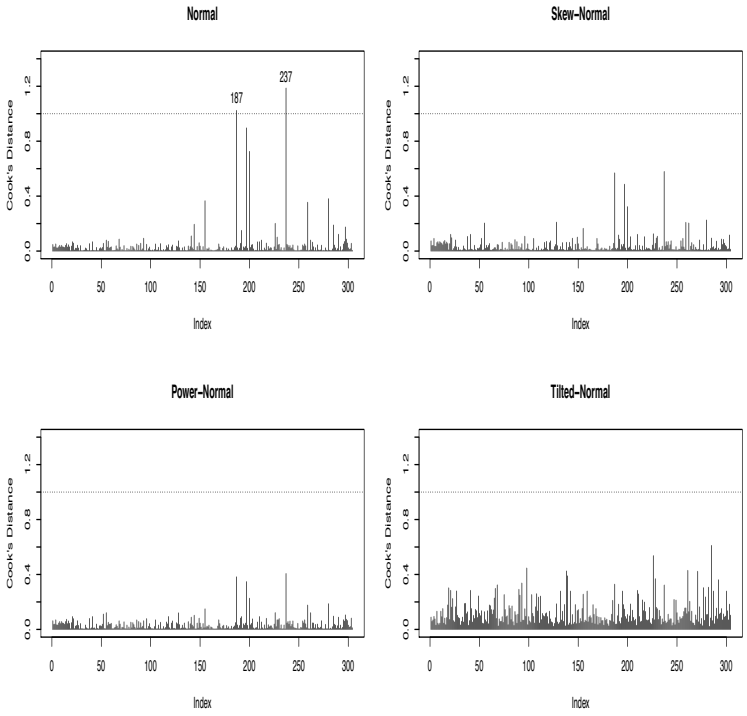

Figure 6. Generalized Cook distance. The influential observations are numbered.

parameter estimates, showing that these models are more robust.

6. Conclusions and Further Research

This paper discussed an asymmetric alternative for the standard tobit model (Tobin, 1958). It was based on the tilted-normal distribution (Maiti and Dey, 2012). The standard tobit model is a special case of the proposed model, which can also be seen as an alternative for the tobit-SN model (Hutton and Stanghellini, 2011) and tobit-PN model (Martínez-Flórez et al., 2013). Parameter estimates were obtained by using the ML method, which was also used for deriving large sample properties for the estimators. All the simulations and statistical analyses were performed using the programming language R version 3.3.1 (R Core Team, 2016). The computational code is available from the authors upon request. Simulation studies indicated good parameter recovery with the estimation approach developed, and appropriateness of the chosen model selection criteria. Since the standard tobit model is a special case of the tobit-TN model, the likelihood ratio statistic can be used for testing the standard tobit model null hypothesis. Application to an American food consumption data set (tomato consumption) indicated that the tobit-TN model can be an useful alternative to the standard tobit model, as well as to some of its asymmetric versions (tobit-SN and tobit-PN models). However, although the tobit-TN model was valid, that is, it has shown an adequate fitting to the data set at hand, we could also have considered a mixture of normal or tilted-normal distributions, for instance, as well as skewed heavy-tailed distributions for the error term, since the tomato consumption data seemed to have a long right tail. More study in this direction is desired.

Future work may also include to consider the use of other flexible distributions with better inferential properties and higher flexibility (e.g. the Families 1 to 4 considered in Jones, 2015) in the tobit framework. Other possible extension of the tobit model considers that the error term follows the centered skew-normal Birnbaum-Saunders distribution proposed by Chaves et al. (2019). Despite of being straightforward, our proposed ML estimation approach performs well, as demonstrated in the simulation results shown in Section 4. However, an interesting alternative to the direct maximization of the log-likelihood function, a procedure that sometimes can be quite cumbersome, is to use the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) or some other extensions like the Monte Carlo EM (MCEM) (Wei and Tanner, 1990), Expectation Conditional Maximization (ECM) (Meng and Rubin, 1993), ECM Either (ECME) (Liu and Rubin, 1994) or the Stochastic Approximation of EM (SAEM) algorithm (Delyon et al., 1999). As stated in Mattos et al. (2018), the EM algorithm is a very popular iterative optimization strategy in models with non-observed or incomplete data, and has many attractive features such as numerical stability, simplicity of implementation and quite reasonable memory requirements. Thus, the EM algorithm provides an interesting setting for the ML estimation of tobit models, including for instance the estimation or prediction of the censored observations. Arellano-Valle et al. (2012), Garay et al. (2016, 2017) and Mattos et al. (2018) developed efficient EM-type algorithms for the ML estimation of their proposed extensions of the standard tobit model (Tobin, 1958). The derivation of an EM-type approach for our proposed tobit-TN model, e.g., by using some general mathematical properties of the Marshall-Olkin family of distributions shown in Cordeiro et al. (2014), will be the subject to our future work. We also intend to develop a Bayesian framework for the tobit-TN model, as similarly as in the works of Garay et al. (2015) and Massuia et al. (2017).

Appendix: Score functions

In this appendix, we show the score functions of the log-likelihood function (6). These quantities are obtained as follows:

$$U(\boldsymbol{\beta}) = \frac{\partial \ell\,(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma}\sum_{i=1}^{n} d_i \left[z_i + 2(1-\gamma)k_i\Phi(z_i)\right]\boldsymbol{x}_i^{\top} - \frac{1}{\sigma}\sum_{i=1}^{n}(1-d_i)\left[w_{0i} - (1-\gamma)k_{0i}\Phi(z_{0i})\right]\boldsymbol{x}_i^{\top},$$

$$U(\sigma) = \frac{\partial \ell\,(\boldsymbol{\theta})}{\partial \sigma} = -\frac{1}{\sigma}\sum_{i=1}^{n} d_i \left[1 - z_i^2 - 2(1-\gamma)k_i\Phi(z_i)z_i\right] - \frac{1}{\sigma}\sum_{i=1}^{n}(1-d_i)z_{0i}\left[w_{0i} - (1-\gamma)k_{0i}\Phi(z_{0i})\right]$$

and

$$U(\gamma) = \frac{\partial \ell\,(\boldsymbol{\theta})}{\partial \gamma} = \frac{1}{\gamma}\sum_{i=1}^{n} d_i - 2\sum_{i=1}^{n} d_i k_i\left[1-\Phi(z_i)\right] - \sum_{i=1}^{n}(1-d_i)k_{0i}\left[1-\Phi(z_{0i})\right],$$

where $z_{0i} = -\boldsymbol{x}_i^{\top}\boldsymbol{\beta}/\sigma$, $z_i = \left(y_i - \boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right)/\sigma$, $k_{0i} = \left[1-(1-\gamma)\left\{1-\Phi(z_{0i})\right\}\right]^{-1}$, $k_i = \left[1-(1-\gamma)\left\{1-\Phi(z_i)\right\}\right]^{-1}$ and $w_{0i} = \phi(z_{0i})/\Phi(z_{0i})$.

Acknowledgements

The research is supported by the Brazilian organization FAPESP.

References

Akaike, H., 1977. On entropy maximization principle. In Krishnaiah, P.R. (Ed.), Applications of Statistics. North-Holland, Amsterdam, pp. 27-41.

Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica, 41, 997-1016.

Amemiya, T., 1984. Tobit models: A survey. Journal of Econometrics, 24, 3-61.

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge.

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, 263-282.

Andrews, D.F., and Mallows, C.L., 1974. Scale mixtures of normal distributions. Journal of the Royal Statistical Society B, 36, 99-102.

Arellano-Valle, R.B., Castro, L.M., González-Farías, G., and Muñoz-Gajardo, K.A., 2012. Student-t censored regression model: Properties and inference. Statistical Methods and Applications, 21, 453-473.

Atkinson, A.C., 1981. Two graphical displays for outlying and influential observations in regression. Biometrika, 68, 13-20.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12, 171-178.

Barros, M., Galea, M., Leiva, V., and Santos-Neto, M., 2018. Generalized tobit models: Diagnostics and application in econometrics. Journal of Applied Statistics, 45, 145-167.

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52, 345-370.

Castro, L.M., Lachos, V.H., Ferreira, G.P., and Arellano-Valle, R.B., 2014. Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach. Statistical Methodology, 18, 14-31.

Chaves, N.L., Azevedo, C.L.N., Vilca-Labra, F., and Nobre, J.S., 2019. A new Birnbaum-Saunders type distribution based on the skew-normal model under a centered parameterization. Chilean Journal of Statistics, 10, 55-76.

Cook, R.D., 1977. Detection of influential observation in linear regression. Technometrics, 19, 15-18.

Cook, R.D., 1986. Assessment of local influence. Journal of the Royal Statistical Society B, 48, 133-169.

Cordeiro, G.M., Lemonte, A.J., and Ortega, E.M.M., 2014. The Marshall-Olkin family of distributions: Mathematical properties and new models. Journal of Statistical Theory and Practice, 8, 343-366.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica, 39, 829-844.

Delyon, B., Lavielle, M., and Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. The Annals of Statistics, 27, 94-128.

Dempster, A., Laird, N., and Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society B, 39, 1-38.

Desousa, M.F., Saulo, H., Leiva, V., and Scalco, P., 2018. On a tobit-Birnbaum-Saunders model with an application to medical data. Journal of Applied Statistics, 45, 932-955.

Efron, B., 1982. The Jackknife, the Bootstrap, and Other Resampling Plans. CBNS-NSF Regional Conference Series in Applied Mathematics, Monograph 38. Philadelphia, PA: SIAM.

Efron, B., and Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, New York.

Fair, R.C., 1977. A note on the computation of the tobit estimator. Econometrica, 45, 1723-1727.

Garay, A.M., Bolfarine, H., Lachos, V.H., and Cabral, C.R.B., 2015. Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. Journal of Applied Statistics, 42, 2694-2714.

Garay, A.M., Lachos, V.H., Bolfarine, H., and Cabral, C.R.B., 2017. Linear censored regression models with scale mixtures of normal distributions. Statistical Papers, 58, 247-278.

Garay, A.M., Lachos, V.H., and Lin, T.-I. 2016. Nonlinear censored regression models with heavy-tailed distributions. Statistics and its Interface, 9, 281-293.

García, V.J., Goméz-Déniz, E., and Vázquez-Polo, F.J., 2010. A new skew generalization of the normal distribution: Properties and applications. Computational Statistics and Data Analysis, 54, 2021-2034.

Goldberger, A.S., 1964. Econometric Theory. Wiley, New York.

Greene, W.H., (2012). Econometric Analysis. Pearson, Boston.

Gupta, D. and Gupta, R.C., 2008. Analyzing skewed data by power normal model. TEST, 17, 197-210.

Hannan, E.J., and Quinn, B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society B, 41, 190-195.

Hubert, M., and Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis, 52, 5186-5201.

Hurvich, C.M., and Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika, 76, 297-307.

Hutton, J.L. and Stanghellini, E., 2011. Modelling bounded health scores with censored skew-normal distributions. Statistics in Medicine, 30, 368-376.

Imon, A.H.M.R., 2005. Identifying multiple influential observations in linear regression. Journal of Applied Statistics, 32, 929-946.

Jones, M.C., 2015. On families of distributions with shape parameters. International Statistical Review, 83, 175-192.

Liu, C., and Rubin, D.B., 1994. The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence. Biometrika, 81, 633-648.

Louzada, F., Shimizu, T.K.O., Suzuki, A.K., Mazucheli, J., and Ferreira, P.H., 2018. Compositional regression modeling under tilted normal errors: An application to a Brazilian super league volleyball data set. Chilean Journal of Statistics, 9, 33-53.

Maiti, S.S., and Dey, M., (2012). Tilted normal distribution and its survival properties. Journal of Data Science, 10, 225-240.

Marshall, A.W. and Olkin, I., 1997. A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika, 84, 641-652.

Martínez-Flórez, G., Bolfarine, H., and Gómez, H.W., 2013. The alpha-power tobit model. Communications in Statistics: Theory and Methods, 42, 633-643.

Massuia, M.B., Garay, A.M., Cabral, C.R.B., and Lachos, V.H., 2017. Bayesian analysis of censored linear regression models with scale mixtures of skew-normal distributions. Statistics and its Interface, 10, 425-439.

Mattos, T.B., Garay, A.M., and Lachos, V.H., 2018. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. Journal of Applied Statistics, 45, 2039-2066.

Meng, X.L., and Rubin, B.D., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika, 80, 267-278.

Monti, A.C., 2003. A note on the estimation of the skew normal and the skew exponential power distributions. Metron, 61, 205-219.

Pewsey, A., Gómez, H.W., and Bolfarine, H., 2012. Likelihood-based inference for power distributions. TEST, 21, 775-789.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. (Version 3.3.1) R Foundation for Statistical Computing, Vienna, Austria.

Rubio, F.J., and Steel, M.F.J., 2012. On the Marshall-Olkin transformation as a skewing mechanism. Computational Statistics and Data Analysis, 56, 2251-2257.

Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461-464.

Sugiura, N., 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics: Theory and Methods, 7, 13-26.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica, 26, 24-36.

USDA, 2000. Continuing survey of food intakes by individuals 1994-1996. CD-ROM. Agricultural Research Service, Washington, DC.

Wei, G.C., and Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association, 85, 699-704.

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 123–129

STOCHASTIC PROCESSES
RESEARCH PAPER

Mixing conditions of conjugate processes

EDUARDO HORTA[1,*] and FLAVIO ZIEGELMANN[1]

[1]Department of Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

(Received: 28 July 2019 · Accepted in final form: 11 December 2019)

Abstract

In this paper, we provide sufficient conditions ensuring that a ψ–mixing property holds for the sequence of empirical cumulative distribution functions associated with a conjugate process. Numerical examples are also provided to illustrate our results.

Keywords: Covariance operator · Functional time series · Random measure.

Mathematics Subject Classification: 60G57 · 60G10 · 62G99 · 62M99.

1. INTRODUCTION

Time series models where the dynamics is driven by a latent, unobservable *state* variable are ubiquitous in the literature – to name a few examples, we mention the ARCH and GARCH models (Engle, 1982; Bollerslev, 1986), the class of hidden Markov models (Baum and Petrie, 1966) and, more recently, the GAS model of Creal et al. (2012). Such models have found widespread use in quantitative finance, economics and other applied sciences, and it is then natural to consider extensions to a framework where the underlying state is infinite dimensional – especially when one takes into account the increasing availability of high dimensional data in the last 20 years. Contributions in that direction have been proposed, among others, by Hörmann et al. (2013) and Aue et al. (2017) who introduce functional versions of the ARCH and GARCH models, respectively. In fact, stochastic differential equations, Bayesian nonparametrics (Ghosal and Van der Vaart, 2017; Quintana, 2010) and many other probabilistic models can be interpreted as pertaining to the class of (infinite dimensional) latent variable models. Exploring such connections is beyond the scope of the present paper..

Also in the setting of an infinite dimensional state variable, Horta and Ziegelmann (2018) introduce the concept of a *conjugate process*, where the latent state is indeed the (random) conditional distribution of the observable continuous-time process. Consistency results are available, and as is common in the framework of Functional Time Series, they rely on imposing a strong mixing condition on the model. However, in this setting some additional difficulties arise because the mixing property is imposed on a functional of observable data, whereas the dynamics is specified in terms of the latent, infinite dimensional state variable. This means that it can be cumbersome to derive the required mixing condition directly from higher level model assumptions (see the discussion in Remark *1*). In this paper, we provide sufficient conditions which ensure that a ψ-mixing property is inherited by the functional of the data whenever the underlying state process is itself ψ-mixing.

*Corresponding author. Email: eduardo.horta@ufrgs.br

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

http://www.soche.cl/chjs

The remainder of the paper is organized as follows. In Section 2, we present the theoretical background, following Horta and Ziegelmann (2018). In Section 3, we state and prove our main results. Section 4 illustrates the theory through a computational example. Section 5 provides some concluding remarks.

2. Theoretical background

In Horta and Ziegelmann (2018) a *conjugate process* is defined to be a pair (ξ, X), where $X := (X_\tau : \tau \geq 0)$ is a real valued, continuous time stochastic process, and $\xi := (\xi_t : t = 0, 1, \dots)$ is a strictly stationary sequence of $M_1(\mathbb{R})$-valued (here $M_1(\mathbb{R})$ denotes the set of Borel probability measures on $\mathbb{R}$) random elements, for which the following condition holds:

$$\mathbb{P}(X_\tau \in B \,|\, \xi_0, \xi_1, \dots) = \xi_t(B), \qquad \tau \in [t, t+1), \tag{1}$$

for each $t = 0, 1, \dots$ and each Borel set B in the real line. From the statistical viewpoint, the sequence ξ is to be understood as a latent (i.e. unobservable) process, and thus all inference must be carried using information attainable from the continuous time, observable process X alone.

A crucial objective in this context is estimation of the operator $R^\mu \colon L^2(\mu) \to L^2(\mu)$ defined by

$$R^\mu f(x) := \int R_\mu(x, y) f(y)\, \mu(\mathrm{d}y), \qquad x \in \mathbb{R}$$

where the kernel R_μ is given by

$$R_\mu(x, y) := \int \mathrm{Cov}(F_0(x), F_1(z))\, \mathrm{Cov}(F_0(y), F_1(z))\, \mu(\mathrm{d}z), \qquad x, y \in \mathbb{R},$$

and where μ is a fixed, arbitrary probability measure on $\mathbb{R}$ equivalent to Lebesgue measure. In the above, $F_t(x) := \xi_t(-\infty, x]$, $x \in \mathbb{R}$, is the (random) cumulative distribution function corresponding to ξ_t. One of the key results in Horta and Ziegelmann (2018) is Theorem 2.1 below, which provides sufficient conditions under which R^μ can be $\sqrt{n}$-consistently estimated. These conditions involve an *a priori* ψ-mixing assumption on the Data Generating Process, and therefore it is of crucial importance to provide tractable conditions which in turn ensure the required ψ-mixing property. Our Theorem 3.1 below provides one such sufficient condition.

Before stating the theorem, we shall shortly introduce the estimator $\widehat{R}^\mu$ which is (as one should expect) a sample analogue of R^μ. Consider, for each $t = 1, \dots, n$, a sample of observations $\{X_{i,t} : i = 1, \dots, q_t\}$ of size q_t from $(X_\tau : \tau \in [t, t+1))$. Typically one has $X_{i,t} = X_{t+(i-1)/q_t}$. Also let $\widehat{F}_t$ denote the empirical cumulative distribution function associated with the sample $X_{1,t}, \dots, X_{q_t,t}$,

$$\widehat{F}_t(x) := \frac{1}{q_t} \sum_{i=1}^{q_t} \mathbb{I}[X_{i,t} \leq x], \qquad x \in \mathbb{R}.$$

Notice that both F_t and $\widehat{F}_t$ are random elements with values in the Hilbert space $L^2(\mu)$, and thus we find ourselves in a framework similar to Horta and Ziegelmann (2016).

In this setting, $\widehat{R}^{\mu}$ is defined to be the operator acting on $L^2(\mu)$ with kernel

$$\widehat{R}_{\mu}(x,y) \coloneqq \int \widehat{C}_1(x,z)\widehat{C}_1(y,z)\mu(\mathrm{d}z), \qquad x,y \in \mathbb{R},$$

where $\widehat{C}_1$ is the sample lag-1 covariance function

$$\widehat{C}_1(x,y) \coloneqq \frac{1}{n-1}\sum_{t=1}^{n-1}\Big(\widehat{F}_t(x) - \bar{F}_0(x)\Big) \times \Big(\widehat{F}_{t+1}(y) - \bar{F}_0(y)\Big), \qquad x,y \in \mathbb{R},$$

with $\bar{F}_0 \coloneqq (1/n)\sum_{t=1}^{n}\widehat{F}_t$.

Last but not least, let $X^{(t)}$ denote the stochastic process $(X_{t+\tau} : \tau \in [0,1))$, so that $X^{(0)}, X^{(1)}, \ldots, X^{(t)}, \ldots$ is a sequence of $\mathbb{R}^{[0,1)}$-valued random elements. We say that a conjugate process $(\xi,\, X)$ is *cyclic independent* if, conditional on ξ, we have that $\big(X^{(t)} : t = 0, 1, \ldots\big)$ is an independent sequence. This means that, for each n and each $(n+1)$-tuple $\mathcal{C}_0, \ldots, \mathcal{C}_n$ of measurable subsets of $\mathbb{R}^{[0,1)}$, it holds that

$$\mathbb{P}(X^{(0)} \in \mathcal{C}_0, \ldots, X^{(n)} \in \mathcal{C}_n \,|\, \xi) = \prod_{t=0}^{n}\mathbb{P}(X^{(t)} \in \mathcal{C}_t \,|\, \xi).$$

We are now ready to state the consistency theorem.

Theorem 2.1 (Horta and Ziegelmann, 2018) Let (ξ, X) be a cyclic–independent conjugate process, and let μ be a probability measure on $\mathbb{R}$ equivalent to Lebesgue measure. Assume that $\big(\widehat{F}_t\colon\ t = 1, 2, \ldots\big)$ is a ψ–mixing sequence, with the mixing coefficients $\Psi(k)$ satisfying $\sum_{k=1}^{\infty} k\,\Psi^{1/2}(k) < \infty$. Then, it holds that

(i) $\|\widehat{R}^{\mu} - R^{\mu}\|_{HS} = O_{\mathbb{P}}\Big(n^{-1/2}\Big)$;

(ii) $\sup_{j\in\mathbb{N}}|\widehat{\theta}_j - \theta_j| = O_{\mathbb{P}}\Big(n^{-1/2}\Big)$.

If moreover the nonzero eigenvalues of R^{μ} are all distinct, then

(iii) $\|\widehat{\psi}_j - \psi_j\|_{L^2(\mu)} = O_{\mathbb{P}}\big(n^{-1/2}\big)$, for each j such that $\theta_j > 0$.

In the above, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm of an (suitable) operator acting on $L^2(\mu)$, $(\theta_j\colon\ j \in \mathbb{N})$ $((\widehat{\theta}_j\colon\ j \in \mathbb{N}))$ denotes the non-increasing sequence of eigenvalues of R^{μ} $(\widehat{R}^{\mu})$, with repetitions if any and, for $j \in \mathbb{N}$, ψ_j $(\widehat{\psi}_j)$ denotes the unique eigenfunction associated with θ_j $(\widehat{\theta}_j)$. Notice that there is some ambiguity in defining things in this manner; to ensure that everything is well defined, we adopt the convention that the sequence (θ_j) contains zeros if and only if R^{μ} is of finite rank. Thus if the range of R^{μ} is infinite dimensional and 0 is one of its eigenvalues, it will not show up in the sequence (θ_j). On the other hand, $\widehat{R}^{\mu}$ is always of finite rank.

3. Main Result

In what follows it will be convenient to assume that the latent process is indexed for $t \in \mathbb{Z}$ and that the continuous time, observable process X is indexed for $\tau \in \mathbb{R}$. That is, we update our definitions so that $\xi \coloneqq (\xi_t : t \in \mathbb{Z})$ and $X \coloneqq (X_\tau : \tau \in \mathbb{R})$. Recall that a strictly stationary sequence $(Z_t : t \in \mathbb{Z})$ of random elements taking values in a measurable

space $\mathcal{Z}$ is said to be ψ-*mixing* if the ψ-mixing coefficient Ψ_Z defined, for $k \in \mathbb{N}$, by

$$\Psi_Z(k) := \sup \left| 1 - \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)\mathbb{P}(B)} \right| \tag{2}$$

is such that $\Psi_Z(k) \to 0$ as $k \to \infty$, where the supremum in (2) ranges over all $A \in \sigma(Z_t \colon t \leq 0)$ and all $B \in \sigma(Z_t \colon t \geq k)$ for which $\mathbb{P}(A)\mathbb{P}(B) > 0$; see Doukhan (1994) and references therein for a thorough treatment of the topic, and also Bradley (2005) for basic properties of mixing conditions.

The ψ–mixing condition in Theorem 2.1 imposes restrictions on the sequence of empirical cumulative distribution functions $(\widehat{F}_t)$ and thus constrains (F_t) and (X_τ) jointly. One could argue that it is more natural to impose a ψ–mixing condition on the latent process (ξ_t) instead, the issue being that it may be the case that a mixing property of the latter sequence is not inherited by $(\widehat{F}_t)$. If a condition slightly stronger than cyclic–independence is imposed, however, then inheritance does hold. This is our main result.

THEOREM 3.1 Let (ξ, X) be a cyclic–independent conjugate process, and let μ be a probability measure on $\mathbb{R}$ equivalent to Lebesgue measure. Assume ξ is ψ–mixing with mixing coefficient sequence Ψ_ξ. If, for each t, the conditional distribution of $X^{(t)}$ given ξ depends only on ξ_t, in the sense that the equality

$$\mathbb{P}\left[X^{(t)} \in \mathcal{C} \,\middle|\, \xi\right] = \mathbb{P}\left[X^{(t)} \in \mathcal{C} \,\middle|\, \xi_t\right] \tag{3}$$

holds for each measurable subset $\mathcal{C}$ of $\mathbb{R}^{[0,1)}$ and each t, then $\left(X^{(t)}\right)$ is ψ–mixing with mixing coefficient sequence $\Psi_X \leq \Psi_\xi$.

COROLLARY 3.2 In the conditions of Theorem 3.1, if $\sum_{k=1}^{\infty} k\Psi_\xi(k)^{1/2} < \infty$, then the ψ–mixing assumption of Theorem 2.1 holds.

Remark 1 Theorem 3.1 and Corollary 3.2 are important as they provide the applied statistician a framework for introducing models in which the ψ-mixing condition of Theorem 2.1 is satisfied, ensuring the possibility of adequate estimation procedures and statistical analyses. A particular setup in which knowledge of ψ-mixing models for multivariate time series is sufficient for obtaining a ψ-mixing sequence of random measures (ξ_t) is the scenario in which the latter sequence is in fact driven by a finite dimensional process. This is the case whenever (ξ_t) satisfies

$$\xi_t(B) = \mathbb{E}\xi_0(B) + \sum_{j=1}^{d} Z_{tj}\lambda_j(B), \qquad t \in \mathbb{Z}, \qquad B \in \text{Borel}(\mathbb{R}),$$

where d is a positive integer, the $Z_{t,j}$ are scalar random variables and the λ_j are signed measures of finite total variation. Indeed, in this setting the dynamic features of (ξ_t) are entirely captured by the multivariate time series $\boldsymbol{Z}_t = (Z_{t1}, \ldots, Z_{tj})$, and it is not difficult to see that if the mixing coefficient sequence $\Psi_{\boldsymbol{Z}}(1), \Psi_{\boldsymbol{Z}}(2), \ldots$ of $(\boldsymbol{Z}_t)$ satisfies the summability condition of Theorem 2.1, then so does Ψ_ξ.

Proof [Proof of Theorem 3.1] For $k \in \mathbb{N}$, let T_1 and T_2 be finite, nonempty subsets of $\{0, -1, -2, \ldots\}$ and $\{k, k+1, k+2, \ldots\}$ respectively, and set $T_0 := T_1 \cup T_2$. Let $\{\mathcal{C}_t, t \in T_0\}$ be a collection of measurable subsets of $\mathbb{R}^{[0,1)}$. By definition, $\sigma(X^{(t)} : t \leq 0)$ coincides with the σ-field generated by the class of sets of the form $\bigcap_{t \in T_1}[X^{(t)} \in \mathcal{C}_t]$ over all finite, nonempty $T_1 \subset \{0, -1, -2, \ldots\}$ and all collections $\{\mathcal{C}_t : t \in T_1\}$ of measurable subsets of $\mathbb{R}^{[0,1)}$, and similarly for $\sigma(X^{(t)} : t \geq k)$.

Notice that by equation (3) and the Doob–Dynkin Lemma (see Kallenberg, 1997, Lemma 1.13) we have $\mathbb{P}\left[X^{(t)} \in \mathcal{C}_t \,|\, \xi\right] = g_t \circ \xi_t$, for some measurable function $g_t \colon M_1(\mathbb{R}) \to \mathbb{R}$. This fact, together with the cyclic–independence assumption, ensures that

$$\mathbb{P}\left\{\bigcap_{t\in T_j}\left[X^{(t)} \in \mathcal{C}_t\right]\right\} = \mathbb{E}\left\{\mathbb{P}\left\{\bigcap_{t\in T_j}\left[X^{(t)} \in \mathcal{C}_t\right] \,\middle|\, \xi\right\}\right\} = \mathbb{E}\left\{\prod_{t\in T_j} g_t \circ \xi_t\right\},$$

$j = 0, 1, 2$ (a similar computation yields strict stationarity of the process $(X^{(t)} : t \in \mathbb{Z})$). Thus, the quantity

$$\left|1 - \frac{\mathbb{P}\left\{\bigcap_{t\in T_0}\left[X^{(t)} \in \mathcal{C}_t\right]\right\}}{\mathbb{P}\{\bigcap_{t\in T_1}[X^{(t)} \in \mathcal{C}_t]\}\mathbb{P}\{\bigcap_{t\in T_2}[X^{(t)} \in \mathcal{C}_t]\}}\right| \tag{4}$$

is seen to be equal to

$$\left|1 - \frac{\mathbb{E}\{\prod_{t\in T_0} g_t \circ \xi_t\}}{\mathbb{E}\{\prod_{t\in T_1} g_t \circ \xi_t\}\mathbb{E}\{\prod_{t\in T_2} g_t \circ \xi_t\}}\right|. \tag{5}$$

Substituting each g_t in (5) by an arbitrary measurable, bounded and positive $g'_t \colon M_1(\mathbb{R}) \to \mathbb{R}$, and taking the supremum over all collections $\{g'_t : t \in T_0\}$ of such g'_t, and over all $T_0 = T_1 \cup T_2$ as above, gives an upper bound to (4). It is easily seen that this supremum yields precisely $\Psi_\xi(k)$. This establishes that $\Psi_X(k) \le \Psi_\xi(k)$ and completes the proof. (By definition $\Psi_\xi(k)$ is obtained by taking the supremum over all collections of g'_t which are indicator functions of measurable subsets of $M_1(\mathbb{R})$.) ■

Proof [Proof of Corollary 3.2] By definition (or using the Doob–Dynkin Lemma) we have that $\widehat{F}_t$ is of the form $\widehat{F}_t = g_t \circ X^{(t)}$ for some measurable $g_t \colon \mathbb{R}^{[0,1)} \to L^2(\mu)$. Since $\mathbb{P}(\widehat{F}_t \in B) = \mathbb{P}(X^{(t)} \in g_t^{-1}(B))$, it follows that the supremum in the LHS over all measurable subsets B of $L^2(\mu)$ is bounded above by $\sup \mathbb{P}(X^{(t)} \in \mathcal{C})$, with $\mathcal{C}$ ranging over all measurable subsets of $\mathbb{R}^{[0,1)}$. An easy adaptation of this argument shows that the mixing coefficient sequence $\Psi_{\widehat{F}}$ is bounded above by Ψ_X. ■

4. Examples

We refer the reader to Horta and Ziegelmann (2018) for an interesting application of the theory of conjugate processes to the problem of financial risk forecasting. Below we provide a simple example to illustrate the theory.

As discussed in Horta and Ziegelmann (2018), the case where (ξ_t) is an independent sequence is of no interest, since in this case R^μ is trivially the zero operator. Consider then an independent identically distributed sequence $(\vartheta_t : t \in \mathbb{Z})$, where ϑ_t is uniformly distributed on $[0, 1]$, and let η_t be the random probability measure defined by (abusing a little on notation) $\eta_t(0) = \vartheta_t$ and $\eta_t(1) = 1-\vartheta_t$. Setting $\xi_t \coloneqq (\eta_t+\eta_{t-1})/2$, we clearly obtain a ψ-mixing sequence which satisfies the summability condition of Theorem 2.1. Indeed, (ξ_t) is 1-dependent. A straightforward computation shows that $\mathrm{Cov}(F_0(x), F_1(y)) = 1/48$ for $x, y \in [0, 1)$ and is identically zero otherwise, and therefore $R_\mu(x, y)$ is a positive constant for $x, y \in [0, 1)$ which only depends on the chosen measure μ.

Now, aside from the assumption that relation (1) holds, the nature of the process $(X_\tau : \tau \in \mathbb{R})$ is rather arbitrary. Below we simulate the case where, conditional on ξ, the process $(X_{t+\tau} : \tau \in [0, 1))$ is a continuous time Markov chain on the state space $\{0, 1\}$ with

stationary distribution $(\xi_t(0), \xi_t(1))$. There is a free parameter in the construction, which is the mean holding time $1/q_0$ of state 0. We set $q_0 = 10$. Thus, conditional on $\xi_t = \lambda_t$, the process $(X_{t+\tau} : \tau \in [0,1))$ is a Markov chain with initial distribution $(\lambda_t(0), \lambda_t(1))$ and generator

$$Q = \begin{pmatrix} -q_0 & q_0 \\ r_t & -r_t \end{pmatrix}$$

where $r_t := q_0\lambda_t(0)/\lambda_t(1)$.

The conjugate process (ξ, X) described above can be informally summarized as follows. At each day, the world finds itself in a (unobservable) *state* which is characterized by a number lying in $[0,1]$. Within each day, given the state of the world, a system can find itself in two distinct (observable) *regimes* (say, regime $\bar{0}$ and regime $\bar{1}$). This system switches between $\bar{0}$ and $\bar{1}$ according to a stationary, continuous time Markov chain, where the state of the world in that day represents the probability of the system being on regime $\bar{0}$ at any given point in time within that day. Figure 1 displays a simulated sample path for the first 4 days of the process just described.

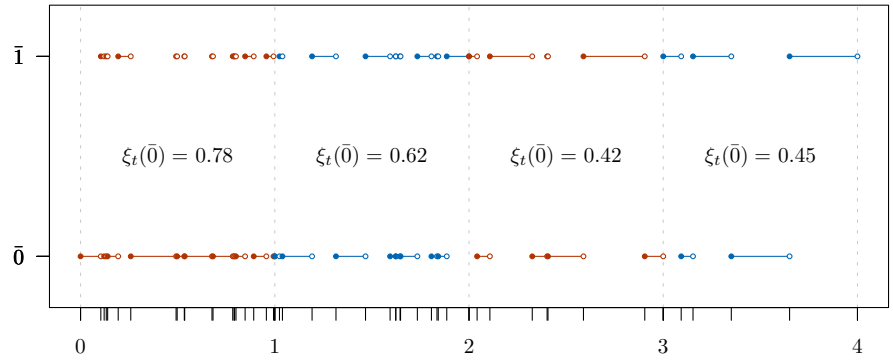

Figure 1. A simulated sample path. Even days are colored in red; odd days in blue.

We also illustrate the consistency result via a Monte Carlo simulation study. For each $t = 1, \ldots, n$, we sample the process $(X_{t+\tau} : \tau \in [0,1))$ once per cycle (that is, we take $q_t = 1$ and $X_{1,t} = X_t$) and compute the corresponding value of $\widehat{C}_1(0,0)$. Figure 2 displays the boxplot of the estimated values of $\widehat{C}_1(0,0)$ across 10000 replications of the above procedure, with the sample size varying in $\{100, 1000, 10000\}$. The blue line indicates the true parameter value $C_1(0,0) = 1/48$.

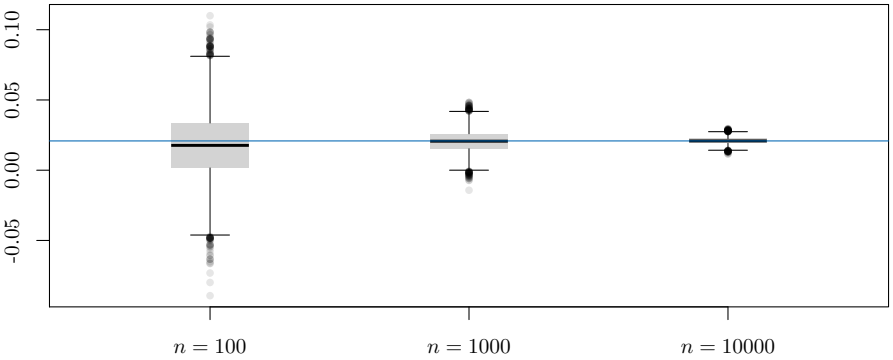

Figure 2. Boxplots of $\widehat{C}_1(0,0)$ values across replications.

5. Concluding remarks

This paper investigated conditions under which a certain ψ-mixing condition is inherited by the empirical cumulative distribution functions associated with a conjugate process (Horta and Ziegelmann, 2018). Our theoretical results, presented in Theorem 3.1 and Corollary 3.2, ensured that whenever the underlying state sequence possesses the required ψ-mixing property, so does the corresponding sequence of empirical cumulative distribution functions. The results are of relevance in settings where the dynamics is governed by an infinite dimensional latent process, as they allow the applied statistician to propose conjugate process models whose parameters can be consistently estimated – thus ensuring the possibility of adequate statistical analyses.

References

Aue, A., Horváth, L., and Pellatt, D., 2017. Functional generalized autoregressive conditional heteroskedasticity. Journal of Time Series Analysis, 38, 3-21.

Baum, L.E. and Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, 37, 1554-1563.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31, 307-327.

Bradley, R.C., 2005. Basic properties of strong mixing conditions. A survey and some open questions. Probability Surveys, 2, 107-144.

Creal, D., Koopman, S.J., and Lucas, A., 2012. Generalized autoregressive score models with applications. Journal of Applied Econometrics, 28, 777-795.

Doukhan, P., 1994. Mixing: Properties and Examples. Springer-Verlag, New York.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom Inflation. Econometrica, 50, 987-1007.

Ghosal, S. and Van der Vaart, A., 2017. Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, Cambridge.

Hörmann, S., Horváth, L., and Reeder, R., 2013. A functional version of the arch model. Econometric Theory, 29, 267-288.

Horta, E. and Ziegelmann, F., 2016. Identifying the spectral representation of Hilbertian time series. Statistics and Probability Letters, 118, 45-49.

Horta, E. and Ziegelmann, F., 2018. Conjugate processes: Theory and application to risk forecasting. Stochastic Processes and their Applications, 128, 727-755.

Kallenberg, O., 1997. Foundations of Modern Probability, Probability and its Applications, Springer, New York.

Quintana, F.A., 2010. Linear regression with a dependent skewed Dirichlet process. Chilean Journal of Statistics, 1, 35-49.

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 131–154

Statistical Quality Control
Research Paper

Performance of Shewhart control charts based on neoteric ranked set sampling to monitor the process mean for normal and non-normal processes

Guilherme Parreira da Silva[1], Cesar Augusto Taconeli[1,*], Walmes Marques Zeviani[1], and Isadora Aparecida Sprengoski do Nascimento

[1]Department of Statistics, Federal University of Paraná, Curitiba, Brazil

(Received: 05 April 2019 · Accepted in final form: 28 May 2019)

Abstract

In this study, we consider the design and performance of control charts using the neoteric ranked set sampling (NRSS) in monitoring industrial processes. NRSS is a recently proposed sampling design, based on the traditional ranked set sampling (RSS). NRSS differs from RSS by constituting, originally, a single set of k^2 sample units, instead of k sets of size k, where k is the final sample size. We evaluate NRSS control charts by average, median and standard deviation of run lengths, based on Monte Carlo simulation results. NRSS control charts perform the best, compared to RSS and some of its extensions, in most simulated scenarios. The impact of imperfect ranking and non normality are also evaluated. An application to concrete strength data serves as an illustration of the proposed method.

Keywords: Generalized normal distribution · Imperfect ranking · Perfect ranking · Run length · Skew-normal distribution

Mathematics Subject Classification: Primary 62D05 · Secondary 62P30

1. Introduction

Nowadays, technological resources are widely available for the real-time monitoring of many industrial processes. Even so, it must be recognized that sampling still plays a fundamental role in statistical quality control. Factors such as high costs, time of inspection and destructive tests may limit the evaluation of a large number of items. In this context, efficient sampling designs, providing more accurate results with smaller sample sizes, are highly useful. Ranked set sampling (RSS) and its extensions have been shown as efficient alternatives to more conventional methodologies (such as simple random sampling - SRS), when ranking sample units, according to their possible values, is substantially cheaper or easier than effectively measuring them. In the area of statistical quality control, RSS and its extensions can be applied, for example, to develop statistical quality control charts.

Originally proposed in 1924 by Walter A. Shewhart, statistical quality control charts (or

*Corresponding author. Email: Email: taconeli@ufpr.br

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

http://www.soche.cl/chjs

simply control charts) constitute a relevant tool for visualizing industrial processes and identifying assignable causes of variation (Shewhart, 1924; Montgomery, 2009). A process is said to be under statistical control when no special or assignable causes are present. Several alternatives to the original control charts were proposed, providing greater speed in detecting out-of-control situations. These alternatives include: the use of additional or alternative decision rules (Koutras et al., 2007); adaptive sampling schemes (Costa and De Magalhaes, 2007; Santore et al., 2019); nonparametric control charts (Qiu, 2018) or even the use of alternative sampling designs to the usual SRS. In this study, we consider a variety of RSS-based designs for constructing control charts.

Proposed by McIntyre (1952), the RSS is an effective sampling design when the variable of interest is expensive or difficult to measure, but it is possible ranking sample units efficiently according to some accessible and cheap criterion (Chen et al., 2003). The ranking process can be performed based, for example, on an expert's judgment or using some concomitant variable. In the first case (personal judgment), the sample units may be ordered based on visual inspection by using photos or videos, among others. In the other case, the sample units are ordered according their possible values for the variable of interest, but based only on values assessed for some correlated and accessible concomitant variable. In both cases, if the ranking criterion is not susceptible to errors, we have the perfect ranking scenario. Errors in the ranking process, however, frequently happen. In this situation, we say that the ranking process is imperfect.

RSS becomes more efficient than SRS as long as a more accurate and accessible ordering criterion is available. Several studies have shown the superiority of RSS over SRS for estimation of different population parameters (see Chen, 2007; Al-Omari and Bouza, 2014; Consulin et al., 2018). Additionally, a large number of sampling designs derived from the original RSS were proposed, such as median ranked set sampling (MRSS) by Muttlak (1997), extreme ranked set sampling (ERSS) by Samawi et al. (1996), and double ranked set sampling (DRSS) by Al-Saleh and AlKadiri (2000), among others.

RSS and its related sampling designs have been studied in the context of statistical quality control. Muttlak and Al-Sabah (2003) considered RSS and two of its modifications, ERSS and MRSS, in the design of Shewhart control charts. The authors have shown, based on an extensive simulation study, that RSS-based control charts dominate their SRS counterpart, requiring, on average, fewer samples to detect a change in the process mean. Additionally, MRSS have showed the best performance among the three sampling designs based on ranked sets. Improved control charts for DRSS schemes were also considered in the design of quality control charts. This class of sampling designs is characterized by the initial selection and ranking of k^3 (instead of k^2) sample units to draw a sample of size k after two ranking cycles. DRSS control charts outperform those based on a single ordering cycle. Recently, Mahdizadeh and Zamanzade (2019) presented a an economic variation of double RSS which reduces the number of training sample units to almost half. Furthermore, memory-based control charts using RSS, as cumulative sum or exponentially weighted moving average chart, were developed and discussed in Abid et al. (2017) and Haq et al. (2015), among others. Al-Omari and Bouza (2014) present a bibliographic review of RSS and control charts based on their related designs.

Zamanzade and Al-Omari (2016) recently proposed neoteric ranked set sampling (NRSS), another sampling design originated from RSS. Technically, its fundamental difference to RSS is the constitution and ordering of a single set of k^2 sample units, instead of k sets of size k like in RSS, MRSS and ERSS. After the ordering process, k units are chosen to compose the final sample, selected according to their specific ranks. The effect of creating a large initial set is the reduction of sample units variance, once the dispersion of order statistics decreases as the sample size increases. This reduction overcomes the covariances induced by sample units selected from the same ranked set. In this way, it was

found, for different sample sizes, correlation levels between the variable of interest and an auxiliary variable and probability distributions that NRSS overcomes RSS and SRS for estimating population mean and variance. As additional studies regarding NRSS and its higher efficiency over RSS and other RSS-based designs we recommend Koyuncu (2018) and Taconeli and Cabral (2019).

NRSS was firstly considered for control charts by Koyuncu and Karagöz (2018) to monitor the mean of bivariate asymmetric distributions. The authors studied the type I error using different RSS designs under perfect ranking (that is, when there are no errors in the ranking process). They considered the Type I Marshall-Olkin bivariate Weibull and bivariate lognormal distributions. They verified that the NRSS and RSS designs have type I error closest to 0.0027, an usual type I error adopted for Shewhart control charts. Moreover, Nawaz and Han (2019) have compared NRSS, RSS, MRSS, and ERSS in the design of homogeneously weighted moving average control charts, registering that NRSS turns out to present the best performance among the considered RSS-based schemes in monitoring the process mean under bivariate normal distribution.

In this paper, we analyze the power of Shewhart-type control charts for monitoring the process mean based on NRSS. The remainder of this article is organized as follows. In Section 2, we briefly describe the RSS-based designs. The Shewhart-type control chart based on NRSS is presented in Section 3. Section 4 covers a simulation study conducted to evaluate the performance of NRSS control charts. A case study is in Section 5, while our concluding remarks are provided in Section 6.

2. Neoteric ranked set sampling and other sampling designs based on ranked sets

In this section, we briefly describe the sampling designs considered in this study. Initially, the original RSS design can be described as presented in Algorithm 1.

Algorithm *1* RSS scheme

1: Selection of k^2 units of the population using SRS, allocating them, randomly, in k sets of size k;
2: Ranking the sample units in each set according to the possible values of the variable of interest, using the pre-established ordering criterion;
3: Selection, for the final sample, of the ith judged unit in the ith set, for $i = 1, \ldots, k$.
4: Steps 1 to 3 can be replicated n times (n cycles) producing a sample of size nk.

We denote the RSS sample by $Y_{[i]j}$, for $i = 1, \ldots, k; j = 1, l \ldots, n$, where $Y_{[i]j}$ represents the observation ranked in the ith position in the jth cycle. In this case, the sample units are independent, but not identically distributed random variables, as a result of the ordering process. Furthermore, in the perfect ranking scenario $Y_{[i].}$ becomes to the ith order statistic from a SRS of size n, which is usually denoted by $Y_{(i).}$. In this work, however, we only use $Y_{[i].}$ for both perfect and imperfect ranking scenarios. When the results are specific to just one of the ranking scenarios, it will be emphasized in the text.

The usual estimator of the population mean using RSS is given by

$$\bar{Y}_{\text{RSS}} = \frac{1}{nk} \sum_{j=1}^{n} \sum_{i=1}^{k} Y_{[i]j},$$

with variance

$$\mathrm{Var}(\bar{Y}_{\mathrm{RSS}}) = \frac{\sigma^2}{nk} - \frac{1}{nk^2}\sum_{j=1}^{n}\sum_{i=1}^{k}(\mu_{[i]} - \mu)^2,$$

where μ and σ^2 are the population mean and variance and $\mu_{[i]} = \mathrm{E}[Y_{[i]j}]$.

The MRSS scheme is detailed in Algorithm 2.

Algorithm 2 MRSS scheme

1: Selection of k^2 units of the population using SRS, allocating them, randomly, into k sets of size k;
2: Ranking the sample units in each set according to the possible values of the variable of interest, using the pre-established ordering criterion;
3: For odd k, selection, for the final sample, of the $(k+1)/2$th judged unit in the each set. For even k, we must select the units judged in position $k/2$ in half of the sets and those judged in position $(k+2)/2$ in the remaining sets;
4: Steps 1 to 3 can be replicated n times (n cycles) producing a sample of size nk.

Next, we present the steps to drawn an ERSS sample in Algorithm 3.

Algorithm 3 ERSS scheme

1: Selection of k^2 units of the population using SRS, allocating them, randomly, into k sets of size k;
2: Ranking the sample units in each set according to the possible values of the variable of interest, using the pre-established ordering criterion;
3: For even k, selection, we must select, for the final sample, the units judged as the minimum in half of the sets and those judged as the maximum in the others. However, if k is odd we must select the units judged as the minimum in $(k-1)/2$ sets; those judged as the maximum in other $(k-1)/2$ sets, and the unit judges as the median (position $(k+1)/2$) in the final set;
4: Steps 1 to 3 can be replicated n times (n cycles) producing a sample of size nk.

Additionally, Zamanzade and Mahdizadeh (2019) proposed the RSS with extreme ranks, which is a more general sampling design including ERSS as a special case. Finally, NRSS scheme (Zamanzade and Al-Omari, 2016) consists of the steps described in Algorithm 4.

Algorithm 4 NRSS scheme

1: Selection of k^2 units of the population using SRS;
2: Ranking the k^2 sample units based on the pre-established ordering criterion;
3: Selection of the $[(i-1)k+l]$-th sample unit for the final sample, for $i = 1, \ldots, k$. If k is odd, then $l = (k+1)/2$; if k is even, then $l = (k+2)/2$ when i is odd and $l = k/2$ when i is even;
4: Again, steps 1-3 can be repeated n times, setting up n cycles and producing a final sample of size nk.

As previously stated, in NRSS the k^2 original sample units must compose (and must be ordered in) a single set, which induces dependence between the observations (differently from the RSS design). The variances of these variables, however, are reduced due to the greater set size, which justifies its higher efficiency. For the sake of illustration, to select a NRSS sample of size $k = 3$, we must select the sample units ranked in positions 2, 5 and

8 from a original ordered sample of size $k^2 = 9$; for a sample of size $k = 4$, we must select those ranked in positions 3, 6, 11 and 14 from a ordered sample of size $k^2 = 16$; and for a sample of size $k = 5$, the sample units ranked in positions 3, 8, 13, 18 and 23 must be selected from a ordered sample of size $k^2 = 25$. These are the sample sizes considered in this study. It is possible to observe that the positions of the selected sample units are, in general, regularly spaced.

The NRSS sample is denoted by $\{Y_{[(i-1)k+l]j}, i = 1, \ldots, k, j = 1, \ldots, n\}$, in which $Y_{[(i-1)k+l]j}$ refers to the unit ranked in position $[(i-1)k+l]$ (of an initial sample of size k^2), in the jth cycle. Under perfect ranking, particularly, $Y_{[(i-1)k+l]j}$ corresponds to the $((i-1)k+l)$th order statistics from a SRS sample of size k^2.

According to Zamanzade and Al-Omari (2016), the NRSS sample mean is an unbiased estimator for the population mean for symmetric distributions, which can be written by:

$$\bar{Y}_{\text{NRSS}} = \frac{1}{nk} \sum_{j=1}^{n} \sum_{i=1}^{k} Y_{[(i-1)k+l]j}, \tag{1}$$

and its variance is given by:

$$\text{Var}(\bar{Y}_{\text{NRSS}}) = \frac{1}{nk^2} \sum_{i=1}^{k} \text{Var}(Y_{[(i-1)k+l]}) + \frac{2}{nk^2} \sum_{1 \le i < i' \le k}^{k} \text{Cov}(Y_{[(i-1)k+l]}, Y_{[(i'-1)k+l]}). \tag{2}$$

3. Statistical quality control charts using NRSS

In this section, the Shewhart-type control chart based on NRSS is presented. Control charts for the process mean based on simple random samples of size k are defined by a central line (CL) and a pair of control limits (LCL and UCL) given by

$$\text{LCL} = \mu_0 - A\sqrt{\text{Var}(\bar{Y}_{\text{SRS}})} = \mu_0 - A\frac{\sigma_0}{\sqrt{k}},$$

$$\text{CL} = \mu_0,$$

$$\text{UCL} = \mu_0 + A\sqrt{\text{Var}(\bar{Y}_{\text{SRS}})} = \mu_0 + A\frac{\sigma_0}{\sqrt{k}},$$

where μ_0 and σ_0 are the in-control process mean and standard deviation, $\bar{Y}_{\text{SRS}}$ the mean of a simple random sample of k units and A the amplitude parameter of the control chart. An observed sample mean beyond the control limits is an indicator of an out-of-control process. It is usual to consider $A = 3$, which, under normal distribution, is associated to a probability of a false alarm (a point outside the control limits for an in-control process) of approximately 0.0027.

We consider control charts for the process mean using NRSS, based on the structure

$$\begin{aligned} \text{LCL} &= \mu_0 - \text{A}\sqrt{\text{Var}(\bar{\text{Y}}_{\text{NRSS}})}, \\ \text{CL} &= \mu_0, \\ \text{UCL} &= \mu_0 + A\sqrt{\text{Var}(\bar{Y}_{\text{NRSS}})}, \end{aligned} \tag{3}$$

where $\bar{Y}_{\text{NRSS}}$ and $\text{Var}(\bar{Y}_{\text{NRSS}})$ are defined in (1) and (2), respectively.

Our proposal constitutes an extension of the conventional SRS control charts, in such a way that the samples are periodically selected using NRSS and the control limits are based on (3). Alternatively, extensions of control charts were previously proposed for some other designs based on RSS. The performance of these control charts are used here as reference to NRSS control charts results.

In our study, to set the values for NRSS control limits, as described in (3), it was firstly necessary to get the values for $\text{Var}(\bar{Y}_{\text{NRSS}})$, for a process under statistical control, for each simulated scenario. Under perfect ranking, $Y_{[(i-1)k+l]}$ is equivalent to the $(i-1)k+l$ order statistic from a SRS sample of size k^2, for $i = 1, \ldots, k$. Thus, in this case we calculated $\text{Var}(\bar{Y}_{\text{NRSS}})$ as presented in (2), by using the properties of order statistics from the normal distribution, presented, for example, in Balakrishnan and Rao (1998).

Under imperfect ranking, due to the ranking errors, the sampling units no longer match to order statistics. In this case, we obtained the values for $\text{Var}(\bar{Y}_{\text{NRSS}})$ by means of a preliminary simulation study. So we simulated $B = 10^6$ NRSS samples from a bivariate normal distribution for different combinations of k and ρ (the correlation between the variable of interest and an auxiliary variable). Bivariate normal distribution is very usual in several industrial applications (Montgomery, 2009). Also, it is largely considered to evaluate the performance of control charts for RSS-based designs. Then, $\text{Var}(Y_{[(i-1)k+l]})$ and $\text{Cov}(Y_{[(i-1)k+l]}, Y_{[(i'-1)k+l]})$ are estimated, respectively, by

$$\text{Var}(Y_{[(i-1)k+l]}) = \frac{\sum_{h=1}^{B} \left(Y_{[(i-1)k+l],h} - \bar{Y}_{[(i-1)k+l]}\right)^2}{B-1}, \quad i = 1, \ldots, k, \tag{4}$$

where

$$\bar{Y}_{[(i-1)k+l]} = \frac{\sum_{h=1}^{B} Y_{[(i-1)k+l],h}}{B},$$

and

$$\begin{aligned} \text{Cov}(Y_{[(i-1)k+l]}, Y_{[(i'-1)k+l]}) = \frac{1}{B-1} \sum_{h=1}^{B} &\left(Y_{[(i-1)k+l],h} - \bar{Y}_{[(i-1)k+l]}\right) \\ &\times \left(Y_{[(i'-1)k+l],h} - \bar{Y}_{[(i'-1)k+l]}\right), \end{aligned} \tag{5}$$

for $1 \leq i < i' \leq k$. Then, we replace (4) and (5) in (2) to obtain the variances, and we used them to set the NRSS control limits under imperfect ranking.

In practice, the true process parameters are rarely (if ever) known. When they are unknown, it is usual to perform the statistical process control in two distinct stages: phase I and phase II (Chakraborti et al., 2008). Phase I consists in selecting a number of samples when the process operates in-control. Their sample units should then be used for estimating the process parameters and calculating the control limits. It is usually recommended the selection of 20-25 samples in phase I, aiming to accurately define the control limits; see Montgomery (2009). Once the control limits were calculated, in phase II the obtained control chart must be used to monitor the process, based on new samples selected over time.

When the process parameters are unknown, we propose the estimation of μ_0 and $\mathrm{Var}(\bar{Y}_{\mathrm{NRSS}})$ based on the results of m independent samples of size k selected from the process in the absence of assignable causes of variation (in-control process), according to

$$\bar{\bar{Y}}_{\mathrm{NRSS}} = \frac{1}{m}\sum_{p=1}^{m} \bar{Y}_{\mathrm{NRSS_P}}$$

and

$$\widehat{\mathrm{Var}}\left(\bar{Y}_{\mathrm{NRSS}}\right) = \frac{1}{k^2}\sum_{i=1}^{k} \widehat{\mathrm{Var}}\left(Y_{[(i-1)k+l]}\right) + \frac{2}{k^2}\sum_{i<i'} \widehat{\mathrm{Cov}}\left(Y_{[(i-1)k+l]}, Y_{[(i'-1)k+l]}\right), \qquad (6)$$

where

$$\widehat{\mathrm{Var}}\left(Y_{[(i-1)k+l]}\right) = \frac{1}{m-1}\sum_{p=1}^{m}\left(Y_{[(i-1)k+l]p} - \bar{Y}_{[(i-1)k+l]}\right)^2,$$

where $\bar{Y}_{[(i-1)k+l]} = (\sum_{p=1}^{m} Y_{[(i-1)k+l]p})/m$ and

$$\widehat{\mathrm{Cov}}(Y_{[(i-1)k+l]}, Y_{[(i'-1)k+l]}) = \frac{1}{m-1}\sum_{p=1}^{m}[(Y_{[(i-1)k+l]p} - \bar{Y}_{[(i-1)k+l]}) \times (Y_{[(i'-1)k+l]p} - \bar{Y}_{[(i'-1)k+l]})], 1 \le i < i' \le k.$$

Thus, in practice the NRSS control charts for the process mean with estimated control limits are defined by substituting, in (3), μ_0 by $\bar{\bar{Y}}_{\mathrm{NRSS}}$ and $\mathrm{Var}(\bar{Y}_{\mathrm{NRSS}})$ by $\widehat{\mathrm{Var}}\left(\bar{Y}_{\mathrm{NRSS}}\right)$

$$\mathrm{LCL} = \bar{\bar{Y}}_{NRSS} - A\sqrt{\widehat{\mathrm{Var}}\left(\bar{Y}_{\mathrm{NRSS}}\right)},$$
$$\mathrm{CL} = \bar{\bar{Y}}_{\mathrm{NRSS}},$$
$$\mathrm{UCL} = \bar{\bar{Y}}_{\mathrm{NRSS}} + A\sqrt{\widehat{\mathrm{Var}}\left(\bar{Y}_{\mathrm{NRSS}}\right)}.$$

In order to investigate the bias of (6) in estimating (2), an additional simulation study was carried out, considering $k = 3, 4, 5$. For each value of k, we simulated 5×10^4 replications of m samples, using NRSS, from a normal standard distribution. For m, values between 5 and 25 were set. At each step, the m simulated samples were considered to estimate $\mathrm{Var}(\bar{Y}_{\mathrm{NRSS}})$. We found that the bias of this estimator is negligible (a relative bias lower than 0.001 was verified for all sample sizes for $m \ge 20$).

4. Monte Carlo evaluation of NRSS-based control charts

In this section we present the run length properties for NRSS-based control charts, and for other RSS-based designs, obtained through a Monte Carlo simulation study. First, we evaluate their performance when the process follows the normal distribution. Thereafter, we analyze how NRSS-based control charts, and its competitors, were affected by different departures from normal distribution, considering models with different levels of skewness and kurtosis. For this purpose, we developed computational routines using `R` language. All simulations were performed using R software (R Core Team, 2019). The packages `MASS` (Venables and Ripley, 2002), `sn` (Azzalini, 2019), and `normalp` (Mineo, 2018) were used to generate samples from normal and non-normal distributions.

To evaluate the performance of NRSS control charts under normal distribution, we simulated samples from a bivariate normal distribution, according to

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

where Y corresponds to the variable of interest and X was the concomitant variable. We assume $\mu_Y = \mu_0 = 0$ as the in-control process mean. The efficiency in ranking the sample units into each set was specified thorough ρ, such that higher levels of imperfect ranking was introduced by decreasing ρ. For the out-of-control scenarios, we consider

$$\mu_Y = \mu_0 + \frac{\delta \sigma_0}{\sqrt{k}},$$

such that δ determines the shift in the process mean:

$$\delta = |\mu_Y - \mu_0| \, \frac{\sqrt{k}}{\sigma_0}, \tag{7}$$

and $\delta = 0$ implies to an in-control process.

As parameters settings for the simulation study we had $k = 3, 4$ and 5; $\delta = 0, 0.1, 0.2, 0.3, 0.4, 0.8, 1.2, 1.6, 2, 2.4$ and 3.2 and $\rho = 0, 0.25, 0.50, 0.75, 0.9$ and 1. To evaluate the performance of control charts we consider the average run length (ARL), defined as the average number of points in a control chart until one exceeds the control limits. Particularly, if we have an in-control process, ARL_0 is the reciprocal of the false alarm error rate; for an out-of-control process, ARL_1 is inversely proportional to the detection probability, representing the average number of samples until the out-of-control state is detected. For each combination of k, and δ we simulated 10^6 independent NRSS samples under perfect ranking and 10^7 under imperfect ranking, for each considered correlation level (ρ value). At this stage, we have to increase the number of simulations, due to some numerical instability in estimating the run length properties. Based on results provided by a previous convergence study (results were not showed), we noticed some additional instability when considering imperfect ranking. This study pointed out that the adopted simulation sizes were satisfactory to achieve satisfactory convergence. The ARL values were calculated as the inverse of the proportion of points (sample means) beyond the control limits. In addition, the simulation results were also summarized by means of standard deviation of the run length (SDRL) and median run length (MRL), since the run length distribution is quite skewed.

The parameters for the simulation study were chosen in such a way to allow the comparison of the ARL values with those presented in other publications, referring to control charts for other sampling designs based on RSS. Moreover, it becomes evident that the considered scenarios (198 in total) comprises a great variety of processes. The sample size was limited to $k = 5$ given the context for application of sampling designs based on RSS (restrictions related to draw big samples, initial selection and ranking of k^2 - or even k^3 or more - sample units, among others). Moreover, the amplitude parameter (A) for the control limits were set, under perfect ranking, so that $\text{ARL}_0 = 370.51$. This is the ARL_0 corresponding to SRS control charts when we set $A = 3$. In this way, we could fairly compare the ARL_1 values for NRSS control charts with those provided by the other sampling designs. The DRSS designs control charts, particularly, produce low values for ARL_0 and, consequently, high false alarms rates when $A = 3$.

Table 1 presents the simulated run length results for RSS-based control charts. Besides NRSS, results obtained by SRS, RSS, ERSS and MRSS are also presented. In this first part of the analysis, we consider perfect ranking ($\rho = 1$), allowing to assess the maximum power provided by each design. Some conclusions drawn from Table 1 are the following:

- The efficiency of NRSS control charts for detecting shifts in process mean increases, as expected, for higher values of δ and k. As an illustration, for $k = 3$ and $\delta = 0.40$ we have ARL = 120.60 compared to ARL = 6.41 for $\delta = 1.20$, while for $k = 5$ and $\delta = 0.40$ we have ARL = 102.60 for $k = 3$ against ARL = 60.14 for $k = 5$;
- The NRSS control charts perform better than SRS control charts in all simulated scenarios. For example, for $k = 3$ and $\delta = 0.80$ we have ARL = 21.25 for NRSS control charts compared to ARL = 71.55 for SRS, while for $k = 5$ and $\delta = 1.60$ we have ARL = 1.46 for NRSS against ARL = 12.38 for SRS;
- The NRSS control charts dominates RSS and ERSS designs in all the simulated scenarios. For example, when compared to RSS, for $k = 3$ and $\delta = 0.80$ we have ARL = 21.25 for NRSS control charts against ARL = 35.43 for RSS, while for $k = 5$ and $\delta = 1.60$ we have ARL = 1.46 for NRSS against ARL = 2.83 for RSS;
- The NRSS control charts overcome the MRSS competitor in all simulated scenarios. This is remarkable, once MRSS is well known by its higher efficiency in estimating the mean, compared to RSS, for symmetric distributions. Additionally, MRSS performs best under both single and DRSS strategies for control charts for the process mean (Mehmood et al., 2013). When $k = 3$ and $\delta = 0.80$ it was verified ARL = 21.25 for NRSS control charts compared to ARL = 29.52 for MRSS, while when $k = 5$ and $\delta = 1.60$ we have ARL = 1.46 for NRSS against ARL = 2.04 for MRSS.

In order to summarize the performance of the different control charts designs, Figure 1 presents the geometric means of the ratios of ARL values for SRS control charts relative to the ones obtained by each of the other sampling designs, for each sample size. The ARL values for SRS control charts were, on average, 2.39 times larger than the corresponding NRSS when $k = 3$; 3 times for $k = 4$ and 3.59 times for $k = 5$. The best performance of NRSS control charts over the RSS, ERSS and MRSS counterparts becomes evident. For MRSS, for example, we have, on average, ARL 1.22 times higher than NRSS for $k = 3$; 1.25 times for $k = 4$ and 1.28 times for $k = 5$.

Table 1. ARL, MRL and SDRL for control charts constructed by SRS and designs based on RSS under perfect ranking

		SRS			RSS			ERSS			MRSS			NRSS		
k	δ	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL
3	0.00	370.40	257	369.90	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01
	0.10	352.93	245	352.43	333.89	232	333.39	340.25	236	339.75	339.67	236	339.17	325.63	226	325.13
	0.20	308.43	214	307.93	266.03	185	265.53	272.18	189	271.68	265.11	184	264.61	234.03	162	233.53
	0.30	253.14	176	252.64	196.93	137	196.43	197.20	137	196.70	186.22	129	185.72	157.23	109	156.73
	0.40	200.08	139	199.58	139.43	97	138.93	137.99	96	137.49	128.12	89	127.62	102.60	71	102.10
	0.80	71.55	50	71.05	35.43	25	34.93	35.35	25	34.85	29.52	21	29.02	21.25	15	20.74
	1.20	27.82	19	27.32	11.54	8	11.03	11.43	8	10.92	9.22	7	8.71	6.41	5	5.89
	1.60	12.38	9	11.87	4.76	3	4.23	4.75	3	4.22	3.80	3	3.26	2.76	2	2.20
	2.00	6.30	5	5.78	2.50	2	1.94	2.49	2	1.93	2.06	2	1.48	1.61	1	0.99
	2.40	3.65	3	3.11	1.61	1	0.99	1.61	1	0.99	1.40	1	0.75	1.20	1	0.49
	3.20	1.73	1	1.12	1.09	1	0.31	1.09	1	0.31	1.04	1	0.20	1.01	1	0.10
4	0.00	370.40	257	369.90	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01
	0.10	352.93	245	352.43	328.08	228	327.58	341.30	237	340.80	318.07	221	317.57	310.56	215	310.06
	0.20	308.43	214	307.93	249.81	173	249.31	266.81	185	266.31	232.45	161	231.95	210.30	146	209.80
	0.30	253.14	176	252.64	174.89	121	174.39	192.64	134	192.14	156.42	109	155.92	126.90	88	126.40
	0.40	200.08	139	199.58	119.36	83	118.86	135.85	94	135.35	100.29	70	99.79	77.86	54	77.36
	0.80	71.55	50	71.05	27.78	19	27.28	33.69	24	33.19	21.42	15	20.91	13.89	10	13.38
	1.20	27.82	19	27.32	8.54	6	8.02	10.70	8	10.19	6.38	5	5.86	4.09	3	3.56
	1.60	12.38	9	11.87	3.55	3	3.01	4.41	3	3.88	2.73	2	2.17	1.89	1	1.30
	2.00	6.30	5	5.78	1.94	1	1.35	2.33	2	1.76	1.59	1	0.97	1.25	1	0.56
	2.40	3.65	3	3.11	1.35	1	0.69	1.53	1	0.90	1.19	1	0.48	1.06	1	0.25
	3.20	1.73	1	1.12	1.03	1	0.18	1.07	1	0.27	1.01	1	0.10	1.00	1	0.00
5	0.00	370.40	257	369.90	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01	370.51	257	370.01
	0.10	352.93	245	352.43	331.68	230	331.18	333.00	231	332.50	329.60	229	329.10	299.58	208	299.08
	0.20	308.43	214	307.93	244.98	170	244.48	254.77	177	254.27	223.41	155	222.91	181.06	126	180.56
	0.30	253.14	176	252.64	165.54	115	165.04	173.73	121	173.23	136.91	95	136.41	104.59	73	104.09
	0.40	200.08	139	199.58	107.88	75	107.38	117.44	82	116.94	85.20	59	84.70	60.14	42	59.64
	0.80	71.55	50	71.05	22.53	16	22.02	26.59	19	26.09	15.56	11	15.05	9.55	7	9.04
	1.20	27.82	19	27.32	6.73	5	6.21	8.13	6	7.61	4.55	3	4.02	2.86	2	2.31
	1.60	12.38	9	11.87	2.83	2	2.28	3.38	2	2.84	2.04	2	1.46	1.46	1	0.82
	2.00	6.30	5	5.78	1.63	1	1.01	1.87	1	1.28	1.31	1	0.64	1.10	1	0.33
	2.40	3.65	3	3.11	1.21	1	0.50	1.32	1	0.65	1.08	1	0.29	1.01	1	0.10
	3.20	1.73	1	1.12	1.01	1	0.10	1.03	1	0.18	1.00	1	0.00	1.00	1	0.00

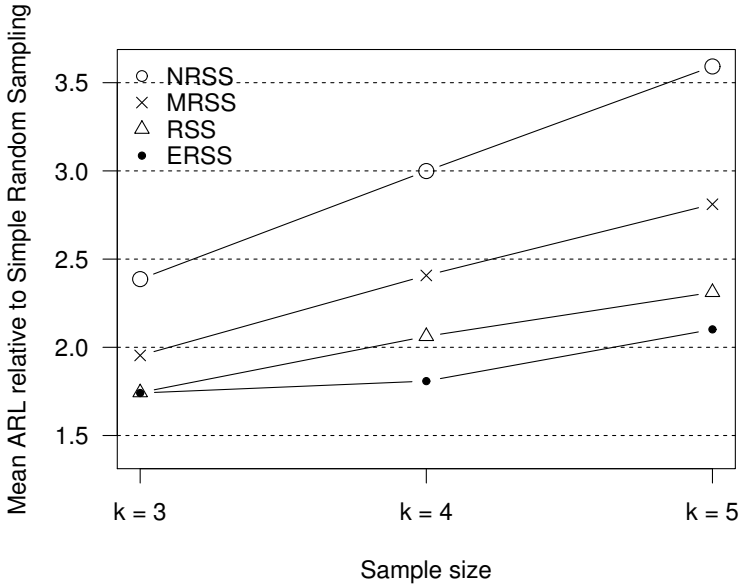

Figure 1. Average relative efficiency from control charts of designs based on RSS compared to SRS under perfect ranking. ARL from RSS, MRSS and ERSS were taken from Al-Omari and Haq (2012).

Table 2 presents the simulation results under imperfect ranking by setting $A = 3$ (3-sigma limits). This is a traditional choice for Shewhart control charts. The ARL values for $\rho = 0$ are identical to the corresponding ones from SRS, once NRSS and SRS are equivalent if the ordering is done completely at random. Based on these results, it is possible to assess the impact of ranking errors in the performance of control charts. Some conclusions from Table 2 are highlighted next:

- Control charts for all RSS based designs lose performance when the correlation between the variables decreases. For example, for NRSS control charts, $k = 3$ and $\delta = 0.8$, ARL $= 21.34$ when $\rho = 1$, ARL $= 31.23$ when $\rho = 0.90$; 44.02 when rho $\rho = 0.75$ and 59.55 when $\rho = 0.50$;
- The ARL values for NRSS control charts are smaller compared to the ones provided by SRS in almost all simulated scenarios with $\delta \neq 0$. NRSS only loses in a few scenarios described by low shifts in process mean and low values for ρ;
- The ARL_0 values from NRSS control charts are around 370.4, as intended. The individual ARL_0 values range from 365.62 when $k = 3$ and $\rho = 0.75$, to 372.04, when $k = 5$ and $\rho = 1$.

Figure 2 shows the geometric means of the ratios of ARL values for SRS control charts relative to the ones obtained by each of the RSS based designs. These results are presented for each sample size and considering the different correlation levels between the auxiliary and the variable of interest. We can notice that NRSS control charts are, in general, more efficient than all other considered sampling designs. Moreover, the superiority of NRSS control charts becomes higher when the correlation between the variables increases. For $\rho = 0.9$ and $\rho = 1$, we have, on average, higher efficiency for the NRSS control charts with $k = 4$ than for the other sampling designs taking $k = 5$, which can reflect in resource savings and lower operational effort.

Table 2. ARL, MRL and SDRL for control charts constructed by NRSS under imperfect ranking

		$\rho = 0$			$\rho = 0.25$			$\rho = 0.50$			$\rho = 0.75$			$\rho = 0.90$			$\rho = 1.00$		
k	δ	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL
3	0.00	370.40	257	369.90	369.11	256	368.61	371.24	257	370.74	365.62	254	365.12	369.37	256	368.87	369.15	256	368.65
	0.10	352.93	245	352.43	349.32	242	348.82	349.84	243	349.34	343.41	238	342.91	335.38	233	334.88	322.07	223	321.57
	0.20	308.43	214	307.93	307.94	214	307.44	297.73	207	297.23	279.09	194	278.59	259.64	180	259.14	233.58	162	233.08
	0.30	253.14	176	252.64	248.44	172	247.94	237.54	165	237.04	211.75	147	211.25	183.29	127	182.79	155.70	108	155.20
	0.40	200.08	139	199.58	195.30	136	194.80	182.48	127	181.98	156.13	108	155.63	127.46	89	126.96	101.62	71	101.12
	0.80	71.55	50	71.05	68.42	48	67.92	59.55	41	59.05	44.02	31	43.52	31.23	22	30.73	21.34	15	20.83
	1.20	27.82	19	27.32	26.29	18	25.79	22.04	15	21.53	15.07	11	14.56	9.92	7	9.41	6.44	5	5.92
	1.60	12.83	9	12.32	11.67	8	11.16	9.55	7	9.04	6.34	5	5.82	4.15	3	3.62	2.76	2	2.20
	2.00	6.30	5	5.78	5.92	4	5.40	4.84	3	4.31	3.26	2	2.71	2.23	2	1.66	1.61	1	0.99
	2.40	3.65	3	3.11	3.43	3	2.89	2.85	2	2.30	2.01	2	1.42	1.49	1	0.85	1.20	1	0.49
	3.20	1.73	1	1.12	1.65	1	1.04	1.45	1	0.81	1.19	1	0.48	1.06	1	0.25	1.01	1	0.10
4	0.00	370.40	257	369.90	369.89	257	369.39	370.89	257	370.39	371.00	257	370.50	368.05	255	367.55	371.43	258	370.93
	0.10	352.93	245	352.43	355.25	246	354.75	345.83	240	345.33	341.05	237	340.55	333.70	231	333.20	310.75	216	310.25
	0.20	308.43	214	307.93	310.98	216	310.48	296.94	206	296.44	274.31	190	273.81	244.20	169	243.70	208.73	145	208.23
	0.30	253.14	176	252.64	246.69	171	246.19	237.34	165	236.84	206.25	143	205.75	167.96	117	167.46	127.24	88	126.74
	0.40	200.08	139	199.58	193.12	134	192.62	178.81	124	178.31	147.41	102	146.91	112.07	78	111.57	76.67	53	76.17
	0.80	71.55	50	71.05	68.61	48	68.11	58.03	40	57.53	39.93	28	39.43	24.92	17	24.41	13.91	10	13.40
	1.20	27.82	19	27.32	26.15	18	25.65	21.15	15	20.64	13.28	9	12.77	7.66	5	7.14	4.10	3	3.57
	1.60	12.83	9	12.32	11.53	8	11.02	9.16	6	8.65	5.57	4	5.05	3.23	2	2.68	1.89	1	1.30
	2.00	6.30	5	5.78	5.86	4	5.34	4.64	3	4.11	2.89	2	2.34	1.82	1	1.22	1.25	1	0.56
	2.40	3.65	3	3.11	3.41	2	2.87	2.74	2	2.18	1.82	1	1.22	1.29	1	0.61	1.06	1	0.25
	3.20	1.73	1	1.12	1.64	1	1.02	1.42	1	0.77	1.14	1	0.40	1.02	1	0.14	1.00	1	0.00
5	0.00	370.40	257	369.90	367.97	255	367.47	368.85	256	368.35	369.72	256	369.22	369.33	256	368.83	372.04	258	371.54
	0.10	352.93	245	352.43	354.70	246	354.20	345.42	240	344.92	341.92	237	341.42	328.10	228	327.60	296.14	205	295.64
	0.20	308.43	214	307.93	308.61	214	308.11	298.38	207	297.88	274.64	191	274.14	235.35	163	234.85	182.99	127	182.49
	0.30	253.14	176	252.64	249.19	173	248.69	234.05	162	233.55	197.91	137	197.41	155.22	108	154.72	104.91	73	104.41
	0.40	200.08	139	199.58	193.77	134	193.27	177.52	123	177.02	141.57	98	141.07	100.78	70	100.28	60.11	42	59.61
	0.80	71.55	50	71.05	67.86	47	67.36	56.81	40	56.31	37.11	26	36.61	21.01	15	20.50	9.65	7	9.14
	1.20	27.82	19	27.32	26.02	18	25.52	20.68	14	20.17	12.20	9	11.69	6.34	5	5.82	2.88	2	2.33
	1.60	12.83	9	12.32	11.48	8	10.97	8.88	6	8.37	5.10	4	4.57	2.72	2	2.16	1.46	1	0.82
	2.00	6.30	5	5.78	5.83	4	5.31	4.52	3	3.99	2.67	2	2.11	1.59	1	0.97	1.10	1	0.33
	2.40	3.65	3	3.11	3.39	2	2.85	2.67	2	2.11	1.71	1	1.10	1.19	1	0.48	1.01	1	0.10
	3.20	1.73	1	1.12	1.64	1	1.02	1.40	1	0.75	1.11	1	0.35	1.01	1	0.10	1.00	1	0.00

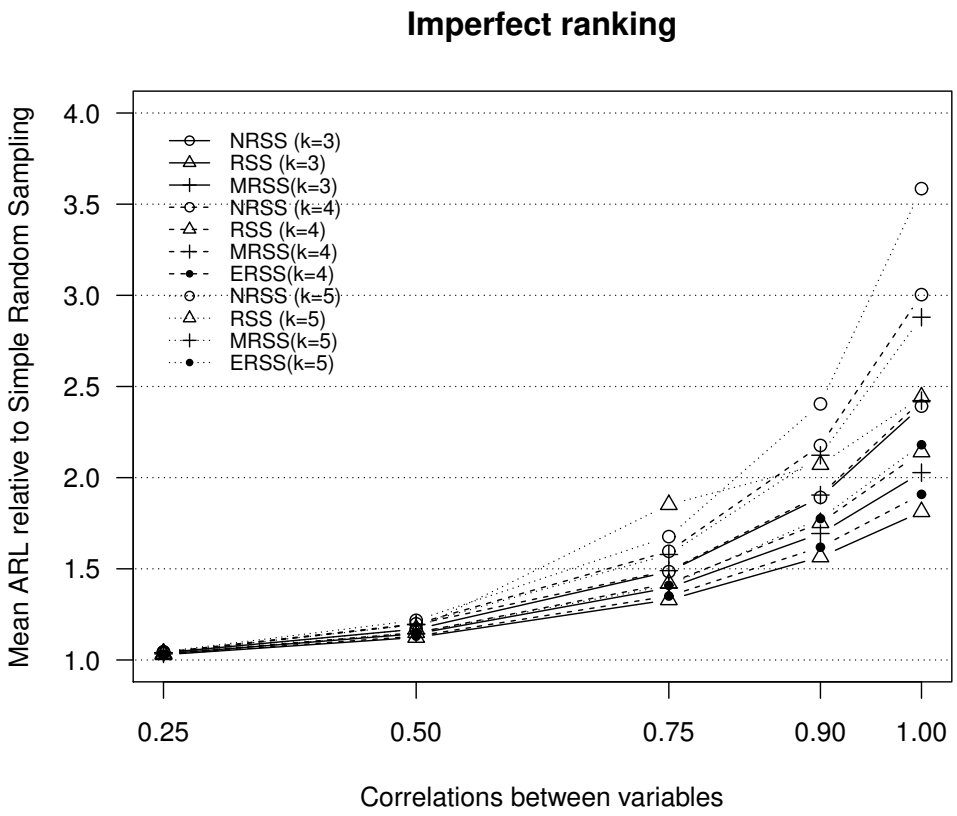

Figure 2. Average relative efficiency from control charts of designs based on RSS compared to SRS under imperfect ranking. ARL from RSS, MRSS and ERSS were taken from Al-Omari and Haq (2012). For $k = 3$ RSS and ERSS provides the same sampling design.

In order to evaluate the effect of non normality on the performance of NRSS control charts, a new simulation study was conducted. Two probability distributions are considered at this point: the skew normal and the generalized normal distributions (Azzalini, 1985; Nadarajah, 2005). Through these models, we were able to evaluate the impact of different levels of skewness and kurtosis on the run length results. The skew normal and the generalized normal models are briefly described in the following paragraphs.

The probability density function of a random variable with skew normal distribution is given by

$$f(y;\epsilon,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{y-\epsilon}{\omega}\right)\Phi\left(\alpha\left(\frac{y-\epsilon}{\omega}\right)\right),$$

where $y \in (-\infty,\infty)$ and $\epsilon \in (-\infty,\infty)$, $\omega > 0$ and $\alpha \in (-\infty,\infty)$ are location, scale and shape parameters, respectively. Additionally, ϕ and Φ represent the probability density function and the cumulative distribution function of the standard normal distribution. The skew normal distribution becomes more asymmetric as $|\alpha|$ increases. When $\alpha > 0$, the distribution is right skewed; left skewed if $\alpha < 0$ and for $\alpha = 0$ we have the normal distribution.

A random variable has generalized normal distribution if its probability density function is given by

$$f(y;\mu,\beta,\alpha) = \frac{1}{2\,\alpha^{1/\alpha}\,\Gamma(1+1/\alpha)\,\beta}\,\mathrm{e}^{-\frac{|y-\mu|^{\alpha}}{\alpha\beta^{\alpha}}},$$

where $y \in (-\infty,\infty)$ and $\mu \in (-\infty,\infty)$, $\beta > 0$ and $\alpha > 0$ are location, scale and shape parameters, respectively. The generalized normal distribution is symmetric around μ and becomes the normal distribution when $\alpha = 2$. In addition, for $\alpha < 2$ it produces leptokurtic (fatter tails) distributions, and platykutics (thinner tails) distributions when $\alpha > 2$. As particular cases of the generalized normal distribution we have, for example, the Laplace ($\alpha = 1$) and uniform ($\alpha \to \infty$) distributions.

We consider four different parameter combinations for each one of the two distributions. For the skew normal model, an increasing sequence of values for α was defined ($\alpha = 1, 2, 3$ and 5), providing distributions with different levels of skewness. Additionally, we set $\omega = 1$ and, for ϵ, we have assigned appropriate values such that the process mean was equal to zero. For the generalized normal model, four different values for α were selected, producing two distributions with heavy tails (for $\alpha = 1$ and 1.5) and two with light tails (for $\alpha = 3$ and 4). Furthermore, for the other model parameters we set $\mu = 0$ and $\beta = 1$. In all cases, the process mean was set at zero since the objective here is to evaluate the robustness of the control charts in maintaining the average (and median) run length for an in-control process ($\mathrm{ARL}_0 = 370.4$). Also, for the sake of brevity we are only considering, at this point, the perfect ranking scenario.

For each one of the eight distributions obtained by combining the two distributions and four specific parameter settings, we have simulated 10^7 samples of sizes $k = 3, 4,$ and 5. Five different sampling designs are considered: NRSS, RSS, MRSS, ERSS and SRS. 3-sigma control limits were properly calculated as described in (3), for the NRSS control charts, and based on the expressions presented in Muttlak and Al-Sabah (2003), for the others. Based on the simulated results, we calculated the corresponding values for ARL, MRL and SDRL, as we can verify in Table 3.

Note in Table 3 that, although all considered sample designs have their respective ARL$'s$ affected by the distribution skewness, NRSS and MRSS provided, in general, the closest values to the nominal $\mathrm{ARL}_0 = 370.4$ for the skew normal distribution. This indicates that these sampling designs are more conservative than their competitors. Table 3 points higher influence in ARL and MRL for the generalized normal if compared with the skew normal distribution in the considered simulated scenarios. This is particularly evident for $\beta = 1$ (Laplace distribution). However, we can also see that the NRSS control charts still dominates all its competitors, producing, in general, ARL_0 values closer to 370.4. Our results are in agreement with those found by Koyuncu and Karagöz (2018), who verified that NRSS control charts present lower type I error when applied to two asymmetric distributions: Type I Marshall-Olkin bivariate Weibull and bivariate lognormal.

Table 3. ARL, MRL and SDRL for control charts constructed by SRS and designs based on RSS under perfect ranking and Non-normal data

			SRS			RSS			MRSS			ERSS			NRSS		
Distribution	k	α	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL	ARL	MRL	SDRL
SN	3	1	351.19	244	350.69	325.09	225	324.59	351.62	244	351.12	325.09	225	324.59	361.73	251	361.23
		2	281.61	195	281.11	285.26	198	284.76	346.92	241	346.42	285.26	198	284.76	348.44	242	347.94
		3	237.08	164	236.58	254.53	177	254.03	352.77	245	352.27	254.53	177	254.03	336.65	234	336.15
		5	203.99	142	203.49	232.87	162	232.37	377.92	262	377.42	232.87	162	232.37	330.02	229	329.52
	4	1	352.94	245	352.44	335.09	232	334.59	356.46	247	355.96	321.09	223	320.59	366.05	254	365.55
		2	298.32	207	297.82	300.11	208	299.61	352.44	244	351.94	237.39	165	236.89	364.65	253	364.15
		3	259.91	180	259.41	277.96	193	277.46	380.36	264	379.86	186.96	130	186.46	378.98	263	378.48
		5	227.08	158	226.58	258.42	179	257.92	423.91	294	423.41	149.55	104	149.05	391.14	271	390.64
	5	1	357.67	248	357.17	339.87	236	339.37	357.17	248	356.67	320.18	222	319.68	368.16	255	367.66
		2	310.86	216	310.36	306.96	213	306.46	340.49	236	339.99	241.84	168	241.34	364.61	253	364.11
		3	274.15	190	273.65	292.93	203	292.43	325.44	226	324.94	191.77	133	191.27	371.25	257	370.75
		5	246.36	171	245.86	274.54	190	274.04	314.66	218	314.16	152.81	106	152.31	383.72	266	383.22
GN	3	1	126.21	88	125.71	118.37	82	117.87	150.16	104	149.66	118.37	82	117.87	189.11	131	188.61
		1.5	237.37	165	236.87	232.05	161	231.55	254.48	177	253.98	232.05	161	231.55	297.96	207	297.46
		3	730.63	507	730.13	466.32	323	465.82	558.22	387	557.72	466.32	323	465.82	429.00	298	428.50
		4	1207.28	837	1206.78	520.61	361	520.11	743.17	515	742.67	520.61	361	520.11	457.01	317	456.51
	4	1	147.39	102	146.89	133.05	92	132.55	176.31	122	175.81	143.60	100	143.10	247.53	172	247.03
		1.5	257.63	179	257.13	251.53	175	251.03	280.31	194	279.81	251.46	174	250.96	328.49	228	327.99
		3	582.92	404	582.42	419.56	291	419.06	465.33	323	464.83	385.33	267	384.83	392.97	273	392.47
		4	798.23	553	797.73	437.79	304	437.29	527.04	365	526.54	371.91	258	371.41	396.34	275	395.84
	5	1	163.59	114	163.09	146.81	102	146.31	211.19	147	210.69	155.41	108	154.91	275.21	191	274.71
		1.5	273.91	190	273.41	266.41	185	265.91	298.77	207	298.27	262.38	182	261.88	345.10	239	344.60
		3	521.44	362	520.94	399.11	277	398.61	441.48	306	440.98	375.56	260	375.06	385.42	267	384.92
		4	635.37	441	634.87	413.40	287	412.90	489.72	340	489.22	377.68	262	377.18	393.67	273	393.17

5. An application to real data

In order to illustrate the application of the NRSS control charts, we used a data set with 1030 observations about the concrete strength to compression (MPa) and the amount of cement (kg) used in the production of concrete blocks (Yeh, 1998). This data set is available in the R package `AppliedPredictiveModeling` (Kuhn and Johnson, 2018). Although this data was not recorded as a case of a quality control process, it serves us, under some assumptions, as a reference population, from which samples were drawn and control charts were constructed. We assumed the concrete strength as the variable of interest and the amount of cement as an auxiliary variable, such that the sample units may be ordered with errors, producing an imperfect ranking scenario. Also, we consider an additional scenario based on perfect ranking. In this case, the sample units were ordered directly from the concrete strength values, and the ranking process did not present any error. Moreover, we assumed the concrete blocks strength distribution in this sample as the natural variability of an industrial process. A square root transformation of the concrete strength was used in order to obtain a better approximation to normal distribution.

In this application, we consider three sampling designs: SRS, RSS and NRSS; two sample sizes: $k = 3$ and $k = 5$, and processes in two different scenarios: in-control ($\delta = 0$) and out-of-control, considering $\delta = 1.2$, as described in (7). Under each sampling design and for each sample size, we selected, with replacement, 25 samples from the original data. These samples are considered for estimating the control limits with A = 3, which corresponds to a probability of a type I error of $\alpha = 0.0027$ (phase 1). Afterwards, 75 new samples were selected for monitoring the process mean (phase 2). For $\delta = 0$, these 75 samples were selected with replacement from the original data; for $\delta = 1.2$, we added to the transformed strength values a normal random variable with mean $1.2\sigma_0/\sqrt{k}$ and standard deviation equals to 0.17 (corresponding to 11.74% of the standard deviation of the transformed concrete strength). This standard deviation value is small enough to characterize the lack of control, predominantly, due to the shift in the process mean, instead of its dispersion (variance).

Figure 3 presents (on the left) the histogram for the distribution of concrete strength, with the estimated normal distribution and kernel density curves. The dispersion plot, on the right, indicates moderate positive linear relationship between the variables. The linear correlation coefficient is $\rho = 0.49$, which points to a moderately favourable scenario for RSS based designs.

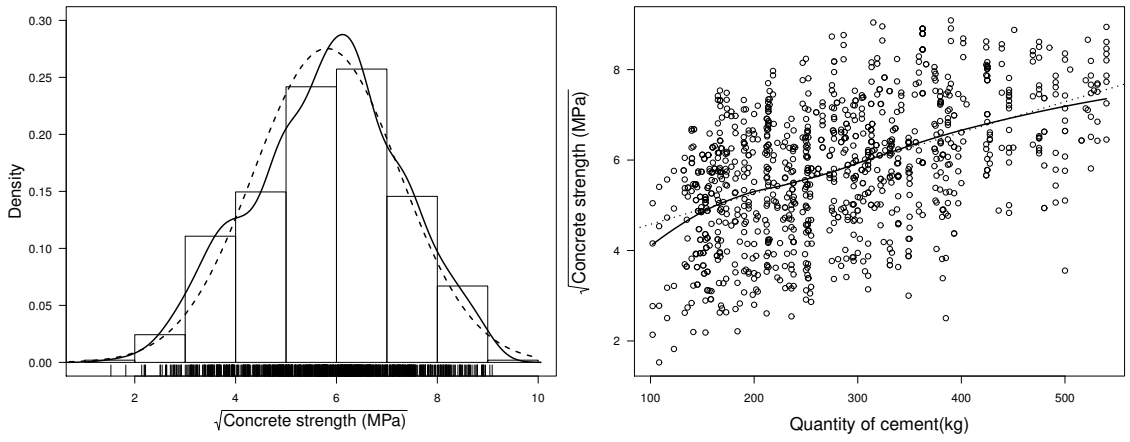

Figure 3. Histogram and scatter plot for concrete strength.

Following, Figures 4 and 5 present the SRS, RSS and NRSS control charts for the process mean considering $k = 3$. The RSS and NRSS control charts were obtained under perfect and imperfect ranking, as previously described. In Figure 4 we have the charts when $\delta = 0$ (in-control process). In all cases, it is possible to notice points randomly distributed around the central line, without any point outside the control limits. This behaviour characterizes an in-control process, as expected. In addition, Figure 5 presents the control charts for $\delta = 1.2$ (out-of-control process). It is possible to observe that the NRSS control chart showed the highest number of points exceeding the control limits (11 points under perfect ranking and 7 under imperfect ranking), followed by RSS (with 9 and 5 points exceeding the control limits, respectively) and SRS control charts (only 2 points outside the limits).

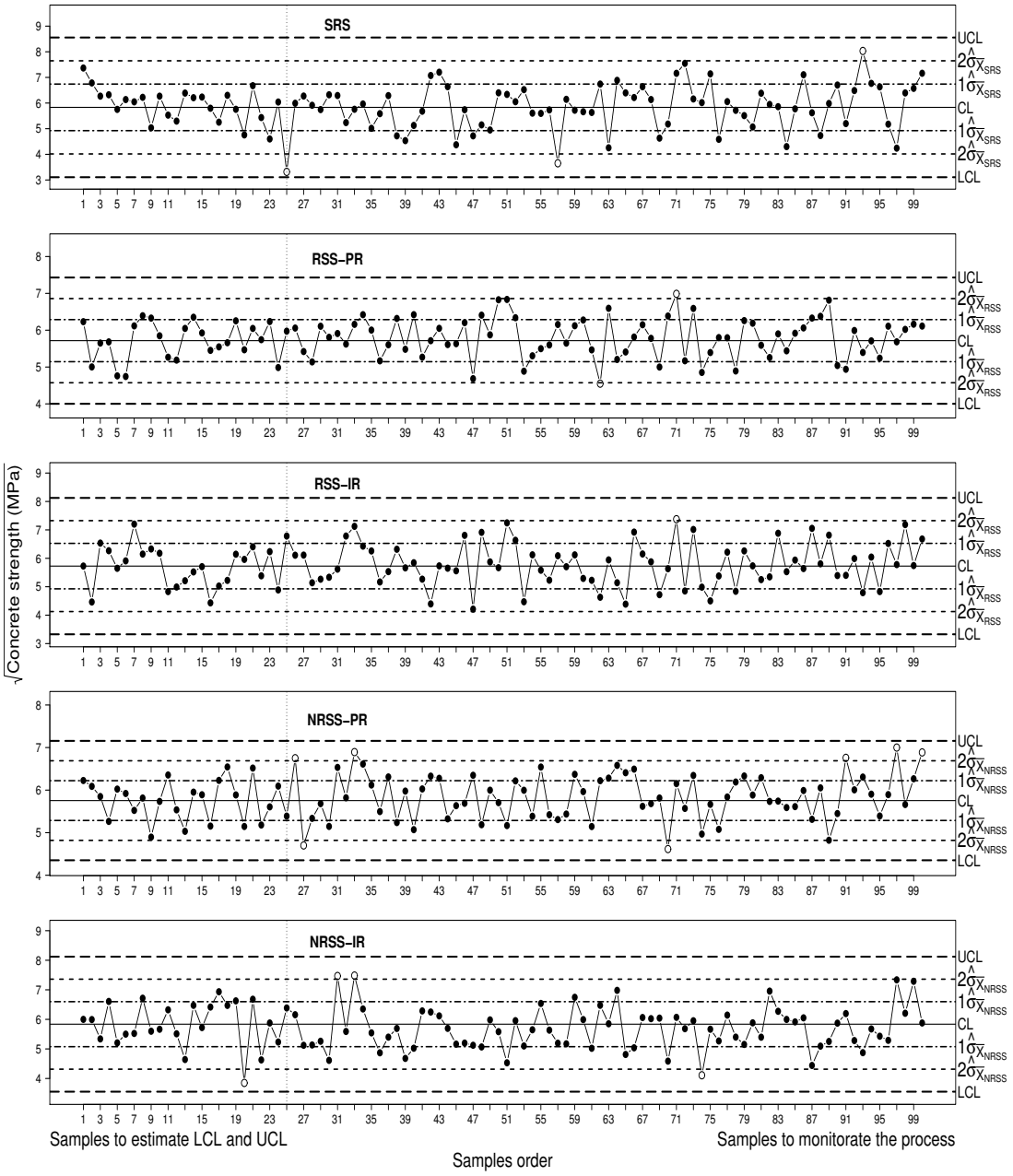

Figure 4. Control charts for concrete strength considering $k = 3$ and an in-control process ($\delta = 0$). Perfect ranking is denoted as PR, and imperfect ranking as IR.

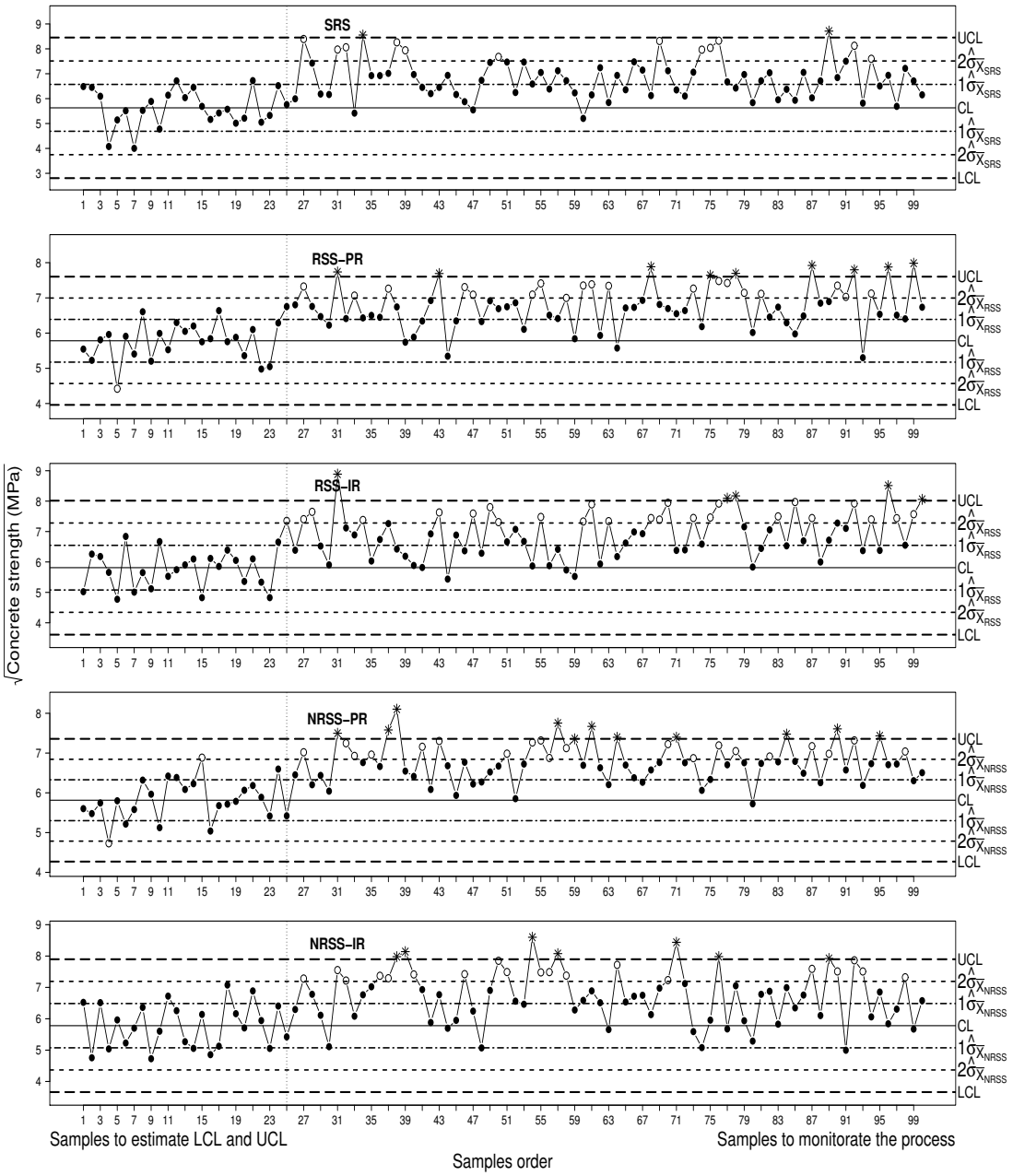

Figure 5. Control charts for concrete strength considering $k = 3$ and an out-of-control process ($\delta = 1.2$). Perfect ranking is denoted as PR, and imperfect ranking as IR.

Figures 6 and 7 present the control charts for $k = 5$, under the same three sampling designs (and five scenarios, when considering perfect and imperfect ranking), simulated, respectively, with $\delta = 0$ and $\delta = 1.2$. It is possible to notice again that NRSS control charts present satisfactory performance, showing randomness and without any point outside the control limits for an in-control process, and also presenting a large number of points exceeding the control limits in the out-of-control scenario (28 under perfect and 9 under imperfect ranking) than RSS (14 and 6 points, respectively) and SRS (with only 3 points outside the limits).

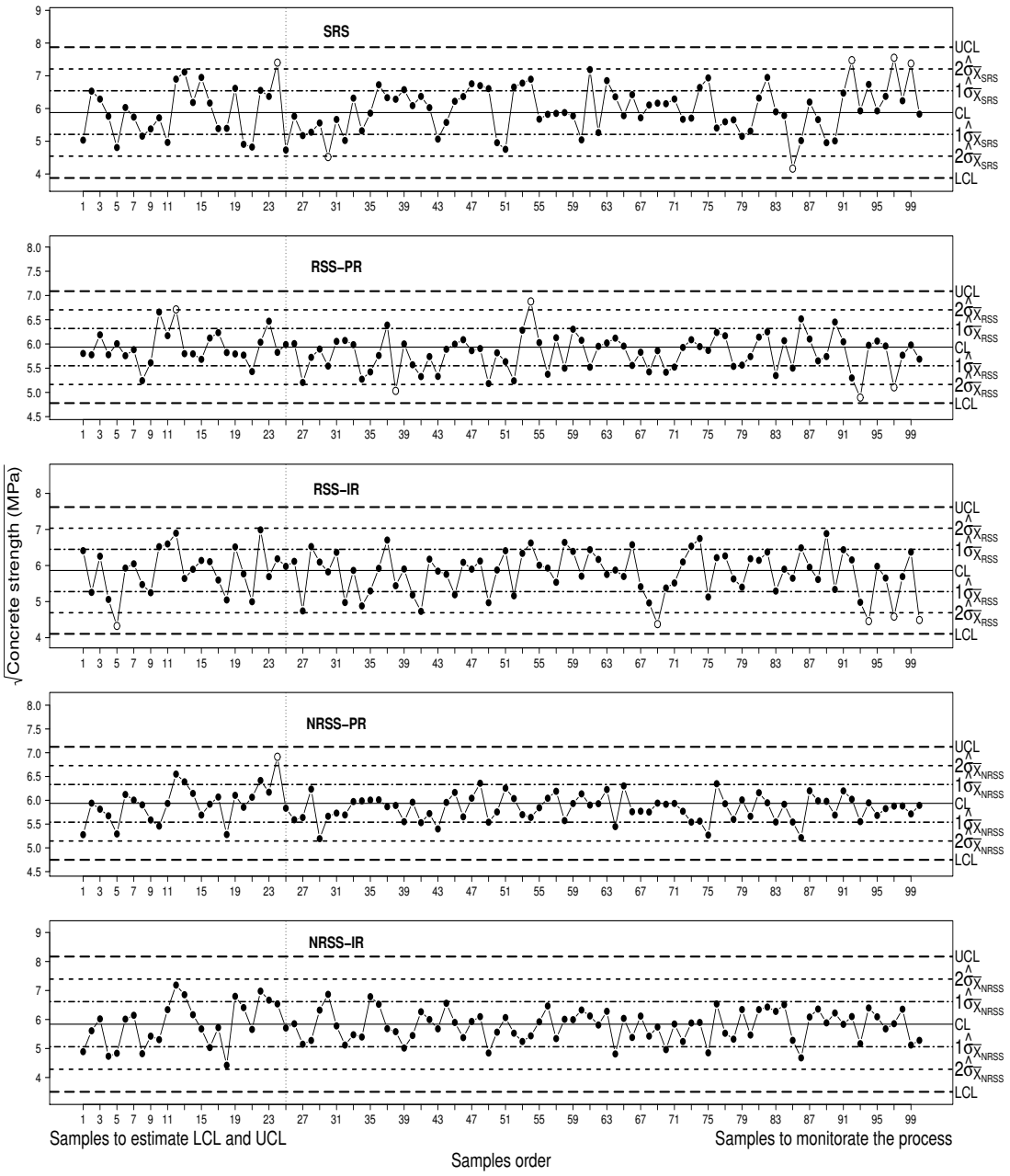

Figure 6. Control charts for concrete strength considering $k = 5$ and an in-control process ($\delta = 0$). Perfect ranking is denoted as PR, and imperfect ranking as IR.

6. Concluding remarks

In this paper, we considered control charts for the mean of a normal distributed process based on NRSS design. These charts were compared to their SRS and RSS based counterparts by means of a simulation study. Under perfect ranking, NRSS control charts overcome all their competitors, providing smaller ARL values for out-of-control process in all simulated scenarios. In addition, the NRSS control charts showed to be competitive when compared to those based on DRSS designs. However, such sampling designs require the initial selection of k^3 sample units for, after two ordering cycles, selecting a final sample of k units. For example, the ARL for NRSS control charts were smaller in all simulated

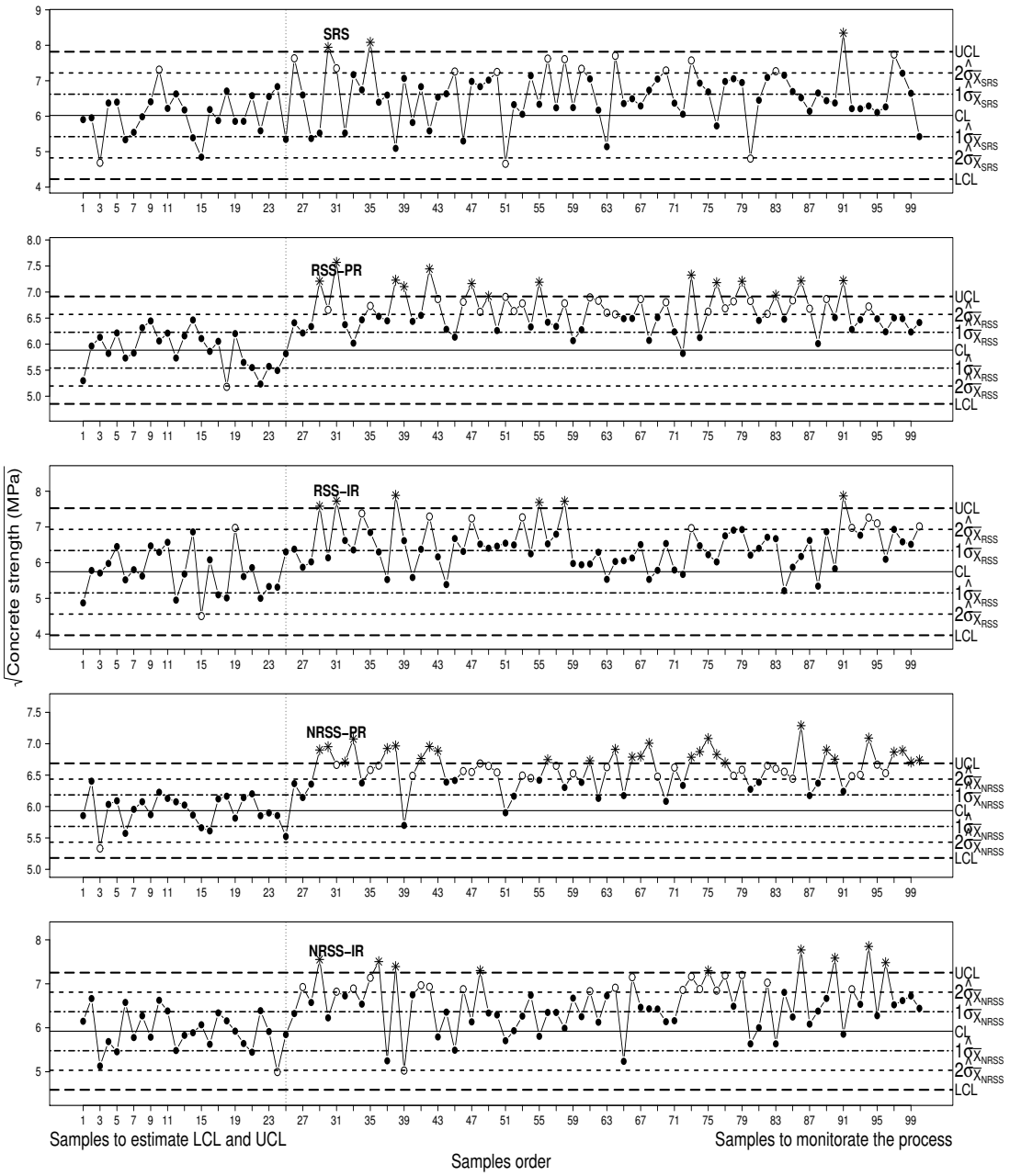

Figure 7. Control charts for concrete strength considering $k = 5$ and an out-of-control process ($\delta = 1.2$). Perfect ranking is denoted as PR, and imperfect ranking as IR.

scenarios when compared to those provided by extreme double ranked set sampling and double extreme ranked set sampling, and surpassed by those provided by double quartile ranked set sampling and quartile double ranked set sampling when $k = 5$ (Abujiya and Muttlak, 2004; Al-Omari and Haq, 2012). Moreover, this superiority is also verified against DRSS control charts for all considered sample sizes. In addition, when considering the double median ranked set sampling and median double ranked set sampling control charts, as can be seen in Abujiya and Muttlak (2004), these designs dominate NRSS, providing lower ARL values. However, it should be considered that double ranked set designs could be expensive, and sometimes infeasible, due to a high operational effort.

Under imperfect ranking, we have shown that the efficiency of NRSS control charts becomes smaller as the correlation between the variables decreases. This is a common fact to other designs based on RSS. Even so, the simulated ARL values for NRSS control charts are predominantly smaller (for out-of-control processes) than the corresponding ones reached by SRS. Additionally, it was possible to verify the superiority of the NRSS control charts with the ones provided by RSS, MRSS and ERSS in most of the simulated scenarios. Also, NRSS was the most robust method for non-normally distributed processes.

In an illustration with real data regarding concrete strength, the SRS, RSS and NRSS control charts presented points randomly distributed around the central line, without any points outside the control limits, when we simulated from a process under statistical control. However, for the out-of-control scenarios, the NRSS control charts performed better when compared to the RSS and the usual control charts based on SRS.

Therefore, based on these results, we recommend NRSS control charts for monitoring the process mean as an efficient alternative to SRS and to other RSS based designs. Under the operational point of view, the ranking of k^2 samples units in a single set (instead of ranking k sets of k units, as it occurs in RSS, MRSS and ERSS designs) may, eventually, become a complicating issue, if the ordering criterion is based, for example, on a visual judgment. However, this will usually not make great difference if the ordering criterion is based, for example, on an auxiliary variable. Finally, the impact of ties in the ranking process should be investigated; see Frey (2012) and Zamanzade and Wang (2018) for some alternatives to overcome the problem of ties in RSS.

Acknowledgement

The authors thank the Editor, an Associate Editor and the reviewers for their valuable comments on an earlier version of this manuscript.

References

Abid, M., Nazir, H.Z., Riaz, M., and Lin, Z., 2017. Investigating the impact of ranked set sampling in nonparametric cusum control charts. Quality and Reliability Engineering International, 33, 203-214.

Abujiya, M. and Muttlak, H., 2004. Quality control chart for the mean using double ranked set sampling. Journal of Applied Statistics, 31, 185-1201.

Al-Omari, A.I. and Bouza, C.N., 2014. Review of ranked set sampling: modifications and applications. Revista Investigación Operacional, 3, 215-240.

Al-Omari, A.I. and Haq, A., 2012. Improved quality control charts for monitoring the process mean, using double-ranked set sampling methods. Journal of Applied Statistics, 39, 745-763.

Al-Saleh, M.F. and AlKadiri, M.A., 2000. Double-ranked set sampling. Statistics and Probability Letters, 48, 205-212.

Azzalini, A., 1985. A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12, 171-178.

Azzalini, A., 2019. The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-t. R package version 1.5-4.

Balakrishnan, N. and Rao, C.R., 1998. Order Statistics: Theory and Methods. Elsevier, Amsterdam.

Chakraborti, S., Human, S., and Graham, M., 2008. Phase i statistical process control charts: an overview and some results. Quality Engineering, 21, 52-62.

Chen, Z., 2007. Ranked set sampling: its essence and some new applications. Environmental and Ecological Statistics, 14, 355-363.

Chen, Z., Bai, Z., and Sinha, B., 2003. Ranked Set Sampling: Theory and Applications. Springer, New York.

Consulin, C.M., Ferreira, D., Rodrigues de Lara, I.A., De Lorenzo, A., di Renzo, L., and Taconeli, C.A., 2018. Performance of coefficient of variation estimators in ranked set sampling. Journal of Statistical Computation and Simulation, 88, 221-234.

Costa, A.F. and De Magalhaes, M.S., 2007. An adaptive chart for monitoring the process mean and variance. Quality and Reliability Engineering International, 23, 821-831.

Frey, J., 2012. Nonparametric mean estimation using partially ordered sets. Environmental and Ecological Statistics, 19, 309-326.

Haq, A., Brown, J., Moltchanova, E., and Al-Omari, A.I., 2015. Effect of measurement error on exponentially weighted moving average control charts under ranked set sampling schemes. Journal of Statistical Computation and Simulation, 85, 1224-1246.

Koutras, M., Bersimis, S., and Maravelakis, P., 2007. Statistical process control using Shewhart control charts with supplementary runs rules. Methodology and Computing in Applied Probability, 9, 207-224.

Koyuncu, N., 2018. Regression estimators in ranked set, median ranked set and neoteri ranked set sampling. Pakistan Journal of Statistics and Operation Research, 14, 89-94.

Koyuncu, N. and Karagöz, D., 2018. New mean charts for bivariate asymmetric distributions using different ranked set sampling designs. Quality Technology and Quantitative Management, 15, 602-621.

Kuhn, M. and Johnson, K., 2018. AppliedPredictiveModeling: Functions and Data Sets for 'Applied Predictive Modeling'. R package version 1.1-7.

Mahdizadeh, M. and Zamanzade, E., 2019. Efficient body fat estimation using multistage pair ranked set sampling. Statistical Methods in Medical Research, 28, 223-234.

McIntyre, G., 1952. A method for unbiased selective sampling, using ranked sets. Australian Journal of Agricultural Research, 3, 385-390.

Mehmood, R., Riaz, M., and Does, R.J., 2013. Control charts for location based on different sampling schemes. Journal of Applied Statistics, 40, 483-494.

Mineo, A.M., 2018. normalp: Routines for Exponential Power Distribution. R package version 0.7.0.1.

Montgomery, D.C., 2009. Statistical Quality Control Wiley, New York.

Muttlak, H., 1997. Median ranked set sampling. Journal of Applied Statistical Sciences, 6, 245-255.

Muttlak, H. and Al-Sabah, W., 2003. Statistical quality control based on ranked set sampling. Journal of Applied Statistics, 30, 1055-1078.

Nadarajah, S., 2005. A generalized normal distribution. Journal of Applied Statistics, 32, 685-694.

Nawaz, T. and Han, D., 2019. Monitoring the process location by using new ranked set sampling-based memory control charts. Quality Technology and Quantitative Management, pages in press. Available online at https://doi.org/10.1080/16843703.2019.1572288.

Qiu, P., 2018. Some perspectives on nonparametric statistical process control. Journal of Quality Technology, 50, 49-65.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.

Samawi, H.M., Ahmed, M.S., and Abu-Dayyeh, W., 1996. Estimating the population mean using extreme ranked set sampling. Biometrical Journal, 38, 577-586.

Santore, F., Taconeli, C.A., and Rodrigues de Lara, I.A., 2019. An adaptive control chart for the process location based on ranked set sampling. Communications in Statistics: Simulation and Computation, pages in press. Available online at https://doi.org/10.1080/03610918.2019.1622722.

Shewhart, W.A., 1924. Some applications of statistical methods to the analysis of physical and engineering data. Bell Labs Technical Journal, 3, 43-87.

Taconeli, C.A. and Cabral, A.D.S., 2019. New two-stage sampling designs based on neoteric ranked set sampling. Journal of Statistical Computation and Simulation, 89, 232-248.

Venables, W.N. and Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.

Yeh, I.C., 1998. Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research, 28, 1797-1808.

Zamanzade, E. and Al-Omari, A.I., 2016. New ranked set sampling for estimating the population mean and variance. Hacettepe Journal of Mathematics and Statistics, 45, 1891-1905.

Zamanzade, E. and Mahdizadeh, M., 2019. Using ranked set sampling with extreme ranks in estimating the population proportion. Statistical Methods in Medical Research, pages in press. Available online at https://doi.org/10.1177/0962280218823793.

Zamanzade, E. and Wang, X., 2018. Proportion estimation in ranked set sampling in the presence of tie information. Computational Statistics, 33, 1349-1366.

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 155–176

Times Series Models
Research Paper

GARCH-in-mean models with asymmetric variance processes for bivariate European option evaluation

Lucas Pereira Lopes[1,*], Vicente Garibay Cancho[1], and Francisco Louzada[1]

[1]Department of Applied Mathematics and Statistics, University of São Paulo, São Carlos, Brazil

(Received: 07 April 2019 · Accepted in final form: 23 June 2019)

Abstract

Options pricing models, which consider asset-objects following a geometric Brownian motion, such as derivations from the traditional Black-Scholes model, assume the volatility of asset-objects to be constant over time. In addition, the normal distribution is the basement of the joint distribution for the case of bivariate options. In this work, we consider GARCH-in-mean models with asymmetric variance specifications to model the volatility of the assets-objects under the risk-neutral dynamics. Moreover, the copula functions model the joint distribution, with the objective of capturing non-linear, linear and tails associations between the assets. We provide a methodology to describe a more realistic pricing option. To illustrate the methodology, we use stocks from two Brazilian companies. Confronting the results obtained with the classic model, which is an extension of the Black-Scholes model, we note that considering constant volatility over time underpricing the options, especially in-the-money options. Overall, the contributions of the proposed methodology are as follows. Using the best copula makes the model more suitable. Extension to marginal models, which consider asymmetry, makes joint modeling more flexible and realistic. Due to the adequate marginal and joint fitting, in addition to the values obtained with the classical consolidated model, there are arguments to believe that the differences obtained between the best models, through the copulas and the extension of the conventional method, are improvements in the calculation of the fair value. The empirical relevance of such alternatives is apparent given the evidence of non-joint-normality in financial emerging markets. In essence, the entire approach may be generalized to any number of time-series of option pricing.

Keywords: Black-Scholes model · Copulas · GARCH models · Pricing.

Mathematics Subject Classification: Primary 62J99 · Secondary 62M20.

1. Introduction

Multivariate options are excellent tools to manage a portfolio's risk. The first works that had as objective the pricing of options in the univariate case were Black and Scholes (1973) and Merton (1973). Through these works, other authors have used the same theory, that is, asset-objects follow a Brownian geometric motion and have proposed bi and multivariate models, such as Stulz (1982), Margrabe (1978), Johnson (1987), Nelsen (2006) and

*Corresponding author. Email: luccas.lopes@usp.br

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

http://www.soche.cl/chjs

Shimko (1994). However, models derived from Brownian geometric motion methods have the assumptions that the volatilities of the assets are constant over time.

To carry out the pricing with more realistic assumptions, researchers have developed other models. For instance, we use the generalized autoregressive conditional heteroskedasticity (GARCH) family of models, because of its ability to incorporate the stylized facts about asset return dynamics. This kind of modeling is popular in economics and finance (Almeida e Hotta, 2014). Furthermore, with Black-Scholes (BS) model assumptions, any contingent claim can be perfectly replicated by its underlying asset and a riskless bond, so the price of a contingent claim is merely the cost of the replicating portfolio. However, using GARCH-type models, it is generally not possible to construct a perfect replicating portfolio, as the volatility of asset returns is permitted to vary over time. It is necessary to define a risk-neutral measure to use the GARCH-type models to consider a general market equilibrium (Liu, Li and Ng, 2015).

The model of Duan (1995) derived a measure of risk-neutral through the standard GARCH model, which the author showed the potential of it concerning the Black-Scholes approach. However, one of the main limitations of the standard GARCH model is the inability to incorporate the effect of asymmetry caused by unplanned returns (Nelsen, 1991). Introduced by Black (1976), this effect implies that volatility tends to grow more when there is an unanticipated drop in returns (that is, bad news) than when there is an unanticipated increase of the same magnitude in returns (that is, good news). This effect, also known as a leverage effect, has been included in the GARCH-type models, such as the exponential GARCH (EGARCH), the non-linear asymmetric GARCH (NGARCH) and the Glosten, Jagannathan, and Runkle GARCH (GJR-GARCH) models. It can be used to price options by deriving their risk-neutral measure.

Furthermore, to understand the price behavior of a multivariate option, it is necessary to use tools that accommodate the co-movements between its underlying processes. A primary tool that is widely used by the methods derived from the traditional Black-Scholes model is the multivariate normal distribution modeling. However, the use of such an approach implies in linear associations as a measure of dependence between the assets. However, empirical evidence presents that a real association between financial series is much more complex (Lopes and Pessanha, 2018).

Therefore, this paper aims to price bivariate options by overcoming two of the above constraints of the classical approach, where asset-objects are modeled marginally by deriving their risk-neutral considering the GARCH, EGARCH, NGARCH and GJR-GARCH models, with copula functions modeling the joint distribution models, with the objective of capturing linear, non-linear and tails dependence. The entire methodology described here may be extended to any multivariate case.

An innovative feature of the present work is the comparison among methodologies, where we consider marginal processes that capture the effect of asymmetry, usually present in financial series. A second point is the performance of a simulation study of the pricing models with the purpose of verifying the good fit of the models used in the literature. It is highlighted as a third point the comparison of the methodology exposed to the standard method, extended from the Black-Scholes model to the bivariate case. The implementation of such methods in the Brazilian stock market, which is characterized as a volatile and unstable market concerning developed markets. Then, compared with the previous papers, the approach in the present paper makes the dynamic pricing more reasonable and tractable. The paper organization follows. Section 2 presents the conceptual framework and the models. In Section 3, we provide the bivariate model methodology and the inference procedures. In Section 4, the results of the proposed method under an artificial and real data sets are illustrated. Finally, Section 5 ends the paper with concluding remarks. Some technical details about different copulas are presented in the appendix.

2. Conceptual Framework and Model Specification

In this section, we present option pricing, the GARCH-in-mean specification and risk-neutral with GARCH-in-mean process.

2.1 Option pricing

A European option call on the maximum of two risky assets (call-on-max) is defined based on the maximum price between two assets. The payoff function of this option is given by

$$g(S(T)) = \max[\max(S_1(T), S_2(T)) - K, 0],$$

where S_i is the price of the ith asset, for $i = 1, 2$, at the maturity date T and K is the strike price or exercise price.

To introduce heteroscedasticity, we use the fundamental theorem of asset pricing (Delbaen and Schachermayer, 1994). This theorem states that once the stock prices $S_1(T)$ and $S_2(T)$ are free from arbitrage and present in a complete market (Hull, 1992), there is a measure of probability $\mathbb{Q}$ such that the discounted price of the payoff function, $\mathrm{e}^{-r(T-t)}g(S_1(T), S_2(T))$, is a martingale under $\mathbb{Q}$ and $\mathbb{Q}$ is equivalent to the real world probability measure $\mathbb{P}$. Therefore, we define the following definition to perform the pricing.

Definition 1. Let S_1 and S_2 be two stocks traded in a complete and free arbitrary market. In addition, be t the present date, T the maturity date and r the fixed risk-free rate yield. Then, the option price considering the payoff function $g(S_1, S_2) = \max[\max(S_1(T), S_2(T)) - K, 0]$ is given by

$$\begin{aligned} v(t, S_1, S_2) &= \mathrm{e}^{-r(T-t)}\mathrm{E}^{\mathbb{Q}}[\max[\max(S_1(T), S_2(T)) - K, 0]|F_t] \\ &= \mathrm{e}^{-r(T-t)} \int_0^\infty \int_0^\infty \max[\max(S_1(T), S_2(T)) - K, 0] f^{\mathbb{Q}}_{S_1,S_2}(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2, \end{aligned}$$

where $f^{\mathbb{Q}}_{S_1,S_2}$ is the joint density function of two measures under neutral risk probability $\mathbb{Q}$, which in this work is modeled by copula functions, and F_t is a filtering containing all information about the assets up to time t.

Now, we express the joint density function using the marginal densities $f_{S_1}(x_1)$ and $f_{S_2}(x_2)$ by means of copula functions expressed as

$$f^{\mathbb{Q}}_{S_1,S_2} = c^{\mathbb{Q}}(F^{\mathbb{Q}}_{S_1}, F^{\mathbb{Q}}_{S_2}) f^{\mathbb{Q}}_{S_1}(x_1) f^{\mathbb{Q}}_{S_2}(x_2),$$

where $c^{\mathbb{Q}} = \partial^2 C^{\mathbb{Q}}(x_1, x_2)/\partial x_1 \partial x_2$ and $C^{\mathbb{Q}}$ is a copula function.

Copulas are useful tools for constructing joint distributions (Sharifonnasabi, Alamatsaz and Kazemi, 2018). That is, copula is a multidimensional distribution function in which the marginal distributions are uniform in $[0, 1]$. A bivariate copula is a function $C : I^2 \longrightarrow I \in [0, 1]$ that satisfies the following conditions: $C(x_1, 0) = C(0, x_1) = 0 \quad and \quad C(x_1, 1) = C(1, x_1) = x_1, \quad x_1 \in I$ and the 2-increasing condition $C(u_2, v_2) - C(u_1, v_2) - C(u_2, v_1) + C(u_1, v_1) \geq 0$, for all u_1, u_2, v_1 and $v_2 \in [0, 1]$ such $u_1 \leq u_2$ and $v_1 \leq v_2$.

One of the most famous theorems in copula theory is the Sklar theorem. According to Sklar's theorem (Sklar, 1959), any bivariate cumulative distribution H_{S_1,S_2} can be represented as a function of the marginal distributions F_{S_1} and F_{S_2}. In addition, whether the

marginal distributions are continuous, the copula exists, is unique and given by

$$H_{S_1,S_2}(x_1, x_2) = C(F_{S_1}(x_1), F_{S_2}(x_2)),$$

where $C(u, v) = P(U \leq u, V \leq v)$, $U = F_{S_1}(x_1)$ and $V = F_{S_2}(x_2)$.

In the case of continuous and differentiable marginal distributions, the joint density function of the copula is expressed as

$$f(x_1, x_2) = f_{S_1}(x_1) f_{S_2}(x_2) c(F_{S_1}(x_1), F_{S_2}(x_2)),$$

where $f_{S_1}(x_1)$ and $f_{S_2}(x_2)$ are the densities for the distribution function $F_{S_1}(x_1)$ and $F_{S_2}(x_2)$, respectively, and

$$c(u, v) = \frac{\partial^2 C(u, v)}{\partial u v}$$

is the density of copula. For further details about copulas, see Nelsen (2006) and Sanfins and Valle (2012). In this work, we consider the normal, Student-t, Gumbel, Frank and Joe copulas. An appendix at the end of this paper provides details about these copula functions. Therefore, to construct a joint process of risk-neutral for the bivariate distribution of the option, the marginal processes are derived first.

2.2 GARCH-in-mean specification under $\mathbb{P}$

Instead of deriving the bivariate risk-neutral distribution directly, each marginal process is proposed to transform separately. Duan (1995) defined an option pricing model considering that the variance of the asset-object is not constant over time. To implement non-constant volatility over the maturity time of the option, we use in this work the generalized autoregressive conditional heteroskedastic (GARCH) models. Bollerslev (1986) introduced the GARCH model by modifying the ARCH model presented by Engle (1982). The use of GARCH models in pricing leads to the correction of some biases in the model of Black and Scholes (1973), including return skewness and leptokurtic behavior.

GARCH-in-mean refers to the inclusion of an extra term m_t in the conditional mean of the model introduced by Bollerslev (1986). An intuitive idea to use these models in derivative pricing is that conditional variance is not constant over time and hence the conditional mean of market returns is a linear function of conditional variance. Another definite reason to work with the GARCH-in-mean models is that these models explain the presence of conditional left skewness observed in stock returns.

Consider a discrete time economy with a risk-free asset. We define a complete filtered probability space $(\Omega, \mathbb{F}, \mathbb{F}_t, \mathbb{P})$ to model uncertainty, where $\mathbb{P}$ is the historical (physical) measure and $\mathbb{F} = \mathbb{F}_t$, for $t = 0, 1, \ldots, T$, is a filtration, or a family of increasing σ-field information sets, representing the resolution of uncertainty based on the information generated by the market prices up to and including time t. We assume the general GARCH-M(p, q) model for the return $y_t = \log(S_t/S_{t-1})$ given by

$$y_t = m_t + \sqrt{h_t}\epsilon_t, \quad h_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i h_{t-i} \phi(\epsilon_{t-i}) + \sum_{i=1}^{q} \beta_i h_{t-i}, \tag{1}$$

where S_t is the stock price at time t and ϵ_t is a sequence of independent and identically distributed random variables with normal distribution; the conditional mean return m_t is assumed to be an F_t-predictable process. In many studies, m_t is assumed to be a function

of the conditional variance h_t of the return and a risk premium quantifier at time t; the function ϕ describes the impact of random shock of return ϵ_t on the conditional variance h_t and $\alpha_0 > 0, \alpha_i$ and $\beta_i \geq 0$.

The conditional mean and variance of y_t are $m_t = \mathrm{E}[y_t|F_{t-1}]$ and $h_t = \mathrm{Var}[y_t|F_{t-1}]$. The effect of past innovations ϵ_{t-1} under the conditional variance h_t have different impacts depending on the function $\phi(\epsilon_{t-1})$, and consequently we have different extensions of the GARCH model. For example, considering $p = q = 1$, when $\phi(\epsilon_{t-1}) = \epsilon_{t-1}^2$, the sign of ϵ_{t-1} there is no effect over h_t, and we have the traditional GARCH proposed by Bollerslev (1986). Thus, the innovations have a symmetric effect on the conditional variance, expressed by

$$h_t = \alpha_0 + \alpha_1 h_{t-1}\epsilon_{t-1}^2 + \beta_1 h_{t-1}. \tag{2}$$

Following Liu, Li and Ng (2015), Duan (1995) and Chiou and Tsay (2008), $m_t = r + \lambda\sqrt{h_t} - k_{\epsilon_t}(\sqrt{h_t})$, where $k_{\epsilon_t}(\sqrt{h_t})$ is the cumulate generating function of the innovation ϵ_t e λ is the premium risk parameter. When ϵ_t follows a normal distribution, we have $k_{\epsilon_t}(\sqrt{h_t}) = h_t/2$. Because standard GARCH models given by equation (1) respond in the same way to positive and adverse events, such models cannot correctly capture the leverage effect. Other forms of the GARCH model, such as EGARCH, NGARCH, and GJR-GARCH, include the asymmetry effect, can thus be used in option pricing and are used in the present work. Nelsen (1991) proposed the exponential GARCH (EGARCH) model. The author assumes that the dynamic of the logarithm of the conditional variance of EGARCH(1,1) is expressed as

$$\log(h_t) = \alpha_0 + \alpha_1(|\epsilon_{t-1}| + \gamma_1\epsilon_{t-1}) + \beta_1 \log(h_{t-1}), \tag{3}$$

where α_0, α_1, β_1 and γ_1 are constant parameters and ϵ forms a sequence of independent standard normal random variables representing random shocks. The EGARCH model does not require such parameter restrictions since the conditional variance is expressed as the exponential of a function. Including the random shock term in absolute value and with a parameter γ_1, the author made volatility a function of both magnitude and sign of the shock.

Engle (1982) introduced the non-linear asymmetric GARCH (NGARCH), which takes into account the leverage effect. In their model, the dynamic of the conditional variance of NGARCH(1,1) is given by

$$h_t = \alpha_0 + \alpha_1 h_{t-1}(\epsilon_{t-1} - \gamma_1)^2 + \beta_1 h_{t-1}, \tag{4}$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$ and γ_1 is a non-negative parameter that captures the negative correlation between return and volatility innovations. Since the parameter α_1 is typically non-negative, a positive γ_1 means that negative random shocks increase volatility more than positive random shockes of similar magnitude. Hence, the NGARCH allows for the levarage through its paramater γ_1.

Another model that takes into account the asymmetry effect of news on volatility is the GJR-GARCH introduced by Glosten, Jagannathan and Runkle (1993). According to this model, the conditional variance dynamic of GJR-GARCH(1,1) is defined as

$$h_t = \alpha_0 + \alpha_1 h_{t-1}\epsilon_{t-1}^2 + \beta_1 h_{t-1} + \gamma_1 h_{t-1} \max(0, -\epsilon_{t-1})^2, \tag{5}$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$ and $\gamma_1 \geq 0$ are constant parameters. This model allows for the leverage effect by adding the extra term $\gamma_1 h_{t-1} \max(0, -\epsilon_{t-1})^2$ when ϵ_t is negative since γ_1 is typically non-negative.

All the models presented above are in the physical measure ($\mathbb{P}$ measure). Now, we discuss their representations in the risk-neutral measure ($\mathbb{Q}$ measure), a prerequisite for pricing options under heteroscedasticity.

2.3 Risk-neutral with GARCH-in-mean process

The concept of risk-neutral valuation relationship (RNVR) has a fundamental role in the process of pricing options. This principle has as the base an asset, which is priced according to the discount of the expected value of a payoff function under a martingale measure, that is, that the economic agents are risk-neutral.

To apply this pricing methodology, we assume that a measure of martingale $\mathbb{Q}$ exists in a discrete economy time, with interest rate and a probability space $(\Omega, \mathbb{F}, \mathbb{F}_t, \mathbb{P})$, where $\mathbb{P}$ is a measure of physical probability and $\mathbb{F}_t$ is a filtering at time t.

Definition 2. A measure of probability $\mathbb{Q}$ is equivalent to a measure of probability $\mathbb{P}$ if:

(1) $\mathbb{Q} \approx \mathbb{P}$, that is, for all event X, $\mathbb{Q}(X) = 0$ and $\mathbb{P}(X) = 0$.

(2) The discounted price process S_t is a martingale under $\mathbb{Q}$, that is, $\mathrm{E}^{\mathbb{Q}}[S_t|F_{t-1}] = S_{t-1}$.

Proposition. Assuming continuously compounded returns, the martingale condition for the discounted stock price can be replaced by

$$\mathrm{E}^{\mathbb{Q}}[\mathrm{e}^{y_t}|F_{t-1}] = \mathrm{e}^{r}.$$

Proof. From second condition in Definition 2, we have

$$\mathrm{E}^{\mathbb{Q}}[S_t|F_{t-1}] = S_{t-1} \Leftrightarrow \mathrm{E}^{\mathbb{Q}}[\mathrm{e}^{-rT}S_t|F_{t-1}] = \mathrm{e}^{r(t-1)}S_{t-1} \Leftrightarrow \mathrm{E}^{\mathbb{Q}}\left[\frac{S_t}{S_{t-1}}|F_{t-1}\right] = \mathrm{e}^{r}$$
$$\Leftrightarrow \mathrm{E}^{\mathbb{Q}}[\mathrm{e}^{y_t|F_{T-1}}] = \mathrm{e}^{r}.$$

Brennan and Schwartz (1979) represented a starting point by providing conditions which ensure the existence of the risk-neutral measure. Duan (1995) proposes an extension of RNVR, referred to as locally risk-neutral valuation relationship (LRNVR) by assuming a conditional Gaussian distribution for the log-returns with unchanged volatility after the change of measure.

Definition 3. A no arbitrage measure $\mathbb{Q}$ equivalent to $\mathbb{P}$ is said to satisfy the local risk-neutral valuation relationship (LRNVR) if the following conditions are satisfied:

(1) $y_t|F_{t-1} \sim \mathrm{N}(m_t, h_t)$ under $\mathbb{P}$, where $\epsilon_t \sim \mathrm{N}(0, 1)$.

(2) $\mathrm{E}^{\mathbb{Q}}[S_t/S_{t-1}|F_{t-1}] = \mathrm{e}^{r}$.

(3) $\mathrm{Var}^{\mathbb{Q}}[\log(S_t/S_{t-1})|F_{t-1}] = \mathrm{Var}^{\mathbb{P}}[\log(S_t/S_{t-1})|F_{t-1}]$.

In the previous definition, the conditional variance under the two measures is required to be equal. This requirement is necessary to estimate the conditional variance under $\mathbb{P}$ and use the framework to obtain the option pricing under $\mathbb{Q}$. This property and the fact of the risk-free rate can replace the conditional mean, yield a well-specified model that does not locally depend on preferences. Duan (1995) proved this latter fact. Here we reduce all preference consideration to the unit risk premium λ. Since $\mathbb{Q}$ is absolutely continuous

for $\mathbb{P}$, the almost certain relationship under $\mathbb{P}$ also holds true under $\mathbb{Q}$. Duan (1995) and Duan et al. (2006) showed that under the risk-neutral measure $\mathbb{Q}$ given by LRNVR, the asset return dynamic becomes

$$y_t = r - \frac{1}{2}h_t + \sqrt{h_t}\tilde{\epsilon}_t, \quad \tilde{\epsilon}_t \sim \mathrm{N}(0,1).$$

In addition:
GARCH(1,1): $h_t = \alpha_0 + \alpha_1 h_{t-1}(\tilde{\epsilon}_{t-1} - \lambda_1)^2 + \beta_1 h_{t-1}$.
EGARCH(1,1): $h_t = \alpha_0 + \alpha_1[|\tilde{\epsilon}_{t-1} - \lambda_1| + \gamma_1(\tilde{\epsilon}_{t-1} - \lambda_1)] + \beta_1 \log(h_{t-1})$.
NGARCH(1,1): $h_t = \alpha_0 + \alpha_1 h_{t-1}(\tilde{\epsilon}_{t-1} - \gamma_1 - \lambda_1)^2 + \beta_1 h_{t-1}$.
GJR-GARCH(1,1): $h_t = \alpha_0 + h_{t-1}[\beta_1 + \alpha_1(\tilde{\epsilon}_{t-1} - \lambda_1)^2 + \gamma_1 \max(0, -\tilde{\epsilon}_{t-1} + \lambda_1)^2]$.
Under LRNVR, the form of m_t just affects the volatility dynamics while the risk-neutralized conditional mean return remains the same, that is, $r - h_t/2$. Now, we have all the variance specification in the risk-neutral measure. According to the equations above, the final asset price is derived from Corollary 1.

Corollary 1. When the locally risk-neutral valuation relationship holds, the terminal price for the ith asset, for $i = 1, 2$, can be expressed as

$$S_{i,T} = S_{i,t}\, \mathrm{e}^{(T-t)r} - \frac{1}{2}\sum_{s=t+1}^{T} h_{i,s} + \sum_{s=t+1}^{T} \sqrt{h_{i,s}}\tilde{\epsilon}_{i,s}].$$

Therefore, under the locally risk-neutral probability measure $\mathbb{Q}$, the option with exercise price K at maturity T has the value

$$v(t, S_1, S_2) = \mathrm{e}^{-r(T-t)}\mathrm{E}^{\mathbb{Q}}[\max[\max(S_1(T), S_2(T)) - K, 0]].$$

Due to the complexity of the GARCH process, analytical solution for the GARCH-in-mean Copula option-pricing model, in general, is not available. Therefore, we work with numerical methods to price the option described in the next section.

3. Methodology and Inference

In this section, we present here the procedure to obtain the price of a bivariate option using the asymmetric variance process by GARCH-in-mean under risk-neutral, copulas theory and Monte Carlo simulations. Chiou and Tsay (2008) and Zhang and Guegan (2008) have inspired this approach.

3.1 Generality

Given y_1 and y_2, two vectors containing the log-returns for the two stocks, we consider the following steps:

(1) For each y_i, with $i = 1, 2$, use quasi-maximum likelihood method described in Subsection 3.2 to estimates the parameters α_0, α_1, β_1 and λ in equation (2) and α_0, α_1, β_1, γ and λ for each marginals given in equations (3), (4) and (5). Thus, the problem is to maximize the function

$$l(\boldsymbol{\theta}, h_t) = -\frac{n}{2}\left[\log(2\pi) + \frac{1}{n}\sum_{t=1}^{n}\left[\log(h_t) + \frac{(y_{it} - m_{it})^2}{h_t}\right]\right],$$

with respect to the parameters, where m_{it} is the mean of GARCH-in-mean given by $r + \lambda\sqrt{h_t} - 1/2h_t$ and r is the fixed risk-free rate yield and h_t corresponds to each variance specification proposed in Subsection 2.2.

(2) Use the estimated parameters to calculate h_t for each specification and ϵ_t in equation (1) with $m_t = r + \lambda\sqrt{h_t} - 1/2h_t$ for each stock.

(3) Therefore, the proposed technique is that the objective copula and the risk-neutral copula are assumed to be the same. To fit the copulas, we transform the data into uniformly distributed random variables. Thus we transform the ϵ_i, for $i = 1, 2$, obtained in Step 2 for each stock into uniformly distributed variables, by $u_i = \Phi(\epsilon_i)$, where Φ is the standard normal cumulative distribution function.

(4) Fit a copula to pairs $[u_1, u_2]$ using maximum likelihood, that is, estimate the copula parameters θ_c

$$\theta_c = \arg\max_{\theta_c} \sum_{t=1}^{n} \log[c((u_{1,t}, u_{2,t}); \theta_c)],$$

where θ_c are the parameters for the specific copula function C and c is the density function for the given copula in the appendix.

(5) Now, using the Monte Carlo simulation, we obtain the option price. In the first step generate a sample $\{u^*_{1,t}, u^*_{2,t}\}_{t=1}^{T}$ from a uniform marginal distribution from one specific copula using the algorithm proposed by Nelsen (2006). Here T is the time to maturity for the option.

(6) For each time step, transform the generated margins to standard normal margins, in the risk-neutral measure, by $\tilde{\epsilon}_{i,t} = \Phi^{-1}(u^*_{i,t})$, for $i = 1, 2$.

(7) Working with $\tilde{\epsilon}_{i,t}$ to calculate the conditional variances under risk-neutral and the parameters estimated in step 1. The two future stock prices at time T are

$$S_{i,T} = S_{i,t}\, \mathrm{e}^{(T-t)r} - \frac{1}{2}\sum_{s=t+1}^{T} h_{i,s} + \sum_{s=t+1}^{T} \sqrt{h_{i,s}}\tilde{\epsilon}_{i,s}].$$

(8) Now, repeat Steps 5 to 7 for N runs. Thus, we obtain the Monte Carlo option price as

$$v(t, S_1, S_2) = \frac{\mathrm{e}^{-r(T-t)}}{N} \sum_{i=1}^{N} \max[\max(S_{1,i}(T), S_{2,i}(T)) - K, 0].$$

3.2 Quasi-maximum likelihood estimation

The assumption of conditional normality is not always appropriate in financial data. However, Weiss (1986) and Bollerslev and Wooldridge (1992) showed that even when normality is inappropriately assumed, maximizing the normalized log-likelihood results in quasi-maximum likelihood (QML) estimates that are consistent and asymptotically normally distributed. In addition, the authors claim that the conditional mean and variance functions of the GARCH models are correctly specified.

In particular, a robust covariance matrix conditional non-normality for the parameter estimates is consistently estimated by $A(\hat{\boldsymbol{\theta}})^{-1}B(\hat{\boldsymbol{\theta}})A(\hat{\boldsymbol{\theta}})^{-1}$, where $A(\hat{\boldsymbol{\theta}})$ and $B(\hat{\boldsymbol{\theta}})$ are the Hessian Matrix and the outer product of the gradients, respectively, calculated for $\boldsymbol{\theta}$. The SEs, computed from the square roots of the diagonal elements, are sometimes called Bollerslev-Wooldridge SE; for more details, see Bollerslev and Wooldridge (1992).

3.3 Model selection

We notice that for each time series we have four specification for variance processes, that is, GARCH(1,1), EGARCH(1,1), NGARCH(1,1) and GJR-GARCH(1,1). Choosing an adequate model is the essence of data analysis, which ultimately returns with good forecasting results.

In this paper, for model selection, we use five different criteria. The first one is the Akaike information criterion (AIC) (Akaike, 1973) given by $\text{AIC} = -2\log(\ell) + 2k$, where ℓ is the maximized value of the likelihood function and k is the number of free parameters in the model. The second one is the Bayesian information criterion (BIC) developed by Schwarz (1978) and given by $\text{BIC} = -2\log(\ell) + k\log(n)$, where n is the number of observations. The third one is the Hannan-Quinn information criterion (HQIC) proposed by Hannan and Quinn (1979) and given by $\text{HQIC} = -2\log(\ell) + 2k\log(\log(n))$. The fourth one is the Akaike information corrected criterion (AICc), developed by Hurvich and Tsai (1989) and given by $\text{AICc} = -2\log(\ell) + 2kn/(n-k-1)$, whereas the fifth one is the consistent Akaike information criterion (CAIC) given by $-2\log(\ell) + k\log(n) + 1$.

Following Genest, Remillard and Beaudoin (2009), we use the goodness-of-fit test, which is based on a comparison of the distance between the estimated and empirical copula by using the Cramer Von Mises test to compare the copula models. The goodness-of-fit statistic is defined as

$$S_n = \int_{[0,1]^d} \mathbb{C}_n(\boldsymbol{u})^2 \mathrm{d}C_n(\boldsymbol{U}),$$

where $C_n(\boldsymbol{U}) = 1/n \sum_{i=1}^{n} \mathbb{I}(U_{i1} \leq u_1; U_{i2} \leq u_2)$ is known as the empirical copula; $\boldsymbol{U}_j = (U_{1j}, \ldots, U_{ij})$ are the pseudo-observations; $\boldsymbol{u} = (u_1, u_2) \in [0,1]^2$; $\mathbb{C}_n = \sqrt{n}(C_n - C_{\theta_n})$ is the empirical process that assess the distance between the empirical copula and the estimation C_{θ_n} and n is the number of observations. Note that testing the null hypothesis that data are fitted by C_{θ_n} can be conducted with this statistic.

We chose this procedure because it can deal with non-linearity, asymmetry, serial dependence and also the well-known heavy-tails of financial assets (Righi and Ceretta, 2011). Furthermore, we make the comparison of the adjusted copula with the empirical copula by the diagonal method Sungur and Yang (1996). In addition, the AIC, AICc, CAIC, BIC and HQICare also used to support decision making in choosing the model.

4. Data Analyses

In this section, we illustrate the proposed methodology under two data sets. We used the software R for implementing the entire methods exposed here. The codes are available from the authors. The first one is artificial data, where we know the parameter values, and then we can verify if the methodology is reliable. The second data set is the Brazilian stock market data.

4.1 Artificial data

We consider here 1000 replications of two correlated time-series for each sample size ($n = 250, 500, 1000$) generated from same parameter structure with the Frank ($\theta = 8$) and marginals as follows:

GARCH(1,1):

$$h_{1,t} = 0.02 + 0.15h_{t-1}(\tilde{\epsilon}_{t-1} - 0.12)^2 + 0.8h_{t-1},$$
$$h_{2,t} = 0.03 + 0.2h_{t-1}(\tilde{\epsilon}_{t-1} - 0.08)^2 + 0.7h_{t-1},$$

EGARCH(1,1):

$$h_{1,t} = -0.3057 + 0.1223[|\tilde{\epsilon}_{t-1} - 0.12| + (-0.5057)(\tilde{\epsilon}_{t-1} - 0.12)] + 0.98ln(h_{t-1}),$$
$$h_{2,t} = -0.3057 + 0.1223[|\tilde{\epsilon}_{t-1} - 0.12| + (-0.5057)(\tilde{\epsilon}_{t-1} - 0.12)] + 0.98ln(h_{t-1}),$$

NGARCH(1,1):

$$h_{1,t} = 0.012 + 0.15h_{t-1}(\tilde{\epsilon}_{t-1} - 0.5 - 0.12)^2 + 0.8h_{t-1},$$
$$h_{2,t} = 0.03 + 0.2h_{t-1}(\tilde{\epsilon}_{t-1} - 0.2 - 0.08)^2 + 0.7h_{t-1},$$

GJR-GARCH(1,1):

$$h_{1,t} = 0.00961 + h_{t-1}[0.93 + 0.024(\tilde{\epsilon}_{t-1} - 0.065)^2 + 0.059max(0, -\tilde{\epsilon}_{t-1} + 0.065)^2],$$
$$h_{2,t} = 0.00961 + h_{t-1}[0.93 + 0.024(\tilde{\epsilon}_{t-1} - 0.065)^2 + 0.059max(0, -\tilde{\epsilon}_{t-1} + 0.065)^2].$$

For each configuration, we calculate the average of the QML estimates, as well as the corresponding robust standard error (SE) , the size of confidence intervals 95% (CI), coverage probability (CP), bias and mean squared error (MSE) of the QML estimators. Tables 1, 2, 3 and 4 report the simulation results for GARCH, NGARCH, EGARCH, and GJR-GARCH, respectively. We observe that the averages of the quasi-maximum likelihood estimates are close to the true values as the sample size increases, as well as decreasing the standard deviations in all the models. We also note low bias and MSEs as the sample size increases. Concerning the size of the confidence interval, we noticed they are getting smaller as the sample size increases. In addition, the empirical coverages are closer to the nominal ones for all four models. With this results, we noticed that all the models have good asymptotic properties.

Table 1. Parameter estimation of both artificial time-series for each GARCH process.

	Parameter	$\alpha_{0,1}$	$\alpha_{1,1}$	β_1	λ_1	$\alpha_{0,2}$	$\alpha_{1,2}$	β_2	λ_2	θ
	Real Value	0.02	0.15	0.8	0.12	0.03	0.2	0.7	0.08	8
$n = 250$	Mean	0.0355	0.1479	0.7511	0.1192	0.0429	0.2020	0.6442	0.0834	7.9349
	SE	0.1695	0.4596	0.9783	0.1798	0.0337	0.0927	0.1983	0.0786	0.5777
	CI size	0.1213	0.2297	0.5095	0.2345	0.1347	0.2784	0.6594	0.1919	2.4122
	CP	0.9880	0.9490	0.9560	0.9760	0.9480	0.9289	0.9480	0.9750	0.9229
	Bias	-0.0155	0.0021	0.0489	0.0009	-0.0129	-0.0020	0.0558	-0.0034	0.0652
	MSE	0.0002	0.0000	0.0024	0.0000	0.0002	0.0000	0.0031	0.0000	0.0042
$n = 500$	Mean	0.0255	0.1492	0.7830	0.1198	0.0347	0.1998	0.6806	0.0825	7.9836
	SE	0.0126	0.0403	0.0615	0.0479	0.0156	0.0562	0.0917	0.0523	0.4092
	CI size	0.0493	0.1462	0.2403	0.1838	0.0541	0.1992	0.3349	0.1572	1.6726
	CP	0.9720	0.9500	0.9570	0.9470	0.9470	0.9269	0.9289	0.9720	0.9399
	Bias	-0.0055	0.0008	0.0170	0.0003	-0.0047	0.0002	0.0194	-0.0025	0.0164
	MSE	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0004	0.0000	0.0003
$n = 1000$	Mean	0.0223	0.1500	0.7931	0.1181	0.0323	0.2004	0.6908	0.0780	8.0054
	SE	0.0089	0.0314	0.0437	0.0403	0.0112	0.0413	0.0659	0.0401	0.2896
	CI size	0.0282	0.0990	0.1442	0.1251	0.0360	0.1308	0.1885	0.1111	1.2808
	CP	0.9600	0.9580	0.9570	0.9530	0.9439	0.9550	0.9550	0.9600	0.9469
	Bias	-0.0023	0.0000	0.0069	0.0019	-0.0023	-0.0004	0.0092	0.0020	-0.0054
	MSE	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000

Table 2. Parameter estimation of both artificial time-series for each NGARCH process.

	Parameter	$\alpha_{0,1}$	$\alpha_{1,1}$	β_1	λ_1	γ_1	$\alpha_{0,2}$	$\alpha_{1,2}$	β_2	λ_2	γ_2	θ
	Real value	0.012	0.15	0.8	0.12	0.5	0.03	0.2	0.7	0.08	0.2	8
$n = 250$	Mean	0.0266	0.1431	0.7656	0.1162	0.5517	0.0416	0.1962	0.6503	0.0798	0.3461	7.9343
	SE	0.1107	0.7069	0.7901	0.5111	6.3509	0.0444	0.1538	0.2345	0.1130	1.1610	0.5773
	CI size	0.0923	0.2061	0.3251	0.2349	0.9699	0.1046	0.2832	0.5296	0.1851	0.9900	2.5200
	CP	0.9860	0.9580	0.9620	0.9820	0.9730	0.9620	0.9429	0.9499	0.9870	0.9960	0.9289
	Bias	-0.0146	0.0069	0.0344	0.0038	-0.0517	-0.0116	0.0038	0.0497	0.0002	-0.0461	0.0657
	MSE	0.0002	0.0000	0.0012	0.0000	0.0027	0.0001	0.0000	0.0025	0.0000	0.0021	0.0043
$n = 500$	Mean	0.0162	0.1435	0.7894	0.1178	0.5345	0.0348	0.1972	0.6796	0.0811	0.3250	7.9601
	SE	0.0277	0.0744	0.1240	0.0964	0.4538	0.0149	0.0510	0.0778	0.0542	0.1790	0.4083
	CI size	0.0296	0.1412	0.1665	0.1859	0.8008	0.0553	0.2005	0.2963	0.1601	0.9699	1.7088
	CP	0.9860	0.9399	0.9520	0.9730	0.9620	0.9540	0.9299	0.9520	0.9740	0.9640	0.9269
	Bias	-0.0042	0.0065	0.0106	0.0022	-0.0345	-0.0048	0.0028	0.0204	-0.0011	-0.0250	0.0399
	MSE	0.0000	0.0000	0.0001	0.0000	0.0012	0.0000	0.0000	0.0004	0.0000	0.0006	0.0016
$n = 1000$	Mean	0.0142	0.1479	0.7934	0.1174	0.5167	0.0323	0.1973	0.6921	0.0789	0.3135	7.9780
	SE	0.0132	0.0425	0.0577	0.0535	0.1983	0.0121	0.0406	0.0625	0.0404	0.1285	0.2890
	CI size	0.0167	0.0971	0.1066	0.1322	0.4957	0.0336	0.1274	0.1810	0.1291	0.6677	1.1894
	CP	0.9520	0.9469	0.9600	0.9640	0.9590	0.9580	0.9479	0.9540	0.9640	0.9590	0.9479
	Bias	-0.0022	0.0021	0.0066	0.0026	-0.0167	-0.0023	0.0027	0.0079	0.0011	-0.0135	0.0220
	MSE	0.0000	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0001	0.0000	0.0002	0.0005

Table 3. Parameter estimation of both artificial time-series for each EGARCH process.

	Parameter	$\alpha_{0,1}$	$\alpha_{1,1}$	β_1	λ_1	γ_1	$\alpha_{0,2}$	$\alpha_{1,2}$	β_2	λ_2	γ_2	θ
	Real Value	-0.3067	0.1223	0.98	0.12	-0.5057	-0.3067	0.1223	0.98	0.12	-0.5057	8
$n = 250$	Mean	-0.2852	0.0876	0.9793	0.1204	-0.5080	-0.2874	0.0891	0.9793	0.1204	-0.5065	7.9388
	SE	0.8405	1.1478	0.0244	0.2555	0.3996	0.6754	0.7820	0.0242	0.2362	0.2318	0.5781
	CI size	0.3182	0.2010	0.0180	0.2353	0.1900	0.3250	0.2215	0.0176	0.2367	0.1876	2.5276
	CP	0.9139	0.8759	0.8829	0.9570	0.9249	0.9149	0.8679	0.8749	0.9640	0.9269	0.9139
	Bias	-0.0205	0.0347	0.0007	-0.0004	0.0023	-0.0183	0.0332	0.0007	-0.0004	0.0008	0.0612
	MSE	0.0004	0.0012	0.0000	0.0000	0.0000	0.0003	0.0011	0.0000	0.0000	0.0000	0.0037
$n = 500$	Mean	-0.2960	0.1070	0.9798	0.1205	-0.5060	-0.2968	0.1082	0.9798	0.1198	-0.5067	7.9560
	SE	0.0649	0.0374	0.0035	0.0515	0.0408	0.0607	0.0493	0.0026	0.0571	0.0437	0.4085
	CI size	0.1790	0.1478	0.0081	0.1756	0.1239	0.1818	0.1592	0.0084	0.1707	0.1248	1.6859
	CP	0.9069	0.9118	0.9009	0.9289	0.9139	0.9179	0.9278	0.9309	0.9379	0.9199	0.9339
	Bias	-0.0097	0.0153	0.0002	-0.0005	0.0003	-0.0089	0.0141	0.0002	0.0002	0.0010	0.0440
	MSE	0.0001	0.0002	0.0000	0.0000	0.0000	0.0001	0.0002	0.0000	0.0000	0.0000	0.0019
$n = 1000$	Mean	-0.3033	0.1171	0.9799	0.1212	-0.5061	-0.3039	0.1176	0.9799	0.1210	-0.5054	7.9865
	SE	0.0391	0.0226	0.0020	0.0362	0.0242	0.0456	0.0233	0.0026	0.0457	0.0448	0.2892
	CI size	0.1116	0.0842	0.0047	0.1231	0.0733	0.1098	0.0838	0.0048	0.1286	0.0707	1.2940
	CP	0.9459	0.9409	0.9591	0.9429	0.9599	0.9449	0.9689	0.9339	0.9419	0.9489	0.9579
	Bias	-0.0024	0.0052	0.0001	-0.0012	0.0004	-0.0018	0.0047	0.0001	-0.0010	-0.0003	0.0135
	MSE	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002

Table 4. Parameter estimation of both artificial time-series for each GJR-GARCH process.

	Parameter	$\alpha_{0,1}$	$\alpha_{1,1}$	β_1	λ_1	γ_1	$\alpha_{0,2}$	$\alpha_{1,2}$	β_2	λ_2	γ_2	θ
	Real Value	0.00961	0.024	0.93	0.065	0.059	0.00961	0.024	0.93	0.065	0.059	8
$n = 250$	Mean	0.0582	0.0306	0.8326	0.0741	0.0548	0.0524	0.0334	0.8346	0.0725	0.0563	7.9335
	SE	3.1908	1.2611	7.8901	1.5276	1.5033	6.3504	7.0343	1.9827	6.7417	7.2567	0.5774
	CI size	0.4059	0.0949	0.9637	0.1953	0.1697	0.3703	0.1123	0.9627	0.1965	0.1697	2.5486
	CP	0.9970	0.9970	0.9880	0.9800	0.9990	0.9870	0.9990	0.9790	0.9840	0.9990	0.9199
	Bias	-0.0486	-0.0066	0.0974	-0.0091	0.0042	-0.0428	-0.0094	0.0954	-0.0075	0.0027	0.0665
	MSE	0.0024	0.0000	0.0095	0.0001	0.0000	0.0018	0.0001	0.0091	0.0001	0.0000	0.0044
$n = 500$	Mean	0.0219	0.0260	0.9045	0.0672	0.0572	0.0224	0.0262	0.9047	0.0682	0.0563	7.9695
	SE	0.2226	0.2865	0.7740	0.2871	0.2983	0.4872	0.4015	1.3715	0.4177	0.4554	0.4087
	CI size	0.0753	0.0611	0.2013	0.1516	0.1141	0.0845	0.0634	0.2057	0.1496	0.1141	1.7545
	CP	0.9790	0.9760	0.9610	0.9730	0.9680	0.9730	0.9670	0.9720	0.9790	0.9440	0.9239
	Bias	-0.0122	-0.0020	0.0255	-0.0022	0.0018	-0.0128	-0.0022	0.0253	-0.0032	0.0027	0.0305
	MSE	0.0001	0.0000	0.0007	0.0000	0.0000	0.0002	0.0000	0.0006	0.0000	0.0000	0.0009
$n = 1000$	Mean	0.0134	0.0251	0.9225	0.0651	0.0564	0.0137	0.0252	0.9211	0.0661	0.0577	7.9674
	SE	0.0160	0.0537	0.0640	0.0981	0.0590	0.0198	0.0521	0.0759	0.0901	0.0618	0.2888
	CI size	0.0299	0.0466	0.0818	0.1209	0.0902	0.0325	0.0476	0.0988	0.1188	0.0875	1.2196
	CP	0.9560	0.9590	0.9410	0.9570	0.9420	0.9510	0.9561	0.9440	0.9520	0.9492	0.9499
	Bias	-0.0037	-0.0011	0.0075	-0.0001	0.0026	-0.0041	-0.0012	0.0089	-0.0011	0.0013	0.0326
	MSE	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0011

4.2 Real data

In principle, price data are not available, since the call-on-max option is typically traded over-the-counter. For this reason, we cannot test the valuation models empirically. However, comparing models with different assumptions can be implemented, as in Zhang and Guegan (2008), Liu, Li and Ng (2015) and Chiou and Tsay (2008). In this section, we carry on the illustration of the proposed methodology on a real data set concerning the two stock prices of Brazilian companies. With the objective of analyzing two companies that could have a high correlation, we choose the companies Bradespar (BRAP4) and Vale S.A. (VALE3) with the aim of investigating two companies that could have a high correlation. The Brazilian company Bradespar admits the shareholdings that the bank Bradesco had in non-financial companies, among them: VCB, Vale, Scopus, and Globo. Thus, Bradespar's stocks price would be directly related to the stocks of Vale S.A., where the company holds the latter's stock control at 17.4 %. The analyzed period is from 07/01/2015 to 07/17/2018, containing 753 observations.

Figure 1 shows the high positive association between the two series, evidencing the requirement subject is financial options using these stocks, given its high correlation. Table 5 reports the similarity between the returns series, both concerning the minimum, mean, median, maximum, standard deviation (SD) and kurtosis, but the VALE3 series has a slightly more pronounced positive asymmetry than the BRAP4 series. As evidenced in section 2, asymmetry is present in financial series, a feature that symmetric GARCH processes have no potential to discriminate between positive and negative asymmetry.

Table 5. Descriptive statistics of returns.

Serie	Minimum	Mean	Median	Maximum	SD	Kurtosis	Skewness
BRAP4	-0.134	0.000	0.000	0.153	0.027	0.050	5.150
VALE3	-0.156	0.000	0.000	0.137	0.026	0.047	5.702

Before presenting the estimated coefficients of time series models, we focus on the analysis of the best model according to the selection criteria. Given the flexibility of the use of models based on copula functions, we select for each marginal the best model according to the selection criteria defined in Section 3.2. According to Table 6, all criteria corroborate that the model GARCH best fit the BRAP4 series, evidencing that there is no asymmetry present in this series, while, the best model for the VALE3 series is the EGARCH (evidencing the asymmetry). This result is in agreement with the statement in Table 5, where the VALE3 stock had an asymmetric coefficient more pronounced than BRAP4.

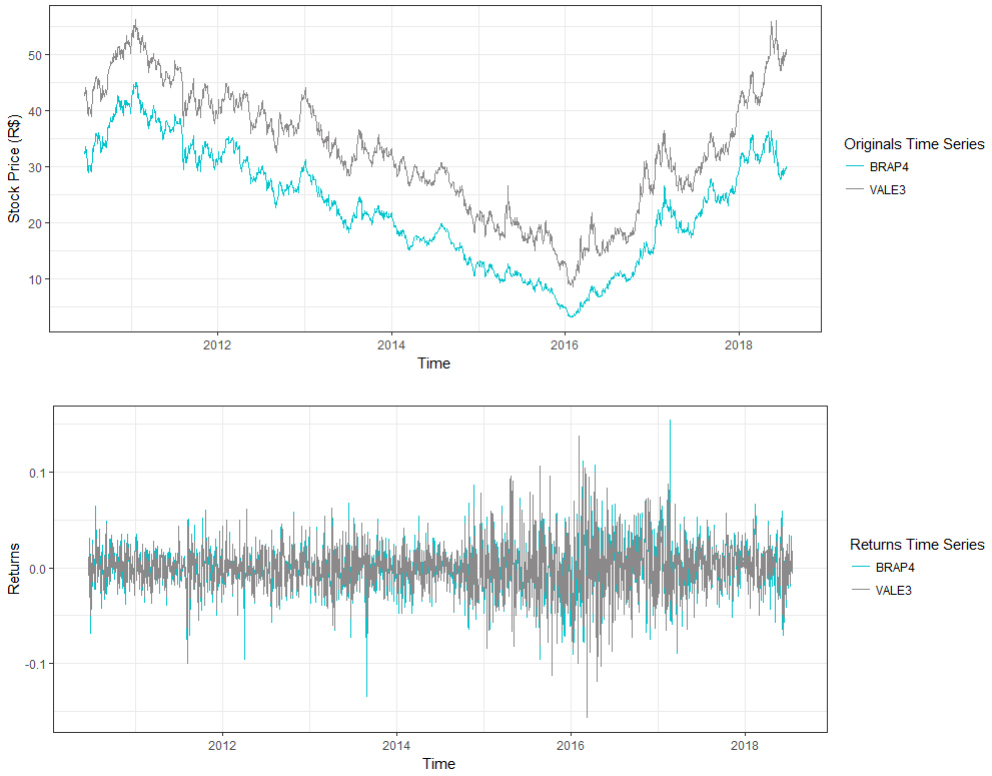

Figure 1. Original Series and Returns.

Table 6. Selection criteria for marginals.

BRAP4	GARCH	NGARCH	EGARCH	GJR-GARCH
AIC	-3071.3128	-3069.3172	-3070.1057	-3069.3678
AICc	-3071.2591	-3069.2367	-3070.0252	-3069.2872
CAIC	-3048.8271	-3041.2102	-3041.9987	-3041.2607
BIC	-3052.8271	-3046.2102	-3046.9987	-3046.2607
HQIC	-3064.1903	-3060.4142	-3061.2027	-3060.4647
VALE3	GARCH	NGARCH	EGARCH	GJR-GARCH
AIC	-3151.2693	-3150.0533	-3153.7289	-3151.7989
AICc	-3151.2156	-3149.9728	-3153.6484	-3151.7183
CAIC	-3123.6918	-3121.9463	-3128.7836	-3125.6219
BIC	-3128.6918	-3126.9463	-3132.7836	-3130.6219
HQIC	-3144.1468	-3141.1503	-3144.8258	-3142.8958

Table 7 reports the coefficients estimated via QML estimates and their respective robust standard errors. According to this result, we noticed that the best model for the BRAP4 series was the GARCH model, where it does not have an asymmetry parameter. We view in this model the high persistence, that is, $\alpha_1 + \beta_1$ very close to one, suggesting that the volatility can be persistent (strong temporal dependence), which opens options of models to analyze series with this feature. The best model for the VALE3 series was the EGARCH, where it presented a parameter of positive asymmetry, that is, a positive shock decreases its volatility.

Table 7. Estimated coefficients and corresponding robust standard errors for marginals.

BRAP4	GARCH	NGARCH	EGARCH	GJR-GARCH
$\hat{\alpha}_0$	6.9191e-06	6.8024e-06	-0.1399	7.0316e-06
	(4.8306e-06)	(4.7823e-06)	(0.0455)	(4.9614e-06)
$\hat{\alpha}_1$	0.0479	0.0477	0.1005	0.0507
	(0.0125)	(0.0127)	(0.0248)	(0.0175)
$\hat{\beta}$	0.9454	0.9457	0.9914	0.9446
	(0.0138)	(0.0140)	(0.0050)	(0.0144)
$\hat{\lambda}$	0.0568	0.0560	0.0560	0.0576
	(0.0358)	(0.0365)	(0.0359)	(0.0364)
$\hat{\gamma}$	-	0.0121	-0.0618	4.2924e-03
		(0.1628)	(0.0235)	(0.0180)
VALE3	GARCH	NGARCH	EGARCH	GJR-GARCH
$\hat{\alpha}_0$	3.7157e-06	3.0711e-06	-0.1081	2.5848e-06
	(2.9974e-06)	(2.9524e-06)	(0.0386)	(2.9020e-06)
$\hat{\alpha}_1$	0.0434	0.0428	0.0969	0.0555
	(0.0116)	(0.0111)	(0.0221)	(0.0152)
$\hat{\beta}$	0.9519	0.9522	0.9957	0.9554
	(0.0121)	(0.0117)	(0.0004)	(0.0113)
$\hat{\lambda}$	0.0579	0.0671	0.0762	0.0679
	(0.0357)	(0.0365)	(0.0387)	(0.0363)
$\hat{\gamma}$	-	0.1771	0.1433	0.0278
		(0.1854)	(0.1438)	(0.0171)

We consider the Kolmogorov-Smirnov, Jarque-Bera, Shapiro-Wilk, and Anderson-Darling tests to verify the assumption of normality of the residuals for the fitted models. Table 8 reports their p-values. All tests did not reject the null hypothesis at 5% that residuals follow a standard normal distribution. In addition, to verify that the increments are independent, Table 8 also reports the result of the Ljung-Box test, where, for all fitted models we do not reject the null hypothesis at 5 % that the residuals are independent.

Table 8. Tests of Normality and Independent Increments for residuals.

BRAP4	GARCH	NGARCH	EGARCH	GJR-GARCH
Kolmogorov-Smirnov	0.9315	0.9403	0.9514	0.9343
Jarque-Bera	0.1159	0.1142	0.2351	0.1225
Shapiro-Wilk	0.2571	0.2572	0.3802	0.2633
Anderson-Darling	0.6680	0.6725	0.6572	0.6652
Ljung-Box	0.4940	0.4938	0.4988	0.4944
VALE3	GARCH	NGARCH	EGARCH	GJR-GARCH
Kolmogorov-Smirnov	0.8737	0.8752	0.8761	0.8733
Jarque-Bera	0.2059	0.1680	0.2548	0.1433
Shapiro-Wilk	0.1752	0.1895	0.2644	0.1718
Anderson-Darling	0.2288	0.2627	0.3426	0.2697
Ljung-Box	0.1927	0.2079	0.2145	0.2152

Figure 2 shows the QQ-plots for the two best models for the series, that is, on the left panel is the GARCH for the BRAP4 series and on the right panel the EGARCH for the VALE3 series, corroborating with the tests in the Table 8, evidencing the non-rejection of the normality of the residuals.

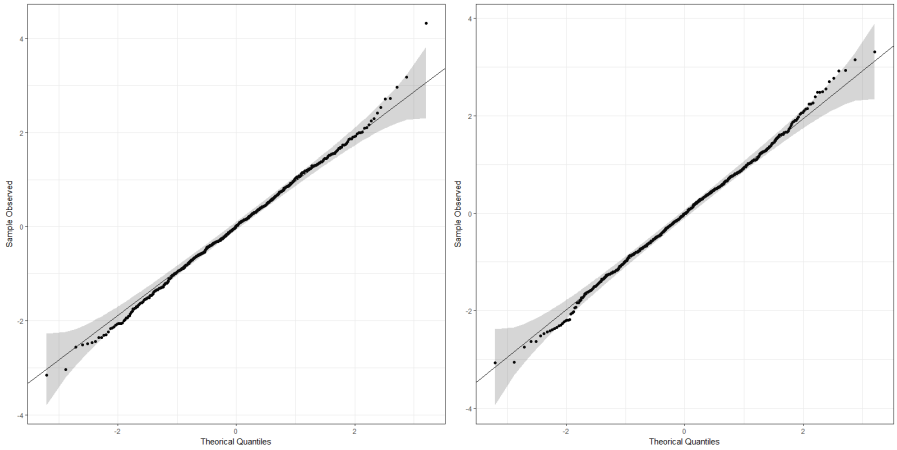

Figure 2. QQ-plots of residuals - GARCH BRAP4 (left panel) and EGARCH VALE3 (right panel).

Figure 3 illustrates the individual behavior of each set of residual fitted through the histograms and the joint behavior through the scatterplot in the center of the figure. As expected, the series has a highly positive association behavior, which is evidenced in the adjustment of the copulas given in Table 9, where the normal and Student-t copulas obtained high and positive values of their parameters ($-1 \leq \theta \leq 1$).

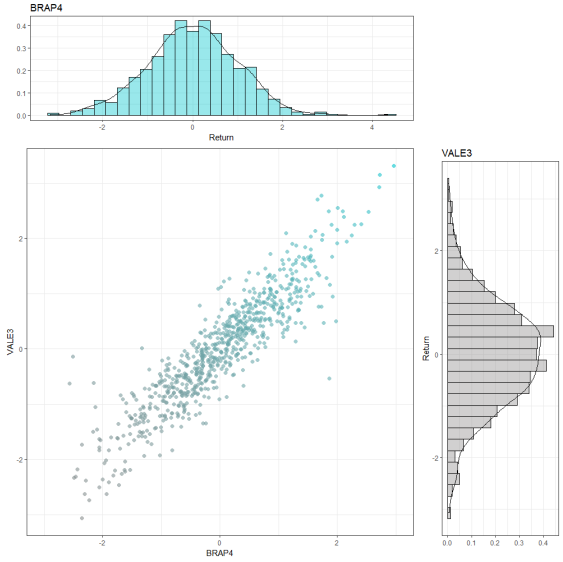

Figure 3. Scatterplot and histograms of residuals - GARCH BRAP4 and EGARCH VALE3.

Table 9. Estimated coefficients and corresponding standard errors (in parentheses) for copulas.

	Normal	Student-t	Gumbel	Frank	Joe
$\hat{\theta}$	0.9059	0.9133	3.4082	14.0430	4.0173
	(0.0048)	(0.0053)	(0.1040)	(0.4965)	(0.1423)

The degree of freedom of the Student-t copula and its respective standard deviation were 7.63401 and 1.7263.

According to Table 10 and the selection criteria adopted, the best copula for this data set was the Student-t copula, though the results found for the Student-t copula are very similar to the one observed for the Frank copula. The empirical copula and the copula adjusted by the diagonal method, where the excellent fit of the two copulas is noted, corroborate this result. The result of the Cramer Von Mises test are 0.0025, 0.0023, 0.0042, 0.0018 and 0.01122, for normal, Student-t, Gumbel, Frank and Joe copulas, respectively. As noted in

Figure 4, the result shows that Frank copula yields the smallest distance between fitted and empirical copula. We note that there is a minimal difference between the Frank and Student-t copulas. Therefore, these two copulas are considered in this work as the best fittings.

Table 10. Selection model of copulas.

	Normal	Student-t	Gumbel	Frank	Joe
AIC	-1290.6171	-1334.1487	-1231.5615	-1310.8730	-970.63696
AICc	-1290.6117	-1334.1327	-1231.5562	-1310.8676	-970.63162
CAIC	-1285.9957	-1324.9059	-1226.9401	-1306.2516	-966.01556
BIC	-1288.8365	-1330.5875	-1229.7809	-1309.0924	-968.85635
HQIC	-1284.9957	-1322.9059	-1225.9401	-1305.2516	-965.01556

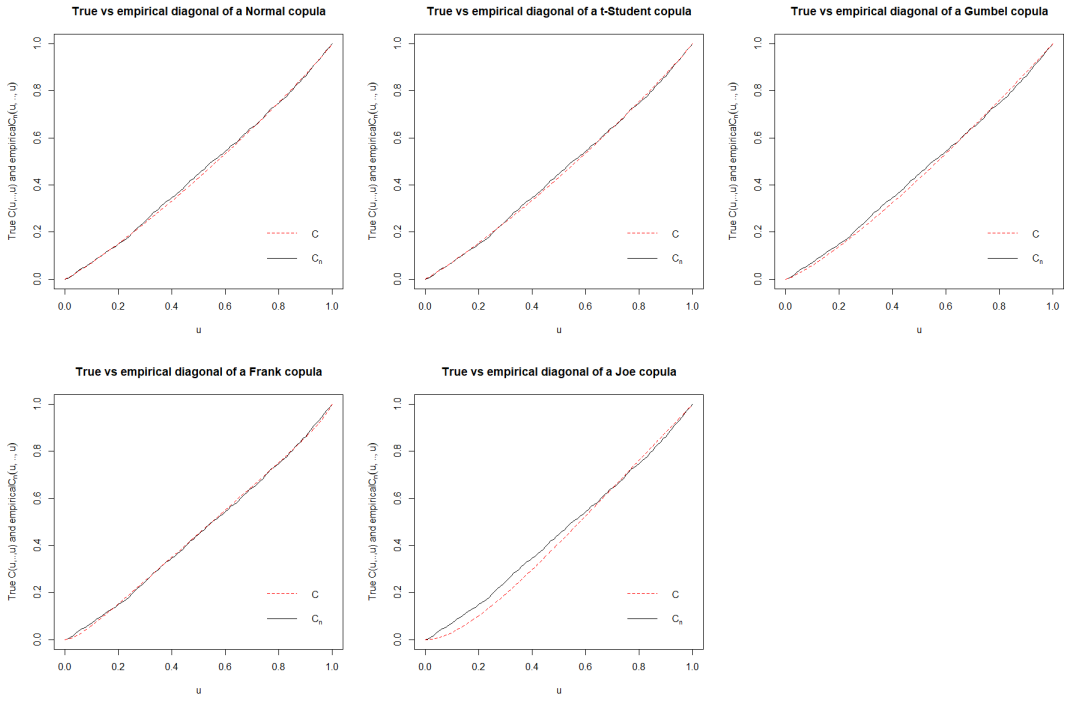

Figure 4. Comparing the empirical copula and the true copula on the diagonal.

Given the good fitting of the marginals obtained via time series models and the good joint fitting via copulas, we now calculate and analyze the option prices considering the call-on-max payoff function. To perform the comparison process, as a benchmark, we compare the results through the methodology proposed with the classical method, which is a Black-Scholes extension for the bivariate case (Haug, 2007), where this model considers the volatility constant over time and the linear dependence structure from the bivariate normal distribution.

The entire study was performed with 100 000 Monte Carlo simulations, 7 % interest rate and maturity time of one year. According to Table 9, as expected, the same behavior is observed for all models, that is, as the strike variable increases it is likely that, in a call option, the price of the option becomes cheaper. We note that the classical model obtained the lowest values for all strike values. Gesk and Roll (1984), Black (1975) and MacBeth and Merville (1980) corroborate this result for the univariate case, where the authors showed that the models that consider constant volatility over time underpricing the options, especially in-the-money (ITM) options. That is, a call option's strike price is

below the market price in the univariate case. In this work we define ITM options when the strike price is less than the minimum between the two assets. Moreover, in Table 11, we can see that the Student-t and Frank copulas have the closest results to each other. The similarity in the excellent fit of the data can explain this result. We noticed the values obtained through normal copula obtained high results. The inability of the normal copula to capture observations in the tails of the distribution, a recurring fact in finances, can explain this result. The copula Joe obtained higher values mainly when the strike was smaller than 40, approaching the model of the normal copula. The Gumbel copula was the one that received the lowest values between the models. Figure 5 shows the behavior of the option price (z-axis) varying the maturity from 1 to 12 months (y-axis, in days) and strike (R$ 40.00 to R$ 60.00). We note that the higher the maturity the values differ little between strike prices, which does not happen when the option has a short maturity, where we indicate that setting at 50 maturity days there is a relatively significant difference varying the price of the strike. For example, Table 12 presents the prices for considering maturity = one month, six months and one year and strike = 20, 40 and 60.

Another fundamental aspect in the management of options risks is to know the levels of dependence between stocks. Therefore, Figure 6 presents the price behavior of the call-on-max option for the Student-t copula by varying its degrees of dependence. This result corroborates with those found by Chiou and Tsay (2008) for the call-on-max option using the American and Taiwanese indices. An intuitive interpretation is: the values of this option tend to be smaller when the underlying assets move in the same direction as when in opposite directions.

Table 11. Prices of a call-on-max option under various strikes values (R$).

Strike	Classic	Normal	Student-t	Gumbel	Frank	Joe
20	31.1182	32.3619	32.2648	32.2532	32.2711	32.5083
22	29.2693	30.5241	30.4283	30.4148	30.4329	30.6440
24	27.4764	28.7327	28.6402	28.6228	28.6425	28.8293
26	25.7468	26.9951	26.9054	26.8845	26.9045	27.0694
28	24.0867	25.3171	25.2295	25.2061	25.2258	25.3714
30	22.5003	23.7025	23.6169	23.5918	23.6110	23.7401
32	20.9906	22.1572	22.0733	22.0457	22.0646	22.1787
34	19.5594	20.6831	20.6028	20.5704	20.5889	20.6899
36	18.2071	19.2840	19.2055	19.1678	19.1867	19.2758
38	16.9331	17.9601	17.8810	17.8399	17.8602	17.9371
40	15.7360	16.7112	16.6316	16.5866	16.6093	16.6745
42	14.6140	15.5377	15.4571	15.4103	15.4349	15.4901
44	13.5643	14.4379	14.3578	14.3081	14.3346	14.3826
46	12.5842	13.4087	13.3304	13.2795	13.3062	13.3498
48	11.6705	12.4495	12.3723	12.3199	12.3477	12.3877
50	10.8198	11.5571	11.4799	11.4270	11.4576	11.4945
52	10.0288	10.7290	10.6522	10.5987	10.6321	10.6678
54	9.2941	9.9621	9.8872	9.8339	9.8675	9.9020
56	8.6122	9.2533	9.1807	9.1278	9.1613	9.1933
58	7.9798	8.6001	8.5283	8.4762	8.5101	8.5392
60	7.3937	7.9981	7.9271	7.8762	7.9117	7.9372

Table 12. Prices (R$) of a call-on-max option varying some Maturity time and Strike (R$).

Maturity\Strike	R$ 20.00	R$ 40.00	R$ 60.00
One Month	30.9671	10.9067	0.5537
Six Months	30.9190	13.6608	4.1864
One Year	30.7311	15.6953	7.0992

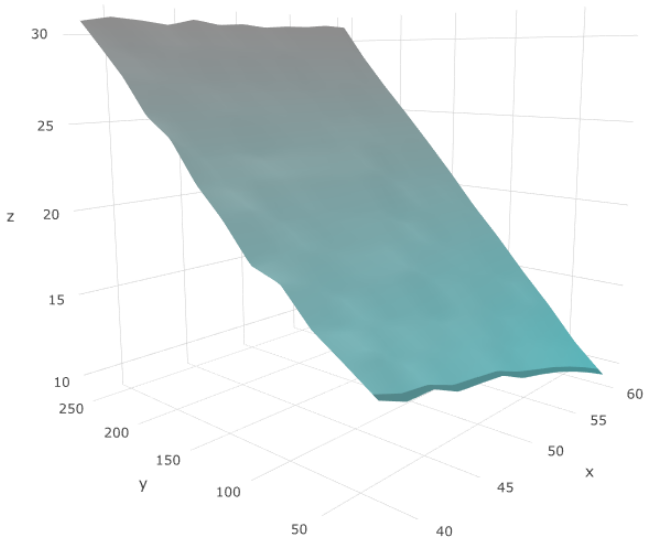

Figure 5. Price (R$) behavior of the call-on-max option ranging from Maturity to Strike.

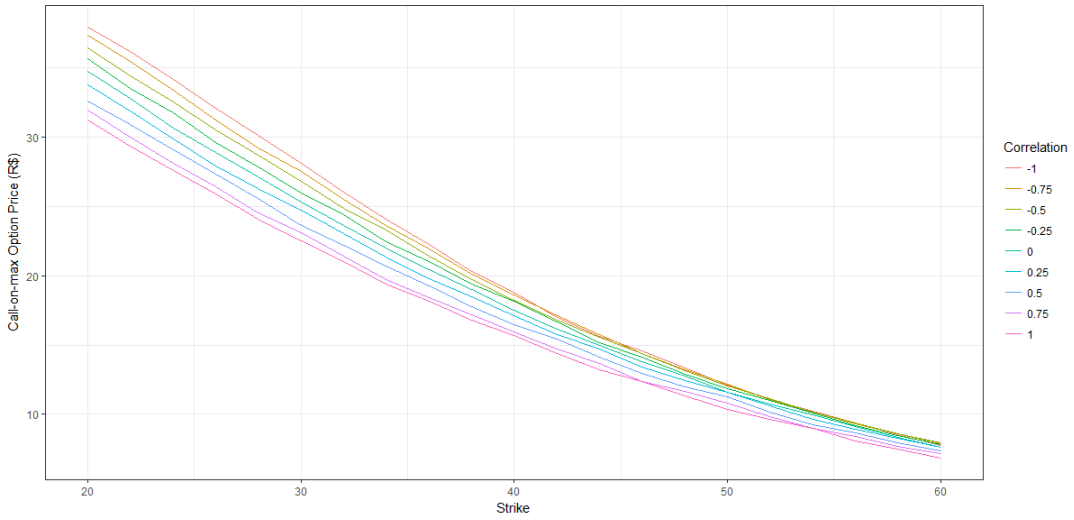

Figure 6. Behavior of the call-on-max option price by varying the copula parameter.

In addition, Figure 6 further shows that in-the-money options have the most substantial differences between dependency levels than out-the-money options (that is, when the strike is higher than the maximum between the two assets). Therefore, it was empirically verified the importance of a good joint fit of the stocks, and above all, the calculation of the correlation between the assets. Moreover, by employing the copulas functions, it is possible to capture linear, non-linear and caudal associations. Recalling, the traditional models derived from a Brownian geometric movement consider bivariate normal to price call-on-max options for two assets, and consequently, the linear correlation coefficient as the measure of association.

5. Concluding Remarks

In this paper, we propose an analysis and comparison among pricing models that consider the volatility of underlying assets and in the presence of dependence between copula framework. The model is an adequate methodology to realize a more realistic pricing option. To consider the modeling of asymmetry present in financial series, we examined three models that are extensions of the GARCH model under the neutral risk measure $\mathbb{Q}$, a pre-requisite to price options (NGARCH, EGARCH, and GJR-GARCH). Therefore, through the flexibility of the copula functions, we chose which marginal processes fit best with each stock and thus proceeded in the joint fitted.

Two databases illustrate the application of the methodology. The first one was an artificial database with the objective of carrying out a simulation study and the second a database of two Brazilian companies. The simulation study showed that all models presented good asymptotic properties. In addition, in the real time-series of two Brazilian stock companies, the model offered a proper fitting and the results obtained were confronted with the classic model, which is an extension of the Black-Scholes model.

The contributions of the proposed method in the present paper are as follows: (i) using the best copula makes the model more suitable; (ii) extension to marginal models that consider asymmetry makes joint modeling more flexible and realistic; (iii) a comparison of methodologies highlights the role of risk management; (iv) due to the good marginal and joint fitted, in addition to the values obtained in relation to the classical consolidated model, there are arguments to believe that the differences obtained between the best models, through the copulas and the extension of the conventional method, are improvements in the calculation of the fair value; and (v) the empirical relevance of such alternatives is apparent given the evidence of non-joint-normality in financial emerging markets.

Finally, we highlight some points for future work. The first one of them, even with extensions to asymmetric models, we often have financial series with heavy tails, which should derive a risk-neutral measure Q for these models, such as considering the non-normality of the residuals. The second point is the adoption of other copula functions, such as power variance function family copulas. The third, we can consider another skew distribution for the errors, such as Arrellano-Valle et al. (2010), Minozzo et al. (2012) and Marcos et al. (2012).

Appendix

The normal copula

The normal copula or commonly known as Gaussian copula receives this name because it comes from the normal density function for $d \geq 2$. A normal bivariate copula is expressed by

$$C(u,v) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{1}{2\pi\sqrt{1-\rho^2}} \mathrm{e}^{-\dfrac{t_1^2 - 2\rho t_1 t_2 + t_2^2}{2(1-\rho^2)}} \mathrm{d}t_1^2 \mathrm{d}t_2^2,$$

where $x_1 = \Phi^{-1}(u)$, $x_2 = \Phi^{-1}(v)$, for $-1 \leq \rho \leq 1$. This type of copula has no dependence on the tails of the distributions and is symmetric.

The Student-t copula

The Student-t copula coincides with the bivariate Student-t distribution function, where its form is defined as

$$C(u,v) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{1}{2\pi\sqrt{1-\rho^2}}\left(1+\frac{t_1^2-2\rho t_1 t_2}{\nu(1-\rho^2)}\right)^{-(\nu+2)/2} \mathrm{d}t_1\mathrm{d}t_2,$$

where ν represents the degrees of freedom of Student-t distribution. As in the case of normal copula, the Student-t marginal copula coincides with the Student-t standard, being $x_1 = t_\nu^{-1}(u)$ e $x_2 = t_\nu^{-1}(v)$. This type of copula does not have independence in the tails, which favors its use in extreme events, such as, for example, unplanned oscillations in the stock market. However, given the symmetry of the function, the degree of dependence on the upper tail is equal to the lower tail.

The Gumbel copula

The Gumbel copula is characterized by the dependence only on the upper tail and is represented as

$$C(u,v) = \mathrm{e}^{-[(-\log(u))^\theta+(-\log(v))^\theta]^{1/\theta}},$$

where $\theta \in [1,\infty]$. When $\theta \to \infty$, dependence is perfectly positive and independent when $\theta = 1$.

The Frank Copula

The form of a Frank copula is expressed through

$$C(u,v) = -\frac{1}{\theta}\log\left(1+\frac{[\mathrm{e}^{-\theta u}-1][\mathrm{e}^{-\theta v}-1]}{\mathrm{e}^{-\theta}-1}\right)$$

where $\theta \neq 0$. When $\theta \to \infty$, we have perfect positive dependence and we have the case of independence when we $\theta \to 0$. This copula has the same dependence on both function tails, such as elliptic copulas.

The Joe copula

The Joe copula is given by

$$C(u,v) = 1 - \left([1-u]^\theta + [1-v]^\theta - [1-u]^\theta[1-v]^\theta\right)^{1/\theta},$$

where $1 \leq \theta \leq \infty$. When $\theta = 1$, we have the case of independence and the case of perfect positive dependence when $\theta \to \infty$.

Acknowledgment

The authors wish to thank the Editors and anonymous referees for their constructive comments on an earlier version of this manuscript, which resulted in this improved version. The study was supported partially by CNPq, CAPES and FAPESP.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. Proc. 2nd International Symposium on Information Theory, In: B N Petrov and F Csaki (Eds.) Akademiai Kiado, Budapest, pp.267-81.

Almeida, D. and Hotta, L.K., 2014. The leverage effect and the asymmetry of the error distribution in Garch-based models: The case of Brazilian Market Related Series. Pesquisa Operacional, 34, 237-250.

Arellano-Valle, R.B and Genton, M.G., 2010. Multivariate unified skew-elliptical distributions. Chilean Journal of Statistics, 1, 17-33.

Black, F., 1975) Fact and fantasy in the use of options. Financial Analysts Journal, 31, 36-41/61-72.

Black, F., 1976. Studies of stock price volatility changes. In Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association, Washington DC, 177-181.

Black, F. and Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy, 81, 637-654.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31, 307-327.

Bollerslev, T. and Wooldridge, J.M., 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews, 11, 143-172.

Brennan, M.J. and Schwartz, F.Z., 1979. A continuous time approach to the pricing of bonds. Journal of Banking and Finance, 3, 133-155.

Chiou, S.C. and Tsay, R.S., 2008. A copula-based approach to option pricing and risk assessment. Journal of Data Science, 6, 273-301.

Delbaen, F. and Schachermayer, W., 1994. A general version of the fundamental theorem of asset pricing. Mathematische Annalen, 300, 463-520.

Duan, J.C., 1995. The GARCH option pricing model. Mathematical Finance, 5, 13-32.

Duan, J., Gauthier, G., Simonato, J.G., and Sasseville, C., 2006. Approximating the GJR-GARCH and EGARCH option pricing models analytically. Journal of Computational Finance, 9, 41-69.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50, 987-1007.

Genest, C., Rmillard, B., and Beaudoin, D., 2009. Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics, 44, 199-214.

Geske, R. and Roll, R., 1984. On valuing American call options with the Black-Scholes European formula. Journal of Finance, 39, 443-455.

Glosten, L. R., Jagannathan, R., and Runkle, D. E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. The journal of Finance, 48, 1779-1801.

Hannan, E.J. and Quinn, B.G., 1979. The determination of the order of an autoregression. Journal of the Royal Statistical Society B, 41, 190-195.

Haug, E.G., 2007. The Complete Guide to Option Pricing Formulas. McGraw-Hill, New York.

Hull, J., 1992. Introduction to Futures and Options Markets. Prentice Hall, New Jersey.

Hurvich, C.M. and Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika, 76, 297-307.

Johnson, H. and David S., 1987. Option pricing when the variance is changing. Journal of Financial and Quantitative Analysis, 22, 143-151.

Liu, Y., Li, J.S.H., and Ng., A.C.Y., 2015. Option pricing under GARCH models with

Hansen's skewed-t distributed innovations. The North American Journal of Economics and Finance, 31, 108-125.

Lopes, L.P. and Pessanha. G.R.G., 2018. Analise de dependencia entre mercados financeiros: uma abordagem do modelo Copula-GARCH. Revista de Financas e Contabilidade da Unimep, 5, 18-38.

MacBeth, J.D. and Merville, L.J., 1980. Tests of the BlackScholes and Cox call option valuation models. The Journal of Finance, 35, 285-301.

Minozzo M. and Ferracuti, L., 2012. On the existence of some skew-normal stationary processes. Chilean Journal of Statistics, 3, 157-170.

Prates, M.O, Dey, D.K., and Lachos, V.H., 2012. A dengue fever study in the state of Rio de Janeiro with the use of generalized skew-normal/independent spatial fields. Chilean Journal of Statistics, 3, 143-155.

Margrabe, W., 1978. The value of an option to exchange one asset for another. The Journal of Finance, 33, 177-186.

Merton, R.C., 1973. Theory of rational option pricing. The Bell Journal of Economics and Management Science, 4, 141-183.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica, 59, 347-370.

Nelsen, R.B., 2006. An Introduction to Copulas. Springer, New York.

Righi, M.B. and Ceretta, P.S., 2011. Extreme values dependence of risk in Latin American markets. Economics Bulletin, 31, 2903-2914.

Sanfins, M.A. and Valle, G., 2012. On the copula for multivariate extreme value distributions. Brazilian Journal of Probability and Statistics, 26, 288-305.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Sharifonnasabi, Z., Alamatsaz, M.H.,, and Kazemi, I., 2018. A large class of new bivariate copulas and their properties. Brazilian Journal of Probability and Statistics, 32, 497-524.

Shimko, D.C., 1994. Options on futures spreads: Hedging, speculation, and valuation. Journal of Futures Markets, 14, 183.213.

Sklar, A., 1959. Fonctions de repartition a n dimensions et leures marges, Publications of the Institute of Statistics of the University of Paris, 8, 229-231

Stulz, R., 1982. Options on the minimum or the maximum of two risky assets: analysis and applications. Journal of Financial Economics, 10, 161-185.

Sungur, E.A. and Yang, Y., 1996. Diagonal copulas of Archimedean class. Communications in Statistics: Theory and Methods, 25, 1659-1676.

Weiss, A.A., 1986. Asymptotic theory for ARCH models: Estimation and testing. Econometric Theory, 2, 107-131.

Zhang, J. and Guegan, D., 2008. Pricing bivariate option under GARCH processes with time-varying copula. Insurance: Mathematics and Economics, 42, 1095-1103.

Chilean Journal of Statistics
Vol. 10, No. 2, December 2019, 177–195

Nonparametric Statistics
Research Paper

Nonparametric estimation of the relative error in functional regression and censored data

Boubaker Mechab[1,*], Nesrine Hamidi[1], and Samir Benaissa[1]

[1]Laboratory of Statistics and Stochastic Processes, University of Djillali Liabes, BP 89, Sidi Bel Abbes 22000, Algeria

(Received: 01 March 2019 · Accepted in final form: 23 June 2019)

Abstract

In this paper, the almost complete consistency and the asymptotic normality of the estimator of the regression operator in the case of a censored response given a functional explanatory variable are investigated under some mild conditions. The latter is constructed from the minimization of the mean squared relative error. The novelty of this work compared to the works found in the literature is that the response variable is censored. A simulation study is carried out to compare the finite sample performance based on mean square error between the classical regression and the relative error regression. Moreover, a real data study is used to illustrate our methodology.

Keywords: Censoring · Functional data analysis · Nonparametric statistics · Relative error regression.

Mathematics Subject Classification: Primary 62G05, 62G20 · Secondary 62F12.

1. Introduction

Functional data analysis is a section of statistics that studies the observation of infinite dimension. More precisely, the observations that are not real or vector variables but random curves. This kind of data appears in many practical situations, and it has been the subject of many works. The first authors who discussed this type of data are Ramsay and Silverman (2005) for the parametric models and monograph of Ferraty and Vieu (2006) for the nonparametric estimation. Recently, many topics concerning the analysis of functional data have been developed and the most recent advances in this field have been collected in the book of Ould-Said et al. (2015). The particularity of the nonparametric estimation consists in estimating an infinite number of parameters whose function is unknown, elements of a certain functional class, such as the density function or the regression function. The latter is one of many methods to predict the link between the response variable Y and the explanatory variable X, assuming the existence of a function $r(X)$ which expresses the relationship between these two variables. The literature concerning this field is widely developed. We refer to Ferraty and Vieu (2004) for more details, where is established the strong consistency of the regression function when the response is scalar given a functional

*Corresponding author. Email: mechaboub@yahoo.fr

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

http://www.soche.cl/chjs

explanatory variable. Usually, to estimate the nonparametric regression model, the authors used the least squares error as a criterion for constructing the predictors (see some details in Louzada et al. (2018)). This method is very sensitive to outliers, and therefore, the presence of large outliers can lead to inappropriate results. For this, the authors developed methods that study robustness of the nonparametric functional regression; see also Attouch et al. (2009) and Gheriballah et al. (2013).

The relative squared error criterion is more convenient as a measure of performance than the previous criterion, since the notion of relative regression is more recent than the others, although the results are still limited. Jones et al. (2008) studied the asymptotic properties of a consistent estimator of this model by using the kernel method. We refer to Mechab and Laksaci (2016) for recent advances, who studied nonparametric relative regression for associated variables. In a functional framework, the paper of Demongeot et al. (2016) brought an extra to the research by studying the almost complete convergence and asymptotic normality of the proposed estimator.

In this paper, we investigate the asymptotic properties of the relative error regression by the kernel method and under censoring data. The literature of this kind of incomplete functional data is quite restricted. We refer to Kohler et al. (2002) and Horrigue and Ould-Said (2011, 2014) for the nonparametric regression quantile estimation under random censorship. Other works have been conducted on this subject for functional data case. We cite for example the work of Khardani et al. (2010). Moreover, our framework was considered by Altendji et al. (2018) for the estimation of the functional relative error regression under random left truncation, where they established the almost complete convergence with rates, as well as the asymptotic normality of the kernel estimator of the functional relative error regression for truncated data. In a more general field, we can see, for example, Hsing and Eubank (2015) and Aneiros et al. (2017). In the present work, we investigate the almost complete convergence and asymptotic normality of our proposed estimator in case of censored functional data.

The organization of this paper is as follows. In Section 2, we construct an estimator of the relative error regression for a censored response. The necessary conditions and main results are presented in Section 3. In Section 4, a numerical study and a real example show the performances of the proposed methodology for finite samples. Also, we establish a confidence interval as an application for the asymptotic normality result. In Section 5, we provide some concluding remarks. The proofs of our results are given in the appendix.

2. Description of the Model and Estimator

2.1 Estimator of the the relative error regression

Let $Z_i = (X_i, Y_i)_{i=1,\ldots,n}$ be a $\mathcal{F} \times \mathrm{R}$ valued measurable strictly stationary process. A common nonparametric modeling of the link between the response variables Y and the explanatory variable X is to suppose that

$$Y = m(X) + \varepsilon, \tag{1}$$

where ε is a random error variable and m is a regression operator usually estimated by minimizing the expected squared loss function given by

$$\mathrm{E}[(Y - m(X))^2 | X].$$

In some situations, this loss function which is considered as a measure of prediction, may not be suitable. Among these situations, the presence of outliers can lead to inappropriate

results since all variables have an equal weight. For this, we overcame this limitation by proposing to estimate the function m with respect to the minimization of the mean squared relative error defined as

$$\mathrm{E}\left[\left(\frac{Y-m(X)}{Y}\right)^2\Big|X\right], \quad Y>0. \tag{2}$$

Obviously, this loss function is a more meaningful measure of prediction performance in the presence of outliers since the range of predicted values is large. Furthermore, the solution of (2) can be expressed by the ratio of first two conditional inverse moments of Y given X. The best predictor of Y given X (as studied in Park and Stefanski (1998)) is given by

$$r(x)=\frac{\mathrm{E}[Y^{-1}|X=x]}{\mathrm{E}[Y^{-2}|X=x]}.$$

We estimate the regression operator r under our relative loss as

$$\widetilde{r}(x)=\frac{\sum_{i=1}^n Y_i^{-1}K(h^{-1}d(x-X_i))}{\sum_{i=1}^n Y_i^{-2}K(h^{-1}d(x-X_i))}, \tag{3}$$

where K is a kernel and $h=h_n$ is a sequence of positive real numbers.

2.2 Estimator of the relative error regression under a random censorship

Let $(X_i,Y_i)_{i=1,\ldots,n}$ be a $\mathcal{F}\times\mathrm{R}$ valued measurable strictly stationary process, where $\mathcal{F}$ is a semi-metric abstract space, denote by d, a semi-metric associated with the space $\mathcal{F}$. We observe the lifetimes Y_n as a sequence of independent and identically distributed random variable (with common unknown absolutely continuous distribution function F with density f).

In censoring case, due to possible withdrawals of items from the study, we observe the censored lifetimes C instead observing the lifetimes Y. Supposing that (C_i) is a sequence of independent and identically distributed censoring random variable (r.v.) with common unknown continuous distribution function G. We remark the pairs (T_i,δ_i) where

$$T_i=Y_i\wedge C_i, \quad \delta_i=\mathrm{I}_{\{Y_i\leq C_i\}}, \quad 1\leq i\leq n,$$

where I_A denotes the indicator of no censoring.

We consider a pseudo estimator of the regression operator r under the censorship and the relative loss given by

$$\widetilde{r}(x)=\frac{\sum_{i=1}^n \delta_i\bar{G}^{-1}(T_i)T_i^{-1}K(h^{-1}d(x-X_i))}{\sum_{i=1}^n \delta_i\bar{G}^{-1}(T_i)T_i^{-2}K(h^{-1}d(x-X_i))}=\frac{\widetilde{g}_1(x)}{\widetilde{g}_2(x)} \tag{4}$$

where $\bar{G}(u)=1-G(u)$ and for $l=1,2$,

$$\widetilde{g}_l(x)=\frac{1}{n\mathrm{E}(K_1(x))}\sum_{i=1}^n \delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x),$$

where $K_i(x)=K(h^{-1}d(x-X_i))$. Since G is unknown in practice, one can estimate it using

the Kaplan and Meier (1958) estimator defined as

$$\bar{G}_n(t) = \begin{cases} \prod_{i=1}^{n} \left(1 - \frac{1-\delta_{(i)}}{n-i+1}\right)^{\mathrm{I}_{\{T_{(i)} \le t\}}}, & \text{if } t < T_{(n)}, \\ 0, & \text{otherwise;} \end{cases}$$

where $T_{(1)} < \cdots < T_{(n)}$ are the order statistics of $(T_i)_{1 \le i \le n}$ and $\delta_{(i)}$ is concomitant with $T_{(i)}$. Thus, an estimator of r is given by

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} \delta_i \bar{G}_n^{-1}(T_i) T_i^{-1} K(h^{-1} d(x - X_i))}{\sum_{i=1}^{n} \delta_i \bar{G}_n^{-1}(T_i) T_i^{-2} K(h^{-1} d(x - X_i))} = \frac{\widehat{g}_{1,n}(x)}{\widehat{g}_{2,n}(x)}, \tag{5}$$

where

$$\widehat{g}_{l,n}(x) = \frac{1}{n\mathrm{E}(K_1(x))} \sum_{i=1}^{n} \delta_i \bar{G}_n^{-1}(T_i) T_i^{-l} K_i(x), \quad l = 1, 2.$$

Let $\tau_F = \sup\{y, \bar{F}(y) > 0\}$ and $\tau_G = \sup\{y, \bar{G}(y) > 0\}$ be a upper endpoints of $\bar{F}$ and $\bar{G}$, respectively. We assume that $\tau_F < \infty$, $\bar{G}(\tau_F) > 0$, which implies that $\tau_F \le \tau_G$) and that $(C_n)_{n \ge 1}$ and $(X_n, Y_n)_{n \ge 1}$ are independent.

3. Assumptions and Main Results

3.1 Consistency: almost complete convergence

We fix a point x in $\mathcal{F}$ and N_x denotes a fixed neighborhood of this point. We will denote by C and $C^{'}$ some strictly positive constants, $g_l(x) = \mathrm{E}[Y^{-l}|X = x]$ for $l = 1, 2$ and we have $B(x, h) = \{x^{'} \in \mathcal{F} | d(x^{'}, x) < h\}$ a ball of center x and a radius h. In what follows, we will need the following assumptions:

(H1) For all $h > 0$, $\mathrm{P}(X \in B(x, h)) =: \phi_x(h) > 0$ and $\lim_{h \to 0} \phi_x(h) = 0$.

(H2) For all $(x_1, x_2) \in N_x^2$ and $l = 1, 2$, we have

$$|g_l(x_1) - g_l(x_2)| \le C d^{k_l}(x_1, x_2) \quad \text{for} \quad k_l > 0.$$

(H3) The kernel K is a measurable function that is supported by (0, 1) and satisfies:

$$0 < C \le K \le C^{'} < \infty.$$

(H4) The bandwidth satisfies:

$$\lim_{n \to +\infty} h = 0 \quad \text{and} \quad \lim_{n \to +\infty} \frac{\log(n)}{n\phi_x(h)} = 0.$$

(H5) The inverse moments of the response variable verify:

$$\mathrm{E}[Y^{-m}|X = x] < C < \infty, \quad \forall m \ge 2.$$

Remark 1 The hypothesis (H1) defines the concentration properties of the probability measures of the explanatory variable X, which is provided by means of a function ϕ_x. This property allows to propose an alternative to the curse of dimensionality problem. (H2) is a regularity condition to facilitate the calculation of the bias part of our estimator. (H3)-(H5) are technical assumptions to ensure the convergence of our results.

THEOREM 3.1 Assume that conditions (H1)-(H5) hold true, we get

$$|\widehat{r}_n(x) - r(x)| = O(h^{k_1}) + O(h^{k_2}) + O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right). \tag{6}$$

LEMMA 3.2 Under assumptions (H1)-(H4), we obtain, for $l = 1, 2$,

$$|\mathrm{E}\left[\widetilde{g}_l(x)\right] - g_l(x)| = O(h^{k_l}). \tag{7}$$

LEMMA 3.3 Under conditions (H1) and (H3)-(H5), we have, for $l = 1, 2$,

$$|\widetilde{g}_l(x) - \mathrm{E}\left[\widetilde{g}_l(x)\right]| = O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right). \tag{8}$$

LEMMA 3.4 Assume hypotheses (H1)-(H5) hold, we have, for $l = 1, 2$,

$$|\widehat{g}_{l,n}(x) - \widetilde{g}_l(x)| = O_{a.s}\left(\sqrt{\frac{\log(\log(n))}{n}}\right). \tag{9}$$

COROLLARY 3.5 Under assumptions of Theorem 3.1,we get

$$|\widehat{g}_{2,n}(x)| \xrightarrow[n\to\infty]{} g_2(x).$$

3.2 ASYMPTOTIC NORMALITY

Here, we establish the asymptotic normality of the estimator $\widehat{r}_n(x)$. To do that, we consider the following assumptions:

(C1) The hypothesis (H1) holds and there exists a function χ_x such that, for all $s \in [0, 1]$, we have $\phi_x(sr)/\phi_x(r) = \chi_x(s) + o(1)$ and $\int_0^1 (K^j)'(s)\chi_x(s)\mathrm{d}s < \infty$, for $j \geq 1$.

(C2) The functions $\Psi_l(u) = \mathrm{E}[g_l(X) - g_l(x)|d(x, X) = u]$ are derivable at 0, for $l = 1, 2$.

(C3) The hypothesis (H3) holds and the kernel K is a differentiable function on $]0, 1[$ and its first derivative function K' satisfies that $C < K' < C'$.

(C4) The small ball probability satisfies:

$$n\phi_x(h) \to \infty.$$

(C5) The inverse moments $g_m(u) = \mathrm{E}[|\bar{G}^{-1}(Y)Y^{-m}||X = u]$ of the censored response variable are continuous in a neighborhood of x, for $m = 1, 2, 3, 4$.

Remark 2 The condition $(C1)$ is realized by several small ball probability functions, there exist many examples, we quote the following (which can be found in Ferraty et al. (2007)):

(i) For some $\gamma > 0$, $\phi_x(h) = C_x h^\gamma$ with $\chi_x(u) = u^\gamma$,

(ii) for some $\gamma > 0$ and $p > 0$, $\phi_x(h) = C_x h^\gamma \exp(-C/h^p)$, with $\chi_x(u) = \delta_1(u)$, where δ_1 is the Dirac function,
(iii) $\phi_x(h) = C_x/|\log(h)|$, with $\chi_x(u) = \mathrm{I}_{[0,1]}(u)$, where I_A is an indicator function of a set A.

THEOREM 3.6 Suppose that conditions (C1)-(C5) hold true, for all $x \in \mathcal{F}$, we have, as $n \to \infty$,

$$\left(\frac{n\phi_x(h)}{\sigma^2(x)}\right)^{\frac{1}{2}} (\widehat{r}_n(x) - r(x)) \xrightarrow{\mathcal{D}} \mathrm{N}(0,1),$$

where $\xrightarrow{\mathcal{D}}$ means the convergence in distribution and

$$\sigma^2(x) = \frac{M_2}{M_1^2}\left(g_2(x) + r^2(x)g_4(x) - 2r(x)g_3(x)\right),$$

with $M_0 = K(1) - \int_0^1 (sK(s))'\chi_x(s)\mathrm{d}s$ and $M_j = K^j(1) - \int_0^1 (K^j)'(s)\chi_x(s)\mathrm{d}s$, for $j = 1, 2$.

PROOF OF THEOREM 3.6. From the decomposition 10, we get the decomposition

$$\begin{aligned}\widehat{r}_n(x) - r(x) = \frac{1}{\widehat{g}_{2,n}(x)g_2(x)} &\left[\left(\widetilde{g}_1(x) - \mathrm{E}\left[\widetilde{g}_1(x)\right]\right) g_2(x) + \left(\mathrm{E}\left[\widetilde{g}_2(x)\right] - \widetilde{g}_2(x)\right) g_1(x)\right.\\ &+ \left(\widehat{g}_{1,n}(x) - \widetilde{g}_1(x)\right) g_2(x) + \left(\widetilde{g}_2(x) - \widehat{g}_{2,n}(x)\right) g_1(x)\\ &\left.+ \left(\mathrm{E}\left[\widetilde{g}_1(x)\right] - g_1(x)\right) g_2(x) + \left(g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]\right) g_1(x)\right].\end{aligned}$$

Then, Theorem 3.6 is a consequence of the following lemmas.

LEMMA 3.7 Under the same conditions of Theorem 3.6, we have

$$\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} \left(\left[\widetilde{g}_1(x) - \mathrm{E}\left[\widetilde{g}_1(x)\right]\right] g_2(x) + \left[\mathrm{E}\left[\widetilde{g}_2(x)\right] - \widetilde{g}_2(x)\right] g_1(x)\right) \xrightarrow{\mathcal{D}} \mathrm{N}(0,1).$$

LEMMA 3.8 Under hypotheses of Theorem 3.6, we get $\widehat{g}_{2,n}(x) \to g_2(x)$, in probability, and

$$\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} \left[\left(\widehat{g}_{1,n}(x) - \widetilde{g}_1(x)\right) g_2(x) + \left(\widetilde{g}_2(x) - \widehat{g}_{2,n}(x)\right) g_1(x)\right] \to 0,$$

in probability.

LEMMA 3.9 Under hypotheses of Theorem 3.6, we obtain

$$\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} \left[\frac{1}{g_2(x)}\left(\mathrm{E}\left[\widetilde{g}_1(x)\right] - g_1(x)\right) g_2(x) + \left(g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]\right) g_1(x)\right] \to 0,$$

in probability.

4. Numerical Studies

4.1 Simulation study on the finite samples

To compare the finite-sample performance of the proposed estimator of $r(x) = \mathrm{E}[Y|X = x]$ to the classical regression, we conducted a small simulation study. We consider a functional regression model defined as

$$Y_i = m(X_i) + \varepsilon,$$

where the random variable ε is normally distributed as $\mathrm{N}(0, 1)$ and

$$m(x) = 4 \exp\left(\frac{1}{1 + \int_0^{\pi} |x(t)|^2 \mathrm{d}t}\right).$$

The functional variable X is chosen as a real-valued function with support $[0, \pi]$, we generate $n = 100$ functional data (see Figure 1) by $X_i(t) = \sin(W_i(t))$, for all $t \in [0, \pi]$ and $i = 1, \ldots, n$, where the random variables W_i are independent and identically distributed and follow the normal distribution $\mathrm{N}(0, 1)$. The curves are discretized on the same grid which is composed of 100 equidistant values in $[0, \pi]$.

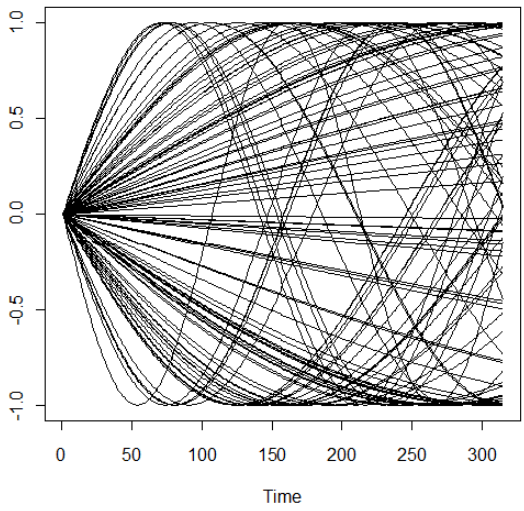

Figure 1. Curves X_i

Our purpose is to compare the mean square error (MSE) of the estimator of relative error regression (RER) with the censored data set and with the classical regression estimator (CR) respectively which are defined as

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^n \delta_i \bar{G}_n^{-1}(T_i) T_i^{-1} K(h^{-1} d(x - X_i))}{\sum_{i=1}^n \delta_i \bar{G}_n^{-1}(T_i) T_i^{-2} K(h^{-1} d(x - X_i))}$$

and

$$\widehat{r}(x) = \frac{\sum_{i=1}^n \delta_i \bar{G}_n^{-1}(T_i) T_i K(h^{-1} d(x - X_i))}{\sum_{i=1}^n \delta_i \bar{G}_n^{-1}(T_i) K(h^{-1} d(x - X_i))}.$$

We choose the quadratic kernel given by

$$K(u) = \frac{3}{4}(1-u^2)\mathrm{I}_{[-1,1]}(u)$$

and the bandwidth h is automatically selected by the procedure of the cross validation.

We give the formula of the MSEs of the both estimators as

$$\mathrm{MSE(RER)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}_{n,i}(X_i))^2$$

and

$$\mathrm{MSE(CR)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{r}_{i}(X_i))^2,$$

where $\widehat{r}_{n,i}$ ($\widehat{r}_i$) is the leave-one-out version of $\widehat{r}_n$ ($\widehat{r}$) computed by removing the ith data from the initial sample.

Table 1. Values of the MSE according to the number of introduced artificial outliers (first line).

Outliers	5	10	20	30	40	50
CR	0.5254138	70.67035	658.129	3702.399	5923.839	14809.60
RER	0.1219565	0.1256098	0.1261814	0.1261834	0.1261834	0.1261834

Note from Table 1 that the MSE values for both kernel methods increase considerably relative to the presence of the outliers, while these errors remain very small in the case of the relative error estimator. In conclusion, the relative error regression performs better than the classical regression, that is, the classical regression is more sensitive to the presence of outliers than the relative error regression.

4.2 Real data application

We apply the theoretical results obtained in the previous section to real data. More specifically, we examine the performance of the relative regression estimator in the presence of outliers than the classical kernel method. For this purpose application, we consider the spectroscopic dataset, are available from http://www.models.kvl.dk/NIRsoil. The data concern spectra of 108 soil samples measured by near infrared reflectance (NIR), in the range 400–2500 nanometre (nm) with a 2 nm resolution (Rinnan and Rinnan, 2007). Thus, the soil samples are obtained during a long-term climate change manipulation experiment at a subarctic fell heath in Abisko, northern Sweden. Moreover, to determine the chemical and microbiological properties of soil, soil organic matter (SOM) was measured as loss on ignition at 550°C and ergosterol concentration was determined through High-Performance Liquid Chromatography (HPLC), which are taken in the following as two response variables. The aim is to analyse relationships between the NIR data (X-variables), and the chemical and microbiological data (Y-variables). For each sample soil, one observes a spectroscopic curve which corresponds to the reflectance at 1050 wavelengths, and its soil organic matter and ergosterol content. Hence, $X_i(t)$ is the reflectance of the i^{th} sample of soil at wavelength t, where $t \in \{400, \ldots, 2500\}$. Let Y_1 and Y_2 be two response variables which correspond to soil organic matter and ergosterol concentration, respectively (see Figures 3 and 4). The functional covariates in Figure 2 shows the 108 NIR reflectance spectra.

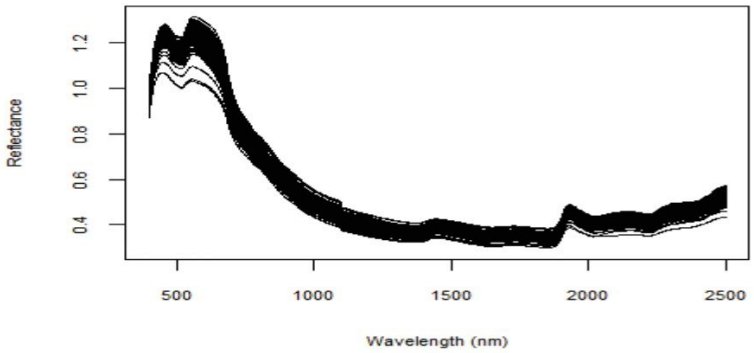

Figure 2. Curves of 108 NIR spectra

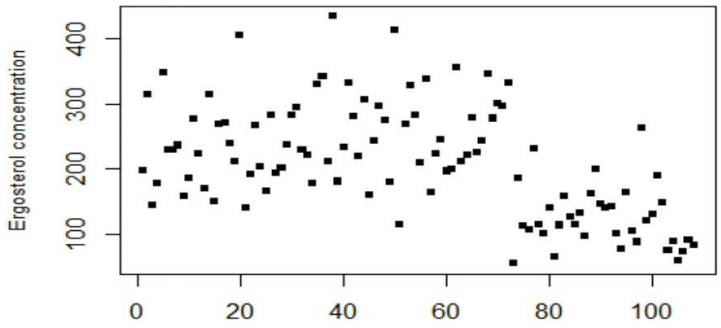

Figure 3. The distribution of 108 values of Y_1 (SOM)

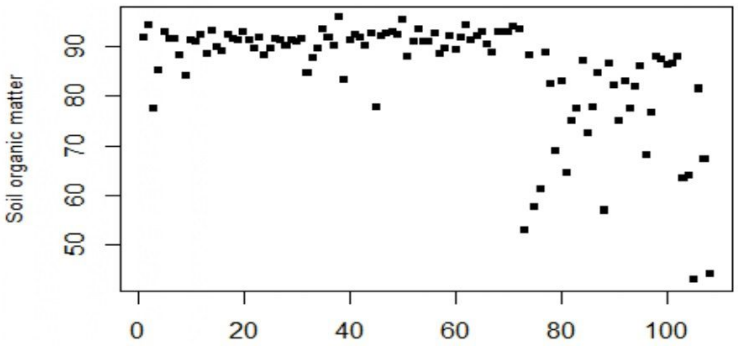

Figure 4. The distribution of 108 values of Y_2 (ergosterol concentration)

Applied to NIR data the MAD-Median method identifies 21 outliers for Y_1 and 1 outlier for Y_2. Recall that we are interested to build two models: $Y_1 = r_1(X) + \varepsilon_1$ and $Y_2 = r_2(X) + \varepsilon_2$, where $r_1(x) = \mathrm{E}(Y_1|X = x)$ and $r_2(x) = \mathrm{E}(Y_2|X = x)$. Furthermore, the dataset was randomly split into a learning sample (72 curves) used to build the estimators, and a testing sample (36 curves) which allows computing the MSE. We note that the result of our simulation study is evaluated over 100 independent replications and its sensitivity to grid sizes or to size of test sample and training sample is not very substantial. Because of the smoothness of the NIR curves, we use the semi-metric based on the second order derivatives, where the curves are replaced by their B-spline expansion. Here, the best results in terms of prediction are obtained for a number of interior knots needed for defining the B-spline basis, equal to 40. Therefore, we chosen the smoothing parameter h via a local cross-validation method on the number of nearest-neighbors. It can be seen that, in the presence of outliers, the relative regression estimator performs better than the classical kernel method. This is confirmed by the MSE obtained respectively in the two cases of study.

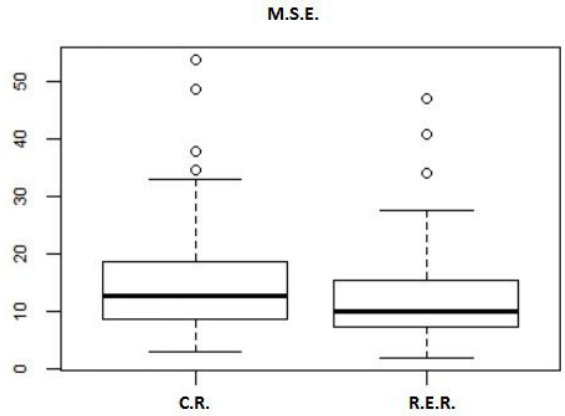

Figure 5. Box plots of the MSE for Y_1

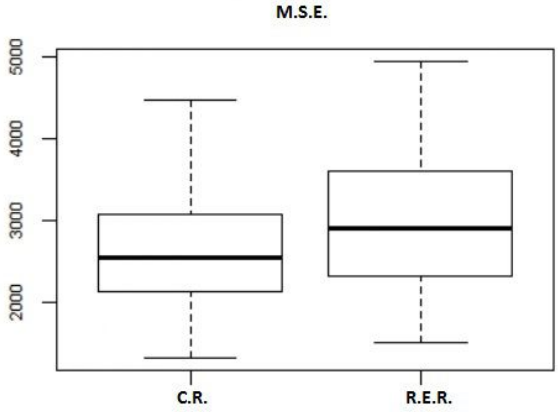

Figure 6. Box plots of the MSE for Y_2

4.3 Confidence bands

A usual application of asymptotic normality is to establish confidence intervals for the true value of the proposed estimator. To determine this band, we need the estimation of the unknown quantity of the asymptotic variance. In our case, we have

$$\sigma^2(x) = \frac{M_2}{M_1^2}\left(g_2(x) + r^2(x)g_4(x) - 2r(x)g_3(x)\right),$$

where M_1, M_2, r and g_l, for $l = 1, 2, 3, 4$, are unknown in practice and have to be estimated. Now a plug-in estimate for the asymptotic standard deviation $\sigma(x)$ can be easily obtained using the estimators $\widehat{M}_1, \widehat{M}_2, \widehat{r}_n$ and $\widehat{g}_{l,n}$ of M_1, M_2, r and g_l respectively. Precisely, we estimate $g_3(x)$ and $g_4(x)$ in the same way as for $g_1(x)$ and $g_2(x)$.

We estimate empirically the constants M_1 and M_2, as

$$\widehat{M_1} = \frac{1}{n\phi_x(h)} \sum_{i=1}^{n} \delta_i \bar{G}_n^{-1}(T_i) K_i(x)$$

and

$$\widehat{M_2} = \frac{1}{n\phi_x(h)} \sum_{i=1}^{n} \delta_i \bar{G}_n^{-1}(T_i) K_i^2(x).$$

Furthermore, we get

$$\widehat{\sigma}(x) = \left(\frac{\widehat{M_2}}{\widehat{M_1^2}} \left(\widehat{g}_{2,n}(x) + \widehat{r}_n^2(x) \widehat{g}_{4,n}(x) - 2\widehat{r}_n(x) \widehat{g}_{3,n}(x) \right) \right)^{\frac{1}{2}}.$$

We have approximate $(1-\zeta)$ confidence bands for $r(x)$ given by

$$\left[\widehat{r}_n(x) - t_{1-\frac{\zeta}{2}} \left(\frac{\widehat{\sigma}^2(x)}{n\phi_x(h)} \right)^{\frac{1}{2}}, \quad \widehat{r}_n(x) + t_{1-\frac{\zeta}{2}} \left(\frac{\widehat{\sigma}^2(x)}{n\phi_x(h)} \right)^{\frac{1}{2}} \right],$$

where $t_{1-\frac{\zeta}{2}}$ denotes the $1-\frac{\zeta}{2} \times 100$th quantile of the standard normal distribution.

5. Concluding Remarks

This paper illustrated the asymptotic properties of the regression operator estimator based on the minimization of the mean squared relative error under censoring data. The resulting relative error regression showed to be consistent and asymptotically distributed normally under appropriate conditions in case of censored functional data. Our theoretical and practical studies confirmed that the relative error regression is more efficient than the classical regression.

Appendix

Proof of Theorem 3.1. This is based on the following decomposition

$$\begin{aligned} |\widehat{r}_n(x) - r(x)| = & \frac{1}{\widehat{g}_{2,n}(x)} \left[|\widehat{g}_{1,n}(x) - \widetilde{g}_1(x)| + |\widetilde{g}_1(x) - \mathrm{E}\left[\widetilde{g}_1(x)\right]| + |\mathrm{E}\left[\widetilde{g}_1(x)\right] - g_1(x)| \right] \\ & + \frac{r(x)}{\widehat{g}_{2,n}(x)} \left[|\widetilde{g}_2(x) - \widehat{g}_{2,n}(x)| + |\mathrm{E}\left[\widetilde{g}_2(x)\right] - \widetilde{g}_2(x)| + |g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]| \right]. \quad (10) \end{aligned}$$

Thus, we prove Theorem 3.1 by the following intermediate results

Proof of Lemma 3.2. We have

$$|\mathrm{E}\left[\widetilde{g}_l(x)\right] - g_l(x)| = \left| \frac{1}{n\mathrm{E}[K_1(x)]} \sum_{i=1}^{n} \mathrm{E}[\delta_i \bar{G}^{-1}(T_i) T_i^{-l} K_i(x) - g_l(x)] \right|.$$

By using a double conditioning with respect to Y_i, we get

$$\begin{aligned}
\mathrm{E}\left[\widetilde{g}_l(x)\right] &= \frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}\mathrm{E}[\mathrm{E}(\delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x)|X_i)] \\
&= \frac{1}{\mathrm{E}[K_1(x)]}\mathrm{E}[K(h^{-1}(x-X_1))\mathrm{E}(\delta_1\bar{G}^{-1}(T_1)T_1^{-l}|X_1)] \\
&= \frac{1}{\mathrm{E}[K_1(x)]}\mathrm{E}[K(h^{-1}(x-X_1))\mathrm{E}(\mathrm{E}[\delta_1\bar{G}^{-1}(T_1)T_1^{-l}|Y_1]|X_1)] \\
&= \frac{1}{\mathrm{E}[K_1(x)]}\mathrm{E}[K(h^{-1}(x-X_1))\mathrm{E}(\bar{G}^{-1}(Y_1)Y_1^{-l}\mathrm{E}[\mathrm{I}_{\{Y_1\leq C_1\}}|Y_1]|X_1)].
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{E}\left[\widetilde{g}_l(x)-g_l(x)\right] &= \frac{1}{\mathrm{E}[K_1(x)]}\mathrm{E}\left[K(h^{-1}(x-X_1))\mathrm{I}_{B(x,h)}(X_1)\left|\mathrm{E}(Y_1^{-l}|X_1)-g_l(x)\right|\right] \\
&= \frac{1}{\mathrm{E}[K_1(x)]}\mathrm{E}\left[K(h^{-1}(x-X_1))\mathrm{I}_{B(x,h)}(X_1)|g_l(X_1)-g_l(x)|\right].
\end{aligned}$$

Thus, under conditions (H2), we get

$$\begin{aligned}
|\mathrm{E}\left[\widetilde{g}_l(x)-g_l(x)\right]| &\leq Ch^{k_l} \\
&= O(h^{k_l}).
\end{aligned}$$

Proof of Lemma 3.3. We have for $l=1,2$

$$\widetilde{g}_l(x)-\mathrm{E}\left[\widetilde{g}_l(x)\right] = \frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}\left[\delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x)-\mathrm{E}[\delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x)]\right].$$

Now, we consider

$$Z_{i,l} = \frac{1}{\mathrm{E}[K_1(x)]}\left[\delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x)-\mathrm{E}[\delta_i\bar{G}^{-1}(T_i)T_i^{-l}K_i(x)]\right].$$

To prove this lemma, we use the exponential inequality given in the monograph of Ferraty and Vieu (2006) (Corollary A.8i). We calculate the quantity of $\mathrm{E}[|Z_{i,l}^m|]$ similarly as in Lemma 6.3 of Ferraty and Vieu (2006). By the Newton binomial expansion, we get

$$\begin{aligned}
\mathrm{E}[|Z_{i,l}^m|] &\leq C\sum_{j=0}^{m}\frac{1}{(\mathrm{E}[K_1])^j}\mathrm{E}\left[\left|\delta_1\bar{G}^{-j}(T_1)T_1^{-jl}K_1^j(x)\right|\right] \\
&\leq C\max_{j=0,\ldots,m}\phi_x^{-j+1}(h) \\
&\leq C\phi_x^{-m+1}(h).
\end{aligned}$$

Then,

$$\mathrm{E}[|Z_{i,l}^m|] = O(\phi_x^{-m+1}(h)).$$

Thus, by applying the mentioned exponential inequality with $a^2 = \phi_x^{-1}(h)$, we have, for all $\varepsilon > 0$,

$$\mathrm{P}\left(\left|\sum_{i=1}^{n} Z_{i,l}\right| > \varepsilon n\right) \leq 2\exp\left(\frac{-\varepsilon^2 n}{2a^2(1+\varepsilon)}\right).$$

We establish

$$\varepsilon = \varepsilon_0\sqrt{\frac{\log(n)}{n\phi_x(h)}}.$$

Hence,

$$\begin{aligned}
\mathrm{P}\left(\left|\sum_{i=1}^{n} Z_{i,l}\right| > \varepsilon n\right) &\leq 2\exp\left(\frac{-\varepsilon_0^2\frac{\log(n)}{n\phi_x(h)}n}{2\frac{1}{\phi_x(h)}(1+\varepsilon_0\sqrt{\frac{\log(n)}{n\phi_x(h)}})}\right) \\
&\leq 2\exp\left(-\frac{\varepsilon_0^2\log(n)}{2(1+\varepsilon_0\sqrt{\frac{\log(n)}{n\phi_x(h)}})}\right) \\
&\leq 2\exp\left(-C\varepsilon_0^2\log(n)\right) \\
&\leq 2n^{-C\varepsilon_0^2}.
\end{aligned}$$

Therefore, an appropriate choice of ε_0 and by Proposition A.4. in Ferraty and Vieu (2006), we deduce that

$$|\widetilde{g}_l(x) - \mathrm{E}\left[\widetilde{g}_l(x)\right]| = O\left(\sqrt{\frac{\log(n)}{n\phi_x(h)}}\right) = o(1).$$

Proof of Lemma 3.4. We have

$$\begin{aligned}
|\widehat{g}_{l,n}(x) - \widetilde{g}_l(x)| &= \left|\frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}\delta_i\bar{G}_n^{-1}(T_i)T_i^{-l}K\left(\frac{x-X_i}{h}\right) - \right. \\
&\qquad\qquad \left. \delta_i\bar{G}^{-1}(T_i)T_i^{-l}K\left(\frac{x-X_i}{h}\right)\right| \\
&= \frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}\left|\mathrm{I}_{\{Y_i\leq C_i\}}\bar{G}_n^{-1}(Y_i)Y_i^{-l}K\left(\frac{x-X_i}{h}\right) - \right. \\
&\qquad\qquad \left. \mathrm{I}_{\{Y_i\leq C_i\}}\bar{G}^{-1}(Y_i)Y_i^{-l}K\left(\frac{x-X_i}{h}\right)\right| \\
&\leq \frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}\left|Y_i^{-l}K\left(\frac{x-X_i}{h}\right)\left(\frac{1}{\bar{G}_n(Y_i)} - \frac{1}{\bar{G}(Y_i)}\right)\right| \\
&\leq \frac{\sup_{t\leq t_F}|\bar{G}_n(t) - \bar{G}(t)|}{\bar{G}_n(t_F)\bar{G}(t_F)}\frac{1}{n\mathrm{E}[K_1(x)]}\sum_{i=1}^{n}Y_i^{-l}K\left(\frac{x-X_i}{h}\right).
\end{aligned}$$

By using conditional expectation, we obtain

$$|\widehat{g}_{l,n}(x) - \widetilde{g}_l(x)| \leq \frac{\sup_{t\leq t_F} |\bar{G}_n(t) - \bar{G}(t)|}{\bar{G}_n(t_F)\bar{G}(t_F)} \frac{1}{n\mathrm{E}[K_1(x)]} \sum_{i=1}^{n} \mathrm{E}\left[Y_i^{-l} K\left(\frac{x - X_i}{h}\right) |X_i\right].$$

Under conditions (H3), (H5) and by taking into account formula (4.28) in Deheuvels and Einmahl (2000), we get

$$|\widehat{g}_{l,n}(x) - \widetilde{g}_l(x)| = O\left(\frac{\log(\log(n))}{n}\right).$$

Proof of Corollary 3.5. We have

$$\mathrm{P}\left(\lim_{n\to\infty} \widehat{g}_{2,n}(x) = g_2(x)\right) = 1.$$

By taking into account the results of Lemmas 3.2-3.4, we prove the corollary.

Proof of Lemma 3.7. We use the same arguments as in Lemma 7 of Demongeot et al. (2016) for censored data.

Let

$$\frac{\sqrt{n\phi_x(h)}}{g_2^2(x)\sigma(x)} \left([\widetilde{g}_1(x) - \mathrm{E}[\widetilde{g}_1(x)]] g_2(x) + [\mathrm{E}[\widetilde{g}_2(x)] - \widetilde{g}_2(x)] g_1(x)\right) = \frac{S_n}{g_2^2(x)\sigma(x)},$$

with $S_n = \sum_{i=1}^{n} (L_i(x) - \mathrm{E}[L_i(x)])$, where

$$L_i(x) = \frac{\sqrt{n\phi_x(h)}}{n\mathrm{E}[K_1]} \delta_i \bar{G}^{-1}(T_i) K_i(x) \left(g_1(x) T_i^{-2} - g_2(x) T_i^{-1}\right).$$

We apply the Lyapunov central limit theorem on $L_i(x)$ for showing the asymptotic normality of S_n. It suffices to show, for some $\delta > 0$, that

$$\frac{\sum_{i=1}^{n} \mathrm{E}\left[|L_i(x) - \mathrm{E}[L_i(x)]|^{2+\delta}\right]}{\left(var\left(\sum_{i=1}^{n} L_i(x)\right)\right)^{\frac{2+\delta}{2}}} \to 0. \tag{11}$$

Clearly,

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{n} L_i(x)\right) &= n\phi_x(h)\mathrm{Var}\left[\widetilde{g}_2(x)g_1(x) - \widetilde{g}_1(x)g_2(x)\right] \\ &= n\phi_x(h)\left[\mathrm{Var}\left(\widetilde{g}_2(x)\right) g_1^2(x) + \mathrm{Var}\left(\widetilde{g}_1(x)\right) g_2^2(x) - 2g_1(x)g_2(x)\mathrm{Cov}(\widetilde{g}_1(x), \widetilde{g}_2(x))\right].\end{aligned}$$

Thus, for $l = 1, 2$, we obtain

$$\begin{aligned}\mathrm{Var}\left(\widetilde{g}_l(x)\right) &= \frac{1}{(n\mathrm{E}[K_1])^2} \sum_{i=1}^{n} \mathrm{Var}\left[\delta_i \bar{G}^{-1}(T_i) T_i^{-l} K_i(x)\right] \\ &= \frac{1}{n(\mathrm{E}[K_1])^2} \mathrm{Var}\left[\delta_1 \bar{G}^{-1}(T_1) T_1^{-l} K_1(x)\right].\end{aligned}$$

By conditioning on the random variable X, using hypotheses $(C1)$ and $(C3)$ and the fact that

$$\mathrm{E}[K_1] = \phi_x(h)\left(K(1) - \int_0^1 K^{'}(s)\chi_x(s)\mathrm{d}s\right) + o(\phi_x(h)),$$

we get

$$\begin{aligned}\mathrm{E}\left[\delta_1\bar{G}^{-2}(T_1)T_1^{-2l}K_1^2(x)\right] &= \mathrm{E}\left[K_1^2(x)\mathrm{E}\left[\bar{G}^{-1}(Y)Y^{-2l}|X=x\right]\right]\\ &= \mathrm{E}\left[\bar{G}^{-1}(Y)Y^{-2l}|X=x\right]\\ &\quad\times\left(\phi_x(h)\left(K^2(1) - \int_0^1 (K^2)^{'}(s)\chi_x(s)\mathrm{d}s\right) + o(\phi_x(h))\right).\end{aligned}$$

By a double conditioning on the random variable X and under conditions (H3) and (H5), we obtain

$$\begin{aligned}\mathrm{E}\left[\delta_1\bar{G}^{-1}(T_1)T_1^{-l}K_1(x)\right] &= \mathrm{E}\left[K_1(x)\mathrm{E}\left[Y_1^{-1}|X=x\right]\right]\\ &\leq C\mathrm{E}[K_1]\\ &\leq C\phi_x(h).\end{aligned}$$

Therefore,

$$\left(\mathrm{E}\left[\delta_1\bar{G}^{-1}(T_1)T_1^{-l}K_1(x)\right]\right)^2 = O(\phi_x(h)^2).$$

Then,

$$\begin{aligned}\mathrm{Var}\left[\delta_1\bar{G}^{-1}(T_1)T_1^{-l}K_1(x)\right] &= \mathrm{E}\left[\bar{G}^{-1}(Y)Y^{-2l}|X=x\right]\\ &\quad\times\left(\phi_x(h)\left(K^2(1) - \int_0^1 (K^2)^{'}(s)\chi_x(s)\mathrm{d}s\right)\right) + O(\phi_x(h)^2).\end{aligned}$$

Thus,

$$\mathrm{Var}\left(\widetilde{g_l}(x)\right) = \frac{\mathrm{E}\left[\bar{G}^{-1}(Y)Y^{-2l}|X=x\right]\left(K^2(1) - \int_0^1 (K^2)^{'}(s)\chi_x(s)\mathrm{d}s\right)}{n\phi_x(h)\left(K(1) - \int_0^1 K^{'}(s)\chi_x(s)\mathrm{d}s\right)^2} \tag{12}$$

$$+\ o\left(\frac{1}{n\phi_x(h)}\right). \tag{13}$$

Now, we calculate the corresponding covariance as

$$\begin{aligned}\mathrm{Cov}(\widetilde{g_1}(x),\widetilde{g_2}(x)) &= \frac{1}{n(\mathrm{E}[K_1])^2}\mathrm{Cov}\left(\delta_1\bar{G}^{-1}(T_1)T_1^{-1}K_1(x), \delta_1\bar{G}^{-1}(T_1)T_1^{-2}K_1(x)\right)\\ &= \frac{1}{n(\mathrm{E}[K_1])^2}\left[\mathrm{E}\left(\delta_1\bar{G}^{-2}(T_1)T_1^{-3}K_1^2(x)\right)\right.\\ &\quad\left.-\ \mathrm{E}\left(\delta_1\bar{G}^{-1}(T_1)T_1^{-1}K_1(x)\right)\mathrm{E}\left(\delta_1\bar{G}^{-1}(T_1)T_1^{-2}K_1(x)\right)\right]\end{aligned}$$

where

$$\mathrm{E}\left(\delta_1\bar{G}^{-2}(T_1)T_1^{-3}K_1^2(x)\right) = \mathrm{E}\left[K_1^2\mathrm{E}\left[\bar{G}^{-1}Y^{-3}|X=x\right]\right]$$
$$= \mathrm{E}\left[\bar{G}^{-1}Y^{-3}|X=x\right]\left(K^2(1) - \int_0^1 (K^2)'(s)\chi_x(s)\mathrm{d}s\right) + o(1).$$

Hence,

$$\mathrm{Cov}(\widetilde{g_1}(x),\widetilde{g_2}(x)) = \frac{\mathrm{E}\left[\bar{G}^{-1}Y^{-3}|X=x\right]\left(K^2(1) - \int_0^1 (K^2)'(s)\chi_x(s)\mathrm{d}s\right)}{n\phi_x(h)\left(K(1) - \int_0^1 K'(s)\chi_x(s)\mathrm{d}s\right)^2} + o\left(\frac{1}{n\phi_x(h)}\right).$$

It follow that

$$\mathrm{Var}\left(\sum_{i=1}^n L_i(x)\right) = g_2^2(x)\sigma + o(1).$$

Therefore, it is sufficient to demonstrate that the numerator of (11) converges to 0 to finish the evidence of this lemma. For that we apply the C_r inequality (see Loeve (1963), p. 155) showing that

$$\sum_{i=1}^n \mathrm{E}\left[|L_i(x) - \mathrm{E}\left[L_i(x)\right]|^{2+\delta}\right] \leq C\sum_{i=1}^n \mathrm{E}\left[|L_i(x)|^{2+\delta}\right] + C'\sum_{i=1}^n |\mathrm{E}\left[L_i(x)\right]|^{2+\delta}.$$

Then, under assumptions (H5) and (H3), we get

$$\sum_{i=1}^n \mathrm{E}\left[|L_i(x)|^{2+\delta}\right] = n^{\frac{-\delta}{2}}(\phi_x(h))^{-1-\frac{\delta}{2}}\mathrm{E}\left[\delta_1^{2+\delta}\bar{G}^{-(2+\delta)}(T_1)K_1^{2+\delta}(x)|g_1(x)T_i^{-2} - g_2(x)T_i^{-1}|^{2+\delta}\right]$$
$$\leq C(n\phi_x(h))^{-1-\frac{\delta}{2}}\left(\mathrm{E}[K_1^{2+\delta}]\right) \to 0.$$

For the second term, we obtain

$$\sum_{i=1}^n |\mathrm{E}\left[L_i(x)\right]|^{2+\delta} \leq n^{\frac{-\delta}{2}}(\phi_x(h))^{-1-\frac{\delta}{2}}\left|\mathrm{E}\left[\delta_1\bar{G}^{-1}(T_1)K_1(x)|g_1(x)T_i^{-2} - g_2(x)T_i^{-1}|\right]\right|^{2+\delta}$$
$$\leq Cn^{\frac{-\delta}{2}}\left(\phi_x(h)\right)^{\frac{1+\delta}{2}} \to 0$$

which finishes the proof.

Proof of Lemma 3.8. For the first term, by taking into account Lemmas 3.2-3.4 and equation (12), we have

$$\mathrm{E}\left[\widetilde{g_2}(x) - g_2(x)\right] \to 0$$

and

$$\mathrm{Var}\left[\widetilde{g_2}(x)\right] \to 0.$$

Then,

$$\widehat{g}_{2,n}(x) - g_2(x) \to 0,$$

in probability. For the second limit, by lemma 3.4 and first limit, we get

$$\mathrm{Var}\left[\widehat{g}_{l,n}(x) - \widetilde{g}_l(x)\right] \to 0.$$

Thus, it follow that

$$\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} \left[\left(\widehat{g}_{1,n}(x) - \widetilde{g}_1(x)\right) g_2(x) + \left(\widetilde{g}_2(x) - \widehat{g}_{2,n}(x)\right) g_1(x)\right] \to 0,$$

in probability.

Proof of Lemma 3.9. We write

$$\begin{aligned}
&\left[\frac{1}{g_2(x)}\left(\mathrm{E}\left[\widetilde{g}_1(x)\right] - g_1(x)\right) g_2(x) + \left(g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]\right) g_1(x)\right] \\
&= \frac{1}{g_2(x)}\left[\mathrm{E}\left[\widetilde{g}_1(x)\right] g_2(x) - g_1(x)g_2(x) + g_1(x)g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right] g_1(x)\right] \\
&= \frac{1}{g_2(x)\mathrm{E}\left[\widetilde{g}_2(x)\right]}\left[\mathrm{E}\left[\widetilde{g}_1(x)\right] g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right] g_1(x)\right]\mathrm{E}\left[\widetilde{g}_2(x)\right] \\
&= A_n \mathrm{E}\left[\widetilde{g}_2(x)\right].
\end{aligned}$$

For A_n, we get

$$A_n = \frac{\mathrm{E}\left[\widetilde{g}_1(x)\right]}{\mathrm{E}\left[\widetilde{g}_2(x)\right]} - \frac{g_1(x)}{g_2(x)},$$

for which suffices to evaluate $\mathrm{E}\left[\widetilde{g}_1(x)\right]$ and $\mathrm{E}\left[\widetilde{g}_2(x)\right]$. By the same arguments used in Lemma 3.2, we obtain

$$\mathrm{E}\left[\widetilde{g}_1(x)\right] = \frac{1}{\mathrm{E}\left[K_1\right]}\mathrm{E}\left[K_1(x)\mathrm{E}\left[Y_1^{-1}|X_1\right]\right]$$

and

$$\mathrm{E}\left[\widetilde{g}_2(x)\right] = \frac{1}{\mathrm{E}\left[K_1\right]}\mathrm{E}\left[K_1(x)\mathrm{E}\left[Y_1^{-2}|X_1\right]\right].$$

By the same ideas used by Ferraty et al. (2007) for regression operator, we demonstrate that

$$\mathrm{E}\left[\widetilde{g}_1(x)\right] = g_1(x) + h\Psi_1^{'}(0)\left[\frac{K(1) - \int_0^1 (sK(s))^{'}\chi_x(s)\mathrm{d}s}{K(1) - \int_0^1 (K)^{'}(s)\chi_x(s)\mathrm{d}s}\right] + o(h)$$

and

$$\mathrm{E}\left[\widetilde{g}_2(x)\right] = g_2(x) + h\Psi_2^{'}(0)\left[\frac{K(1) - \int_0^1 (sK(s))^{'}\chi_x(s)\mathrm{d}s}{K(1) - \int_0^1 (K)^{'}(s)\chi_x(s)\mathrm{d}s}\right] + o(h).$$

Thus,

$$A_n = \frac{\mathrm{E}\left[\widetilde{g}_1(x)\right]}{\mathrm{E}\left[\widetilde{g}_2(x)\right]} - r(x) = hB_n(x) + o(h),$$

where

$$B_n = \frac{(\Psi_1^{'}(0) - r(x)\Psi_2^{'}(0))M_0}{M_1 g_2(x)}.$$

For the second term, we have

$$\mathrm{E}\left[\widetilde{g}_2(x)\right] = g_2(x) + h\Psi_2^{'}(0)\left[\frac{K(1) - \int_0^1 (sK(s))^{'}\chi_x(s)\mathrm{d}s}{K(1) - \int_0^1 (K)^{'}(s)\chi_x(s)\mathrm{d}s}\right] + o(h).$$

Then,

$$\mathrm{E}\left[\widetilde{g}_2(x)\right] - g_2(x) = O(h).$$

Hence, to show that Lemma 3.9 converges to 0 in probability, we have

$$\mathrm{E}\left[\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} A_n\left(|g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]|\right)\right] = 0$$

and

$$\mathrm{Var}\left[\left(\frac{n\phi_x(h)}{g_2^2(x)\sigma^2(x)}\right)^{\frac{1}{2}} A_n\left(|g_2(x) - \mathrm{E}\left[\widetilde{g}_2(x)\right]|\right)\right] = O(A_n^2) = O(h^2) \to 0,$$

which complete the proof.

References

Altendji, B., Demongeot, J., Laksaci, A., and Rachdi, M., 2018. Functional data analysis: estimation of the relative error in functional regression under random left truncation model. Journal of Nonparametric Statistics, 30, 1-19.

Aneiros, G., Bongiorno, E.G., Cao, R., and Vieu, P., 2017. Functional Statistics and Related Fields. Springer, Cham.

Attouch, M., Laksaci, A., and Ould-Said, E., 2009. Asymptotic distribution of robust estimator for functional nonparametric models. Communications in Statistics: Theory and Methods, 38, 1317-1335.

Deheuvels, P. and Einmahl, J.H.J., 2000. Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications. The Annals of Probability, 28, 1301-1335.

Demongeot, J., Hamie, A., Laksaci, A., and Rachdi, M., 2016. Relative-error prediction in nonparametric functional statistics: Theory and practice. Journal of Multivariate Analysis, 146, 261-268.

Ferraty, F., Mas, A., and Vieu, P., 2007. Nonparametric regression on functional data: Inference and practical aspects. Australian and New Zealand Journal of Statistics, 49, 267-286.

Ferraty, F. and Vieu, P., 2004. Nonparametric models for functional data, with application in regression times series prediction and curves discrimination. Journal of Nonparametric Statistics, 16, 111-127.

Ferraty, F. and Vieu, P., 2006. Nonparametric Functional Data Analysis. Theory and Practice. Springer, New York.

Gheriballah, A., Laksaci, A., and Sekkal, S., 2013. Nonparametric M-regression for functional ergodic data. Statistics and Probability Letters, 83, 902-908.

Horrigue, W. and Ould-Said, E., 2011. Strong uniform consistency of a nonparametric estimator of a conditional quantile for censored dependent data and functional regressors. Random Operators and Stochastic Equations, 19, 131-156.

Horrigue, W. and Ould-Said, E., 2014. Nonparametric regression quantile estimation for dependent functional data under random censorship: Asymptotic normality. Communications in Statistics: Theory and Methods, 44, 4307-4332.

Hsing, T. and Eubank, R., 2015. Theoretical foundations of functional data analysis with an introduction to linear operators. Wiley, Chichester.

Jones, M.C., Park, H., Shin, K-Il., Vines, S.K., and Jeong, S.O., 2008. Relative error prediction via kernel regression smoothers. Journal of Statistical Planning and Inference, 138, 2887-2898.

Kaplan, E.L. and Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of American Statistical Association, 53, 457-481.

Khardani, S., Lemdani, M., and Ould-Said, E., 2010. Some asymptotic properties for a smooth kernel estimator of the conditional mode under random censorship. Journal of the Korean Statistical Society, 39, 455-469.

Kohler, M., Màthé K., and Pintér, M., 2002. Prediction from randomly right censored data. Journal of Multivariate Analysis 80, 73-100.

Loeve, M., 1963. Probability Theory. Van Nostrand, Princeton.

Louzada, F., Shimizu, T.K.O., Suzuki, A.K., Mazucheli, J., and Ferreira, P.H., 2018. Compositional regression modeling under tilted normal errors: An application to a Brazilian super league volleyball data set. Chilean Journal of Statistics, 9, 33-53.

Mechab, W. and Laksaci, A., 2016. Nonparametric relative regression for associated random variables. Metron, 74, 75-97.

Ould-Said, E., Ouassou, I., and Rachdi, M., 2015. Functional Statistics and Applications. Springer, Switzerland.

Park, H. and Stefanski, L.A., 1998. Relative-error prediction. Statistics and Probability Letters, 40, 227-236.

Ramsay, J.O. and Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.

Rinnan, R. and Rinnan, A.A., 2007. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. Soil Biology and Biochemistry, 39, 1664-1673.

Information for authors

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF electronic version of the manuscript in a free format to Víctor Leiva, Editor-in-Chief of the ChJS (E-mail: `chilean.journal.of.statistics@gmail.com`). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a"cover letter" presenting their manuscript and mentioning: "We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere".

Preparation of accepted manuscripts

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at `http://chjs.mat.utfsm.cl/files/ChJS.zip`. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at `http://www.ams.org/mathscinet/msc/`. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

References in the text must be given by the author's name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

Copyright