ISSN 1645-6726

REVSTAT

Statistical Journal

vol. 20 - no. 4 - July 2022

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT-Statistical Journal, vol.20, n. 4 (July 2022)

vol.1, 2003-   . - Lisbon : Statistics Portugal, 2003-   .

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Publisher – *Statistics Portugal*

Layout-Graphic Design – *Carlos Perpétuo* | **Cover Design*** – *Helena Nogueira*

Edition - 140 copies | **Legal Deposit Registration** - 191915/03 | **Price** [VAT included] - € 9,00

*image: stain glass window by Abel Manta (1888-1982)

Editorial

In memoriam - Julio Singer [1950-2022]

We regretfully announce the passing of **Julio da Motta Singer**, 71, of São Paulo - Brazil, on May 25th 2022. He was full professor of the Department of Statistics in the Institute of Mathematics and Statistics (IME) at the University of Sâo Paulo (USP) and our colleague at REVSTAT as Associate Editor since 2014. He holds a degree in Production Mechanical Engineering from the University of São Paulo (1973), a MSc in Statistics from the University of São Paulo (1977) and a PhD in Biostatistics from the University of North Carolina (1983).

His scientific works focus on Statistics, especially on research topics involving categorized data, longitudinal data, linear models, large sample theory and mixed models. **Julio Singer** contributed to both the training of several undergraduate and graduate students and the development of Statistics in Brazil, and beyond.

One of **Julio Singer**'s passions was to develop statistical methodology to deal with applied studies involving real data. Hence, he contributed decisively to the consolidation of the Center for Applied Statistics at IME-USP. On his website, he has always made some of these datasets available to motivate the statistical community to produce and improve their methods with motivations of general interest, as well as their corresponding computational codes/routines.

In addition to his involvement in the teaching of Statistics at USP, **Julio Singer** left several theoretical and applied statistics books and articles published in international journals, namely with our collaboration, and several statistical advisory works carried out for researchers and companies from different scientific areas. He usually liked to spread his findings at scientific meetings, including those promoted by Statistical Brazilian Association, from which he received the Career Award in 2018, and Statistical Portuguese Society.

Finally, here is our thanks for **Julio Singer**'s legacy left to Statistics and particularly his work with REVSTAT, always done in a serious and robust way but with his perceptive humour!

May 25, 2022

Carlos Daniel Paulino (*Associate Editor*)
Giovani L. Silva (*Co-Editor*)

INDEX

Abstracted/indexed in: *Journal Citation Reports™ – JCR (Clarivate), Current Index to Statistics, Google Scholar, Mathematical Reviews® (MathSciNet®), Zentralblatt für Mathematic, SCImago Journal & Country Rank* and *Scopus®*.

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 405–426
https://doi.org/10.57805/revstat.v20i4.379

Time Series Analysis for Longitudinal Survey Data under Informative Sampling and Nonignorable Missingness

Authors: Zhan Liu ✉
– Hubei Key Laboratory of Applied Mathematics, School of Mathematics and Statistics, Hubei University,
Wuhan, 430062, China
20170066@hubu.edu.cn

Chun Yip Yau
– Department of Statistics, Chinese University of Hong Kong,
Hong Kong, China
cyyau@sta.cuhk.edu.hk

Received: January 2020 Revised: June 2020 Accepted: August 2020

Abstract:

- The analysis of longitudinal survey data is often complicated when informative sampling or nonignorable missing data exists. Existing methods that can handle both informative sampling and nonignorable missing data are only limited to the situation of no time dependence in the data. In this paper, we develop a sample likelihood based approach for estimation of time series model in longitudinal survey data under informative sampling and nonignorable missingness. In particular, some informative sampling models and a response model are proposed to describe the mechanisms of informative sampling and nonignorable missingness. A sample likelihood is derived based on the conditional distribution of the observed measurements. Also, an effective computation algorithm is developed to compute the sample likelihood. Simulation studies are carried out to investigate the performance of the proposed estimator. A real data example based on data from AIDS Clinical Trial Group 193A Study is presented to illustrate the proposed method.

Keywords:

- *autoregressive model; exponential model; probit model; logistic model; sample likelihood.*

AMS Subject Classification:

- 62D05, 62F10.

✉ Corresponding author.

1. INTRODUCTION

Longitudinal surveys are designed to measure a sample of respondents repeatedly over time, and have been extensively applied in various fields such as clinical studies, biological research and social sciences. Longitudinal surveys are prevalence in studying human's behaviors, health, and mortality because they provide efficient means to estimate the change in the population, evaluate interventions, test causal hypotheses, and reduce the cost of data collection [35]. Since longitudinal surveys are conducted at different points of time, the serial observations obtained from a given unit usually show time dependence. Therefore, a time series model can be employed to analyze longitudinal survey data [12].

Informative sampling, which refers to sampling design in which the sampling probabilities are correlated with the response variable (conditional on covariates), is often encountered in longitudinal surveys, see, e.g., Fuller [15]. However, studies ignoring informative sampling can lead to seriously biased results (Pfeffermann [27], [26]; Eideh and Nathan [12]; Eideh [9]; Sverchkov and Pfeffermann [33]). To handle informative sampling, Pfeffermann *et al.* [25] derived the sample distribution from the population distribution and the sampling probabilities under informative sampling, which can permit the use of classical inference methods. Chambers and Skinner [7], and Pfeffermann and Sverchkov [24] discussed the sample likelihood approach, the pseudo-likelihood approach and the estimating equations approach for fitting generalized linear models under informative sampling, based on the sample distribution of Pfeffermann *et al.* [25]. In fact, the sample likelihood approach has been explored in many different directions including small area estimation (Pfeffermann and Sverchkov [22]; Eideh and Nathan [11]; Verret *et al.* [37]), general linear modelling (Chambers and Skinner [7]; Pfeffermann and Sverchkov [22]; Eideh [9]), and multi-level model analysis (Pfeffermann *et al.* [23]; Cai [6]). Recently, Bonnery *et al.* [4] established the asymptotic properties of the sample likelihood approach under informative sampling. Other proposed methods include the inverse probability weighting method (Boudreau and Lawless [5]; Kim and Skinner [17]) and calibration adjustments (Moser *et al.* [20]). However, most of the above studies explored the informative sampling problem in the non-longitudinal survey context. Informative sampling in longitudinal surveys was considered in Eideh and Nathan [12], [13], and Eideh [9]. Eideh and Nathan [12], [13] discussed the sample likelihood and pseudo-likelihood methods in fitting time series models for longitudinal survey data under informative sampling. Eideh [9] explored further the sample likelihood, pseudo-likelihood likelihood and estimating equations methods in fitting general linear model for longitudinal survey data under informative sampling.

In addition to informative sampling, another major issue in longitudinal surveys is the missing data problem. Following Little and Rubin [18], the mechanisms of missing data can be classified into three types: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). In particular, missing completely at random and missing at random are called ignorable missingness, whereas not missing at random is called nonignorable missingness. Under nonignorable missingness, the missing probability depends on the response variable, and thus will lead to unreliable estimation results (Eideh [9]; Schlomer *et al.* [30]; Taisir and Islam [34]). A solution to this problem is the modeling of nonignorable missing data, which has been applied to general linear models (Bahari *et al.* [2]), generalized linear mixed models (Stubbendick and Ibrahim [32]; Sabry *et al.* [29];

Almohisen *et al.* [1]), quantile regression models (Yuan and Yin [38]), latent random effects models (Tseng *et al.* [36]; Bhuyan [3]), and Markov chain models (Cole *et al.* [8]; Taisir and Islam [34]).

When informative sampling and nonignorable missingness occur in longitudinal surveys simultaneously, the joint treatment of the two problems becomes a key issue. Pfeffermann [21] proposed a unified approach to handle the two problems by combining the observed data model with the missing data model and the target population model based on the Bayes theorem. Sverchkov and Pfeffermann [33] extended the approach in Pfeffermann and Sverchkov [22] in small area estimation under informative sampling to the case that both informative sampling and nonignorable missingness exist. However, these approaches only considered data measured at a certain time point and are not applicable to longitudinal data. Eideh and Nathan [10], and Farahania *et al.* [14] considered methods to handle informative sampling and nonignorable missingness simultaneously in longitudinal data analysis. However, their discussions focus mainly on general regression models.

In this paper, we study time series modeling for longitudinal survey data under informative sampling and nonignorable missingness. Treating informative sampling and nonignorable missingness simultaneously becomes especially challenging in time series models due to the serial correlation of the response variable at various time points. We consider models to explore the effect of each of informative sampling and nonignorable missingness. For informative sampling, a variety of models, including exponential, probit, and logistic models are considered to capture the dependence between the selection probability and the response variable. For nonignorable missingness, we consider a logistic model to relate the response probability to the response variables. Based on these models, we derive a sample likelihood for parameter estimation under informative sampling and nonignorable missingness. To compute the sample likelihood function efficiently, an approximation to the integrals in the sample likelihood based on series expansions is proposed. Simulation studies and real data application are provided to illustrate the effectiveness of the proposed method.

The remainder of the paper is organized as follows. Section 2 describes time series models and parameter estimation methods for longitudinal survey data. Section 3 discusses informative sampling and nonignorable missingness in longitudinal surveys. In Section 4, the sample likelihood is derived for conducting time series analysis in longitudinal survey data under informative sampling and nonignorable missingness. Simulations studies and real data analysis are performed in Sections 5 and 6, respectively.

2. TIME SERIES MODEL FOR LONGITUDINAL SURVEY DATA

Let $U = \{1, ..., N\}$ be the index set of a finite population U of size N. Let $y_{i,t}$ $(i = 1, ..., N,$ $t = 1, ..., T)$ be the value of a response variable y of unit i at time t, and x_i be the values of the covariates of unit i, which are always observed and remain constant over time. A random sample S of size n is then selected from the finite population at time 1 $(t = 1)$ and measured independently from time 1 to time T. Suppose that $y_{i,t}$ is correlated with the past values $y_{i,t'}$, $1 \leq t' < t \leq T$, for each T. A time series model can then be fitted to analyze this longitudinal survey data. Typically, time series models with short-range dependence are often applied in decision-making and policymaking [12]. For simplicity, we consider the first-order autoregressive (AR(1)) model

$$y_{i,t} - \mu = \phi(y_{i,t-1} - \mu) + \varepsilon_{i,t}\,, \qquad i = 1, ..., N, t = 1, ..., T\,, \tag{2.1}$$

where μ is the mean level of the data, the errors $\varepsilon_{i,t} \overset{\text{iid}}{\sim} N(0, \sigma^2)$, and $|\phi| < 1$. The model parameter $\theta = (\mu, \phi, \sigma)$ is of our interest. Note that unit i in the AR(1) model will fall into the set $\{1, ..., n\}$ when the sample data is used to estimate the model parameters.

Usually, the maximum likelihood estimation approach is employed to obtain the model parameter estimators. Let $\mathbf{y}_i = (y_{i,1}, ..., y_{i,T})'$ be the vector of T measurements on unit i ($i = 1, ..., N$). Then, the density function of $\mathbf{y}_i$ can be expressed as $f(\mathbf{y}_i; \theta) = f(y_{i,1}; \theta) \cdot \cdot \prod_{t=2}^{T} f(y_{i,t}|y_{i,t-1}; \theta)$. For the AR(1) model, we have $y_{i,1} \sim N(\mu, \sigma^2/(1 - \phi^2))$ and $f(y_{i,t}|y_{i,t-1}; \theta) = (2\pi\sigma^2)^{-1/2} \exp\{-[y_{i,t} - \phi(y_{i,t-1} - \mu) - \mu]^2/(2\sigma^2)\}$. Thus, the log-likelihood function of θ can be written as

$$\log L(\theta) \;=\; \sum_{i=1}^{n} \log f(y_{i,1}; \theta) + \sum_{i=1}^{n} \sum_{t=2}^{T} \log f(y_{i,t}|y_{i,t-1}; \theta)\,. \tag{2.2}$$

It follows that the maximum likelihood estimator of θ can be obtained by maximizing the log-likelihood function in (2.2).

3. INFORMATIVE SAMPLING AND NONIGNORABLE MISSINGNESS IN LONGITUDINAL SURVEYS

3.1. Informative sampling

Analytic inference from longitudinal survey data usually fails to account for the complex sampling design, such as informative sampling. A sampling design is called informative when the sample selection probabilities are related to the response variable y, even after conditioning on the covariates. In practice, selection probabilities may be correlated with the response variable, the covariates and possibly, design variables used for sampling. For simplicity, we consider the case that selection probabilities depend only on the response variable.

Let I_i be the sample indicator variable, taking values of 1 if unit $i \in U$ is selected to the sample S and 0 if otherwise. The selection probabilities can then be denoted by $\pi_i = P(I_i = 1|y_i)$. Let $f_s(y_i)$ and $f_p(y_i)$ denote the sample density and the population density of y_i, respectively. In fact, the density functions $f(y_{i,1}; \theta)$ and $f(y_{i,t}|y_{i,t-1}; \theta)$ in Section 2 are the population densities, which can also be denoted by $f_p(y_{i,1}; \theta)$ and $f_p(y_{i,t}|y_{i,t-1}; \theta)$, respectively. Following Pfeffermann *et al.* [25] as well as Sikov and Stern [31], the sample density $f_s(y_i)$ is given by

$$\begin{aligned} f_s(y_i) \;&=\; f(y_i|I_i = 1) \;=\; \frac{f(y_i, I_i = 1)}{P(I_i = 1)} \\ &=\; \frac{P(I_i = 1|y_i) f_p(y_i)}{P(I_i = 1)} \;=\; \frac{E_p(\pi_i|y_i) f_p(y_i)}{E_p(\pi_i)}\,, \end{aligned} \tag{3.1}$$

where $\pi_i = P(I_i = 1|y_i)$, $E_p(\pi_i|y_i) = \int P(I_i = 1|y_i, \pi_i) f_p(\pi_i|y_i) d\pi_i = P(I_i = 1|y_i)$, and $E_p(\pi_i)$ $= \int P(I_i = 1|y_i) f_p(y_i) dy_i = P(I_i = 1)$. Under informative sampling, the selection probability $\pi_i = P(I_i = 1|y_i)$ depends on y_i. Hence, $E_p(\pi_i|y_i) \neq E_p(\pi_i)$ and $P(I_i = 1|y_i) \neq P(I_i = 1)$, yielding $f_s(y_i) \neq f_p(y_i)$ in general. That is, the sample distribution is different from the population distribution. However, the sample distribution is viewed as the same as the population distribution in many analysis under informative sampling, which have resulted in false inferences (Pfeffermann [27], [26]).

In order to access the sample density, $E_p(\pi_i|y_i) = P(I_i = 1|y_i)$ can be modeled to explore the relationship between the selection probabilities π_i and the response variable values y_i. Pfeffermann *et al.* [25] and Eideh and Nathan [12] considered

$$\text{Exponential model:}\quad E_p(\pi_i|y_i) = \exp(a_0 + a_1 y_i)\,, \tag{3.2}$$

where a_0 and a_1 are unknown model parameters. Besides, the probit model and logistic model, which are less common in longitudinal surveys under informative sampling, can also be explored to explain the informative sampling mechanism:

$$\text{Probit model:}\quad E_p(\pi_i|y_i) = \Phi(b_0 + b_1 y_i)\,, \tag{3.3}$$

$$\text{Logistic model:}\quad E_p(\pi_i|y_i) = \frac{\exp(c_0 + c_1 y_i)}{1 + \exp(c_0 + c_1 y_i)}\,, \tag{3.4}$$

where b_0, b_1, c_0, c_1 are unknown model parameters.

3.2. Nonignorable missingness

Missing data is another problem which often arises in longitudinal surveys. Here, we assume that there exists nonignorable missingness in longitudinal surveys. In particular, the values $y_{i,1}$ at time 1 are complete and some of $y_{i,2}, ..., y_{i,T}$ suffer from missingness for $i = 1, ..., n$. Denote the response indicator variable by

$$\delta_{i,t} = \begin{cases} 1 & \text{if } y_{i,t} \text{ is observed}\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.5}$$

The nonignorable missingness implies that missingness depends on the response variable. In other words, the response probability is related to the response variable. Under the AR(1) model, we model the response mechanism using a logistic model

$$\begin{aligned} P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}) &=: \pi(x_i, y_{i,t-1}, y_{i,t}; \eta) \\ &= \frac{\exp(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 y_{i,t})}{1 + \exp(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 y_{i,t})}\,, \end{aligned} \tag{3.6}$$

where $\eta = (\eta_1, \eta_2, \eta_3)$ is the unknown parameter. Equation (3.6) asserts that the response probability $P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t})$ at time t depends not only on the value $y_{i,t}$ at time t and the covariate x_i, but also on its past value $y_{i,t-1}$. Clearly, the response mechanism is nonignorable missingness. Note that (3.6) extends the nonignorable response mechanism in Qin *et al.* [28] by incorporating the effect of past observations into the response probability.

For notational simplicity, only one covariate x is considered in the response model. The extension to multiple covariates $x_1, ..., x_p$ in the response model is straightforward.

If we ignore the informative sampling and nonignorable missingness, using the complete case (CC) analysis (Farahania *et al.* [14]), the log-likelihood function of θ in the AR(1) model based on the observed data is rewritten as

$$
\begin{aligned}
\log L(\theta) &= \sum_{i=1}^{n} \log f(y_{i,1};\theta) + \sum_{t=2}^{T}\sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t} \log f(y_{i,t}|y_{i,t-1};\theta) \\
&= \sum_{i=1}^{n}\left\{-\frac{1}{2}\log\left(\frac{2\pi\sigma^2}{1-\phi^2}\right) - \frac{(1-\phi^2)(y_{i,1}-\mu)^2}{2\sigma^2}\right\} \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t}\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}[y_{i,t}-\phi(y_{i,t-1}-\mu)-\mu]^2\right\}.
\end{aligned}
\tag{3.7}
$$

Then, we can get the maximum likelihood estimator $\hat{\theta}$ of θ via maximizing the log-likelihood function in (3.7). However, the obtained estimator $\hat{\theta}$ is obviously biased because it ignores the informative sampling and nonignorable missingness (Pfeffermann *et al.* [25]; Little and Rubin [18]; Farahania *et al.* [14]). In fact, the observed sample distribution is different from the population distribution under both informative sampling and nonignorable missingness, which cannot guarantee that the log-likelihood function in (3.7) gives the correct estimates.

4. SAMPLE LIKELIHOOD AND ESTIMATION UNDER INFORMATIVE SAMPLING AND NONIGNORABLE MISSINGNESS

4.1. Sample likelihood under informative sampling

The sample distribution differs from the population distribution under informative sampling. Therefore, the sample likelihood will be different from the general likelihood under noninformative sampling. Because the sample is only selected from the finite population at time 1 in longitudinal surveys, the sample distribution at time 1 can be obtained by replacing y_i in (3.1) with $y_{i,1}$ in longitudinal surveys. In what follows, the sample density function $f_s(\mathbf{y}_i)$ of $\mathbf{y}_i$ in longitudinal surveys under informative sampling can be expressed as

$$
\begin{aligned}
f_s(\mathbf{y}_i) &= f_s(y_{i,1};\theta)\prod_{t=2}^{T} f_p(y_{i,t}|y_{i,t-1};\theta) \\
&= \frac{E_p(\pi_i|y_{i,1})f_p(y_{i,1};\theta)}{E_p(\pi_i)}\prod_{t=2}^{T} f_p(y_{i,t}|y_{i,t-1};\theta).
\end{aligned}
\tag{4.1}
$$

Then, the log-likelihood function becomes

$$
\begin{aligned}
\log L &= \sum_{i=1}^{n}\log E_p(\pi_i|y_{i,1}) - \sum_{i=1}^{n}\log E_p(\pi_i) \\
&\quad + \sum_{i=1}^{n}\log f(y_{i,1};\theta) + \sum_{i=1}^{n}\sum_{t=2}^{T}\log f(y_{i,t}|y_{i,t-1};\theta).
\end{aligned}
\tag{4.2}
$$

4.2. Sample likelihood under informative sampling and nonignorable missingness

When nonignorable missingness also exists in longitudinal surveys under informative sampling, $\sum_{i=1}^{n}\sum_{t=2}^{T}\log f(y_{i,t}|y_{i,t-1};\theta)$ in (4.2) needs to be modified since $f(y_{i,t}|y_{i,t-1};\theta)$ is not available when $y_{i,t}$ or $y_{i,t-1}$ is missing. Taking the response mechanism (3.6) into account, we propose to replace $f(y_{i,t}|y_{i,t-1};\theta)$ by the conditional densities based on the observed response, namely $f(y_{i,t}|x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)$ or $f(y_{i,t}|x_i,y_{i,t-1},\delta_{i,t-1}=1,\delta_{i,t}=1)$, depending on whether $y_{i,t-1}$ is missing or not. It follows that the log-likelihood function under informative sampling and nonignorable missingness can be rewritten as

$$
\begin{aligned}
\log L =& \sum_{i=1}^{n}\log E_p(\pi_i|y_{i,1}) - \sum_{i=1}^{n}\log E_p(\pi_i) + \sum_{i=1}^{n}\log f(y_{i,1};\theta)\\
&+\sum_{i=1}^{n}\sum_{t=2}^{T}\delta_{i,t-1}\delta_{i,t}\log f(y_{i,t}|x_i,y_{i,t-1},\delta_{i,t-1}=1,\delta_{i,t}=1)\\
&+\sum_{i=1}^{n}\sum_{t=2}^{T}(1-\delta_{i,t-1})\delta_{i,t}\log f(y_{i,t}|x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)\,.
\end{aligned}
\tag{4.3}
$$

Next, we derive the expressions for $f(y_{i,t}|x_i,y_{i,t-1},\delta_{i,t-1}=1,\delta_{i,t}=1)$ and $f(y_{i,t}|x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)$ in the following lemma. The proof is given in the Appendix.

Lemma 4.1. *The conditional density $f(y_{i,t}|x_i,y_{i,t-1},\delta_{i,t-1}=1,\delta_{i,t}=1)$ satisfies*

$$
f(y_{i,t}|x_i,y_{i,t-1},\delta_{i,t-1}=1,\delta_{i,t}=1) = \frac{\pi(x_i,y_{i,t-1},y_{i,t})f(y_{i,t}|y_{i,t-1})}{\int\pi(x_i,y_{i,t-1},y_t)f(y_t|y_{i,t-1})dy_t}\,,
\tag{4.4}
$$

and $f(y_{i,t}|x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)$ satisfies

$$
\begin{aligned}
&f(y_{i,t}|x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)\\
&=\frac{\iint f(y_{t-2})f(y_{t-1}|y_{t-2})f(y_{i,t}|y_{t-1})\pi(x_i,y_{t-1},y_{i,t})[1-\pi(x_i,y_{t-2},y_{t-1})]dy_{t-2}dy_{t-1}}{f(x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)}.
\end{aligned}
\tag{4.5}
$$

Substituting (4.4) and (4.5) into (4.3) yields the following log-likelihood function under informative sampling and nonignorable missingness

$$
\begin{aligned}
\log L =& \sum_{i=1}^{n}\log E_p(\pi_i|y_{i,1}) - \sum_{i=1}^{n}\log E_p(\pi_i) + \sum_{i=1}^{n}\log f(y_{i,1};\theta)\\
&+\sum_{t=2}^{T}\sum_{i=1}^{n}\delta_{i,t-1}\delta_{i,t}\Big\{\log f(y_{i,t}|y_{i,t-1};\theta)+\log\pi(x_i,y_{i,t-1},y_{i,t};\eta)\\
&\quad-\log\int\pi(x_i,y_{i,t-1},y_t;\eta)f(y_t|y_{i,t-1};\theta)dy_t\Big\}\\
&+\sum_{t=2}^{T}\sum_{i=1}^{n}(1-\delta_{i,t-1})\delta_{i,t}\Big\{\log\iint f(y_{t-2})f(y_{t-1}|y_{t-2})f(y_{i,t}|y_{t-1})\pi(x_i,y_{t-1},y_{i,t})\\
&\quad\cdot[1-\pi(x_i,y_{t-2},y_{t-1})]dy_{t-2}dy_{t-1}-\log f(x_i,\delta_{i,t-1}=0,\delta_{i,t}=1)\Big\}\,.
\end{aligned}
\tag{4.6}
$$

Using (3.2), (3.3) and (3.4), the log-likelihood functions under nonignorabe missingness and the three informative sampling models can be expressed as

Exponential model:

$$
\begin{aligned}
(4.7)\quad & \log L(\theta, \eta, a_1) \\
&= a_1 \sum_{i=1}^{n} y_{i,1} - n[a_1\mu + \sigma^2 a_1^2/(2(1-\phi^2))] + \sum_{i=1}^{n} \log f(y_{i,1};\theta) \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t}\Big\{\log f(y_{i,t}|y_{i,t-1};\theta) + \log \pi(x_i, y_{i,t-1}, y_{i,t};\eta) \\
&\qquad - \log \int \pi(x_i, y_{i,t-1}, y_t;\eta) f(y_t|y_{i,t-1};\theta)dy_t\Big\} \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n}(1-\delta_{i,t-1})\delta_{i,t}\Big\{\log \iint f(y_{t-2})f(y_{t-1}|y_{t-2})f(y_{i,t}|y_{t-1})\pi(x_i, y_{t-1}, y_{i,t}) \\
&\qquad \cdot [1-\pi(x_i, y_{t-2}, y_{t-1})]dy_{t-2}dy_{t-1} - \log f(x_i, \delta_{i,t-1}=0, \delta_{i,t}=1)\Big\},
\end{aligned}
$$

Probit model:

$$
\begin{aligned}
(4.8)\quad & \log L(\theta, \eta, b_0, b_1) \\
&= \sum_{i=1}^{n} \log \Phi(b_0 + b_1 y_{i,1}) - \sum_{i=1}^{n} \log \int \Phi(b_0 + b_1 y_{i,1}) f(y_{i,1})dy_{i,1} + \sum_{i=1}^{n} \log f(y_{i,1};\theta) \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t}\Big\{\log f(y_{i,t}|y_{i,t-1};\theta) + \log \pi(x_i, y_{i,t-1}, y_{i,t};\eta) \\
&\qquad - \log \int \pi(x_i, y_{i,t-1}, y_t;\eta) f(y_t|y_{i,t-1};\theta)dy_t\Big\} \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n}(1-\delta_{i,t-1})\delta_{i,t}\Big\{\log \iint f(y_{t-2})f(y_{t-1}|y_{t-2})f(y_{i,t}|y_{t-1})\pi(x_i, y_{t-1}, y_{i,t}) \\
&\qquad \cdot [1-\pi(x_i, y_{t-2}, y_{t-1})]dy_{t-2}dy_{t-1} - \log f(x_i, \delta_{i,t-1}=0, \delta_{i,t}=1)\Big\},
\end{aligned}
$$

Logistic model:

$$
\begin{aligned}
(4.9)\quad & \log L(\theta, \eta, c_0, c_1) \\
&= -\sum_{i=1}^{n} \log[1+\exp(-c_0 - c_1 y_{i,1})] - \sum_{i=1}^{n} \log \int [1+\exp(-c_0 - c_1 y_{i,1})]^{-1} f(y_{i,1})dy_{i,1} \\
&\quad + \sum_{i=1}^{n} \log f(y_{i,1};\theta) + \sum_{t=2}^{T}\sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t}\Big\{\log f(y_{i,t}|y_{i,t-1};\theta) \\
&\qquad + \log \pi(x_i, y_{i,t-1}, y_{i,t};\eta) - \log \int \pi(x_i, y_{i,t-1}, y_t;\eta) f(y_t|y_{i,t-1};\theta)dy_t\Big\} \\
&\quad + \sum_{t=2}^{T}\sum_{i=1}^{n}(1-\delta_{i,t-1})\delta_{i,t}\Big\{\log \iint f(y_{t-2})f(y_{t-1}|y_{t-2})f(y_{i,t}|y_{t-1})\pi(x_i, y_{t-1}, y_{i,t}) \\
&\qquad \cdot [1-\pi(x_i, y_{t-2}, y_{t-1})]dy_{t-2}dy_{t-1} - \log f(x_i, \delta_{i,t-1}=0, \delta_{i,t}=1)\Big\}.
\end{aligned}
$$

Therefore, the maximum likelihood estimators of θ, η, a_1, b_0, b_1, c_0, and c_1 can be obtained by maximizing the log-likelihood functions in (4.7), (4.8) or (4.9).

4.3. Computations of the likelihood function

Note that computing the log-likelihood functions in (4.7), (4.8) and (4.9) involves the density $f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)$, as well as the integrals $\int \pi(x_i, y_{i,t-1}, y_t; \eta) f(y_t|y_{i,t-1}; \theta) dy_t$, $\int [1 + \exp(-c_0 - c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$, $\iint f(y_{t-2}) f(y_{t-1}|y_{t-2}) f(y_{i,t}|y_{t-1}) \pi(x_i, y_{t-1}, y_{i,t}) \cdot \cdot [1 - \pi(x_i, y_{t-2}, y_{t-1})] dy_{t-2} dy_{t-1}$ and $\int \Phi(b_0 + b_1 y_{i,1}) f(y_{i,1}) dy_{i,1}$. In this section we discuss effective computations for these quantities.

First, $f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)$ can be approximated by the empirical distribution

$$f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1) \approx \sum_{\substack{i, \delta_{i,t}=1; \\ \delta_{i,t-1}=0}} (1 - \delta_{i,t-1}) \delta_{i,t} \Big/ n \,.$$

Next, the following lemma provides a series expansion for the integral $\int \pi(x_i, y_{i,t-1}, y_t; \eta) \cdot \cdot f(y_t|y_{i,t-1}; \theta) dy_t$. The proof is provided in the Appendix.

Lemma 4.2. *The integral $\int \pi(x_i, y_{i,t-1}, y_t) f(y_t|y_{i,t-1}) dy_t$ satisfies*

$$\int \pi(x_i, y_{i,t-1}, y_t) f(y_t|y_{i,t-1}) dy_t$$

$$= \begin{cases} \displaystyle\sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2) \Phi(\gamma - \beta k) \\ \displaystyle\qquad + \frac{1}{c} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp(\beta^2 (k+1)^2/2) [1 - \Phi(\gamma + \beta k + \beta)]\,, & \beta > 0\,, \\ \displaystyle\sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2) [1 - \Phi(\gamma - \beta k)] \\ \displaystyle\qquad + \frac{1}{c} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp(\beta^2 (k+1)^2/2) \Phi(\gamma + \beta k + \beta)\,, & \beta < 0\,, \\ \displaystyle\frac{1}{1+c}\,, & \beta = 0\,, \end{cases} \tag{4.10}$$

where $c = \exp[-(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 \tilde{\mu})]$, $\tilde{\mu} = \mu + \phi(y_{i,t-1} - \mu)$, $\beta = -\eta_3 \sigma$, $\gamma = -\log c/\beta$ *and Φ is the distribution function of standard normal distribution.*

In practice, the infinite series in (4.10) has to be approximated by a finite truncated sum. Simulation studies show that the truncation of summing up to $k = 10$ gives a good approximation to the infinite series in most cases.

Based on Lemma 4.2, the following corollary gives a similar series expansion for the integral $\int [1 + \exp(-c_0 - c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$ in (4.9). The proof is presented in the Appendix.

Corollary 4.1. *The integral* $\int [1+\exp(-c_0 - c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$ *satisfies*

$$\int [1+\exp(-c_0 - c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$$

$$= \begin{cases} \displaystyle\sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2) \Phi(\gamma - \beta k) \\ \qquad \displaystyle + \frac{1}{c} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp(\beta^2 (k+1)^2/2)[1 - \Phi(\gamma + \beta k + \beta)]\,, & \beta > 0\,, \\ \displaystyle\sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2)[1 - \Phi(\gamma - \beta k)] \\ \qquad \displaystyle + \frac{1}{c} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp(\beta^2 (k+1)^2/2) \Phi(\gamma + \beta k + \beta)\,, & \beta < 0\,, \\ \displaystyle\frac{1}{1+c}\,, & \beta = 0\,, \end{cases} \tag{4.11}$$

where $c = \exp(-c_0 - c_1 \mu)$, $\beta = -c_1 \sigma / \sqrt{1-\phi^2}$, $\gamma = -\log c/\beta$ *and* Φ *is the distribution function of standard normal distribution.*

Lastly, for the double integral $\iint f(y_{t-2}) f(y_{t-1}|y_{t-2}) f(y_{i,t}|y_{t-1}) \pi(x_i, y_{t-1}, y_{i,t}) \cdot \cdot [1 - \pi(x_i, y_{t-2}, y_{t-1})] dy_{t-2} dy_{t-1}$, the series expansion approach is not applicable. Thus, it is necessary to consider other numerical methods for computing the double integral. Here, we adopt the Gauss-Hermite quadrature (Liu and Pierce [19]) to approximate it. Similarly, the Gauss-Hermite quadrature can also be employed to approximate the integral $\int \Phi(b_0 + b_1 y_{i,1}) f(y_{i,1}) dy_{i,1}$ in (4.8). In `R`, the function `gauss.quad` under the package `statmod` can be employed. Simulations show that the choice of 9 nodes gives satisfactory performance. In summary, the computation of maximum likelihood function based on Lemma 4.2, Corollary 4.1 and the Gauss-Hermite quadrature has higher efficiency than that based on direct integration.

5. SIMULATION STUDIES

To evaluate the performance of the estimators obtained by dealing with informative sampling and nonignorable missingness in longitudinal surveys, we conduct a simulation study to compare the estimators under informative sampling and/or nonignorable missingness. In the simulation, $N = 1000$ univariate normal values of $y_{i,1}$ are independently generated from $y_1 \sim N(\mu, \sigma^2/(1-\phi^2))$ for the first time period $(t = 1)$, where $\mu = 0.8$, $\phi = 0.3$ and $\sigma = 0.5$. Then, we generate $N = 1000$ population values of $y_{i,t}$ $(i = 1, ..., N)$ at time $t = 2, ..., T$ with $T = 10, 20$ and 40 from the AR(1) model, $y_{i,t} - \mu = \phi(y_{i,t-1} - \mu) + \varepsilon_{i,t}$, where $\varepsilon_{i,t} \sim N(0, 1)$ is independent error term. The AR(1) model parameters μ, ϕ and σ are of our interest.

For the sample selection, samples of size $n = 10$, 20 and 40 are selected from the population via probability proportional to size (PPS) systematic sampling with size variable z. The size variable z values are generated in the following ways, which produce various sampling methods:

(**1**) Exponential sampling: $z_i = \exp(0.9 + 0.3y_{i,1} + \mu_i)$, $\mu_i \sim U(0,1)$.

(**2**) Probit sampling: $z_i = \Phi(0.72 + 0.09y_{i,1} + \mu_i)$, $\mu_i \sim U(0,2)$.

(**3**) Logistic sampling: $z_i = [1 + \exp(0.6 - 0.5y_{i,1} - \mu_i)]^{-1}$, $\mu_i \sim U(0,5)$.

(**4**) Noninformative sampling: $z_i = \exp(1.5\mu_i)$, $\mu_i \sim U(0,4)$.

Note that exponential sampling, probit sampling and logistic sampling are informative. Under the above sampling approaches, selection probabilities are defined as $\pi_i = nz_i \Big/ \sum_{i=0}^{N} z_i$.

For the missingness mechanism, the population value of the covariate is generated from $x_i \sim N(0,1)$, $i = 1, ..., N$. We assume that the covariate x_i and the response variable $y_{i,1}$ at time $t = 1$ are always observed, but $y_{i,t}$ at time $t = 2, ..., T$ may subject to missingness. The response or missing indicator $\delta_{i,t}$ of $y_{i,t}$ are independently generated from a Bernoulli distribution with the response probabilities $\pi_{it}(\eta) = P(\delta_{i,t} = 1 | x_i, y_{i,t-1}, y_{i,t}; \eta)$ specified by $\pi_{it}(\eta) = [1 + \exp(-\eta_1 x_i - \eta_2 y_{i,t-1} - \eta_3 y_{i,t})]^{-1}$, where $\eta_1 = 0.2, \eta_2 = 0.4, \eta_3 = -0.5$. The average response rates under exponential sampling, probit sampling, logistic sampling and non-informative sampling are about 50% for the above nonignorable missing mechanism.

For samples under exponential sampling, probit sampling and logistic sampling, we compute the model parameter estimates by maximizing the sample likelihood under informative sampling and nonignorable missingness. For the sample under noninformative sampling, the model parameter estimators is obtained by maximizing the following log-likelihood function.

$$
\begin{aligned}
(5.1) \qquad \log L \;=\; & \sum_{i=1}^{n} \log f(y_{i,1};\theta) \\
& + \sum_{t=2}^{T} \sum_{i=1}^{n} \delta_{i,t-1}\delta_{i,t} \Big\{ \log \pi(x_i, y_{i,t-1}, y_{i,t}; \eta) + \log f(y_{i,t} | y_{i,t-1}; \theta) \\
& \quad - \log \int \pi(x_i, y_{i,t-1}, y_t; \eta) f(y_t | y_{i,t-1}; \theta) dy_t \Big\} \\
& + \sum_{t=2}^{T} \sum_{i=1}^{n} (1 - \delta_{i,t-1})\delta_{i,t} \\
& \quad \cdot \Big\{ \log \iint f(y_{t-2}) f(y_{t-1} | y_{t-2}) f(y_{i,t} | y_{t-1}) \pi(x_i, y_{t-1}, y_{i,t}) \\
& \quad \cdot [1 - \pi(x_i, y_{t-2}, y_{t-1})] dy_{t-2} dy_{t-1} - \log f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1) \Big\}.
\end{aligned}
$$

For comparison, we also compute the naive estimators, which ignore informative sampling and nonignorable missingness, and are obtained by maximizing the log-likelihood function (3.7). Moreover, the estimators obtained by ignoring informative sampling or nonignorable missingness under exponential sampling, probit sampling and logistic sampling are computed. The estimation procedure is repeated $B = 500$ times. For each estimator, the Monte Carlo biases (Bias), standard deviations (SD) under various n and T are presented. Besides, we also compute the estimation error $\|\hat{\theta} - \theta\|_2$ of the parameter $\theta = (\mu, \phi, \sigma)$, denoted by ER, and the standard deviation of ER to further measure the performance of θ. The results are provided in Tables 1, 2 and 3.

Table 1: Monte Carlo biases, standard deviations and estimation errors of the point estimators under $n = 10$ and $T = 10$.

Sampling	**Estimate**	**Naive**		**Proposed**		**Ignore Sampling**		**Ignore Missingness**	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
Exponential	$\widehat{\mu}$	−0.0103	0.1129	−0.0001	0.0735	0.0186	0.0732	−0.0947	0.1712
	$\widehat{\phi}$	−0.0241	0.2032	−0.0093	0.1196	−0.0096	0.1139	−0.0256	0.2003
	$\widehat{\sigma}$	−0.0232	0.0648	−0.0032	0.0507	−0.0003	0.0521	−0.0355	0.0638
Missing	$\widehat{\eta}_1$			−0.0135	0.0592	−0.0103	0.0636		
46.28%	$\widehat{\eta}_2$			0.0425	0.0548	0.0461	0.0527		
	$\widehat{\eta}_3$			0.0206	0.0526	0.0131	0.0559		
	$\widehat{a}_1$			0.0203	0.0657			0.4578	1.0580
	ER (SD)	0.2134 (0.1176)		0.1147 (0.0958)		0.1147 (0.0912)		0.2529 (0.1425)	
Probit	$\widehat{\mu}$	−0.7105	14.8879	0.0011	0.0743	0.0035	0.0777	−0.0919	0.2255
	$\widehat{\phi}$	−0.0420	0.4807	−0.0163	0.1061	−0.0170	0.1226	−0.0029	0.2371
	$\widehat{\sigma}$	−0.0184	0.1993	−0.0043	0.0500	−0.0051	0.0505	−0.0190	0.0725
Missing	$\widehat{\eta}_1$			−0.0092	0.0504	−0.0061	0.0616		
46.76%	$\widehat{\eta}_2$			0.0337	0.0462	0.0475	0.0645		
	$\widehat{\eta}_3$			0.0179	0.0452	0.0260	0.0509		
	$\widehat{b}_0$			0.0210	0.0517			8.3838	135.7257
	$\widehat{b}_1$			0.0176	0.0485			5.4545	114.8679
	ER (SD)	0.9055 (14.8865)		0.1061 (0.0910)		0.1175 (0.1007)		0.2855 (0.1988)	
Logistic	$\widehat{\mu}$	−0.0412	0.1105	−0.0021	0.0492	0.0032	0.0764	−0.0625	0.1091
	$\widehat{\phi}$	−0.0361	0.2065	0.0113	0.0460	−0.0150	0.1152	0.0188	0.0801
	$\widehat{\sigma}$	−0.0300	0.0612	0.0015	0.0425	−0.0033	0.0510	−0.0118	0.0555
Missing	$\widehat{\eta}_1$			−0.0055	0.0323	−0.0048	0.0600		
46.55%	$\widehat{\eta}_2$			0.0134	0.0252	0.0454	0.0536		
	$\widehat{\eta}_3$			0.0061	0.0217	0.0228	0.0561		
	$\widehat{c}_0$			0.0190	0.0241			0.0478	0.0497
	$\widehat{c}_1$			0.0289	0.0223			0.0656	0.0569
	ER (SD)	0.2183 (0.1213)		0.0631 (0.0499)		0.1145 (0.0938)		0.1335 (0.0892)	
Noninform	$\widehat{\mu}$	−0.0404	0.1103	0.0032	0.0779				
	$\widehat{\phi}$	−0.0411	0.2348	−0.0230	0.1312				
	$\widehat{\sigma}$	−0.0258	0.0660	−0.0029	0.0516				
Missing	$\widehat{\eta}_1$			−0.0052	0.0645				
46.29%	$\widehat{\eta}_2$			0.0516	0.0721				
	$\widehat{\eta}_3$			0.0213	0.0680				
	ER (SD)	0.2327 (0.1463)		0.1249 (0.1042)					

From Table 1, it can be seen that the proposed method that deals with informative sampling and nonignorable missingness simultaneously generally has smaller biases in comparison with the others under the four sampling mechanisms. As expected, the parameter estimation error of the proposed method is the smallest among all methods under various sampling schemes, followed by the estimators handling nonignorable missingness but ignoring informative sampling, whereas the estimation errors of the naive estimators and the estimators dealing with informative sampling but ignoring nonignorable missingness are relatively large among the four methods under exponential sampling, probit sampling and logistic sampling. Moreover, it is obvious that the proposed estimators of the parameters μ, ϕ, σ in AR(1) model have smaller biases than the naive estimators when the sampling design is noninformative.

All of these indicate that the proposed method has a good performance in handling nonignorable missingness. Besides, the proposed method generally yields the smallest standard deviations of the four methods for the estimation of the parameters μ, ϕ, σ under different sampling approaches. Similar results can be found in Table 2 and 3 which focus on different sample sizes. From Tables 1, 2 and 3, it can be seen as well that the estimation error of the proposed method decreases with the increase in the sample size n and the time period T for the four sampling schemes. It is noteworthy that the differences between the estimation errors of the proposed estimators and the estimators ignoring informative sampling but handling nonignorable missingness become smaller under various sampling schemes as n and T increase. This is reasonable because the sampling at time 1 may have a smaller effect on the estimation of the AR(1) model parameters as the time period T becomes larger. In conclusion, the proposed method performs best in the estimation of parameters.

Table 2: Monte Carlo biases, standard deviations and estimation errors of the point estimators under $n = 20$ and $T = 20$.

Sampling	**Estimate**	**Naive**		**Proposed**		**Ignore Sampling**		**Ignore Missingness**	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
Exponential	$\widehat{\mu}$	−0.0374	0.0625	0.0043	0.0439	0.0140	0.0395	−0.0904	0.0735
	$\widehat{\phi}$	−0.0069	0.0976	−0.0139	0.0710	−0.0115	0.0671	−0.0077	0.0963
	$\widehat{\sigma}$	−0.0053	0.0332	0.0039	0.0269	0.0048	0.0261	−0.0113	0.0331
Missing	$\widehat{\eta}_1$			−0.0322	0.0507	−0.0319	0.0507		
49.36%	$\widehat{\eta}_2$			0.0533	0.1096	0.0541	0.0399		
	$\widehat{\eta}_3$			0.0117	0.1068	0.0162	0.0381		
	$\widehat{a}_1$			0.0263	0.0452			0.3980	0.5521
	ER (SD)	0.1148 (0.0529)		0.0669 (0.0585)		0.0701 (0.0467)		0.1419 (0.0629)	
Probit	$\widehat{\mu}$	−0.0612	0.0629	0.0008	0.0381	0.0004	0.0400	−0.0880	0.0806
	$\widehat{\phi}$	−0.0020	0.0977	−0.0116	0.0621	−0.0063	0.0658	0.0061	0.1022
	$\widehat{\sigma}$	−0.0082	0.0341	0.0037	0.0262	0.0035	0.0273	−0.0052	0.0365
Missing	$\widehat{\eta}_1$			−0.0337	0.0508	−0.0308	0.0543		
49.40%	$\widehat{\eta}_2$			0.0422	0.0319	0.0553	0.0406		
	$\widehat{\eta}_3$			0.0157	0.0335	0.0232	0.0386		
	$\widehat{b}_0$			0.0218	0.0311			−1.4934	14.4169
	$\widehat{b}_1$			0.0197	0.0358			10.9502	95.3024
	ER (SD)	0.1233 (0.0571)		0.0636 (0.0457)		0.0675 (0.0465)		0.1472 (0.0662)	
Logistic	$\widehat{\mu}$	−0.0570	0.0617	−0.0010	0.0288	0.0036	0.0395	−0.0661	0.0638
	$\widehat{\phi}$	−0.0035	0.1012	0.0095	0.0331	−0.0083	0.0688	0.0170	0.0585
	$\widehat{\sigma}$	−0.0074	0.0329	0.0056	0.0209	0.0031	0.0251	0.0001	0.0301
Missing	$\widehat{\eta}_1$			−0.0190	0.0322	−0.0285	0.0489		
49.27%	$\widehat{\eta}_2$			0.0178	0.0198	0.0555	0.0426		
	$\widehat{\eta}_3$			0.0095	0.0215	0.0199	0.0366		
	$\widehat{c}_0$			0.0193	0.0191			0.0438	0.0394
	$\widehat{c}_1$			0.0308	0.0191			0.0601	0.0382
	ER (SD)	0.1222 (0.0592)		0.0433 (0.0246)		0.0677 (0.0491)		0.0984 (0.0579)	
Noninform	$\widehat{\mu}$	−0.0699	0.0622	0.0014	0.0423				
	$\widehat{\phi}$	0.0012	0.0985	0.0002	0.0641				
	$\widehat{\sigma}$	−0.0093	0.0331	0.0020	0.0258				
Missing	$\widehat{\eta}_1$			−0.0391	0.0540				
49.39%	$\widehat{\eta}_2$			0.0575	0.0398				
	$\widehat{\eta}_3$			0.0216	0.0393				
	ER (SD)	0.1253 (0.0625)		0.0681 (0.0438)					

Table 3: Monte Carlo biases, standard deviations and estimation errors of the point estimators under $n = 40$ and $T = 40$.

Sampling	Estimate	Naive		Proposed		Ignore Sampling		Ignore Missingness	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
Exponential	$\widehat{\mu}$	−0.0614	0.0342	0.0014	0.0297	0.0078	0.0282	−0.0890	0.0377
	$\widehat{\phi}$	0.0012	0.0504	−0.0097	0.0677	−0.0090	0.0613	0.0023	0.0504
	$\widehat{\sigma}$	−0.0040	0.0194	0.0039	0.0160	0.0051	0.0175	−0.0066	0.0190
Missing	$\widehat{\eta}_1$			−0.0746	0.0666	−0.0745	0.0660		
50.69%	$\widehat{\eta}_2$			0.0891	0.1987	0.1009	0.1384		
	$\widehat{\eta}_3$			−0.0025	0.1928	−0.0081	0.1423		
	$\widehat{a}_1$			0.0344	0.0425			0.3001	0.3354
	ER (SD)	0.0828 (0.0318)		0.0462 (0.0608)		0.0485 (0.0517)		0.1048 (0.0360)	
Probit	$\widehat{\mu}$	−0.0742	0.0316	0.00081	0.0224	0.0027	0.0346	−0.0904	0.0390
	$\widehat{\phi}$	0.0002	0.0474	−0.0085	0.0373	−0.0107	0.0813	0.0045	0.0503
	$\widehat{\sigma}$	−0.0051	0.0180	0.0043	0.0151	0.0052	0.0206	−0.0038	0.0189
Missing	$\widehat{\eta}_1$			−0.0788	0.0548	−0.0769	0.0784		
50.67%	$\widehat{\eta}_2$			0.0663	0.0372	0.1129	0.2268		
	$\widehat{\eta}_3$			0.0098	0.0325	−0.0103	0.2049		
	$\widehat{b}_0$			0.0301	0.0282			−0.7884	7.4087
	$\widehat{b}_1$			0.0261	0.0303			5.5026	38.9418
	ER (SD)	0.0905 (0.0302)		0.0394 (0.0256)		0.0501 (0.0765)		0.1059 (0.0372)	
Logistic	$\widehat{\mu}$	−0.0716	0.0344	−0.0029	0.0190	0.0029	0.0332	−0.0693	0.0358
	$\widehat{\phi}$	0.0061	0.0496	0.0094	0.0298	−0.0073	0.0741	0.0191	0.0392
	$\widehat{\sigma}$	−0.0040	0.0166	0.0070	0.0112	0.0062	0.0173	−0.0004	0.0160
Missing	$\widehat{\eta}_1$			−0.0389	0.0391	−0.0704	0.0632		
50.58%	$\widehat{\eta}_2$			0.0278	0.0179	0.1123	0.2690		
	$\widehat{\eta}_3$			0.0080	0.0243	−0.0142	0.2514		
	$\widehat{c}_0$			0.0245	0.0175			0.0494	0.0320
	$\widehat{c}_1$			0.0357	0.0170			0.0524	0.0323
	ER (SD)	0.0892 (0.0337)		0.0340 (0.0190)		0.0481 (0.0684)		0.0830 (0.0368)	
Noninform	$\widehat{\mu}$	−0.0754	0.0347	0.0013	0.0256				
	$\widehat{\phi}$	0.0004	0.0477	−0.0079	0.0465				
	$\widehat{\sigma}$	−0.0043	0.0169	0.0049	0.0160				
Missing	$\widehat{\eta}_1$			−0.0725	0.0551				
50.56%	$\widehat{\eta}_2$			0.0939	0.0724				
	$\widehat{\eta}_3$			0.0049	0.0746				
	ER (SD)	0.0915 (0.0331)		0.0451 (0.0335)					

6. REAL DATA ANALYSIS

The longitudinal data examined in this section comes from AIDS Clinical Trial Group 193A Study (Henry *et al.* [16]). It concerns AIDS patients with advanced immune suppression which is measured with CD4 counts. A total of 1309 patients were randomized to one of the four treatment groups including (1) 600mg zidovudine alternating monthly with 400mg didanosine, (2) 600mg zidovudine plus 2.25mg of zalcitabine, (3) 600mg zidovudine plus 400mg of didanosine, and (4) 600mg zidovudine plus 400mg of didanosine plus 400mg of nevirapine. The numbers of patients in the four treatment groups are $n = 325$, 324, 330 and 330, respectively. Treatments started at the time of week 0 (baseline), and were measured before the treatments and every 8 weeks. That is, data is collected on the 0, 8, 16, 24, 32, 40th weeks. Here, we denote the six follow-up time points by $t = 1, 2, 3, 4, 5, 6$. The measured outcome variable log(CD4 count + 1) is of our interest, whose values in six time intervals $(0, 4]$, $(4, 12]$, $(12, 20]$, $(20, 28]$, $(28, 36]$, $(36, 40]$ are viewed as y_t for $t = 1, 2, 3, 4, 5, 6$.

Note that the last record of the variable log(CD4 count + 1) in the interval is adopted as y_t if there are more than one values of log(CD4 count + 1) in a time interval. The covariates related to the response variable include Age (years) and Gender (Male=1, Female=0). Details on the data set can be found at `https://content.sph.harvard.edu/fitzmaur/ala/cd4.txt`.

In the longitudinal survey, the covariates are completely observed, whereas the response variable y_t (CD4 counts) is subject to missingness due to skipping visits or dropouts. In fact, a low CD4 count implies that HIV has damaged a patient's immune system to an extent that they are at risk of serious illnesses or even deaths. Thus, a lower CD4 count increases the chance of dropouts due to serious illnesses or deaths. As the patients' dropouts are related to the CD4 count, the missing process is potentially nonignorable. The missing rates under the four treatments are approximately 37.79%, 37.19%, 37.93% and 35.86%, respectively. Let $\delta_{i,t}$ be the indicator variable for $y_{i,t}$. Define

$$\delta_{i,t} = \begin{cases} 1 & \text{if } y_{i,t} \text{ is observed}, \\ 0 & \text{otherwise}, \end{cases} \tag{6.1}$$

for $i = 1, 2, ..., n$ and $t = 1, 2, 3, 4, 5, 6$. We are interested in estimating the response probability $P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t})$. We fit the response model using the age variable x_1 and the gender variable x_2 in the following logistic model:

$$P(\delta_{i,t} = 1|x_{i1}, x_{i2}, y_{i,t-1}, y_{i,t}) = \frac{\exp(\eta_1 x_{i1} + \eta_2 x_{i2} + \eta_3 y_{i,t-1} + \eta_4 y_{i,t})}{1 + \exp(\eta_1 x_{i1} + \eta_2 x_{i2} + \eta_3 y_{i,t-1} + \eta_4 y_{i,t})}, \tag{6.2}$$

where $\eta_1, \eta_2, \eta_3, \eta_4$ are the unknown parameters. This missing mechanism is obviously nonignorable. For comparison, we also consider the following working model for the response probability under ignorable missing mechanism:

$$P(\delta_{i,t} = 1|x_{i1}, x_{i2}) = \frac{\exp(\eta'_1 x_{i1} + \eta'_2 x_{i2})}{1 + \exp(\eta'_1 x_{i1} + \eta'_2 x_{i2})}, \tag{6.3}$$

where η'_1 and η'_2 are the unknown parameters. The response probability in equation (6.3) only depends on the covariates x_1 and x_2, implying that the missing mechanism is ignorable.

Assume that the sampling design is exponential sampling, probit sampling and logistic sampling, respectively. For comparison, we consider two models, the AR(1) model (2.1) and the following mean model.

$$y_{i,t} = \mu + \varepsilon_{i,t}, \quad i = 1, ..., n, \quad t = 1, ..., 6, \tag{6.4}$$

where $\varepsilon_{i,t} \sim N(0, \sigma^2)$. In fact, the mean model has no time dependence and been considered by Zhao *et al.* [39]. The estimates of model parameters μ, ϕ, σ under different missing models, sampling schemes and treatments, together with the mean squares of the model residuals (MSE), are presented in Tables 4 and 5.

As shown in Tables 4 and 5, Treatment 4 presents greater estimated values of μ than other Treatments regardless of models, missing mechanisms or sampling approaches. Also, the estimates of μ under Treatment 1 are the lowest among all treatments for all sampling methods and two missing models. That is, patients under Treatment 4 are superior to those under other Treatments in terms of the average number of CD4 counts, and the average number of patients' CD4 counts under Treatment 1 is relatively low. In fact, a high CD4 counts indicates a strong immune system, which suggests that the patient lives longer. This may reduce the possibility to drop outs for patients, which in turn reduces the differences between the parameter estimates under nonignorable missingness and ignorable missingness.

Table 4: Estimates for the AIDS clinical trial group 193A study data under nonignorable missingness.

Sampling	Estimate	Treatment 1		Treatment 2		Treatment 3		Treatment 4	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
		AR(1) Model	Mean Model	AR(1) Model	Mean Model	AR(1) Model	Mean Model	AR(1) Model	Mean Model
Exponential	$\widehat{\mu}$	2.5268	2.7442	2.6766	2.7326	2.6609	2.7989	2.8167	2.8772
	$\widehat{\phi}$	0.7124		0.6561		0.7228		0.7730	
	$\widehat{\sigma}$	0.7076	0.9504	0.7618	1.0893	0.7739	1.1018	0.7203	1.1377
	MSE	0.5539	0.6781	0.5848	0.8760	0.7674	1.0883	0.9008	1.3400
Probit	$\widehat{\mu}$	2.9169	2.7406	2.8934	2.8550	2.8528	2.9042	2.9490	3.1211
	$\widehat{\phi}$	0.6963		0.7092		0.7470		0.7591	
	$\widehat{\sigma}$	0.7202	0.9300	0.7641	1.0827	0.7526	1.1261	0.7392	1.1644
	MSE	0.5265	0.6784	0.5761	0.8511	0.7504	1.0439	0.8805	1.2657
Logistic	$\widehat{\mu}$	2.6969	2.7452	2.9060	2.7831	2.8900	2.7952	2.9263	2.9543
	$\widehat{\phi}$	0.6276		0.6951		0.7671		0.7809	
	$\widehat{\sigma}$	0.7544	0.9577	0.7740	1.0982	0.7597	1.1136	0.7288	1.1028
	MSE	0.5182	0.6780	0.5717	0.8621	0.7538	1.0903	0.8903	1.3036

Table 5: Estimates for the AIDS clinical trial group 193A study data under ignorable missingness.

Sampling	Estimate	Treatment 1		Treatment 2		Treatment 3		Treatment 4	
		Bias	SD	Bias	SD	Bias	SD	Bias	SD
		AR(1) Model	Mean Model	AR(1) Model	Mean Model	AR(1) Model	Mean Model	AR(1) Model	Mean Model
Exponential	$\widehat{\mu}$	2.5349	2.6818	2.6518	2.7504	2.7867	2.9202	3.1894	3.0855
	$\widehat{\phi}$	0.6718		0.6961		0.7288		0.7639	
	$\widehat{\sigma}$	0.7002	0.9481	0.7563	1.0625	0.7701	1.1311	0.7490	1.1440
	MSE	0.5428	0.6880	0.5938	0.8705	0.7523	1.0391	0.8573	1.2691
Probit	$\widehat{\mu}$	2.7210	2.7339	2.7974	2.7847	2.8407	2.8982	3.2598	3.1054
	$\widehat{\phi}$	0.6775		0.6994		0.7286		0.7692	
	$\widehat{\sigma}$	0.7065	0.9519	0.7614	1.0698	0.7728	1.1334	0.7365	1.1449
	MSE	0.5289	0.6792	0.5806	0.8617	0.7461	1.0458	0.8535	1.2669
Logistic	$\widehat{\mu}$	2.8759	2.7102	2.8401	2.7172	2.7586	2.9382	2.8827	2.9391
	$\widehat{\phi}$	0.6661		0.7182		0.7373		0.7777	
	$\widehat{\sigma}$	0.7525	0.9815	0.7634	1.0796	0.7772	1.0921	0.7411	1.1267
	MSE	0.5184	0.6825	0.5822	0.8812	0.7579	1.0343	0.8941	1.3098

This point is in line with the fact that the estimates of the key model parameter ϕ under nonignorable missingness are very close to those under ignorable missingness in the same Treatment 4 for various sampling approaches, whereas there is a clear difference between the parameter estimates of ϕ under nonignorable missingness and ignorable missingness in Treatment 1 for different sampling schemes. Moreover, the estimator of ϕ in the AR(1) model under Treatment 4 is the largest among all treatments under each informative sampling model for each missing mechanism, suggesting that the number of CD4 counts of Treatment 4 keeps decreasing more slowly in comparison with the others. Therefore, we conclude that Treatment 4 has better effect on the AIDS disease than other treatments. Besides, in terms of the variance estimators $\hat{\sigma}^2$ of residuals and MSE, the AR(1) model yields lower $\hat{\sigma}^2$ and MSE than the mean model. Thus, it seems very reasonable to use the AR(1) model over the mean model to analyze this data set.

A. APPENDIX

Proof of Lemma 4.1: First, the conditional density $f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1}{=}1, \delta_{i,t}{=}1)$ can be obtained, similar to Pfeffermann *et al.* [25], as

$$\text{(A.1)} \qquad f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1, \delta_{i,t} = 1) \\ = \frac{P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}, \delta_{i,t-1} = 1) f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1)}{\int P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_t, \delta_{i,t-1} = 1) f(y_t|x_i, y_{i,t-1}, \delta_{i,t-1} = 1) dy_t}.$$

The term $P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}, \delta_{i,t-1} = 1)$ on the right side of (A.1) can be written as

$$\text{(A.2)} \qquad P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}, \delta_{i,t-1} = 1) \\ = \frac{P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}) P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}, y_{i,t}, \delta_{i,t} = 1)}{P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}, y_{i,t})} \\ = P(\delta_{i,t} = 1|x_i, y_{i,t-1}, y_{i,t}) \\ = \pi(x_i, y_{i,t-1}, y_{i,t}).$$

The term $f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1)$ on the right side of (A.1) can be written as

$$\text{(A.3)} \qquad f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1) = \frac{P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}, y_{i,t}) f(y_{i,t}|y_{i,t-1})}{P(\delta_{i,t-1} = 1|x_i, y_{i,t-1})},$$

where $f(y_{i,t}|y_{i,t-1}) = \exp\{-[y_{i,t} - \mu - \phi(y_{i,t-1} - \mu)]^2/2\sigma^2\}/\sqrt{2\pi}\sigma$.

Next, the two conditional probabilities of $\delta_{i,t-1}$ in (A.3) can be expressed as

$$\text{(A.4)} \qquad P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}, y_{i,t}) \\ = \int P(\delta_{i,t-1} = 1|x_i, y_{t-2}, y_{i,t-1}) f(y_{t-2}|y_{i,t-1}, y_{i,t}) dy_{t-2} \\ = \int \pi(x_i, y_{t-2}, y_{i,t-1}) f(y_{t-2}|y_{i,t-1}, y_{i,t}) dy_{t-2},$$

and

$$\text{(A.5)} \qquad P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}) \\ = \int P(\delta_{i,t-1} = 1|x_i, y_{t-2}, y_{i,t-1}) f(y_{t-2}|y_{i,t-1}) dy_{t-2} \\ = \int \pi(x_i, y_{t-2}, y_{i,t-1}) f(y_{t-2}|y_{i,t-1}) dy_{t-2},$$

respectively, where $\pi(x_i, y_{t-2}, y_{i,t-1})$ is defined in (3.6).

According to the AR(1) model, we can easily prove $f(y_{t-2}|y_{i,t-1}, y_{i,t}) = f(y_{t-2}|y_{i,t-1})$. Then, we have $P(\delta_{i,t-1} = 1|x_i, y_{i,t-1}, y_{i,t}) = P(\delta_{i,t-1} = 1|x_i, y_{i,t-1})$. Moreover, $f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1) = f(y_{i,t}|y_{i,t-1})$ holds. Thus, the conditional density in (A.1) can be written as

$$\text{(A.6)} \qquad f(y_{i,t}|x_i, y_{i,t-1}, \delta_{i,t-1} = 1, \delta_{i,t} = 1) = \frac{\pi(x_i, y_{i,t-1}, y_{i,t}) f(y_{i,t}|y_{i,t-1})}{\int \pi(x_i, y_{i,t-1}, y_t) f(y_t|y_{i,t-1}) dy_t}.$$

Therefore, (4.4) in Lemma 4.1 holds.

Now we derive the results for $f(y_{i,t}|x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)$. Based on the definition of the conditional density, we have

$$f(y_{i,t}|x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1) = \frac{f(x_i, y_{i,t}, \delta_{i,t-1} = 0, \delta_{i,t} = 1)}{f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)}, \tag{A.7}$$

where $f(x_i, y_{i,t}, \delta_{i,t-1} = 0, \delta_{i,t} = 1)$ can be given by

$$\begin{aligned}
&f(x_i, y_{i,t}, \delta_{i,t-1} = 0, \delta_{i,t} = 1)\\
&= \iint f(x_i, y_{t-2}, y_{t-1}, y_{i,t}) f(\delta_{i,t-1} = 0, \delta_{i,t} = 1|x_i, y_{t-2}, y_{t-1}, y_{i,t}) dy_{t-2} dy_{t-1}\\
&= \iint f(x_i, y_{t-2}) f(y_{t-1}|x_i, y_{t-2}) f(y_{i,t}|x_i, y_{t-2}, y_{t-1}) P(\delta_{i,t} = 1|x_i, y_{t-2}, y_{t-1}, y_{i,t})\\
&\quad \cdot P(\delta_{i,t-1} = 0|x_i, y_{t-2}, y_{t-1}, y_{i,t}, \delta_{i,t} = 1) dy_{t-2} dy_{t-1}\\
&= \iint f(y_{t-2}) f(y_{t-1}|y_{t-2}) f(y_{i,t}|y_{t-1}) \pi(x_i, y_{t-1}, y_{i,t})[1 - \pi(x_i, y_{t-2}, y_{t-1})] dy_{t-2} dy_{t-1}.
\end{aligned} \tag{A.8}$$

Thus, we can obtain

$$\begin{aligned}
&f(y_{i,t}|x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)\\
&= \frac{\iint f(y_{t-2}) f(y_{t-1}|y_{t-2}) f(y_{i,t}|y_{t-1}) \pi(x_i, y_{t-1}, y_{i,t})[1 - \pi(x_i, y_{t-2}, y_{t-1})] dy_{t-2} dy_{t-1}}{f(x_i, \delta_{i,t-1} = 0, \delta_{i,t} = 1)}.
\end{aligned} \tag{A.9}$$

It follows that (4.5) in Lemma 4.1 holds. □

Proof of Lemma 4.2: According to $\pi(x_i, y_{i,t-1}, y_{i,t}) = \exp(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 y_{i,t})/[1 + \exp(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 y_{i,t})] = 1/[1 + \exp(-\eta_1 x_i - \eta_2 y_{i,t-1} - \eta_3 y_{i,t})]$ and $f(y_{i,t}|y_{i,t-1}) = (2\pi\sigma^2)^{-1/2} \exp\{-[y_{i,t} - \phi(y_{i,t-1} - \mu) - \mu]^2/(2\sigma^2)\}$, we have

$$\begin{aligned}
&\int \pi(x_i, y_{i,t-1}, y_t) f(y_t|y_{i,t-1}) dy_t\\
&= \frac{1}{\sqrt{2\pi}\sigma} \int \frac{1}{1 + \exp[-(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 y_t)]} \exp\left\{-\frac{[y_t - \phi(y_{i,t-1} - \mu) - \mu]^2}{2\sigma^2}\right\} dy_t.
\end{aligned} \tag{A.10}$$

Let $\tilde{\mu} = \mu + \phi(y_{i,t-1} - \mu)$ and $c = \exp[-(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 \tilde{\mu})]$, we can obtain

$$\begin{aligned}
&\int \pi(x_i, y_{i,t-1}, y_t) f(y_t|y_{i,t-1}) dy_t\\
&\quad = \frac{1}{\sqrt{2\pi}\sigma} \int \frac{1}{1 + \exp\{-[\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3 \tilde{\mu} + \eta_3 (y_t - \tilde{\mu})]\}} \exp\left[-\frac{(y_t - \tilde{\mu})^2}{2\sigma^2}\right] dy_t\\
&\quad = \frac{1}{\sqrt{2\pi}\sigma} \int \frac{1}{1 + c \cdot \exp(-\eta_3 x)} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx\\
&\quad = \frac{1}{\sqrt{2\pi}} \int \frac{1}{1 + c \cdot \exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy,
\end{aligned} \tag{A.11}$$

where $\beta = -\eta_3 \sigma$.

When $\beta > 0$ and $0 < c \cdot \exp(\beta y) < 1$, we have $y < \gamma = -\log c/\beta$. Further, we can write

$$
\begin{aligned}
\text{(A.12)} \quad & \int \pi(x_i, y_{i,t-1}, y_t) f(y_t | y_{i,t-1}) dy_t \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\gamma} \frac{1}{1 + c \cdot \exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy + \int_{\gamma}^{\infty} \frac{1}{1 + c \cdot \exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy \right] \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\gamma} \sum_{k=0}^{\infty} [-c \cdot \exp(\beta y)]^k \exp\left(-\frac{y^2}{2}\right) dy \right. \\
&\quad \left. + \frac{\exp(\beta^2/2)}{c} \int_{\gamma}^{\infty} \sum_{k=0}^{\infty} [-1/(c \cdot \exp(\beta y))]^k \exp\left[-\frac{(y+\beta)^2}{2}\right] dy \right] \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\gamma} \sum_{k=0}^{\infty} (-c)^k \exp\left(\frac{\beta^2 k^2}{2}\right) \exp\left[-\frac{(y-\beta k)^2}{2}\right] dy \right. \\
&\quad \left. + \frac{1}{c} \int_{\gamma}^{\infty} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp\left[\frac{\beta^2 (k+1)^2}{2}\right] \exp\left\{-\frac{[y + \beta(k+1)]^2}{2}\right\} dy \right] \\
&= \sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2) \Phi(\gamma - \beta k) + \frac{1}{c} \sum_{k=0}^{\infty} (-\frac{1}{c})^k \exp[\beta^2 (k+1)^2/2][1 - \Phi(\gamma + \beta k + \beta)] \,.
\end{aligned}
$$

Similarly, when $\beta < 0$ and $0 < c \cdot \exp(\beta y) < 1$, we have $y > \gamma = -\log c/\beta$. Then we can obtain

$$
\begin{aligned}
\text{(A.13)} \quad & \int \pi(x_i, y_{i,t-1}, y_t) f(y_t | y_{i,t-1}) dy_t \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{\gamma} \frac{1}{1 + c \cdot \exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy + \int_{\gamma}^{\infty} \frac{1}{1 + c \cdot \exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy \right] \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{\gamma}^{\infty} \sum_{k=0}^{\infty} (-c)^k \exp\left(\frac{\beta^2 k^2}{2}\right) \exp\left[-\frac{(y-\beta k)^2}{2}\right] dy \right. \\
&\quad \left. + \frac{1}{c} \int_{-\infty}^{\gamma} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp\left[\frac{\beta^2 (k+1)^2}{2}\right] \exp\left\{-\frac{[y + \beta(k+1)]^2}{2}\right\} dy \right] \\
&= \sum_{k=0}^{\infty} (-c)^k \exp(\beta^2 k^2/2) [1 - \Phi(\gamma - \beta k)] + \frac{1}{c} \sum_{k=0}^{\infty} \left(-\frac{1}{c}\right)^k \exp[\beta^2 (k+1)^2/2] \Phi(\gamma + \beta k + \beta) \,.
\end{aligned}
$$

Specially, when $\beta = 0$, we get

$$
\int \pi(x_i, y_{i,t-1}, y_t) f(y_t | y_{i,t-1}) dy_t = \frac{1}{\sqrt{2\pi}} \int \frac{1}{1+c} \exp\left(-\frac{y^2}{2}\right) dy = \frac{1}{1+c} \,.
$$

Thus, Lemma 4.2 holds. ☐

Proof of Corollary 4.1: Note that the results in Lemma 4.2 can also be used to compute the integral $\int [1+\exp(-c_0-c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$ in (4.9). Similar to the proof of Lemma 4.2, the integral $\int [1+\exp(-c_0-c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$ can be written as

$$\begin{aligned}\text{(A.14)} \qquad & \int [1+\exp(-c_0-c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1} \\ &= \frac{\sqrt{1-\phi^2}}{\sqrt{2\pi}\sigma} \int \frac{1}{1+\exp(-c_0-c_1 y_{i,1})} \exp\left\{-\frac{(1-\phi^2)(y_{i,1}-\mu)^2}{2\sigma^2}\right\} dy_{i,1}\,.\end{aligned}$$

Let $y = \sqrt{1-\phi^2}(y_{i,1}-\mu)/\sigma$, we have

$$\begin{aligned}\text{(A.15)} \qquad & \int [1+\exp(-c_0-c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1} \\ &= \frac{1}{\sqrt{2\pi}} \int \frac{1}{1+\exp[-c_0-c_1(\sigma y/\sqrt{1-\phi^2}+\mu)])} \exp\left(-\frac{y^2}{2}\right) dy \\ &= \frac{1}{\sqrt{2\pi}} \int \frac{1}{1+c\cdot\exp(\beta y)} \exp\left(-\frac{y^2}{2}\right) dy\,,\end{aligned}$$

where $c = \exp(-c_0-c_1\mu)$ and $\beta = -c_1\sigma/\sqrt{1-\phi^2}$. Thus, we can compute the integral $\int [1+\exp(-c_0-c_1 y_{i,1})]^{-1} f(y_{i,1}) dy_{i,1}$ by replacing $c = \exp[-(\eta_1 x_i + \eta_2 y_{i,t-1} + \eta_3\tilde{\mu})]$ and $\beta = -\eta_3\sigma$ in Lemma 4.2 with $c = \exp(-c_0-c_1\mu)$ and $\beta = -c_1\sigma/\sqrt{1-\phi^2}$. It follows that Corollary 4.1 holds. □

ACKNOWLEDGMENTS

The authors thank the Editor, the Associate Editor and referees for their constructive comments. The collaborative work described in this paper was supported by HKSAR-RGC-GRF Nos. 14305517, 14601015 and 14302719 (Yau) and National Social Science Foundation of China, No. 18BTJ022 (Liu).

REFERENCES

[1] Almohisen, A.; Henderson, R. and Alshingiti, A.M. (2019). An alternative sensitivity approach for longitudinal analysis with dropout, *Journal of Probability and Statistics*, `https://doi.org/10.1155/2019/1019303`.

[2] Bahari, F.; Parsi, S. and Ganjali, M. (2019). Empirical likelihood inference in general linearmodel with missing values in response and covariates by MNAR mechanism, *Statistical Papers*, `https://doi.org/10.1007/s00362-019-01103-0`.

[3] Bhuyan, P. (2019). Estimation of random-effects model for longitudinal data with nonignorable missingness using Gibbs sampling, *Computational Statistics*, `https://doi.org/10.1007/s00180-019-00887-x`.

[4] Bonnery, D.; Breidt, F.J. and Coquet, F. (2018). Asymptotics for the maximum sample likelihood estimator under informative selection from a finite population, *Bernoulli*, **24**(2), 929–955.

[5] Boudreau, C. and Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data, *The Canadian Journal of Statistics*, **34**(2), 203–216.

[6] Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses, *Sociological Methodology*, **43**(1), 178–219.

[7] Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley and Sons, Ltd., New York.

[8] Cole, B.F.; Bonetti, M.; Zalavasky, A.M. and Gelber, R.D. (2005). A multistate markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness, *Statistics in Medicine*, **24**(15), 2317–2334.

[9] Eideh, A.A.H. (2010). Fitting general linear model for longitudinal survey data under informative sampling, *Statistics in Transition-new series*, **11**(3), 517–538.

[10] Eideh, A.A.H. and Nathan, G. (2009a). *Joint treatment of nonignorable dropout and informative sampling for longitudinal survey data*. In "Methodology of Longitudinal Surveys" (P. Lynn, Ed.), pp. 251-264, John Wiley and Sons, Ltd., New York.

[11] Eideh, A.A.H. and Nathan, G. (2009b). Two-stage informative cluster sampling — estimation and prediction with applications for small-area models, *Journal of Statistical Planning and Inference*, **139**(9), 3088–3101.

[12] Eideh, A.A.H. and Nathan, G. (2006a). Fitting time series models for longitudinal survey data under informative sampling, *Journal of Statistical Planning and Inference*, **136**(9), 3052–3069.

[13] Eideh, A.A.H. and Nathan, G. (2006b). The analysis of data from sampling surveys under informative sampling, *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **10**, 41–51.

[14] Farahania, Z.S.M.; Khorrama, E.; Ganjalib, M. and Baghfalakic, T. (2019). Longitudinal data analysis in the presence of informative sampling: weighted distribution or joint modelling, *Journal of Applied Statistics*, **46**(12), 2111–2127.

[15] Fuller, W.A. (2009). *Sampling Statistics*, John Wiley and Sons, Inc., New Jersey.

[16] Henry, K.; Erice, A.; Tierney, C.; Balfour, J.H.; Fischl, M.A.; Kmack, A.; Liou, S.H.; Kenton, A.; Hirsch, M.S.; Phair, J.; Martinez, A. and for the AIDS Clinical Trial Group 193A Study Team (1998). A randomized, controlled, double-blind study comparing the survival benefit of four different reverse transcriptase inhibitor therapies (three-drug, two-drug, and alternating drug) for the treatment of advanced AIDS, *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **19**(4), 339–349.

[17] Kim, J.K. and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling, *Biometrika*, **100**(2), 385–398.

[18] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis With Missing Data* (Second Edition), John Wiley and Sons, Inc., New Jersey.

[19] Liu, Q. and Pierce, D.A. (1994). A note on Gauss-Hermite quadrature, *Biometrika*, **81**(3), 624–629.

[20] Moser, A.; Bopp, M.; Zwahlen, M. and Swiss National Cohort Study Group (2018). Calibration adjustments to address bias in mortality analyses due to informative sampling — a census-linked survey analysis in Switzerland, `https://doi.org/10.7717/peerj.4376`.

[21] Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: a unified approach, *Calcutta Statistical Association Bulletin*, **69**(1), 35–63.

[22] Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas, *Journal of the American Statistical Association*, **102**(480), 1427–1439.

[23] Pfeffermann, D.; Moura, F.A.D.S. and Silva, P.L.D.N. (2006). Multi-level modelling under informative sampling, *Biometrika*, **93**(4), 943–959.

[24] Pfeffermann, D. and Sverchkov, M. (2003). *Fitting generalized linear models under informative sampling*. In "Analysis of Survey Data" (Southampton, 1999), Wiley Series in Survey Methodology, pp. 175–195, Wiley, Chichester.

[25] Pfeffermann, D.; Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling, *Statistica Sinica*, **8**, 1087–1114.

[26] Pfeffermann, D. (1996). The use of sampling weights for survey data analysis, *Statistical Methods in Medical Research*, **5**(3), 239–261.

[27] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *International Statistical Review*, **61**, 317–337.

[28] Qin, J.; Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling, *Journal of the American Statistical Association*, **97**(457), 193–200.

[29] Sabry, M.Y.; Kholy, R.B.E. and Gad, A.M. (2016). Generalized linear mixed models for longitudinal data with missing values: a monte carlo EM approach, *International Journal of Probability and Statistics*, **5**(3), 82–88.

[30] Schlomer, G.L.; Bauman, S. and Card, N.A. (2010). Best practices for missing data management in counseling psychology, *Journal of Counseling Psychology*, **57**(1), 1–10.

[31] Sikov, A. and Stern, J.M. (2019). Application of the full Bayesian significance test to model selection under informative sampling, *Statistical Papers*, **60**(1), 89–104.

[32] Stubbendick, A.L. and Ibrahim, J.G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data, *Statistica Sinica*, **16**, 1143–1167.

[33] Sverchkov, M. and Pfeffermann, D. (2018). Small area estimation under informative sampling and not missing at random non-response, *Journal of the Royal Statistical Society: Series A*, **181**(4), 981–1008.

[34] Taisir, R. and Islam, M.A. (2014). EM algorithm for longitudinal data with non-ignorable missing values: an application to health data, *Bangladesh Journal of Scientific Research*, **27**(2), 133–142.

[35] Thompson, M.E. (2015). Using longitudinal complex survey data, *Annual Review of Statistics and Its Application*, **2**, 305–320.

[36] Tseng, C.H.; Elashoff, R.; Li, N. and Li, G. (2016). Longitudinal data analysis with non-ignorable missing data, *Statistical Methods in Medical Research*, **25**(1), 205–220.

[37] Verret, F.; Rao, J.N.K. and Hidiroglou, M.A. (2015). Model-based small area estimation under informative sampling, *Survey Methodology*, **41**(2), 333–347.

[38] Yuan, Y. and Yin, G.S. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data, *Biometrics*, **66**, 105–114.

[39] Zhao, P.Y.; Wang, L. and Shao, J. (2018). Analysis of longitudinal data under nonignorable nonmonotone nonresponse, *Statistics and Its Interface*, **11**(2), 265–279.

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 427–447
https://doi.org/10.57805/revstat.v20i4.382

Impact of Academic Authorship Characteristics on Article Citations*

Authors: Philipp Otto ✉
– European University Viadrina,
Germany
otto@europa-uni.de

Philipp Otto
– European University Viadrina,
Germany
potto@europa-uni.de

Received: March 2020 Revised: August 2020 Accepted: September 2020

Abstract:

- Scientific self-evaluation practices are increasingly built on citation counts. Citation practices for the top journals in economics, psychology, and statistics illustrate article characteristics that influence citation frequencies. Citation counts differ between the investigated disciplines, with economics attracting the most citations and statistics the least. Although articles in statistics are cited less frequently, its proportion of uncited articles is the smallest of all three disciplines. Academic authorship characteristics clearly influence the number of citations. Having authors alphabetically ordered, a practice differently present in the investigated disciplines, increases citations. Further, the more authors there are, the more the article is cited, and a first author with a common surname has positive effects on citation counts, whereas two or more authors sharing a surname attracts fewer citations. In addition, the shorter the article's title, the higher the number of citations.

Keywords:

- *scientometrics; publication index; citation characteristics; popular author names; alphabetical authorship.*

AMS Subject Classification:

- 62C25, 91C05.

✉ Corresponding author.

*To the best of our knowledge, this is the first published article where both authors share their surname and given name, while working at the same university. Thus, the two authors are largely indistinguishable, which highlights the importance of individual author identifiers like ORCID.

"If men define situations as real, they are real in their consequences."
(William Isaac Thomas & Dorothy Swaine Thomas 1859, p. 572)

1. INTRODUCTION

Being cited is typically good news for the author(s) of a paper. However, the reference made could be rather critical. In any case, the number of citations reflects the academic impact of an article, and citation counts often provide an initial estimate of the quality of the cited publication, its author(s), and the publishing journal. Because journal rankings and, therefore, academic success are increasingly based on citation counts, the central aim of journal editors appears to be to select articles with the highest citation count expectation (cf. Bornmann *et al.* 2011 [4]). Whereas the practice of quantifying the number of achieved citations in published work is widespread and appears rather useful, citation criteria are manifold and can potentially be self-supporting.

Generally, citation rates are difficult to predict. In this paper, potential drivers are investigated on an exemplary basis for the highest SCImago-ranked journals in economics, psychology, and statistics. Even after ten years, a large proportion (12.4%) of articles were not cited, and half of the articles in the top-ranked journals remained below 20 citations, whereas the total number of citations is slightly above 200 on average. Considering average citations per year, the maximum increase in citations is reached somewhere after 11 years (see Figure 1). This leads to the question of whether there are any identifiable criteria that can explain higher citation counts?

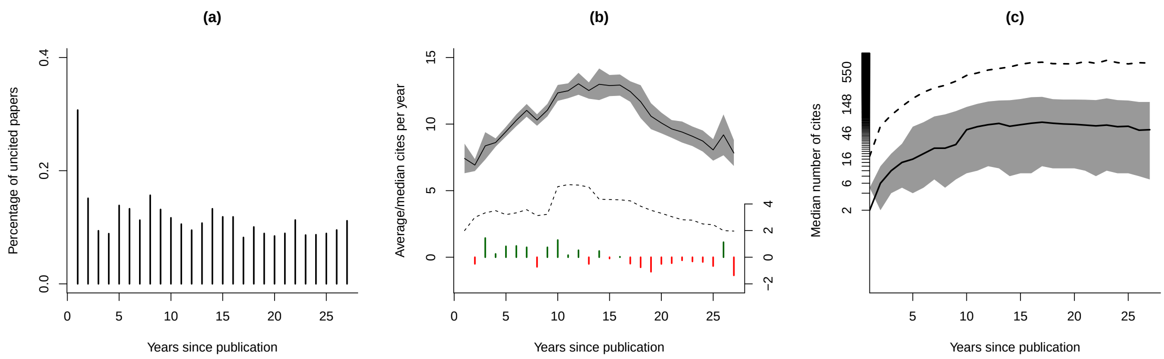


Figure 1: Temporal dynamics of the total number of citations per year since publication.
(**a**) Percentage of uncited articles.
(**b**) Average citations (solid line, with 95%-confidence intervals as shaded area) and median citations (dashed line) per year depicted for papers, which have been published $1, 2, \ldots, 27$ years ago, with absolute temporal differences per year as red/green-colored bars.
(**c**) Median total number of citations after $1, 2, \ldots, 27$ years with the shaded area representing the interquartile range and the 95% quantile as a dashed line.

The most common dependency is that the more an article has been cited in the past, the more it will be cited in the future (cf. Stegehuis *et al.* 2015 [30]). Furthermore, a typical article citation curve describes a steady increase over its life cycle. Within approximately three years, an article typically gains momentum (or lack thereof), then reaches a top level of citations somewhere between 10 and 15 years. Thereafter, the majority of articles are cited less frequently.[1] Various factors can be investigated to compare the above-median cited articles against those below. We quantify some easily available article differentials, with a concentration on authorship characteristics, namely *research discipline*, *years since publication*, *title length*, *number of authors*, *alphabetically ordered authors*, *author name-sharing*, and *common author name*). Beginning with a specification of the potential influences and postulating canonical regularities, we provide an empirical analysis using a freely-available data source with an accordingly adapted statistical model and present the results for the investigated dependencies. In the conclusion, the postulated regularities are critically evaluated, how these results relate to other regularities reported in the literature is discussed, and an outlook on the future development of applicable article quality criteria is provided.

2. CITATION CRITERIA AND POSTULATED DEPENDENCIES

The hereby proposed citation criteria introduce alternative measures for explaining citation counts, which are derived historically, structurally, or purely descriptively. All the tested criteria are easily quantifiable and can be divided into the following two categories: structural regularities, or purely authorship-related characteristics. This shifts the focus from quality or relevance toward other criteria as the ones being responsible for citation counts. As an implicit test, it refutes the discussion on the usefulness of derived empirical indicators for academic success, such as the Hirsch (2005) [20] index and others (compare for example Lindsey 1989 [23]), but also illustrates potential regularities as to the ways researchers are citing each other's work.

2.1. Structural regularities

Differences in academic disciplines provide a starting point in the evaluation of article characteristics to find regularities in citing practices. Here, economic, psychology, and statistics publications were used to study discipline-specific differences, as well as broader influences on citation frequencies.

The following exemplary regularities were provided ad hoc: psychological publications would be cited more often (mainly in other disciplines) due to a generally larger public interest in their research topic and strong interdisciplinary focus (compare interdisciplinary citations in Jacobs 2013 [21]). Statistics is the smallest discipline and, therefore, citations were expected to be less frequent, although statistics are used for empirical analyses in all disciplines. This postulates a regularity that can be summarized as

Hypothesis 1. *Citation frequencies vary over research disciplines with being:*

(**a**) *higher for psychology publications;*
(**b**) *lower for statistics publications.*

[1] A more general description of citation changes over time, with more profound numbers on passing critical thresholds to develop a momentum, would require time-series data. Investigations that account for other temporal influences, such as citation density or prolonging increases in citations are provided by Quandt (1976) [28] or Parolo *et al.* (2015) [26].

Other characteristics can be article specific and illustrate a direct structural dependency with citation frequencies. Two discipline-independent influences were proposed with opposing regularities: citation frequencies increase with the *years since publication* and decrease with the *title length of the article*. Naturally, it takes time for articles to be cited and for the academic community to acknowledge new work. However, one could also expect a slowdown several years after the time of publication, due to decreased novelty. Another issue that was included is simplicity. An anticipated effect is based on information processing and recall. The *title length of the article* serves as an indicator to investigate this kind of influence. Bounded rationality, in the form of limitations when recalling more complex article titles, could lead to lower citation counts. These two apparent article characteristics needed to be controlled, in addition to the differences between the research disciplines, when investigating the following influences.

2.2. Authorship characteristics

Authorship characteristics might also affect citation frequencies. These characteristics could result from academic practices or other easily identifiable article differentials. Thus, the guiding question was, how much variance in citation frequencies can be explained by extrinsic article characteristics related to authorship. This would be in addition to structural influences and the article's quality as the fundamental value.

The first source for identifiable article differentials is academic differences based on the cultural and historical development of respective research disciplines. A prominent example in this regard would be how authors are ordered in a joint publication. Some disciplines prefer purely alphabetical order, whereas others strictly list the author names in the order of the contributed amounts of work. This difference in approach for author listing is exploited by Van Praag and van Praag (2008) [33] and Einav and Yariv (2006) [11], who postulate a positive correlation between the surname initials and the scientific success of the author. The influence of the initial letter of the first author can, thus, be seen as a random characteristic independent of the article's quality.

Our three investigated research fields differ with regard to author listing order. Author listings could be either alphabetical or organized by their respective shares of work (i.e., the first author would be the main author of the article). However, it is not always feasible to distinguish between these two kinds of author listings. A non-alphabetically sorted list of authors does not automatically imply that the first author contributed the most, and in an alphabetically sorted list of authors, the first author could still be the main contributing author. For simplification, Figure 2 illustrates this relation for articles with two authors. Plot (a) shows the percentage of articles in which the authors are listed alphabetically. Van Praag and van Praag (2008) [33] computed the probability of an alphabetical ordering for uniformly distributed first letters. However, the chance of having a surname with the initial letter being 'A' differs from that of having the initial letter 'Z'. Hence, in our data set, we used the observed frequencies of the first letters of all surnames as a proxy for the natural distribution of initial letters. The ratio between the observed percentages of alphabetically ordered authors, and this baseline probability can be seen as the percentage of authors intentionally sorted by the first letter of their surnames. This further implies that the authors of the remaining articles are listed in a non-alphabetical way — potentially to reflect the amount of contributed work.

The accordingly estimated proportions of intentionally alphabetically ordered authors are shown in Figure 2(b), which were strictly lower in psychology when compared with economics and statistics. One can conclude that the first author is most likely to be the main author for articles published in the top psychology journals, whereas in economics and statistics, both authorship orderings coexist.[2] Note that only the first letter of the surname is compared. Names with the same first letter are considered as being alphabetically ordered, although this includes the curiosity that, if all authors have the same surname, they are considered as being alphabetically ordered, although these are at the same time non-alphabetically ordered.

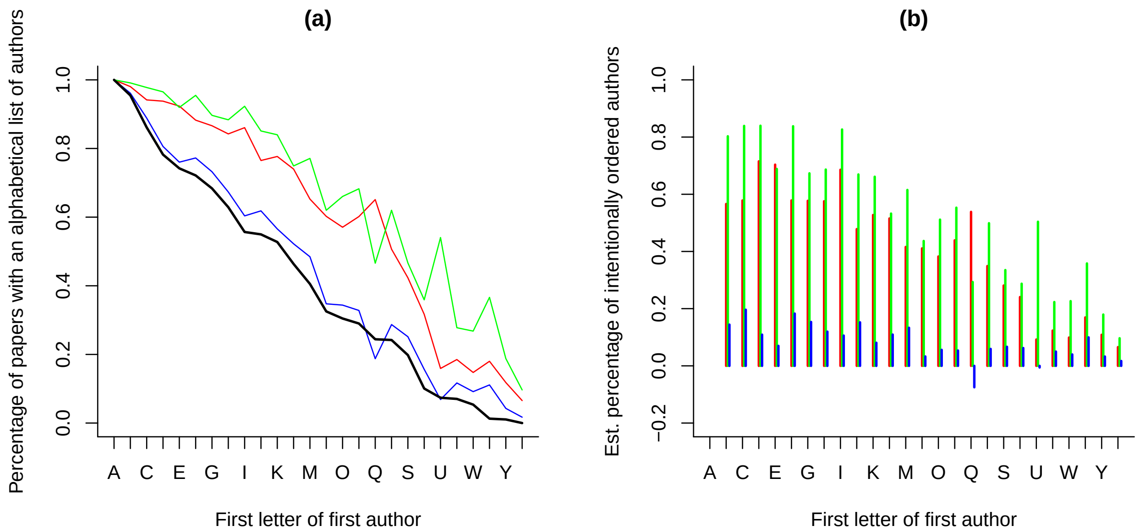


Figure 2: Percentage of articles with two authors having alphabetically ordered names separated by the initial letter of the first author.
(**a**) Percentage of ordered lists of authors in economics (red), statistics (green), and psychology (blue). The bold line depicts the probability of two random surnames being in alphabetical order.
(**b**) Ratio between observed frequencies and the expected base probability (black baseline) illustrates the proportion of intentionally alphabetically ordered authors.

In addition to the citation differences between the three investigated disciplines, publication practices could affect an article's citation count. The two different ways of ordering authors might directly influence its number of citations because the main author is not easily identifiable with *alphabetically ordered authors*, and the allocation of the main work to one specific versus various researchers might influence its citation.

Hypothesis 2. *Citation frequencies change when the main author is listed as the first author of the article.*

[2]As the estimated frequencies from our data set could be biased, Appendix A provides a comparison of these results to the distribution of UK surnames, as reported by Gray (1958) [15], and for the top 100 surnames in the United States of America (provided by the U.S. Census Bureaus for the year 2000), thereby confirming these regularities.

The relation between citation counts and surname familiarity is included in the analysis as another test for the influence of recall simplicity. The top 100 U.S. surnames served as a proxy for *common author names*.[3]

Hypothesis 3. *Citation frequencies increase with the first author having a common surname.*

Another simplicity-related claim goes back to Goodman *et al.* (2015) [14], who investigated a descriptive curiosity of authors sharing surnames. Sources for name doubling, or more generally *author name-sharing*, could be for various reasons and could also directly link to citation counts. Without knowing why the same name occurs twice (or even more often), we argue that these articles are easier to remember and to recall.

Hypothesis 4. *Citation frequencies increase when authors share their surnames.*

A more universal relationship is hypothesized for authorship with regard to the number of people involved with the published research. The *number of authors* is expected to show a direct relationship with citation counts.

Hypothesis 5. *Citation frequencies increase with the number of listed authors for an article.*

With more authors, the new information spreads faster and can be expected to be better connected within the respective scientific communities — not to mention direct (or reciprocal) self-citations.

3. EMPIRICAL DATA ANALYSIS

The systematic rating of evoked citations increasingly influences the scientific evaluation process, ranging from the rankings of individual publications to that of authors and journals. A practical advantage is that citations can easily be retrieved, in addition to diverse article characteristics.[4] The predictive variables of interest are the *research discipline*, *years since publication*, *title length*, *number of authors*, *alphabetically ordered authors*, *author name-sharing*, and *common author name*.

3.1. Data and descriptive statistics

The data analysis was based on $196,365$ journal articles that were published in 115 journals from 1990 to 2016. For each, we observed the current citation count as well as various article characteristics. To be precise, the focus was on the highest-ranked journals

[3]This list also includes popular surnames from other nationalities (e.g., Lee, Nguyen, or Rodriquez). In addition, we considered the soundex of all names to account for different spellings such as Li, Lee, or Liu, but this opposes a unique author identification and, thereby, the postulate of recall simplicity.

[4]Different elicitation methods are described more broadly in Ball (2014) [3].

in three scientific fields, namely economics, psychology, and statistics. The definition of journals belonging to the top journals, to be included in the following analysis, is based on the SCImago journal ranking within the respective subject areas:

- "Economics, Econometrics and Finance": top ten journals of each subcategory (except "Science" as not being a mainly economic journal);
- "Psychology": top ten journals of each subcategory;
- "Statistics, Probability and Uncertainty": top quartile journals (as already a "subcategory").

All included journals are listed in Table 1 (31 from economics, 57 from psychology, and 27 from statistics), with the number of articles, the average SCImago journal ranking index (SJR), the average Hirsch index (H), and the average citations per document for each of the three investigated research areas. The number of total citations recorded until November 2017 serves as a performance measure of each article. To be more specific, citation counts reported by Microsoft Academic Search (MAS) are used as the dependent variable. These counts partly incorporate statistical models based on network data to provide more accurate citation counts; a more detailed discussion of the data set and the MAS citation count is provided in Appendix A.

For the empirical analysis of the postulated hypotheses, we use the current citation counts of all papers published within these journals and the above-mentioned time period. Hence, the citation counts are cumulated values for each individual paper, but independent across time because each paper appears only once in the sample. Figure 1(a) depicts the percentage of uncited articles with respect to the elapsed *years since publication* (in full years). This ratio decreases from thirty percent for all publications in the year of publication (i.e., 2016) to approximately twelve percent within the first three years. The proportion of articles not cited remains stable thereafter, whereas the total number of citations increases over time. The positive growth rate lasts for about 11 years after publication.

The annual average and median citations depicted in plot (b) of Figure 1 have their peaks after 11 years, which implies declining growth rates afterward. However, it is important to note that we have independent samples over time, such that the downslope is partly due to the generally increasing number of citations. For comparison, we also depict the lower quartiles, medians, upper quartiles, and 95%-quantiles of the total citation counts over the elapsed time since publication on a log-scale in Figure 1(c). This supports the assumption that the number of new citations increases in the beginning but reduces with decreasing novelty, and the latter effect seems to be strengthened by an overall increase in the number of citations over the years since 1990 (i.e., older articles are cited less often over their citation life-span). Moreover, Table 2 summarizes the descriptive statistics for the central variables of the regression: the number of citations, percentage of uncited articles, average years since publication, and number of authors (36.7% with one author and 25.7% with two authors). In addition, the average title length is included as the number of characters in the title of the article. Author name-sharing occurred in 0.2% of all included articles.

Table 1: List of the included journals.

Field	Journals	Number of journals	Total number of articles	Average SJR	Average H	Average citations per document
Economics	Academy of Management Journal, Academy of Management Review, Accounting Review, Administrative Science Quarterly, American Economic Journal: Applied Economics, American Economic Journal: Economic Policy, American Economic Journal: Macroeconomics, American Economic Journal: Microeconomics, American Economic Review, Annual Review of Financial Economics, Econometrica, Experimental Economics, Journal of Accounting and Economics, Journal of Accounting Research, Journal of Economic Literature, Journal of Finance, Journal of Financial and Quantitative Analysis, Journal of Financial Economics, Journal of International Economics, Journal of Management, Journal of Monetary Economics, Journal of Political Economy, Journal of Supply Chain Management, Journal of the European Economic Association, Management Science, Quantitative Marketing and Economics, Quarterly Journal of Economics, Review of Economic Studies, Review of Financial Studies, Structural Equation Modeling, Theoretical Economics	31	44192	8.825	115.258	227.89
Psychology	Accounting, Organizations and Society, Annual Review of Psychology, Behaviour Research and Therapy, Biological Psychology, Child Development, Child Development Perspectives, Clinical Psychological Science, Clinical Psychology Review, Cognition, Cognitive Psychology, Current Directions in Psychological Science, Depression and Anxiety, Developmental Review, Developmental Science, Educational Psychologist, European Journal of Personality, European Review of Social Psychology, Evolution and Human Behavior, Frontiers in Behavioral Neuroscience, Frontiers in Human Neuroscience, Health Psychology Review, Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Child Psychology and Psychiatry and Allied Disciplines, Journal of Clinical Child and Adolescent Psychology, Journal of Consulting and Clinical Psychology, Journal of Consumer Psychology, Journal of Educational Measurement, Journal of Experimental Psychology: General, Journal of Memory and Language, Journal of Organizational Behavior, Journal of Personality and Social Psychology, Journal of Research in Crime and Delinquency, Journal of the American Academy of Child and Adolescent Psychiatry, Journal of Youth and Adolescence, Learning and Instruction, Learning and Memory, Memory and Cognition, Neuropsychology Review, Neuroscience and Biobehavioral Reviews, Organizational Behavior and Human Decision Processes, Personality and Social Psychology Bulletin, Personality and Social Psychology Review, Personnel Psychology, Perspectives on Psychological Science, Political Psychology, Psychological Bulletin, Psychological Medicine, Psychological Methods, Psychological Review, Psychological Science, Psychological Science in the Public Interest, Supplement, Psychotherapy and Psychosomatics, Research on Language and Social Interaction, Social Cognitive and Affective Neuroscience, Social Issues and Policy Review, Trends in Cognitive Sciences	57	106406	3.515	116.860	111.1802
Statistics	Annales de l'institut Henri Poincare (B) Probability and Statistics, Annals of Applied Probability, Annals of Applied Statistics, Annals of Mathematics, Annals of Probability, Annals of Statistics, Annual Review of Statistics and Its Application, Biometrika, Biostatistics, Electronic Journal of Probability, Finance and Stochastics, Journal of Business and Economic Statistics, Journal of Computational and Graphical Statistics, Journal of Multivariate Analysis, Journal of Statistical Planning and Inference, Journal of Statistical Software, Journal of the American Statistical Association, Journal of the Royal Statistical Society. Series A: Statistics in Society, Journal of the Royal Statistical Society. Series B: Statistical Methodology, Journal of the Royal Statistical Society. Series C: Applied Statistics, Probability Theory and Related Fields, Scandinavian Journal of Statistics, Scientific Data, Statistica Sinica, Statistical Science, Statistics and Computing, Test	27	45767	2.848	63.963	50.28357

Table 2: Descriptive statistics of selected covariates.

Variable	Freq. of 0	Min.	L.Q.	Median	Mean	U.Q.	Max.	St. Dev.
Citations	0.124	0	5	21	123.25	108	56424	482.86
Citations (> 0)	—	1	9	29	140.76	126	56424	513.63
Years since publ.	—	1	5	10	11.71	18	27	7.68
Title length	—	9	57	77	81.20	100	567	33.47
Number of authors	—	1	1	2	2.74	3	50	2.26
Single author	0.742	—	—	—	—	—	—	—
Author name-sharing	0.998	—	—	—	—	—	—	—

3.2. Model

Because more than ten percent of the articles were not cited within the investigated time frame, the statistical model needs to account for this excess of non-citations. For our data, a zero-inflated negative binomial model was used because it provided a comparatively better fit than other models (e.g., a zero-inflated Poisson model), which is further supported by the Ord plot (see Ord 1967 [25]). Please see Appendix B for a more detailed discussion of this distributional choice.

To define the statistical model, we introduce a random variable Y for the citation counts. The observations of Y are denoted by y. Then, the conditional probability of Y is given by

$$\begin{aligned} P\Big(Y=y \,\Big|\, \mathbf{X}_z, \mathbf{X}_c, \boldsymbol{\beta}_z, \boldsymbol{\beta}_c\Big) &= \\ &= P_z\Big(Y=0 \,\Big|\, \mathbf{X}_z, \boldsymbol{\beta}_z\Big)\, I_{\{0\}}(y) + \Big(1 - P_z\Big(0 \,\Big|\, \mathbf{X}_z, \boldsymbol{\beta}_z\Big)\Big)\, P_c\Big(Y=y \,\Big|\, \mathbf{X}_c, \boldsymbol{\beta}_c\Big)\,, \end{aligned} \tag{3.1}$$

where $\mathbf{X}_z$ and $\mathbf{X}_c$ are the matrices of explanatory variables for the probability of $Y=0$ (index z) and $Y=y\geq 0$ (index c). The respective coefficients for these regressors are $\boldsymbol{\beta}_c$ and $\boldsymbol{\beta}_z$. Moreover, $I_A(x)$ stands for the indicator function on a set A. Whereas P_z describes the conditional probability for $Y=0$, the probability density of P_c defines the number of citations. For our analysis, we assume that P_c is a negative binomial distribution, i.e.,

$$P_c\Big(Y=y \,\Big|\, \mathbf{X}_c, \boldsymbol{\beta}_c\Big) = \frac{\Gamma(\theta+y)}{\Gamma(y+1)\,\Gamma(\theta)}\, r^y (1-r)^\theta \qquad \text{with} \quad r = \frac{\exp(\mathbf{X}_c\boldsymbol{\beta}_c)}{\exp(\mathbf{X}_c\boldsymbol{\beta}_c)+\theta}\,.$$

Due to the methodological separation of articles into cited and uncited, it is possible to distinguish two different effects: the predictive variable $\mathbf{X}_z$, influencing the fact of an article being cited at all, and $\mathbf{X}_c$, influencing the number of citations of a particular work. Corresponding regression coefficients are obtained as maximum-likelihood estimators of a generalized linear model, which is computationally implemented as in Zeileis *et al.* (2008) [40]. The starting values of the iterative maximization of the likelihood function have been chosen by an expectation maximization algorithm.

3.3. Results

All articles were searched for characteristics that explained, firstly, if it was cited at all and, secondly, the number of citations reached.[5] Table 3 shows the results of the zero-inflated negative binomial model with parameters estimated by the maximum-likelihood approach (cf. Greene 2003 [16]; Zeileis *et al.* 2008 [40]). For this, we included all variables introduced in Section 2 that have a potential influence on citation counts. For simplicity of interpretation of the results, we omit potential interactions between the regressors, which are reported in Appendix C. To allow for a more intuitive interpretation of the regression coefficients, we report the corresponding odds ratios r_i for the count and zero component of the model.

Table 3: Estimated coefficients $\hat{\beta}_i^z$ and $\hat{\beta}_i^c$ as well as odds ratios $\hat{r}_i^z$ and incident risk ratios $\hat{r}_i^c$ of a zero-inflated negative binomial regression model for citation counts. The zero-inflated effect as well as the count effect are significant for all introduced regressors and p-values are given in parentheses.

Variable	i	Zero-inflation coefficients		Count coefficients	
		$\hat{\beta}_i^z$	$\hat{r}_i^z$	$\hat{\beta}_i^c$	$\hat{r}_i^c$
Regressors					
Intercept	0	2.760 (< 0.0001)		3.589 (< 0.0001)	
Field of research: Psychology	1	0.368 (< 0.0001)	1.445	−0.256 (< 0.0001)	0.774
Field of research: Statistics	2	−0.662 (< 0.0001)	0.516	−1.095 (< 0.0001)	0.334
Years since publication: in full years	3	−0.052 (< 0.0001)	0.949	0.072 (< 0.0001)	1.074
Title length: number of characters in title	4	0.015 (< 0.0001)	1.015	−0.001 (< 0.0001)	0.999
Number of authors	5	−4.638 (< 0.0001)	0.010	0.027 (< 0.0001)	1.027
Alphabetically ordered authors: true	6	0.539 (0.214)	1.714	0.201 (< 0.0001)	1.222
Author name-sharing: existent	7	1.450 (0.031)	4.264	−0.220 (0.001)	0.803
Common author name: first author within top 100 surnames	8	−0.227 (< 0.0001)	0.797	0.043 (< 0.0001)	1.044
$\log(\hat{\theta})$				−0.835 (< 0.0001)	
Summary Statistics					
AIC	1771146				
$\exp(\log(\hat{\theta}))$	0.547				
LR (null model)	22313.31				

[5] Articles with total citations that were above the 95% quantile are neglected to avoid anomalies due to outlying observations.

These ratios depict the factor by which the expected citation count or probability of being cited changes if the corresponding dummy variable is present or the independent variable is increased by one unit (see Table 3).

3.3.1. Structural regularities

Citation existence and level are highly influenced by the amount of time passed since an article has been published. The older the publication, the higher the likelihood that the publication does not belong to the class of not cited articles, while its citation count is expected to be higher. Thus, *years since publication* increase the likelihood of being cited (negative zero-inflation coefficient $\hat{\beta}_3^z$), as well as the number of citations (positive count coefficient $\hat{\beta}_3^c$). Further, the expected regularities for *title length* are fully confirmed. The longer the title, the more likely it belongs to the uncited articles category and the lower the citation counts. These strong and clear intrinsic influences fully confirm the first two expected regularities, that citation frequencies are indeed determined by the *years since publication* as well as by its *title length*.

Mixed results are observed concerning the differences in the three research disciplines because partly opposite patterns were noted. For Statistics, both coefficients $\hat{\beta}_2^z$ and $\hat{\beta}_2^c$ are negative, which indicates opposite effects. Whereas Statistics has fewer uncited articles when compared with Economics, these articles gather fewer citations. Examining the count model, we see that citation counts were lower in both Psychology (contradicting Hypothesis 1(a) and Statistics (supporting Hypothesis 1(b). Consequently, Economics attracted the most citations compared to the two other disciplines. Given that an article is cited, Statistics articles were cited less frequently when compared to Economics and Psychology. This fully supports Hypothesis 1(b) because the respective coefficients of the count model confirm this order, i.e., $0 > \hat{\beta}_1^c > \hat{\beta}_2^c$. Articles in Statistics were cited less often than articles in Psychology ($p < 0.0001$) and articles in Economics ($p < 0.0001$). Moreover, citations in Psychology were lower than in Economics ($p < 0.0001$). These pairwise relations are also supported by Mann–Whitney-U tests on all cited articles (citations > 0). Thus, the postulated order of the disciplines concerning citation frequencies when being cited is confirmed only when comparing Statistics with Psychology or Economics, but not when comparing Psychology with Economics. The research discipline has a strong influence on the number of citations, but the relations postulated under Hypothesis 1 are only partially confirmed.

3.3.2. Authorship characteristics

Authorship characteristics generally remain influential for citation frequencies, when controlling for structural regularities. However, the empirical findings were not always as hypothesized. Articles having *alphabetically ordered authors* show an opposing effect; these are more inflated by uncited articles, but they are cited more often (i.e., $\hat{\beta}_5^z$ and $\hat{\beta}_5^c$ are positive). Hypothesis 2 is only partially supported. Having the first author as the main author is more likely to attract at least one citation, but this effect is insignificant. Articles where the main author appears as the first author are, in fact, cited significantly less than articles with purely alphabetical ordering.[6]

[6] Although this effect of alphabetically ordered authors is largely reduced in Psychology, it still has a positive influence across all the considered research disciplines. Interactions with research discipline and their cultural differences in sorting authors is further discussed in Appendix C.

In contrast, Hypothesis 3 is fully supported. Having a *common author name*, as a first author surname characteristic consistently related to citation likelihood and frequency. Having a common surname increases the probability of being cited. Important here is that judging whether the surname is a common name based on the exact spelling, rather than on its soundex, leads to a better model fit. Thus, the unique spelling of the name seems to be crucial for its recall simplicity. Another unexpected result was observed regarding the influence of *author name-sharing*. For both cases of being cited and the frequency of citations, the relation is in the opposite direction than postulated under Hypothesis 4. Articles that have (for some authors) the same surnames were significantly less likely to be cited, and in cases where they were cited, they are cited significantly less often. Hence, our hypotheses concerning authorship simplicity are only partly confirmed: having a common name has a positive effect, but when authors share the same surname, this is negatively related to citation frequencies. Note that authors randomly sharing a surname is more frequent for popular names.

The strongest influence on citations was the *number of authors*, which increases the likelihood of being cited as well as the number of citations. The negative zero-inflation coefficient ($\hat{\beta}_3^z$) and the positive count coefficient ($\hat{\beta}_3^c$) clearly support Hypothesis 5.

4. DISCUSSION AND CONCLUSION

Influences on citation counts has received little attention besides noting its fundamental and growing importance for evaluating scientific productivity. Everyday practice simply assumes a direct relation between the gained citations and the importance of the research. This does neglect alternative influences on citation counts. In this regard, various authorship characteristics were evaluated for three research disciplines in social sciences. Without claiming any kind of prominence, systematic regularities can be observed in the data. The *time since publication* is possibly the most important structural component, for which a monotonic increasing relationship is confirmed. To determine an article's citation life (possibly with a critical growth period), however, time series of the citation counts of each article would be required. Although it naturally takes time to acknowledge quality, the duration or speed of this process remains uncertain. Broader issues, such as an overall increase in publications and citations, further complicate this analysis. In addition, fashionable trends are difficult to isolate, particularly in cases where quality intertwines with the novelty of the research topic (compare Van Dalen and Henkens 2001 [32]; Webster *et al.* 2009 [38]; Chen 2012 [8]). Our empirical results show that the *title length* decreases the likelihood and frequency of being cited. Simplicity might help recognition. A positive relation between an article having a short title and citation counts has already been claimed for economic articles (Bramoullé and Ductor 2018 [6]; Gnewuch and Wohlrabe 2017 [13]). These results are confirmed here, whereas recognition not only decreases the chance of belonging to the class of uncited articles, but it also increases the number of attracted citations. However, simplicity and recall probability can oppose uniqueness, which might play a role as well. Naturally, the predictive power of such content-free characteristics needs to be investigated in more detail to be applicable because, for example Didegah and Thelwall (2013) [9] claim in a broader study of research disciplines that the length of the title has no significant influence on citation counts.

Differences between the *field of research* (Hypothesis 1) illustrate a more specific regularity in citation frequencies. This potentially originates from other sources than research quality. These differences could have historical reasons or be confounded with the other expected regularities as well as authorship characteristics. We compared articles in Psychology, Economics, and Statistics, where the popularity was expected to decrease in this order (also due to the size of the (sub-)discipline in the case of Statistics). The postulated relationship is not fully reflected in the citation count data. Articles published in the top journals in Psychology are less frequently cited than those in Economics, but publications in Statistics were cited the least. Interestingly, our regression analysis provides a more profound picture. Articles in Statistics are cited less often, but there were also fewer nil citations. These seemingly opposing effects might be due to a flatter distribution pattern, which might also be responsible for the advantage of Economics over Psychology. It is worth noting that only the top journals of each subject are included in the analysis. A broader sample, of course, might reveal different relations. The proportion of uncited articles can be expected to be more profound and the concentration of citations on fewer articles (such as those in top journals) to be more pronounced in Economics. This is because Economics is more concentrated on a smaller number of leading publications along with a higher impact factor of the top economics journals. This tendency toward the top journals seems to be prolonged (Card and DellaVigna 2013 [7]; Heckman and Moktan 2018 [19]). Fourcade *et al.* (2015) [12] claim that Economics is generally more hierarchically organized. Why the pattern of citation counts in Statistics shows a flatter distribution requires further investigation, possibly in comparison to a larger and more diverse number of research fields. In general, explanations for the variety in citation counts has to be searched and accounted for as has been stressed by Varin *et al.* (2016) [34] regarding cross-citations among highly ranked statistics journals or by Aksnes (2006) [1] for subfields of research in Norway. Radicchi *et al.* (2008) [29] and Albarrán *et al.* (2011) [2] provide first approaches to correct citation count evaluations with respect to the field of research.

A central idea put forward here is to isolate various authorship characteristics that can explain part of the observed variation in citations. This could not only lead to a better understanding of the relationship between quality and being cited but also illustrates the potential pitfalls of not being cited. Not all of the included characteristics have a strong effect, and the results do sometimes point in the opposing direction. If articles have *alphabetically ordered authors* (Hypothesis 2), this actually increased the number of citations but reduced the likelihood of being cited at all. This kind of academic tradition, which is more prominent in Economics and Statistics, could represent things other than quality (dominance, conservatism, etc.). Although indirect and only in terms of citation frequencies, this confirms the claim made by Van Praag and van Praag (2008) [33] that authors with names toward the beginning of the alphabet tend to be more successful (under the assumption that an author's future citations directly depend on previous citations).

Author names can also have an influence in terms of their popularity, especially under the expectation of recognition simplicity (Hypothesis 3); namely, that the first author having a *common author name* increases the number of citations, an occurrence that is confirmed by the data. Note that this expectation equally applies to how having an uncommon name (below the 100 most common names benchmark) leads to fewer citations, possibly because it is more difficult to recall unpopular names. Other demographic or personal author characteristics might help to further elaborate upon this kind of relationship. Naturally, author influences that are not investigated here, such as reputation (as for example *author eminence* as in

Haslam *et al.* 2008 [18]) or connectivity (as for example *number of references* as in Haslam *et al.* 2008 [18]; Vieira and Gomes 2010 [35]; Bornmann *et al.* 2012 [5]; Chen 2012 [8]; Didegah and Thelwall 2013 [9]), could play a central role for citation counts. Along the lines of research embedding, the strongest authorship influence on citation counts is the *number of authors* (Hypothesis 5). This is not only the result of self-citations, which have not been distinguished here; rather, it is attributed to the fact that the more authors there are, the better the interconnectivity and the higher the potential of the paper to be discovered. Thus, the research output is better represented in the respective scientific community, and connections to neighboring fields become more likely. Systematic self- or cross-citations can clearly oppose quality concerns, but dependencies are manifold. For example, collocation effects in the citation networks of authors and institutions can be observed (see Yan and Ding 2012 [39]). Still, a larger number of authors can positively affect the quality of an article, due to increased awareness or a more sophisticated cross-checking, for example, but negative effects of co-authorship can also result from this self-selection process (cf. Ductor 2015 [10]). Also note that for natural sciences, Onodera and Yoshikane (2015) [24] report only a weak and Bornmann *et al.* (2012) [5] a negative effect of the number of authors on citation counts. In summary, a better understanding of the different effect strengths of the investigated authorship characteristics is required to be more conclusive here.

Initially most surprising for us was that *author name-sharing* appears to have the opposite effect than expected (Hypothesis 4) because it negatively influences citation counts. Authorship recognition does not appear to be the driving influence. Possibly, this influence of recognizing an article is largely covered by the popularity of the first author's surname because more frequent names already result more often in coauthors sharing their surnames. Further, the list of reasons for authors sharing the name (given by Goodman *et al.* 2015 [14]) provides a plausible answer here. The sources for people having the same name and publishing an article together (i.e., marriage or other family relations) might reduce the quality of its content. However, name-sharing could also be fully coincidental (as in the case of the "Goodmen"). Furthermore, name-sharing might represent narrowness, and internationality has been reported as a factor strongly increasing citations. Documented positive influences are international collaboration (Didegah and Thelwall 2013 [9]), authors not sharing the same department (Vieira and Gomes 2010 [35]), as well as the article being published in English (Van Dalen and Henkens 2001 [32]; Bornmann *et al.* 2012 [5]). This further illustrates the need for systematically distinguishing behavioral influences from those that represent and acknowledge the quality of an article.

Citation indices have been proposed as a heuristic method for informing decision-making on various levels (see for example Perry and Reny 2016 [27]; Hamermesh 2018 [17]). With diverse drivers influencing citation frequencies, these must be treated even more cautiously. Little has been done to better understand citation behavior, despite it being increasingly crucial in determining academic success. Although it is reasonable to argue that all the articles included in our analysis are of substantial quality because they are published in the top journals of their respective research field, a large proportion are still rarely or not cited at all, whereas other articles strongly pull citations. If specific authorship characteristics are influencing this process, and various data sources exist to evaluate the dependencies here, then these can easily be detected and controlled to better inform decisions. Complementary proxies for research quality are, thus, required to supplement citation indices and journal ranks, both of which are currently solely based on citation count data.

APPENDICES

A. DATA SOURCE

Figure 3 visualizes the distribution of the observed counts by a so-called rootogram, depicting the histogram bars pinned to the best-fitting density curve. In this case, we plot the counts against a negative binomial distribution. This figure shows two major issues that need to be addressed. First, uncited articles are excessive because articles cited between one and three times are less frequently observed than expected by a negative binomial distribution. Consequently, we observed such an excess of zero citations that small counts were overestimated. Second, there is a substantial gap in articles for the area between 33 and 50 citation counts. This lack is due to the specific counting approach of Microsoft Academic Search. In particular, the software uses a statistical model based on citation graphs to estimate citation counts, from which the accuracy is lower for all publications just below 50 citations (confirmed by Microsoft Academic Search). Thus, they reported the true citation count only for the remaining publications, for which the predicted count is less than 50. The resulting anomalous pattern for articles cited between 33 and 50 times is rather unsatisfactory. However, the observed effects should not substantially differ, with the main influence on goodness-of-fit measures being based on residuals.

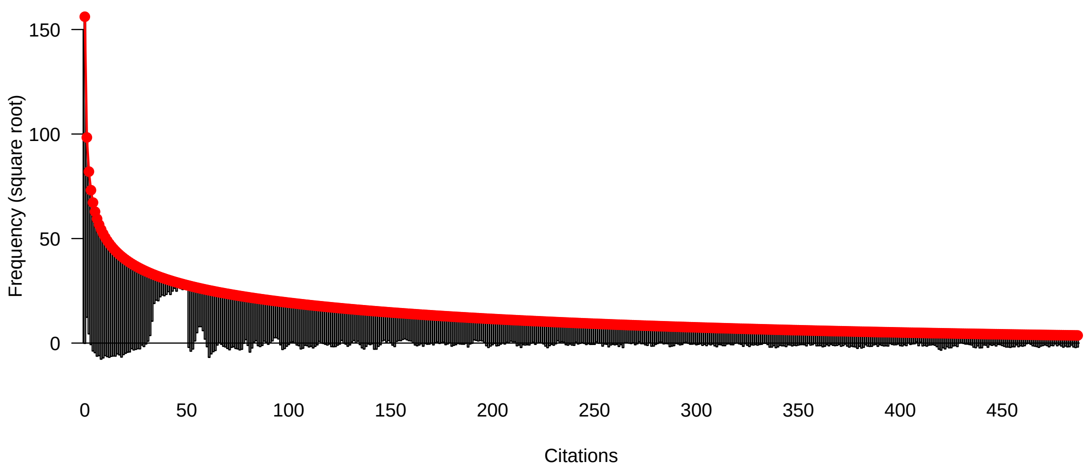


Figure 3: Rootogram (hanging histogram bars) and best fitting negative binomial distribution colored in red. The gap between 33 and 50 citations is due to the specific reporting of the Microsoft Academic Search program. The total number of citations is shown on a square-root scale.

Section 2.2 includes the likelihood of two authors in an article being in alphabetical order, to estimate the proportion of intentionally ordered author lists. The reasons for this calculation would be the observed empirical frequencies of the initial letters, thus resulting in the included articles of the top journals of Economics, Statistics, and Psychology. However, this could be a biased proxy for the true distribution of the first letters of surnames. Hence, we compared these frequencies to the frequency table published by Gray (1958) [15].

In contrast, Gray (1958) [15] reports the distribution for UK surnames only, which might differ from the frequency distribution of first letters of surnames globally. To further justify the results, we also compared our estimated distributions from the data against the top 100 U.S. surnames from the census in 2002. Figure 4 depicts these empirical distributions.

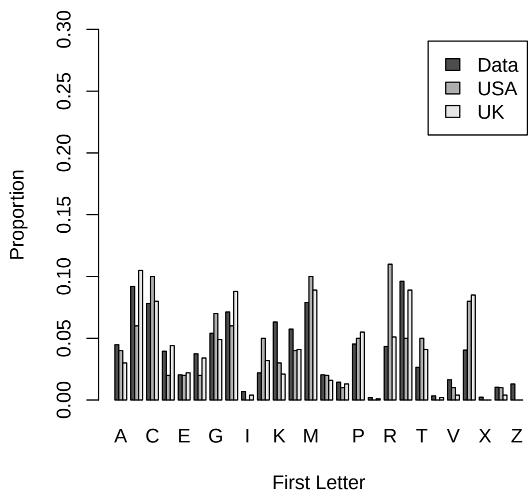


Figure 4: Empirical distribution of the first letter of surnames for our data set (dark-gray), the top 100 U.S. surnames (gray), and the UK surnames by Gray (1958) [15] (light-gray).

There are no large differences between the estimated probabilities, aside for some letters (e.g., 'R' or 'W') where we observe fewer authors in our data than one would expect when looking at the top 100 U.S. surnames or the results of Gray (1958) [15]. However, this did not affect the main findings. Differences in the resulting ratios are small, as shown in Figure 5 (analogously to Figure 2), based on the empirical distribution of UK surnames (also not different for the 100 U.S. surnames census data).

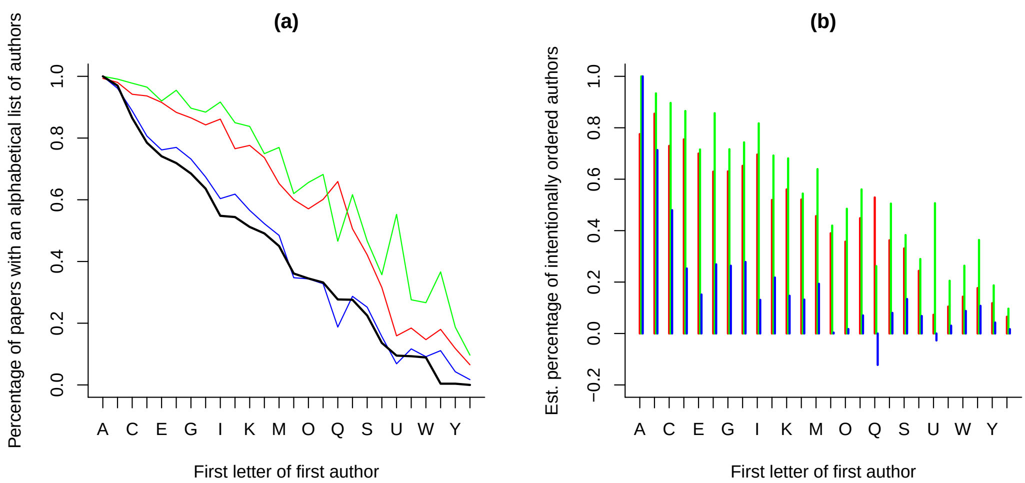


Figure 5: In contrast to Figure 2, we chose the empirical distribution of UK surnames reported by Gray (1958) [15] as a benchmark, i.e., the bold line first plot (**a**) depicts the probability for two random surnames being in alphabetical order according to this empirical distribution. In the second plot (**b**), we computed the ratio between the observed frequencies of ordered authors and the estimated probability (black baseline) as an estimate for the percentage of articles that were being intentionally set in alphabetical order.

B. MODEL SELECTION

First impressions of the underlying discrete probability of the citation counts can be obtained by the so-called Ord's plot (cf. Ord 1967 [25]). For our data, the plot indicates that the data are generated by a negative binomial distribution, which is also supported by the histogram, or rootogram (e.g. Wainer 1974 [37]). Lee *et al.* (2007) [22] observed similar behavior for patent citation counts. Comparing a zero-inflated negative binomial and zero-inflated Poisson model by a Vuong test (cf. Vuong 1989 [36]), the negative binomial model is significantly preferred, with a test statistic of $|z| = 308.4410$ (uncorrected). Less-complex models, such as the negative binomial model without zero inflation, can be ruled out due to their larger information criteria (the Akaike information criterion (AIC) is 1,787,653 for the negative binomial model and 1,771,146 for the zero-inflated model).

Moreover, the zero-inflated model allows for the comparison of the probability for being cited and the citation counts across the fields, whereas count data models without zero inflation measure the overall effect. For instance, the estimated coefficients for the indicator variables of research field are -0.235 ($\hat{\beta}_5^c$, Psychology) and -1.055 ($\hat{\beta}_6^c$, Statistics) for a negative binomial, without modeling the inflation of uncited articles. This confirms our results, namely that articles in Psychology are more often cited than in Statistics and that the latter articles are cited the least (in this particular group of the three research disciplines). However, it does not allow for interpretations regarding the excess of uncited articles.

Furthermore, the reported model results (in Table 3 of Section 3.3) include all introduced potential characteristics from Section 2 influencing citation counts as main effects. To provide a model with the best data fit, we also selected covariates and their interactions by stepwise minimizing AIC. The resulting model is discussed next as "model extensions" (in Appendix C).

C. MODEL EXTENSIONS

All results were obtained by a simple regression model, which meant an easier interpretation because we only focused on the direction of the main effects, despite the possibility that there could be interactions between the regressors. For instance, alphabetically sorted authors could have different implications for each research discipline. Although it is sometimes common to sort authors alphabetically (66.1% of all the included articles with more than one author in statistics), authors were less often sorted alphabetically in Psychology (24.7%) or Economics (77.1%).

Including interaction terms for the above-mentioned effects, the interpretation of the results does not change. We report the estimated coefficients and ratios for this more complex model in Table 4. All included interaction terms were found to have a significant influence. Moreover, the AIC is smaller compared to the model reported in Table 3.

To control for the fact that the probability for name-sharing authors is increased with an increasing number of authors, we estimated a further model with only partial data.

In particular, we only included articles that had exactly two authors. For this model (B), parameter estimates and ratios were shown in an analogous manner in Table 5. The results are in line with the results of the model described in Section 3.3, with a negative impact of authors sharing the same surnames, as well as more uncited articles of authors sharing the same surnames.

Table 4: Estimated parameters $\hat{\beta}_{A,i}$ with odds ratios $\hat{r}^z_{A,i}$ or incidence risk ratios $\hat{r}^c_{A,i}$ of the zero-inflated negative binomial model for the first alternative model (A) with p-values in parentheses.

Variable	i	Zero-inflation coefficients		Count coefficients	
		$\hat{\beta}^z_{A,i}$	$\hat{r}^z_{A,i}$	$\hat{\beta}^c_{A,i}$	$\hat{r}^c_{A,i}$
Regressors					
Intercept	0	2.139 (< 0.001)		3.721 (< 0.001)	
Field of research: Psychology	1	0.364 (< 0.001)	1.439	−0.214 (< 0.001)	0.807
Field of research: Statistics	2	−0.731 (< 0.001)	0.481	−1.092 (< 0.001)	0.336
Years since publication: in full years	3	−0.052 (< 0.001)	0.950	0.067 (< 0.001)	1.076
Title length: number of characters in title	4	0.016 (< 0.001)	1.016	−0.001 (< 0.001)	0.999
Number of authors	5	−4.085 (< 0.001)	0.017	0.011 (0.161)	1.011
Single author (additional effect)	6	—	—	−0.288 (< 0.001)	0.750
Alphabetically ordered authors: true	7	−0.018 (0.955)	0.982	0.157 (< 0.001)	1.170
Author name-sharing: existent	8	1.476 (0.029)	4.377	−0.211 (0.001)	0.810
Common author name: first author within top 100 surnames	9	−0.238 (< 0.001)	0.788	0.042 (0.002)	1.042
Interaction: number of authors in Psychology	10	—	—	−0.014 (0.068)	0.986
Interaction: number of authors in Statistics	11	—	—	0.010 (0.229)	1.010
Interaction: alph. ordered authors in Psychology	12	—	—	−0.139 (< 0.001)	0.870
Interaction: alph. ordered authors in Statistics	13	—	—	−0.071 (0.001)	0.932
$\log(\hat{\theta})$				−0.604 (< 0.0001)	
Summary Statistics					
AIC		1770226			
$\exp(\log(\hat{\theta}))$		0.547			
LR (null model)		22778.4			

Table 5: Estimated parameters $\hat{\beta}_{B,i}$ with odds ratios $\hat{r}^z_{B,i}$ and incidence risk ratios $\hat{r}^c_{B,i}$ of the zero-inflated negative binomial model for all articles of only two authors (alternative model B) and with p-values in parentheses.

Variable	i	Zero-inflation coefficients		Count coefficients	
		$\hat{\beta}^z_{B,i}$	$\hat{r}^z_{B,i}$	$\hat{\beta}^c_{B,i}$	$\hat{r}^c_{B,i}$
Regressors					
Intercept	0	−4.743 (< 0.001)		3.708 (< 0.001)	
Field of research: Psychology	1	−2.410 (< 0.001)	0.090	−0.126 (< 0.001)	0.881
Field of research: Statistics	2	−2.937 (< 0.001)	0.053	−0.997 (< 0.001)	0.369
Years since publication: in full years	3	−0.057 (0.003)	0.944	0.064 (< 0.001)	1.066
Title length: number of characters in title	4	0.030 (< 0.001)	1.031	−0.001 (< 0.001)	0.999
Number of authors	5	—	—	—	—
Alphabetically ordered authors: true	6	−1.044 (< 0.001)	0.352	0.205 (< 0.001)	1.228
Author name-sharing: existent	7	1.080 (0.072)	2.945	−0.176 (0.006)	0.839
Common author name: first author within top 100 surnames	8	−1.460 (0.051)	0.232	0.038 (0.086)	1.039
Interaction: number of authors in Psychology	9	—	—	—	—
Interaction: number of authors in Statistics	10	—	—	—	—
Interaction: alph. ordered authors in Psychology	11	—	—	−0.211 (< 0.001)	0.810
Interaction: alph. ordered authors in Statistics	12	—	—	−0.121 (0.004)	0.886
$\log(\hat{\theta})$				−0.561 (< 0.0001)	
Summary Statistics					
AIC		591036.7			
$\exp\big(\log(\hat{\theta})\big)$		0.571			
LR (null model)		5106.93			

REFERENCES

[1] Aksnes, D.W. (2006). Citation rates and perceptions of scientific contribution, *Journal of the American Society for Information Science and Technology*, **57**(2), 169–185.

[2] Albarrán, P.; Crespo, J.A.; Ortuño, I. and Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates, *Scientometrics*, **88**(2), 385–397.

[3] Ball, R. (2014). *Bibliometrie: Einfach-verständlich-nachvollziehbar*, Walter de Gruyter.

[4] Bornmann, L.; Mutz, R.; Marx, W.; Schier, H. and Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**(4), 857–879.

[5] Bornmann, L.; Schier, H.; Marx, W. and Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, **6**(1), 11–18.

[6] Bramoullé, Y. and Ductor, L. (2018). Title length, *Journal of Economic Behavior & Organization*, **150**, 311–324.

[7] Card, D. and DellaVigna, S. (2013). Nine facts about top journals in economics, *Journal of Economic Literature*, **51**(1), 144–61.

[8] Chen, C. (2012). Predictive effects of structural variation on citation counts, *Journal of the American Society for Information Science and Technology*, **63**(3), 431–449.

[9] Didegah, F. and Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties, *Journal of Informetrics*, **7**(4), 861–873.

[10] Ductor, L. (2015). Does co-authorship lead to higher academic productivity? *Oxford Bulletin of Economics and Statistics*, **77**(3), 385–407.

[11] Einav, L. and Yariv, L. (2006). What's in a surname? The effects of surname initials on academic success, *Journal of Economic Perspectives*, **20**(1), 175–187.

[12] Fourcade, M.; Ollion, E. and Algan, Y. (2015). The superiority of economists, *Journal of Economic Perspectives*, **29**(1), 89–114.

[13] Gnewuch, M. and Wohlrabe, K. (2017). Characteristics and citations in economics, *Scientometrics*, **110**(3), 1573–1578.

[14] Goodman, A.C.; Goodman, J.; Goodman, L. and Goodman, S. (2015). A few Goodmen: surname-sharing economist coauthors, *Economic Inquiry*, **53**(2), 1392–1395.

[15] Gray, P.G. (1958). Initial letters of surnames, *Applied Statistics*, **7**, 58–59.

[16] Greene, W.H. (2003). *Econometric Analysis*, Pearson Education India.

[17] Hamermesh, D.S. (2018). Citations in economics: measurement, uses, and impacts, *Journal of Economic Literature*, **56**(1), 115–56.

[18] Haslam, N.; Ban, L.; Kaufmann, L.; Loughnan, S.; Peters, K.; Whelan, J. and Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology, *Scientometrics*, **76**(1), 169–185.

[19] Heckman, J.J. and Moktan, S. (2018). *Publishing and promotion in economics: the tyranny of the top five*, Technical Report, National Bureau of Economic Research.

[20] Hirsch, J.E. (2005). An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences*, **102**(46), 16569–16572.

[21] Jacobs, J.A. (2013). *In Defense of Disciplines: Interdisciplinarity and Specialization in the Research University*, University of Chicago Press.

[22] Lee, Y.-G.; Lee, J.-D.; Song, Y.-I. and Lee, S.-J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: the case of KIST, *Scientometrics*, **70**(1), 27–39.

[23] Lindsey, D. (1989). Using citation counts as a measure of quality in science measuring what's measurable rather than what's valid, *Scientometrics*, **15**(3–4), 189–203.

[24] Onodera, N. and Yoshikane, F. (2015). Factors affecting citation rates of research articles, *Journal of the Association for Information Science and Technology*, **66**(4), 739–764.

[25] Ord, J. (1967). Graphical methods for a class of discrete distributions, *Journal of the Royal Statistical Society. Series A (General)*, 232–238.

[26] Parolo, P.D.B.; Pan, R.K.; Ghosh, R.; Huberman, B.A.; Kaski, K. and Fortunato, S. (2015). Attention decay in science, *Journal of Informetrics*, **9**(4), 734–745.

[27] Perry, M. and Reny, P.J. (2016). How to count citations if you must, *American Economic Review*, **106**(9), 2722–2741.

[28] Quandt, R.E. (1976). Some quantitative aspects of the economics journal literature, *Journal of Political Economy*, **84**(4, Part 1), 741–755.

[29] Radicchi, F.; Fortunato, S. and Castellano, C. (2008). Universality of citation distributions: toward an objective measure of scientific impact, *Proceedings of the National Academy of Sciences*, **105**(45), 17268–17272.

[30] Stegehuis, C.; Litvak, N. and Waltman, L. (2015). Predicting the long-term citation impact of recent publications, *Journal of Informetrics*, **9**(3), 642–657.

[31] Thomas, W.I. and Thomas, D.S. (1928). *The methodology of behavior study.* In "The Child in America: Behavior Problems and Programs" (A.A. Knopf, Ed.), chapter 13, pp. 553–576, New York, online: `https://brocku.ca/MeadProject/Thomas/Thomas_1928_13.html`.

[32] Van Dalen, H. and Henkens, K. (2001). What makes a scientific article influential? The case of demographers, *Scientometrics*, **50**(3), 455–482.

[33] Van Praag, C.M. and van Praag, B. (2008). The benefits of being economics professor A (rather than Z), *Economica*, **75**(300), 782–796.

[34] Varin, C.; Cattelan, M. and Firth, D. (2016). Statistical modelling of citation exchange between statistics journals, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **179**(1), 1–63.

[35] Vieira, E.S. and Gomes, J.A. (2010). Citations to scientific articles: its distribution and dependence on the article features, *Journal of Informetrics*, **4**(1), 1–13.

[36] Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica: Journal of the Econometric Society*, 307–333.

[37] Wainer, H. (1974). The suspended rootogram and other visual displays: an empirical validation, *The American Statistician*, **28**(4), 143–145.

[38] Webster, G.D.; Jonason, P.K. and Schember, T.O. (2009). Hot topics and popular papers in evolutionary psychology: analyses of title words and citation counts in evolution and human behavior, 1979–2008, *Evolutionary Psychology*, **7**(3), 147470490900700301.

[39] Yan, E. and Ding, Y. (2012). Scholarly network similarities: how bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other, *Journal of the American Society for Information Science and Technology*, **63**(7), 1313–1326.

[40] Zeileis, A.; Kleiber, C. and Jackman, S. (2008). Regression models for count data in R, *Journal of Statistical Software*, **27**(8), 1–25.

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 449–462
https://doi.org/10.57805/revstat.v20i4.381

On Uniform and α-Monotone Discrete Distributions

Authors: M.C. Jones
– Department of Mathematics and Statistics, The Open University, U.K.
m.c.jones@open.ac.uk

Received: October 2019 Revised: August 2020 Accepted: September 2020

Abstract:

- In this partly expository article, I am concerned with some simple yet fundamental aspects of discrete distributions that are either uniform or have α-monotone probability mass functions. In the univariate case, building on work of F.W. Steutel published in 1988, I look at Khintchine's theorem for discrete monotone distributions in terms of mixtures of discrete uniform distributions, along with similar results for discrete α-monotone distributions. In the multivariate case, I develop a new general family of multivariate discrete distributions with uniform marginal distributions associated with copulas and consider families of multivariate discrete distributions with α-monotone marginals associated with these.

Keywords:

- *Khintchine's theorem; multivariate geometric distribution; multivariate discrete uniform distribution; multivariate Poisson distribution.*

AMS Subject Classification:

- Primary 62E10, Secondary 62H05.

1. INTRODUCTION

In this partly expository article, I am concerned with some simple yet fundamental aspects of distributions on $\mathbb{N}_0 \equiv 0, 1, ...,$ whose probability mass functions (p.m.f.'s) p are uniform or more generally monotone nonincreasing or even more generally α-monotone (see below), together with certain extensions of these distributions to $\mathbb{N}_0^d \equiv \mathbb{N}_0 \times \cdots \times \mathbb{N}_0$, especially $\mathbb{N}_0^2$, and subsets thereof. As a prime example of a univariate distribution with a non-uniform monotone nonincreasing p.m.f. — a 'monotone p.m.f.' for short — think of the geometric distribution; the Poisson distribution turns out to be an example of an α-monotone distribution.

The main topics to be considered in this article, by section, are:

- **§2**. Khintchine's theorem for monotone distributions on $\mathbb{N}_0$, re-interpreted in terms of mixtures of discrete uniform distributions, and a consequent variance inequality for univariate discrete monotone distributions.
- **§3**. A general family of multivariate discrete distributions with uniform marginal distributions associated in an attractive yet novel way with copulas;
- **§4**. Univariate α-monotone distributions on $\mathbb{N}_0$ which, for $0 < \alpha < 1$, are a 'stronger' subset of monotone distributions, and which are of interest for $\alpha > 1$ also, when they can be non-monotone and include many familiar distributions. Originally introduced by Steutel (1988) [21], I pursue further interpretation and properties.
- **§5**. Families of multivariate discrete distributions with α-monotone marginals associated with the distributions of Sections 3 and 4. Their correlation structures are explicit and relatively straightforward.

Potential Bayesian applications of Khintchine's theorem for discrete distributions (§2) are to the provision of monotone prior distributions for discrete-valued parameters and of nonparametric priors for α-monotone discrete distributions (similar to e.g. Brunner & Lo, 1989 [5], in the continuous case). Families of multivariate discrete distributions with separation between marginal and dependence parameters (§3 and especially §5) can, as in the continuous case, form good test-beds for simulation studies; in particular, as a referee suggests, the opportunity arises to simulate correlated discrete variables with a given correlation matrix and univariate margins. Distributions with monotone and especially α-monotone marginals can be used as models for appropriate data too, of course. I look briefly at alternative multivariate geometric and Poisson distributions to those in e.g. Davy & Rayner (1996) [7] and Bermúdez & Karlis (2011) [3], respectively, while alternatives to existing multivariate binomial (e.g. Westfall & Young, 1989 [23]) and multivariate negative binomial (e.g. Shi & Valdez, 2014 [20]) distributions are also readily available but not developed explicitly.

All mathematical manipulations made in this article have the major benefit of being simple and direct. As I go along, it will often be useful to point out analogies and connections with results for continuous data which have uniform or α-monotone probability density functions (p.d.f.'s) f on $\mathbb{R}^+$, and their multivariate extensions.

2. DISCRETE KHINTCHINE'S THEOREM

Let f be a monotone p.d.f. on $\mathbb{R}^+$. Then, the renowned Khintchine's Theorem (Khintchine, 1938 [16], Feller, 1971 [9]) says that $X \sim f$ can be written as a uniform scale mixture, either as $X = UY$, where U and Y are independent, $U \sim \text{Uniform}(0,1)$ and $Y \sim G$ for some cumulative distribution function (c.d.f.) G on $\mathbb{R}^+$, or equivalently as $X|Y = y \sim \text{Uniform}(0,y)$, $Y \sim G$. If f is differentiable, then typically G has a p.d.f. g such that $g(x) = -xf'(x)$. (The distribution of Y is not absolutely continuous if f has support $(0,b)$ say, when $b < \infty$ and $f(b) > 0$; see Section 4.)

Implicit in Steutel's (1988) [21] paper on "discrete α-monotonicity" — of which, more in Section 4 — is a corresponding result to Khintchine's theorem in the discrete case. (See also the earlier work of Medgyessy, 1972 [17].) It is framed in terms of binomial thinning, as first proposed by Steutel and van Harn (1979) [22]. For values of $\theta \in [0,1]$, the random variable $N_{m,\theta}$ is the binomially thinned version of the count $m \in \mathbb{N}_0$ if

$$N_{m,\theta} \equiv \theta \circ m \equiv \sum_{j=1}^{m} B_j$$

where the sum is understood to be zero if $m = 0$. Here, $B_1, ..., B_m$ are independent Bernoulli(θ) random variables. (Note that if $\theta = 1$, $N_{m,\theta} = m$ and if $\theta = 0$, $N_{m,\theta} = 0$.) A useful equivalent way of expressing $N_{m,\theta} = \theta \circ m$ is as

$$N_{m,\theta} = \theta \circ m \sim \text{Binomial}(m,\theta)$$

where Binomial$(0,\theta)$ is understood to be the degenerate distribution at zero.

The above is binomial thinning for fixed θ and m, extensions to which are to mix over distributions for their random variable versions, Θ and/or M. So, consider the distribution of $N = \Theta \circ M \sim p$ on $\mathbb{N}_0$ where $\Theta \sim h$ on $(0,1)$, independently of $M \sim q$ on $\mathbb{N}_0$. This distribution can be expressed as

$$N|M = m \sim \text{BinMix}(m), \qquad M \sim q,$$

with the binomial mixture distribution 'BinMix' defined as follows: $N_m \equiv \Theta \circ m \sim \text{BinMix}(m)$ if

$$N_m|\Theta = \theta \sim \text{Binomial}(m,\theta), \qquad \Theta \sim h. \tag{2.1}$$

Steutel's (1988) [21] observation is that taking $\Theta \sim \text{Uniform}(0,1)$ is equivalent to p being a monotone p.m.f. on $\mathbb{N}_0$. I now note that in that case, where $h(\theta) = I(0 < \theta < 1)$ and $I(\cdot)$ denotes the indicator function,

$$N_m = \Theta \circ m \sim \text{Uniform}\{0, ..., m\},$$

that is, the binomial mixture distribution reduces to the uniform distribution on $\{0, ..., m\}$.

To see this, note that, for each $x \in \{0, ..., m\}$,

$$\int_0^1 \binom{m}{x} \theta^x (1-\theta)^{m-x}\, d\theta = \binom{m}{x} B(x+1, m-x+1) = \frac{1}{m+1}$$

(here, $B(\cdot,\cdot)$ is the beta function). This is, of course, a very special case of the beta-binomial distribution (see Johnson, Kemp and Kotz, 2005 [13], Section 6.9.2).

The discrete analogue of Khintchine's theorem can therefore be given most simply — and not unexpectedly given its continuous analogue — as a discrete uniform mixture, as in Result 2.1:

Result 2.1. *A p.m.f. p on $\mathbb{N}_0$ is monotone if and only if $N \sim p$ can be written as*

$$N|M = m \sim \text{Uniform}\{0, ..., m\}, \qquad M \sim q,$$

where q is any p.m.f. on $\mathbb{N}_0$. In fact, the p.m.f.s p and q are related by

$$p(n) = \sum_{m=n}^{\infty} \frac{q(m)}{m+1}, \qquad q(m) = (m+1)\,\{p(m) - p(m+1)\}\,. \tag{2.2}$$

Also, the corresponding c.d.f.s P and Q are related by

$$Q(n) = P(n) - (n+1)p(n+1).$$

Example 2.1.

(**a**) Let $N \sim \text{Geometric}(p)$, $0 < p < 1$, which has strictly decreasing p.m.f. In this case,

$$q(m) = (m+1)\,p^2(1-p)^m,$$

that is, $M \sim \text{NegativeBinomial}(2, p)$, which is the distribution of the sum of two independent Geometric(p) random variables.

(**b**) Let $N \sim \text{Poisson}(\mu)$ with $0 < \mu \le 1$. Then, p is monotone on $\mathbb{N}_0$, and Result 2.1 applies with

$$q(m) = (m+1-\mu)\,p(m).$$

One of a number of ways of interpreting q is that it is the distribution of $M_0 + B$ where $B \sim \text{Bernoulli}(\mu)$, independent of $M_0 \sim \text{Poisson}(\mu)$.

(**c**) Now let $M \sim \text{Poisson}(\lambda)$, $\lambda > 0$. Then, N has the strictly decreasing p.m.f.

$$p(n) = \frac{e^{-\lambda}}{\lambda} \sum_{j=n+1}^{\infty} \frac{\lambda^j}{j!} = \frac{1}{\lambda}\,\Gamma(\lambda; n+1)$$

where $\Gamma(\cdot;\cdot)$ is the incomplete gamma function ratio. From (2.3) below, $\mathbb{E}(N) = \lambda/2$ and $\mathbb{V}(N) = \lambda(6+\lambda)/12$, so p is overdispersed as well as decreasing.

(**d**) The distribution of part (c) is a special case of taking $q(m) = (m+1)\,r(m+1)/\mu_r$ where r is an arbitrary p.m.f. on $\mathbb{N}_0$ with finite mean μ_r. Then, $p(n) = \overline{R}(n)/\mu_r$ where $R(n) = P(R > n)$ and $R \sim r$, so p is clearly monotone.

(**e**) There is no distribution satisfying $p = q$. If there were, p must satisfy $p(m+1)/p(m) = m/(m+1)$, $m = 0, 1, ...$, and this was shown by Leo Katz in the 1940s not to correspond to a valid distribution (see Johnson *et al.*, 2005 [13], Section 2.3.1).

Either directly or as a consequence of more general results for mixed binomial thinning, it is easy to show that

$$\mathbb{E}(N) = \mathbb{E}(M)/2, \qquad \mathbb{V}(N) = \left[4\mathbb{V}(M) + 2\mathbb{E}(M) + \{\mathbb{E}(M)\}^2\right]/12. \tag{2.3}$$

Since $\mathbb{V}(M) \geq 0$ and $\mathbb{E}(M) = 2\mathbb{E}(N)$, the following variance-mean inequality arises.

Result 2.2. *If N follows a monotone p.m.f. on $\mathbb{N}_0$, then*

$$\mathbb{V}(N) \geq \mathbb{E}(N)\{1 + \mathbb{E}(N)\}/3,$$

and any monotone distribution is overdispersed if $\mathbb{E}(N) > 2$.

This inequality and observation arose in Jones and Marchand (2019) [15] from a different perspective. The inequality is the discrete analogue of the inequality $\mathbb{V}(X) \geq \{\mathbb{E}(X)\}^2/3$ of Johnson and Rogers (1951) [14] in the continuous monotone case.

3. MULTIVARIATE DISCRETE UNIFORM DISTRIBUTIONS

Write c and C for the p.d.f. and c.d.f. of an absolutely continuous copula on $(0,1)^d$ (e.g. Nelsen, 2006 [18], Joe, 1997 [11], 2014 [12]). This section and the next can be seen as an investigation of a role for such multivariate continuous uniform distributions in providing the dependence properties of certain multivariate discrete distributions, starting in this section with multivariate discrete distributions with discrete uniform marginal distributions, referred to from here on as multivariate discrete uniform distributions. Note that this is quite different from the use of a copula in conjunction with the discontinuous c.d.f.'s and quantile functions of discrete marginals, a common practice but with a number of "dangers and limitations", as discussed by Genest and Nešlehová (2007) [10]. That said, a multivariate discrete uniform distribution does *not* fulfil the same role for multivariate discrete distributions as a copula does for multivariate continuous distributions because univariate discrete c.d.f.'s, when considered as functions of their random variable, are not distributed as discrete uniforms i.e., if X has distribution F, and F is discrete, then $F(X)$ is not uniform. In contrast, $F(X)$ is (continuous) uniform when F is continuous.

The fact that a binomial distribution mixed over a continuous uniform distribution for its probability parameter is itself a discrete uniform distribution suggests that a multivariate discrete uniform distribution can be defined as the distribution of $(N_1, ..., N_d)$ on $\{0, ..., m_1\} \times \cdots \times \{0, ..., m_d\}$ such that

$$\begin{aligned} N_i | \Theta_i = \theta_i \;&\sim\; \text{Binomial}(m_i, \theta_i) \quad \text{independently for} \quad i = 1, ..., d, \\ \Theta^{(d)} \equiv \{\Theta_1, ..., \Theta_d\} \;&\sim\; c(\theta_1, ..., \theta_d). \end{aligned}$$

The joint p.m.f. of $(N_1, ..., N_d)$ is

$$\begin{aligned} &p_U(n_1, ..., n_d \,|\, m_1, ..., m_d) \\ &\quad = \left\{\prod_{i=1}^d \binom{m_i}{n_i}\right\} \int_0^1 \!\cdots\! \int_0^1 \left\{\prod_{i=1}^d \theta_i^{n_i}(1-\theta_i)^{m_i-n_i}\right\} c(\theta_1, ..., \theta_d)\; d\theta_1 \cdots d\theta_d\,. \end{aligned} \tag{3.1}$$

Its univariate marginal distributions are discrete uniform by construction because those of the copula are continuous uniform.

Moments of this construction are readily available and, in particular, correlations are determined by those of the copula as follows. Since $\mathrm{Cov}(N_i, N_j|\Theta^{(d)} = \theta^{(d)}) = 0$, it is the case that

$$\text{(3.2)} \qquad \mathrm{Cov}(N_i, N_j) = \mathrm{Cov}\{\mathbb{E}(N_i|\Theta^{(d)} = \theta^{(d)}), \mathbb{E}(N_j|\Theta^{(d)} = \theta^{(d)})\} = m_i m_j \mathrm{Cov}(\Theta_i, \Theta_j).$$

Also, since $\mathbb{V}(N_i) = m_i(m_i+2)/12$, $\mathbb{V}(N_j) = m_j(m_j+2)/12$, it is the case that

$$\text{(3.3)} \qquad \mathrm{Corr}(N_i, N_j) = \frac{m_i m_j \mathrm{Corr}(\Theta_i, \Theta_j)/12}{\sqrt{m_i(m_i+2)m_j(m_j+2)}/12} = \sqrt{\frac{m_i}{m_i+2}}\sqrt{\frac{m_j}{m_j+2}}\,\mathrm{Corr}(\Theta_i, \Theta_j).$$

So, while the correlation of N_i and N_j has the same sign as that of Θ_i and Θ_j, it reduces to one-third that of the copula in the binary case, and increases, tending to a factor of 1, as the marginal supports grow larger. Note that $\mathrm{Corr}(\Theta_i, \Theta_j)$ is Spearman's rho.

The existence of this simple relationship between discrete and continuous uniform correlations is a reason for preferring the current construction to discretisations of the copula, although the two can be very similar, as the following simple example shows.

Example 3.1. Consider the bivariate Farlie–Gumbel–Morgenstern (FGM) copula given by

$$C(u,v) = uv\{1 + \phi(1-u)(1-v)\}, \qquad c(u,v) = 1 + \phi(1-2u)(1-2v),$$

on $0 < u, v < 1$ with $-1 \le \phi \le 1$. Entering this into (3.1) when $d = 2$ gives

$$p_{FGM}(n_1, n_2) = \frac{1}{(m_1+1)(m_2+1)}\left\{1 + \phi\,\frac{(2n_1 - m_1)(2n_2 - m_2)}{(m_1+2)(m_2+2)}\right\};$$

its correlation, from (3.3) and e.g. Example 2.4 of Joe (1997) [11], is

$$\sqrt{\frac{m_1}{m_1+2}}\sqrt{\frac{m_2}{m_2+2}}\,\frac{\phi}{3}.$$

A natural discretisation of any C in the bivariate case is

$$\begin{aligned} p'(n_1, n_2) = {} & C\left(\frac{n_1+1}{m_1+1}, \frac{n_2+1}{m_2+1}\right) + C\left(\frac{n_1}{m_1+1}, \frac{n_2}{m_2+1}\right) \\ & - C\left(\frac{n_1+1}{m_1+1}, \frac{n_2}{m_2+1}\right) - C\left(\frac{n_1}{m_1+1}, \frac{n_2+1}{m_2+1}\right) \end{aligned}$$

which turns out in the FGM case to equate to

$$\text{(3.4)} \qquad p'_{FGM}(n_1, n_2) = \frac{1}{(m_1+1)(m_2+1)}\left\{1 + \phi\,\frac{(2n_1 - m_1)(2n_2 - m_2)}{(m_1+1)(m_2+1)}\right\};$$

this differs just a little from p_{FGM}. The correlation associated with this model, calculated directly from (3.4), is similar to that of p_{FGM}, but a little larger; it is

$$\sqrt{\frac{m_1(m_1+2)}{(m_1+1)^2}}\sqrt{\frac{m_2(m_2+2)}{(m_2+1)^2}}\,\frac{\phi}{3}.$$

Formula (3.1) is a particular way of constructing multivariate distributions with uniform univariate marginals. If a multivariate discrete uniform distribution is specified by other means, there is not necessarily a copula leading to it via construction (3.1). Even when there is, as with copula discretisation, there is not generally a unique copula leading to that distribution. The following simple, if extreme, example makes this clear.

Example 3.2. Let $d = 2$ and $m_1 = m_2 = 1$. In this case, the elements of the joint p.m.f. of (N_1, N_2) depend only on $p_U(0,0) \le 1/2$, since $p_U(0,1) = \{1-2p_U(0,0)\}/2$, $p_U(1,0) = \{1-2p_U(0,0)\}/2$ and $p_U(1,1) = p_U(0,0)$. Write $\mathbb{E}_C$ for expectation under the copula. Then, from (3.1), we have

$$\begin{aligned}
p_U(0,0) &= \mathbb{E}_C\{(1-\Theta_1)(1-\Theta_2)\} = \mathbb{E}_C(\Theta_1\Theta_2),\\
p_U(0,1) &= \mathbb{E}_C\{(1-\Theta_1)\Theta_2\} = \tfrac{1}{2} - \mathbb{E}_C(\Theta_1\Theta_2),\\
p_U(1,0) &= \mathbb{E}_C\{\Theta_1(1-\Theta_2)\} = \tfrac{1}{2} - \mathbb{E}_C(\Theta_1\Theta_2),\\
p_U(1,1) &= \mathbb{E}_C(\Theta_1\Theta_2).
\end{aligned}$$

Therefore, any copula with $\mathbb{E}_C(\Theta_1\Theta_2) = p_U(0,0)$ will give rise to this bivariate binary uniform distribution. (In fact, the uniform marginals of the copula are not required for this argument: the copula can be replaced by any distribution on $(0,1)\times(0,1)$ with marginal means equal to $1/2$ and $\mathbb{E}(\Theta_1\Theta_2) = p_U(0,0)$.) However, the product moment requirement translates to $\mathrm{Corr}(\Theta_1,\Theta_2) = 12p_U(0,0) - 3$, which restricts the existence of such a mixing distribution to when $1/6 \le p_U(0,0) \le 1/3$.

4. DISCRETE α-MONOTONICITY

I now return to the univariate domain. To set the scene, I first describe the situation in the continuous case. There, α-monotonicity was introduced by Olshen and Savage (1970) [19] (see also Dharmadhikari and Joag-Dev, 1988 [8], and Bertin, Cuculescu and Theodorescu, 1997 [4]): the distribution of a continuous random variable X is said to be α-monotone if and only if the distribution of X^α is monotone, $\alpha > 0$. Then, X can be written in the form $X = A_\alpha Y$ say, where $A_\alpha \sim \mathrm{Beta}(\alpha, 1)$, independently of $Y \sim g$ on $\mathbb{R}^+$, in a similar manner to Khintchine's theorem; equivalently, $X = U^{1/\alpha}Y$ where $U \sim \mathrm{Uniform}(0,1)$. Clearly $\alpha = 1$ corresponds to ordinary monotonicity. By construction, if a distribution is α_0-monotone say, then is it also α-monotone for all $\alpha > \alpha_0$. In particular, α-monotone distributions with $\alpha < 1$ are also ordinary monotone.

Providing an alternative view of an equivalent formulation of Abouammoh (1987/1988) [1], Steutel (1988) [21] first put forward discrete α-monotonicity in the following manner: for $\alpha > 0$, $N \sim p$ is discrete α-monotone if $N = A_\alpha \circ M_\alpha = U^{1/\alpha} \circ M_\alpha$, where $A_\alpha \sim \mathrm{Beta}(\alpha,1)$, $U \sim \mathrm{Uniform}(0,1)$ and either of these is independent of $M_\alpha \sim q_\alpha$ on $\mathbb{N}_0$. The distribution of N can now be recognized, from Section 2, as being that of

$$N|M_\alpha = m_\alpha \sim \mathrm{BetaBinomial}(m_\alpha, \alpha, 1), \qquad M_\alpha \sim q_\alpha, \tag{4.1}$$

where the BetaBinomial$(m_\alpha, \alpha, 1)$ distribution has p.m.f.

$$p_{BB1}(x) = \frac{\alpha\, m_\alpha!\,\Gamma(x+\alpha)}{x!\,\Gamma(m_\alpha+\alpha+1)} \tag{4.2}$$

for $x \in \{0, ..., m_\alpha\}$. This is because now $h(\theta) = \alpha\theta^{\alpha-1} I(0 < \theta < 1)$ in (2.1) so that the binomial mixture distribution becomes

$$\alpha \int_0^1 \binom{m_\alpha}{x} \theta^{x+\alpha-1}(1-\theta)^{m_\alpha - x}\, d\theta = \alpha \binom{m_\alpha}{x} B(x+\alpha, m_\alpha - x + 1) = p_{BB1}(x).$$

(4.1) and (4.2) lead directly to confirmation of Steutel's (1988) [21] formula

$$p(n) = \alpha \frac{\Gamma(n+\alpha)}{n!} \sum_{m=n}^{\infty} \frac{m!\, q_\alpha(m)}{\Gamma(m+\alpha+1)}.$$

Steutel then observes that

$$(n+\alpha)p(n) - (n+1)p(n+1) = \alpha q_\alpha(n) \tag{4.3}$$

from which it can be concluded that discrete α-monotonicity corresponds to p having the simple property that

$$(n+\alpha)p(n) \geq (n+1)p(n+1).$$

Here, the inequality is strict except when $q_\alpha(n) = 0$. The corresponding c.d.f.s P and Q_α are related by

$$\alpha Q_\alpha(n) = \alpha P(n) - (n+1)p(n+1),$$

which can be readily checked to give rise to (4.3). Comments above on continuous α-monotonicities for various values of α continue to hold in the discrete case.

It can be added that (4.3) can also be written

$$q(n) = (1-\alpha)p(n) + \alpha\, q_\alpha(n) \tag{4.4}$$

where $q = q_1$ is as at (2.2) in Result 2.1. To corroborate and interpret (4.4) in the case that $0 < \alpha \leq 1$, an alternative way of expressing α-monotonicity arises from writing $A_\alpha = UV$ where $U \sim \text{Uniform}(0,1)$ independently of some appropriate V; this is possible when $0 < \alpha \leq 1$ because then Beta$(\alpha, 1)$ is monotone (nonincreasing). Moreover, Beta$(\alpha, 1)$ is then a distribution on a finite interval with non-zero density at its upper endpoint. As signposted at the start of Section 2, the density of V is not $-xf'(x)$ if f has support $(0, b)$ and $f(b) > 0$; in fact,

$$V \sim \begin{cases} Y & \text{with probability } 1-\alpha, \\ b & \text{with probability } \alpha, \end{cases}$$

where $Y \sim -xf'(x)/\{1 - f(b)\}$ on $(0, b)$. When $b = 1$ and $h(x) = \alpha x^{\alpha-1}$ so that $h(1) = \alpha$, it turns out that $-xh'(x)/\{1 - h(1)\} = h(x)$. In the case of discrete α-monotonicity with $0 < \alpha \leq 1$, it follows that $N = A_\alpha \circ M = (UV) \circ M = U \circ (V \circ M)$ so that $N = U \circ N_0$ where $U \sim \text{Uniform}(0,1)$ and

$$N_0 \sim \begin{cases} N & \text{with probability } 1-\alpha, \\ M & \text{with probability } \alpha, \end{cases}$$

which is immediately seen to be equivalent to (4.4).

By any of a number of routes, it can be shown that, for α-monotone distributions for any $\alpha > 0$,

$$\mathbb{E}(N) = \frac{\alpha \mathbb{E}(M_\alpha)}{\alpha + 1}, \quad \mathbb{V}(N) = \frac{\alpha\left[(\alpha+1)^2 \mathbb{V}(M_\alpha) + (\alpha+1)\mathbb{E}(M_\alpha) + \{\mathbb{E}(M_\alpha)\}^2\right]}{(\alpha+1)^2(\alpha+2)}.$$

Since $\mathbb{V}(M_\alpha) \geq 0$ and $\mathbb{E}(M_\alpha) = (\alpha+1)\mathbb{E}(N)/\alpha$, the following variance-mean inequality ensues.

Result 4.1. *If N follows an α-monotone p.m.f. on $\mathbb{N}_0$ for all $\alpha \geq \alpha_{min}$ say, then*

$$\mathbb{V}(N) \; \geq \frac{\mathbb{E}(N)\{\alpha_{min} + \mathbb{E}(N)\}}{\alpha_{min}(\alpha_{min}+2)} \geq \frac{\mathbb{E}(N)\{\alpha + \mathbb{E}(N)\}}{\alpha(\alpha+2)}.$$

The 'outside' inequality is essentially Theorem 3.1 of Abouammoh, Ali and Mashhour (1994) [2] with $a = 0$ and Corollary 5.3.21 of Bertin *et al.* (1997) [4]. An α-monotone distribution is thereby guaranteed to be overdispersed if $\mathbb{E}(N) > \alpha_{min}(\alpha_{min}+1)$. Of course, the outside inequality in Result 4.1 reduces to Result 2.2 when $\alpha = 1$.

Example 4.1.

(**a**) $N \sim \text{Geometric}(p)$ is α-monotone for $\alpha \geq 1-p \equiv \alpha_{min}$. Using (4.3), the corresponding p.m.f. of M_α is

$$q_\alpha(m) = \{(m+1)p - (1-\alpha)\}p(1-p)^m/\alpha.$$

As noted in Example 2.1(a), $M_1 \sim \text{NegativeBinomial}(2,p)$ while it can now also be observed that M_{1-p} has the distribution of $M_1 + 1$. The dispersion inequality for α-monotone distributions confirms the overdispersion of the geometric distribution for all $0 < p < 1$.

(**b**) Let $N \sim \text{Poisson}(\mu)$ with $0 < \mu \leq \alpha$. Then, the Poisson p.m.f. p is α-monotone on $\mathbb{N}_0$, and formula (4.3) applies to give

$$q_\alpha(m) = (m + \alpha - \mu)\, p(m)/\alpha.$$

Now, q_α is the distribution of $M_0 + B$ where $B \sim \text{Bernoulli}(\mu/\alpha)$, independent of $M_0 \sim \text{Poisson}(\mu)$. In particular, q_μ is the length-biased form of the Poisson distribution which is, in fact, the distribution of $M_0 + 1$. The dispersion inequality is, of course, not satisfied for any $\mu > 0$.

(**c**) Both of the above examples together with binomial and negative binomial distributions are covered by the Katz family, for which

$$(1+n)\, p(n+1) = (a + bn)\, p(n);$$

see Section 2.3.1 of Johnson *et al.* (2005) [13]. In general, $a > 0$ and $b < 1$, but α-monotonicity restricts the range of a to $0 < a \leq \alpha$. For any Katz distribution,

$$q_\alpha(m) = \{(\alpha - a) + (1-b)m\}p(m)$$

reducing to $q_a(m) = (1-b)mp(m)/a$ when $\alpha = a$. Let $K_{a,b}$ be a random variable following the Katz distribution with parameters a and b. Then, the latter length-biased distribution is also the distribution of $K_{a+b,b} + 1$. Since $\mathbb{E}(K_{a,b}) = a/(1-b)$ and $\mathbb{V}(K_{a,b}) = a/(1-b)^2$, the dispersion inequality yields overdispersion if $(a+1)(1-b) < 1$ while a Katz distribution is actually overdispersed for $0 < b < 1$. The general results reduce to those of part (a) when $a = b = 1-p$ and part (b) when $a = \mu$, $b = 0$. They give results for the $\text{Binomial}(k,p)$ distribution when $a = kp/(1-p)$, $b = -p/(1-p)$, and to the $\text{NegativeBinomial}(k,p)$ distribution when $a = k(1-p)$, $b = (1-p)$.

5. MULTIVARIATE DISCRETE DISTRIBUTIONS WITH α-MONOTONE UNIVARIATE MARGINALS

Combining Sections 2 and 3 further, it is natural to develop discrete distributions on $\mathbb{N}_0^d$ with monotone univariate marginals as the distribution of $N^{(d)} \equiv (N_1, ..., N_d)$ where

$$\begin{aligned} N_i|M_i = m_i, \Theta_i = \theta_i \;&\sim\; \text{Binomial}(m_i, \theta_i) \quad \text{independently for} \quad i = 1, ..., d, \\ M^{(d)} \equiv \{M_1, ..., M_d\} \;&\sim\; q(m_1, ..., m_d), \\ \Theta^{(d)} \equiv \{\Theta_1, ..., \Theta_d\} \;&\sim\; c(\theta_1, ..., \theta_d), \end{aligned}$$

where q is now an arbitrary p.m.f. on $\mathbb{N}_0^d$ and $M^{(d)}$ is independent of $\Theta^{(d)}$. This is, of course, equivalent to mixing the multivariate discrete uniform distribution of Section 3 over q:

$$N^{(d)}|M^{(d)} = \{m_1, ..., m_d\} \sim p_U(n_1, ..., n_d|m_1, ..., m_d), \qquad M^{(d)} \sim q(m_1, ..., m_d).$$

To additionally fold in the work of Section 4, to provide multivariate discrete distributions with α-monotone marginal distributions (more properly $\alpha^{(d)}$-monotone marginal distributions where $\alpha^{(d)} \equiv \{\alpha_1, ..., \alpha_d\}$), the key is to replace $\Theta^{(d)}$ by $\Theta_\alpha^{(d)} \equiv \{\Theta_1^{1/\alpha_1}, ..., \Theta_d^{1/\alpha_d}\}$. Let the resulting random variable be $N_\alpha^{(d)}$. The joint p.m.f. of $N_\alpha^{(d)}$ is

$$\begin{aligned} p_D(n_1, ..., n_d; \alpha_1, ...\alpha_d) \;=\; & \sum_{m_1=n_1}^{\infty} \cdots \sum_{m_d=n_d}^{\infty} q(m_1, ..., m_d) \left\{ \prod_{i=1}^d \binom{m_i}{n_i} \right\} \\ & \times \int_0^1 \cdots \int_0^1 \left\{ \prod_{i=1}^d \theta_i^{n_i/\alpha_i} (1 - \theta_i^{1/\alpha_i})^{m_i - n_i} \right\} c(\theta_1, ..., \theta_d)\; d\theta_1 \cdots d\theta_d \,. \end{aligned} \tag{5.1}$$

Its univariate marginal distributions have the α_1-monotone, α_2-monotone, ..., α_d-monotone p.m.f.'s of Section 4 by construction. The form of (5.1) involves d infinite sums and integrals but, as will be seen below, certain special cases simplify considerably. Moments remain readily available and correlations are as follows. Using (2.3) and (3.2),

$$\text{Cov}(N_i, N_j) = \mathbb{E}(M_i M_j)\, \text{Cov}(\Theta_i^{1/\alpha_i}, \Theta_j^{1/\alpha_j}) + \frac{\alpha_i}{\alpha_i + 1} \frac{\alpha_j}{\alpha_j + 1} \text{Cov}(M_i, M_j)$$

so that

$$\begin{aligned} & \text{Corr}(N_i, N_j) \\ = & \frac{\mathbb{E}(M_i M_j)\, \text{Corr}(\Theta_i^{1/\alpha_i}, \Theta_j^{1/\alpha_j}) + \sqrt{\alpha_i(\alpha_i + 2)\alpha_j(\alpha_j + 2)}\, \text{Cov}(M_i, M_j)}{\sqrt{\left[(\alpha_i + 1)^2\, \mathbb{V}(M_i) + (\alpha_i + 1)\, \mathbb{E}(M_i) + \{\mathbb{E}(M_i)\}^2\right] \left[(\alpha_j + 1)^2\, \mathbb{V}(M_j) + (\alpha_j + 1)\, \mathbb{E}(M_j) + \{\mathbb{E}(M_j)\}^2\right]}} \,. \end{aligned} \tag{5.2}$$

In the following two subsections, I will take a brief look at two major particular cases of this in terms of the form of distribution for M. These distributions and their properties are analogues of those given in Section 3 of Bryson and Johnson (1982) [6] in the continuous case when $d = 2$. They are theoretically interesting but for the most part may prove to have limited practical applicability.

5.1. When $M_1, ..., M_d$ are mutually independent

Let $M_i \sim q_i$, independently for $i = 1, ..., d$. This allows the dependence structure of p_D to depend only on that of C ameliorated by the value of $\alpha^{(d)}$. The joint p.d.f. of $N_\alpha^{(d)}$ is given by the obvious small change to (5.1). The correlation of N_i and N_j, given by (5.2), reduces to

$$\begin{aligned}\mathrm{Corr}(N_i, N_j) &= \sqrt{\frac{\mathbb{E}(M_i)}{(\alpha_i+1)^2\mathbb{D}(M_i)+\mathbb{E}(M_i)+\alpha_i+1}} \\ &\quad\times\sqrt{\frac{\mathbb{E}(M_j)}{(\alpha_j+1)^2\mathbb{D}(M_j)+\mathbb{E}(M_j)+\alpha_j+1}}\ \mathrm{Corr}(\Theta_i^{1/\alpha_i}, \Theta_j^{1/\alpha_j}).\end{aligned} \tag{5.3}$$

where $\mathbb{D}(M) = \mathbb{V}(M)/\mathbb{E}(M)$ is the index of dispersion of M. Again, this has the same sign as the correlation associated with the copula and is always a reduction of the absolute value of the correlation compared with that of the copula, sometimes considerably so.

Example 5.1. This example concerns a family of multivariate distributions with geometric marginal distributions. Following Example 2.1(a), let $q_i(m) = (m+1)\,p_i^2(1-p_i)^m$ with $\mathbb{E}(M_i) = 2(1-p_i)/p_i$ and $\mathbb{V}(M_i) = 2(1-p_i)/p_i^2$, $i = 1, ..., d$. The corresponding multivariate geometric distribution arises by taking $\alpha_1 = \cdots = \alpha_d = 1$. Reduction of (5.1) in this case requires simplification of terms of the form $\sum_{m=n}^\infty (m+1)p^2(1-p)^m\binom{m}{n}\theta^n(1-\theta)^{m-n}$ which is achieved by noting that, with $0 < \psi \equiv (1-p)(1-\theta) < 1$,

$$\begin{aligned}\sum_{m=n}^\infty (m+1)\binom{m}{n}\psi^{m-n} &= (n+1)\sum_{m=n}^\infty \binom{m+1}{n+1}\psi^{m-n} \\ &= (n+1)\sum_{j=0}^\infty \binom{n+j+1}{j}\psi^j = \frac{n+1}{(1-\psi)^{n+2}}.\end{aligned}$$

This results in the joint p.m.f.

$$\begin{aligned}&p_G(n_1, ..., n_d; p_1, ..., p_d) \\ &= \prod_{i=1}^d (n_i+1)p_i^2(1-p_i)^{n_i}\int_0^1\cdots\int_0^1\left[\prod_{i=1}^d \frac{\theta_i^{n_i}}{\{1-(1-p_i)(1-\theta_i)\}^{n_i+2}}\right]c(\theta_1, ..., \theta_d)\, d\theta_1\cdots d\theta_d\end{aligned}$$

with correlations

$$\mathrm{Corr}(N_i, N_j) = \frac{1}{3}\sqrt{(1-p_i)(1-p_j)}\,\mathrm{Corr}(\Theta_i, \Theta_j).$$

The correlations associated with this family of multivariate geometric distributions are therefore limited to the range $-1/3 < \mathrm{Corr}(N_i, N_j) < 1/3$, although the range of correlations decreases as the p_i's increase.

Example 5.2. In a similar manner to Example 5.1, this example concerns a family of multivariate distributions with Poisson marginals. It arises by taking $q_i(m) = \mu_i^{m-1}e^{-\mu_i}/(m-1)!$, $m = 1, 2, ...$, and $\alpha_j = \mu_j$, $j = 1, ..., d$ (cf. Example 4.1(b)). In this case, simplification of (5.1) requires simplification of sums of the form $\sum_{m=n}^\infty e^{-\mu}\mu^{m-1}\binom{m}{n}\theta^{n/\mu}(1-\theta^{1/\mu})^{m-n}/(m-1)!$. Now, with $\Omega \equiv \mu(1-\theta^{1/\mu}) > 0$,

$$\sum_{m=n}^\infty m\frac{\Omega^{m-n}}{(m-n)!} = \sum_{m=n}^\infty (m-n)\frac{\Omega^{m-n}}{(m-n)!} + n\sum_{m=n}^\infty \frac{\Omega^{m-n}}{(m-n)!} = (\Omega+n)e^\Omega.$$

The corresponding joint p.m.f. is

$$p_P(n_1,...,n_d;\mu_1,...,\mu_d)$$
$$= \prod_{i=1}^d \frac{\mu^{n_i}}{n_i!}\int_0^1\cdots\int_0^1 \left\{\prod_{i=1}^d \theta_i^{n_i/\mu_i}\left(1-\theta_i^{1/\mu_i}+\frac{n_i}{\mu_i}\right)e^{-\mu_i\theta_i^{1/\mu_i}}\right\} c(\theta_1,...,\theta_d)\; d\theta_1\cdots d\theta_d\,.$$

Since $\mathbb{E}(M_i)=\mu_i+1, \mathbb{V}(M_i)=\mu_i,\ i=1,...,d$, the correlations associated with these distributions are

$$\mathrm{Corr}(N_i,N_j) = \frac{1}{\sqrt{(\mu_i+2)(\mu_j+2)}}\,\mathrm{Corr}(\Theta_i^{1/\mu_i},\Theta_j^{1/\mu_j})$$

so that $-1/2 < \mathrm{Corr}(N_i,N_j) < 1/2$. In this case, the range of correlations decreases as the mean parameters increase.

5.2. When $M_1,...,M_d$ are equal or most strongly dependent

Let $M_1=\cdots=M_d=M$ say, $i=1,...,d$, with $M\sim q_0$. This particular comonotonicity also allows the dependence structure of p_D to depend on that of C, but with an opportunity for higher correlations. Let $n_{\max}=\max(n_1,...,n_d)$. The joint p.d.f. of $N_\alpha^{(d)}$ is given by

$$p_D(n_1,...,n_d;\alpha_1,...,\alpha_d)$$
$$= \sum_{m=n_{\alpha,\max}}^{\infty} q_0(m)\left\{\prod_{i=1}^d\binom{m}{n_i}\right\}\int_0^1\cdots\int_0^1\left\{\prod_{i=1}^d \theta_i^{n_i/\alpha_i}(1-\theta_i^{1/\alpha_i})^{m-n_i}\right\} c(\theta_1,...,\theta_d)\; d\theta_1\cdots d\theta_d\,.$$

Its correlations are, from (5.2),

$$\rho_{ij} \equiv \mathrm{Corr}(N_i,N_j)$$
$$= \frac{\{\mathbb{D}(M)+\mathbb{E}(M)\}\,\mathrm{Corr}(\Theta_i^{1/\alpha_i},\Theta_j^{1/\alpha_j})+\sqrt{\alpha_i(\alpha_i+2)\alpha_j(\alpha_j+2)}\,\mathbb{D}(M)}{\sqrt{\left[(\alpha_i+1)^2\mathbb{D}(M)+\mathbb{E}(M)+\alpha_i+1\right]\left[(\alpha_j+1)^2\mathbb{D}(M)+\mathbb{E}(M)+\alpha_j+1\right]}}\,, \tag{5.4}$$

which are all equal if $\alpha_1=\cdots=\alpha_d$. If r_{ij} denotes the correlation at (5.3) when both M_i and M_j have the distribution of M, then

$$\rho_{ij} = r_{ij} + \frac{\mathbb{D}(M)\left\{\mathrm{Corr}(\Theta_i^{1/\alpha_i},\Theta_j^{1/\alpha_j})+\sqrt{\alpha_i(\alpha_i+2)\alpha_j(\alpha_j+2)}\right\}}{\sqrt{\left[(\alpha_i+1)^2\mathbb{D}(M)+\mathbb{E}(M)+\alpha_i+1\right]\left[(\alpha_j+1)^2\mathbb{D}(M)+\mathbb{E}(M)+\alpha_j+1\right]}}$$

which is typically greater than r_{ij}, certainly whenever $\alpha_i(\alpha_i+2)\alpha_j(\alpha_j+2)>1$.

Example 5.3. While in Sections 3 and 5.1 the independence copula with density $c(\theta_1,...,\theta_d)=\prod_{i=1}^d I(0<\theta_i<1)$ results in distributions with independent marginals, this is not the case here because of the commonality of M. In fact, using the independence copula, the joint p.m.f. of $N_\alpha^{(d)}$ depends only on $n_{\max}$ and is given by

$$p_I(n_1,...,n_d;\alpha_1,...,\alpha_d) = \sum_{m=n_{\max}}^{\infty} q_0(m)(m!)^d\prod_{i=1}^d\frac{\alpha_i\,\Gamma(n_i+\alpha_i)}{n_i!\,\Gamma(m+1+\alpha_i)},$$

reducing to

$$p_I(n_1, ..., n_d; 1, ..., 1) \;=\; \sum_{m=n_{1,\max}}^{\infty} \frac{q_0(m)}{(m+1)^d}\,.$$

The corresponding correlations are, in general,

$$\mathrm{Corr}(N_i, N_j) \;=\; \sqrt{\frac{\alpha_i(\alpha_i+2)}{(\alpha_i+1)^2\,\mathbb{D}(M)+\mathbb{E}(M)+\alpha_i+1}}\;\sqrt{\frac{\alpha_j(\alpha_j+2)}{(\alpha_j+1)^2\,\mathbb{D}(M)+\mathbb{E}(M)+\alpha_j+1}}\;\mathbb{D}(M)\,,$$

which are all positive. When $\alpha_1 = \cdots = \alpha_d = 1$,

$$0 \;<\; \mathrm{Corr}(N_i, N_j) \;=\; \frac{3\mathbb{D}(M)}{4\mathbb{D}(M)+\mathbb{E}(M)+2} \;<\; \frac{3}{4}\,.$$

Example 5.4. For a general copula, let us contrast the correlation structure associated with the specific multivariate geometric and Poisson distributions of Examples 5.1 and 5.2 when $M_1, ..., M_d$ are independent with the corresponding distributions when $M_1 = \cdots = M_d = M$.

(**a**) Let $\alpha_1 = \cdots = \alpha_d = 1$ and $M \sim \mathrm{NegativeBinomial}(2, p)$. Then, the corresponding family of multivariate distributions with Geometric(p) marginals has correlations

$$\mathrm{Corr}(N_i, N_j) = \frac{1}{2} + \frac{(3-2p)\,\mathrm{Corr}(\Theta_i, \Theta_j)}{6}.$$

In this case, $0 < \mathrm{Corr}(N_i, N_j) < 1$, contrasting with a range of $(-1/3, 1/3)$ in Example 5.1. In fact, these correlations are always greater than those when $p_i = p_j = p$ in the independent M's case because $\alpha(\alpha+2) = 3 > 1$. In the case of the independence copula as in Example 5.3, $\mathrm{Corr}(N_i, N_j) = 1/2$.

(**b**) Let $\alpha_1 = \cdots = \alpha_d = \mu$ and $M = M_1 + 1$ where $M_1 \sim \mathrm{Poisson}(\mu)$, as in Example 5.2. Then, the corresponding family of multivariate Poisson distributions has correlations

$$\mathrm{Corr}(N_i, N_j) = \left(\frac{\mu}{\mu+1}\right)^2 + \frac{(\mu^2+3\mu+1)\,\mathrm{Corr}(\Theta_i^{1/\mu}, \Theta_j^{1/\mu})}{(\mu+1)^2(\mu+2)}.$$

It is certainly the case that $-1/2 < \mathrm{Corr}(N_i, N_j) < 1$ (contrasting with $(-1/2, 1/2)$ in Example 5.2) although slightly more negative correlation is possible for certain very small μ. The correlation is greater than that when $\mu_i = \mu_j$ in Example 5.2 whenever $\mathrm{Corr}(\Theta_i^{1/\mu}, \Theta_j^{1/\mu}) > -\mu(\mu+2)$. In the case of the independence copula, $0 < \mathrm{Corr}(N_i, N_j) = \mu^2/(\mu+1)^2 < 1$.

Finally, if $M_1, ..., M_d$ are not the same, then the strongest dependence is comonotonicity or the Fréchet upper bound. The expression for p_D does not simplify but the pair $\{N_i, N_j\}$ can be more highly correlated in comparison to Section 5.1.

ACKNOWLEDGMENTS

I am very grateful to the referees for their helpful comments which led to a much improved paper.

REFERENCES

[1] Abouammoh, A.M. (1987/1988). On discrete α-unimodality, *Statistica Neerlandica*, **41**, 239–244. Correction, **42**, 141.

[2] Abouammoh, A.M.; Ali, A.M. and Mashhour, A.F. (1994). On characterizations and variance bounds of discrete α-unimodality, *Statistical Papers*, **35**, 151–161.

[3] Bermúdez, L. and Karlis, D. (2011). Bayesian multivariate Poisson models for insurance ratemaking, *Insurance: Mathematics and Economics*, **48**, 226–236.

[4] Bertin, E.M.J.; Cuculescu, I. and Theodorescu, R. (1997). *Unimodality of Probability Measures*, Kluwer, Dordrecht.

[5] Brunner, L.J. and Lo, A.Y. (1989). Bayes methods for a symmetric unimodal density and its mode, *Annals of Statistics*, **17**, 1550–1566.

[6] Bryson, M.C. and Johnson, M.E. (1982). Constructing and simulating multivariate distributions using Khintchine's theorem, *Journal of Statistical Computation and Simulation*, **16**, 129–137.

[7] Davy, P.J. and Rayner, J.C.W. (1996). Multivariate geometric distributions, *Communications in Statistics – Theory and Methods*, **25**, 2971–2987.

[8] Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, Convexity, and Applications*, Academic Press, Boston, MA.

[9] Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. 2, Wiley, New York.

[10] Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data, *Astin Bulletin*, **37**, 475–515.

[11] Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

[12] Joe, H. (2014). *Dependence Modeling with Copulas*, Chapman & Hall/CRC Press, Boca Raton, FL.

[13] Johnson, N.L.; Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*, 3rd ed., Wiley, Hoboken, NJ.

[14] Johnson, N.L. and Rogers, C.A. (1951). The moment problem for unimodal distributions, *Annals of Mathematical Statistics*, **22**, 433–439.

[15] Jones, M.C. and Marchand, É. (2019). Multivariate discrete distributions via sums and shares, *Journal of Multivariate Analysis*, **171**, 83–93.

[16] Khintchine, A.Y. (1938). On unimodal distributions, *Izvestiya Nauchno-Issledovatel'skogo Instituta Matematiki i Mekhaniki*, **2**, 1–7.

[17] Medgyessy, P. (1972). On the unimodality of discrete distributions, *Periodica Mathematica Hungarica*, **2**, 245–257.

[18] Nelsen, R.B. (2006). *An Introduction to Copulas* 2nd ed., Springer, New York.

[19] Olshen, R.A. and Savage, L.J. (1970). A generalized unimodality, *Journal of Applied Probability*, **7**, 21–34.

[20] Shi, P. and Valdez, E.A. (2014). Multivariate negative binomial models for insurance claim counts, *Insurance: Mathematics and Economics*, **55**, 18–29.

[21] Steutel, F.W. (1988). Note on discrete α-unimodality, *Statistica Neerlandica*, **42**, 137–140.

[22] Steutel, F.W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability, *Annals of Probability*, **7**, 893–899.

[23] Westfall, P.H. and Young, S.S. (1989). P value adjustments for multiple tests in multivariate binomial models, *Journal of the American Statistical Association*, **84**, 780–786.

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 463–479
https://doi.org/10.57805/revstat.v20i4.383

Smooth PLS Regression for Spectral Data*

Authors: Athanasios Kondylis
– Philip Morris International R&D, Philip Morris Products S.A.,
Quai Jeanrenaud 5, 2000, Neuchâtel, Switzerland
athanasios.kondylis@pmi.com

Received: March 2020 Revised: August 2020 Accepted: October 2020

Abstract:

- Partial least squares (PLS) regression reduces the regression problem from a large-p number of interrelated predictors to a small-m number of extracted factors. These use information for predicting the response making PLS regression models extremely good for prediction purposes. The PLS regression coefficient vector is determined by the PLS factor loadings which drive the dimension reduction process; it should therefore be smooth, especially when the factor subspace dimension is small. We explore smooth alternatives for PLS regression revisiting a topic that triggered the research interest over the last two decades. We use for this the discrete wavelet transform focusing on PLS regression applications in near infra-red spectroscopy.

Keywords:

- *PLS regression; Krylov subspaces; discrete wavelet transform; spectroscopy.*

AMS Subject Classification:

- 62J07.

*The opinions expressed in this text are those of the author and do not necessarily reflect the views of Philip Morris International.

1. INTRODUCTION

Spectral data are characterized by a large number of interrelated measurements, intensities and absorptions, which are regularly recorded across a range of wavelengths. They are recorded by means of modern instruments and are often used as predictors in regression problems. In near infra-red (NIR) spectroscopy, in the food industry, for instance, samples of meat are analyzed for their fat content, and their NIR spectra are then used to predict fat concentration. Similar applications may be found in agriculture for the determination of properties of grains, in oil industry, in the analysis of pharmaceuticals, etc.

Using the spectral measurements as predictors in a regression problem limits traditional regression methods and implies the use of high-dimensional regression techniques. Partial least squares (PLS) regression has been for a long time implemented to deal with such regression problems, see [1]. PLS methods are based on reducing the dimension of the regression problem to a small-m number of factors rather than a large-p number of variables. This is achieved using information on the response variable, making PLS regression models excellent for prediction purposes.

More than twenty years have passed since the first smooth PLS regression has been presented in [2]. The authors have been motivated by non-parametric regression techniques in [3], and established the link between PLS regression and functional data analysis. This link resulted in numerous publications on PLS regression for functional data; see [4, 5, 6, 7, 8, 9]. The increasing interest in using functional data techniques for spectral applications stems from the fact that spectral data are indeed functional. NIR spectra, for example, are discrete instances of the chemical spectrum of a sample on a range of different wavelengths. This is illustrated in Figure 1 for 60 gasoline samples for which their spectral measurements are recorded at every two nanometers (nm) from 900 to 1700 nm. They are discrete values of continuous functions which are also smooth. Following [2] the extracted factor loadings should resemble to the spectra, and therefore should exhibit some degree of smoothness; the same holds for the regression solution. The gasoline samples data together with other two spectral data sets will be used in the examples that follow.

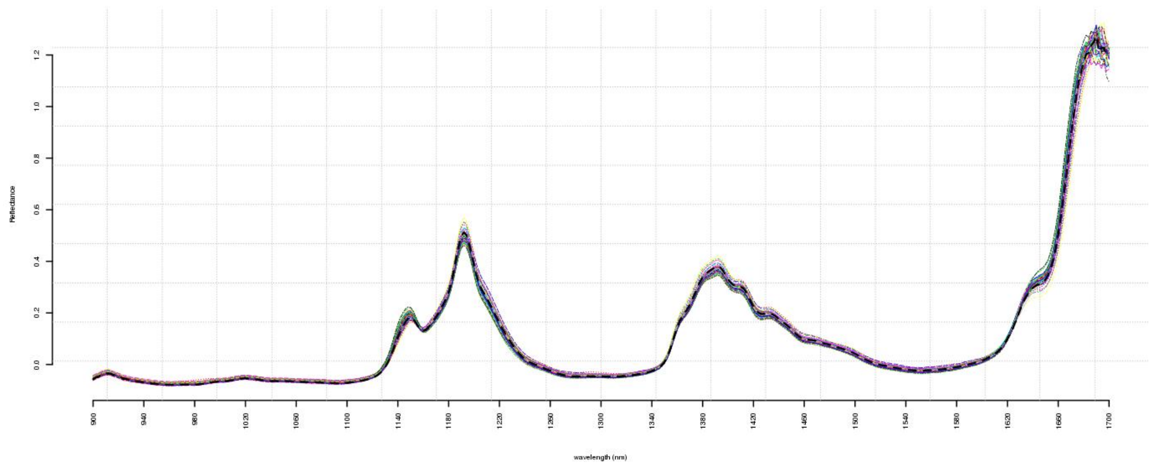


Figure 1: Gasoline data: Spectral data for 60 gasoline samples measured from 900 to 1700 nanometers (nm). The spectral data are registered every two nanometers.

We revisit smooth PLS regression after a short overview on PLS regression given in Section 2. Two smooth PLS regression using wavelets are presented in Section 3 and Section 4. Their theoretical properties are investigated in Section 5; proofs are given in the Appendix. In Section 6 three well-known NIR data sets are revisited in order to illustrate smooth PLS regression. Focus is mainly given on NIR applications. Nevertheless, the presented smooth PLS regression alternative naturally applies to other spectral data, as well. Conclusions are given in Section 7.

Throughout the paper bold face lower and upper case letters are used for vectors and matrices, respectively. The number of samples will be denoted by n while the number of predictors by p. The subscript m is used to denote the dimension of the PLS regression models, while the hat suffix is used for least squares fitted vectors. Further notations are introduced when needed.

2. PLS REGRESSION

Working within a linear model framework for regression problems the following linear model is assumed:

$$y_i = \mu + \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i\,, \qquad i = 1, ..., n\,, \tag{2.1}$$

where y_i is the observed response for sample i, $\boldsymbol{x}_i$ are p-vectors of explanatory variables, $\boldsymbol{\beta}$ is the unknown p-vector of regression parameters, and ϵ_i the error term of the regression model. Without loss of generality we assume data to be centred to zero and therefore we freely assume $\mu = 0$. Using matrix notation: $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ stands for the data matrix with predictors in its columns, $\boldsymbol{y}$ is the response vector, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown regression coefficient vector commonly estimated using least squares.

When the number of predictors (p) is large relative to the sample size (n) and/or the predictors are correlated, the least squares solution, when it exists, is highly variable due to rank deficiency of the data matrix $\boldsymbol{X}$. When $n < p$ the least squares solution doesn't even exist. In such cases, PLS regression offers an alternative by solving the regression problem after reducing its dimension; from hundreds of correlated predictors $\boldsymbol{x}_j$, $j = 1, ..., p$, to a small set of orthogonal components $\boldsymbol{t}_m$ with $m << p$. These are linear combination of the original predictors, and are used in the final regression on the response. PLS regression, therefore, iteratively approximates the least squares solution from a sequence of subspaces indexed by $m \leq p$. Using m orthogonal components in the final model, PLS regression lets for bias to decrease variance, and allows for a low mean square error for the final regression solution.

The restriction of orthogonal components may be relaxed in order to get PLS regression on orthogonal loadings. This has given rise to two different implementations of PLS regression, see [10] and [11]. The two algorithms are equivalent for prediction purposes; for a proof see [12]. Both PLS regression algorithms deflate data at each iteration, and $\boldsymbol{X}$-residuals and $\boldsymbol{y}$-residuals are used instead of $\boldsymbol{X}$ and $\boldsymbol{y}$ when $m > 1$. These are least squares residuals and will be denoted hereafter by $\boldsymbol{E}_m$ and $\boldsymbol{f}_m$, respectively, while we let $\boldsymbol{E}_0 = \boldsymbol{X}$ and $\boldsymbol{f}_0 = \boldsymbol{y}$. An important simplification when the response is a vector is the following: deflating $\boldsymbol{y}$ is not necessary; see [1]. More efficient computational algorithms for PLS regression without $\boldsymbol{X}$-data deflation have been proposed in [13] and [14]. We provide in Algorithm 1 a sketch of the PLS regression on orthogonal loadings; see [11]. This implementation will be used in the PLS regression calculations throughout the rest of the paper.

Algorithm 1 – Partial least squares regression on orthogonal loadings.

Input: For $i = 1,..,n$ and $j = 1,...,p$, $\boldsymbol{E}_0 = \boldsymbol{X}$ and $\boldsymbol{f}_0 = \boldsymbol{y}$.

For $m = 1, 2, ..., k \leq p$

1. Compute $\boldsymbol{p}_m$ according to: $\boldsymbol{p}_m = \boldsymbol{E}'_{m-1}\boldsymbol{f}_{m-1}$.
2. Derive $\boldsymbol{t}_m = \boldsymbol{E}_{m-1}\,\boldsymbol{p}_m/\boldsymbol{p}'_m\,\boldsymbol{p}_m$ and store in $\boldsymbol{T}_m = (\boldsymbol{t}_1, ..., \boldsymbol{t}_m)$.
3. $\boldsymbol{E}_m = \boldsymbol{E}_{m-1} - \boldsymbol{t}_m\,\boldsymbol{p}'_m$.
4. $\boldsymbol{f}_m = \boldsymbol{y} - \sum_{a=1}^{m} \boldsymbol{t}_a\,\widehat{q}_{ma}$ where
   $\widehat{\boldsymbol{q}}_m = (\widehat{q}_{m1}, ..., \widehat{q}_{ma}, ..., \widehat{q}_{mm})' = (\boldsymbol{T}'_m\,\boldsymbol{T}_m)^{-1}\,\boldsymbol{T}'_m\,\boldsymbol{y}$.

Output: Give the resulting sequence of the fitted vectors $\widehat{\boldsymbol{y}}_m = \boldsymbol{T}_m\,\widehat{\boldsymbol{q}}_m$.

The PLS regression coefficient vector $\widehat{\boldsymbol{\beta}}_m^{\text{pls}}$ is determined by the matrix $\boldsymbol{P}_m$ containing in its columns the orthogonal loading vectors $\boldsymbol{p}_1, ..., \boldsymbol{p}_m$. It is derived according to:

$$\widehat{\boldsymbol{\beta}}_m^{\text{pls}} = \boldsymbol{P}_m\,\widehat{\boldsymbol{q}}_m\,, \tag{2.2}$$

where $\widehat{\boldsymbol{q}}_m$ is defined in Algorithm 1. Similar to principal components; see [15] the dimension reduction process of PLS implies a change of basis from the p-dimensional unit basis to a subspace of reduced dimension $m < p$. For principal components this corresponds to the subspace generated by a small set of selected eigenvectors. For PLS regression the new basis corresponds to the Krylov subspace of dimension up to m, defined as follows:

Definition 2.1. For matrix $\mathsf{A} = \boldsymbol{X}'\boldsymbol{X}$ and vector $\mathsf{b} = \boldsymbol{X}'\boldsymbol{y}$ the Krylov subspace of dimension $m \leq p$ is given by:

$$\mathcal{K}_m(\mathsf{b}, \mathsf{A}) = \mathsf{span}(\mathsf{b}, \mathsf{A}^1\mathsf{b}, ..., \mathsf{A}^{m-1}\mathsf{b})\,. \tag{2.3}$$

The loading vectors in $\boldsymbol{P}_m$ (see Algorithm 1) span the Krylov subspace $\mathcal{K}_m(\mathsf{b}, \mathsf{A})$. The same holds for the PLS regression solution; see [12]. The PLS regression coefficient based on m components is given as the solution to:

$$\widehat{\boldsymbol{\beta}}_m^{\text{pls}} = \operatorname{argmin}_{\boldsymbol{\beta}}\big\{(\boldsymbol{y} - \widehat{\boldsymbol{y}})'\,(\boldsymbol{y} - \widehat{\boldsymbol{y}})\big\} \qquad \text{where} \quad \widehat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathcal{K}_m(\mathsf{b}, \mathsf{A})\,. \tag{2.4}$$

Krylov spaces are location and scale invariant (see [16], chapter 12) and they further benefit from the following property:

Remark 2.1. For an orthogonal basis change in $\mathcal{K}_m(\mathsf{b}, \mathsf{A})$ induced by an orthogonal matrix $\mathcal{Q}$ we get an orthogonal similarity transformation of A, that is:

$$\mathcal{K}_m(\mathcal{Q}\,\mathsf{b},\, \mathcal{Q}\,\mathsf{A}\,\mathcal{Q}') = \mathcal{Q}\,\mathcal{K}_m(\mathsf{b}, \mathsf{A})\,, \qquad \text{for} \quad m \leq p\,. \tag{2.5}$$

The last property becomes even more interesting given that the Discrete Wavelet Transform (DWT), to be used in the following section, is such an orthogonal matrix.

3. SMOOTH PLS REGRESSION ON WAVELET TRANSFORMED DATA

Spectral data are discrete values of continuous functions. Wavelets are used to approximate such functional data by means of the so-called mother and father wavelet, at different scales ℓ and locations k according to:

$$f(\boldsymbol{x}) = \sum_{k \in Z} c_{\ell_0,k}\, \phi_{\ell_0,k}(\boldsymbol{x}) + \sum_{\ell_0 \leq \ell,\, k \in Z} d_{\ell,k}\, \psi_{\ell,k}(\boldsymbol{x}) \,, \tag{3.1}$$

where $c_{\ell,k}$ and $d_{\ell,k}$ are the scaling and detail wavelet coefficients, respectively. The father wavelet coefficient at scale zero (ℓ_0) reflects the global average of the spectrum, and when the data are centered it is equal to zero. The wavelet transform can be expressed as a matrix multiplication using the Discrete Wavelet Transform (DWT) matrix; see [17], Chapter 12 as well as [18], paragraph 4.3. This allows changing coordinates system from the original to the wavelet domain forwards and backwards. The operation is fast ([19]) and safe given that DWT is orthogonal. Each row spectrum $\boldsymbol{x}_i$ is mapped into a vector of wavelet coefficients $\widetilde{\boldsymbol{x}}_i$ by means of matrix multiplication according to: $\widetilde{\boldsymbol{x}}_i = \mathcal{W}\boldsymbol{x}_i$, where $\mathcal{W}$ is the DWT orthogonal matrix of dimension $p \times p$. Note that for a spectral data matrix $\boldsymbol{X}$ the DWT is given by postmultiplying the spectral data by $\mathcal{W}'$, to get:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{X}\, \mathcal{W}' . \tag{3.2}$$

PLS regression on transformed data has been presented in [5]. It is run on the wavelet domain instead of the original spectra. The regression solution is then approximated on the wavelet domain as:

$$\widehat{\widetilde{\boldsymbol{\beta}}}{}^{\text{pls}}_{m,\ell} = \arg\min_{\widetilde{\boldsymbol{\beta}}} \left\{ (\boldsymbol{y} - \widehat{\boldsymbol{y}})' \, (\boldsymbol{y} - \widehat{\boldsymbol{y}}) \right\} \qquad \text{where} \quad \widehat{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\, \widetilde{\boldsymbol{\beta}}, \quad \widetilde{\boldsymbol{\beta}} \in \mathcal{K}_m(\widetilde{\mathsf{b}}, \widetilde{\mathsf{A}}) \,, \tag{3.3}$$

with $\widetilde{\mathsf{A}} = \mathcal{W}_\ell\, \mathsf{A}\, \mathcal{W}_\ell'$ and $\widetilde{\mathsf{b}} = \mathcal{W}_\ell\, \mathsf{b}$. The matrix $\mathcal{W}_\ell$ denotes the truncated DWT matrix of dimension $2^\ell \times p$. The use of the subscript ℓ for the coefficient vector in the transformed coordinates is used to highlight the wavelet truncation. Mother wavelet coefficients associated to the finest scales and very often the noisy part of the spectrum are truncated to zero. The final regression solution is recovered in original coordinates by means of the inverse DWT, denoted hereafter as iDWT. Using matrix multiplication this is the transpose of the DWT matrix. The PLS regression solution is smooth and given according to:

$$\widehat{\boldsymbol{\beta}}{}^{\text{spls.1}}_{m} = \mathcal{W}_\ell'\, \widehat{\widetilde{\boldsymbol{\beta}}}{}^{\text{pls}}_{m,\ell} \,. \tag{3.4}$$

The authors in [5] used the term 'wavelet compressed data' to describe their algorithm motivated by the wavelet's outstanding performance to retain spectral information in a few wavelet coefficients. They truncated wavelet coefficients based on their variance spectrum, retaining most often the largest ones. Our motivation is smoothness. We truncate to zero wavelet coefficients associated to the finest resolution level scales. Other truncation strategies could be based upon other rules such as the universal threshold or using adaptive thresholding rules at each different resolution level; see [20], [21] and the references therein.

The smooth PLS regression algorithm based on wavelet transformed data is implemented using the orthogonal loadings PLS regression algorithm. It is similar to Algorithm 1, and therefore will not be given here. It uses all vectors and matrices z transformed in the wavelet domain and denoted $\widetilde{z}$. For instance, the loading vector $\boldsymbol{p}_m$ is replaced by $\widetilde{\boldsymbol{p}}_m$.

The same holds for all data and residual data matrices, for the score vectors, and for the coefficient vectors $\boldsymbol{q}$ and $\boldsymbol{\beta}$. Expression (3.4) is used in the end to recover the final regression solution back in the original coordinates system. The choice of ℓ is an additional argument in the algorithm's input.

4. PLS REGRESSION ON SMOOTH LOADINGS

Transforming data to the wavelet domain is not the only one way to obtain a smooth PLS regression solution. Smoothness may be embedded directly on the loadings. This is done here by means of a PLS regression algorithm on smooth loadings. Wavelets are used on the loading vectors and data aren't transformed. At each iteration m the loading vector is reconstructed using a subset of the wavelet coefficients. The resulting loading vectors are both orthogonal and smooth. They are orthogonal due to the PLS algorithm, and smooth due to wavelet truncation. In terms of matrix multiplication we truncate the DWT matrix $\mathcal{W}$ to its first ℓ rows, that is, $\mathcal{W}_\ell$ which correspond to the coarsest scales. The resulting reconstructed smooth loading vector is given as: $\boldsymbol{p}^\star_m = \mathcal{W}'_\ell\, \ddot{\boldsymbol{p}}_m$, with

$$\ddot{\boldsymbol{p}}_m = \sum_{\check{r},\check{k}\in Z} d_{\check{r},\check{k}}\, \psi_{\check{r},\check{k}}(\boldsymbol{p}_m)\,, \tag{4.1}$$

being the approximated loading vector using all the detail wavelet coefficients for scales up to $\check{r}$ and their associated locations $\check{k}$. The smooth loadings $(\boldsymbol{p}^\star_1, ..., \boldsymbol{p}^\star_m)$ are stored in the matrix $\boldsymbol{P}^\star_m$. Similarly the regression coefficients $\widehat{q}^\star_{ma}$ are stored in the vector $\widehat{\boldsymbol{q}}^\star_m = (\widehat{q}^\star_{m1}, ..., \widehat{q}^\star_{ma}, ..., \widehat{q}^\star_{mm})'$. The final regression solution is given according to Expression (2.2) with matrix $\boldsymbol{P}^\star_m$ taking over $\boldsymbol{P}_m$. The algorithm for PLS regression on smooth loadings is sketched in Algorithm 2.

Algorithm 2 – PLS regression on smooth loadings.

Input: For $i = 1, .., n$ and $j = 1, ..., p$, $\boldsymbol{E}_0 = \boldsymbol{X}$ and $\boldsymbol{f}_0 = \boldsymbol{y}$.

Select ℓ such that $2^\ell < p$ and compute $\mathcal{W}_\ell$.

For $m = 1, 2, ..., k \le p$

1. Compute $\boldsymbol{p}^\star_m$ according to: $\boldsymbol{p}^\star_m = \mathcal{W}'_\ell\, \ddot{\boldsymbol{p}}_m$,
   where $\ddot{\boldsymbol{p}}_m$ as in Expression (4.1) with $\boldsymbol{p}_m = \boldsymbol{E}'_{m-1} \boldsymbol{f}_{m-1}$.
2. Derive $\boldsymbol{t}^\star_m = \boldsymbol{E}_{m-1}\, \boldsymbol{p}^\star_m / \boldsymbol{p}^{\star\prime}_m \boldsymbol{p}^\star_m$ and store in $\boldsymbol{T}^\star_m = (\boldsymbol{t}^\star_1, ..., \boldsymbol{t}^\star_m)$.
3. $\boldsymbol{E}_m = \boldsymbol{E}_{m-1} - \boldsymbol{t}^\star_m\, \boldsymbol{p}^{\star\prime}_m$.
4. $\boldsymbol{f}_m = \boldsymbol{y} - \sum_{a=1}^m \boldsymbol{t}^\star_a\, \widehat{q}^\star_{ma}$ where
   $\widehat{\boldsymbol{q}}^\star_m = (\widehat{q}^\star_{m1}, ..., \widehat{q}^\star_{ma}, ..., \widehat{q}^\star_{mm})' = (\boldsymbol{T}^{\star\prime}_m \boldsymbol{T}^\star_m)^{-1}\, \boldsymbol{T}^{\star\prime}_m\, \boldsymbol{y}$.

Output: Give the resulting sequence of the fitted vectors $\widehat{\boldsymbol{y}}^{\text{spls}}_m = \boldsymbol{X}\, \widehat{\boldsymbol{\beta}}^{\text{spls.2}}_m$,
where $\widehat{\boldsymbol{\beta}}^{\text{spls.2}}_m = \boldsymbol{P}^\star_m\, \widehat{\boldsymbol{q}}^\star_m$ for $\boldsymbol{P}^\star_m = (\boldsymbol{p}^\star_1, ..., \boldsymbol{p}^\star_m)$.

The PLS regression on smooth loadings algorithm is computationally much faster than the algorithm for smooth PLS regression on wavelet transformed data. In the former algorithm the data are not transformed and only a few matrix-vector multiplications are required.

In Algorithm 2 the wavelet expansion and truncation is done once for each loading vector. Normally the number of the extracted loadings is much smaller than the number of data samples. Moreover, the regression solution resulting from Algorithm 2 is on the original coordinates system and there is no need to be transformed back from the wavelet to the original domain. It turns out that the relation between the two algorithms is far more interesting from a theoretical point of view. This is further explored in the following section.

5. THEORETICAL ASPECTS OF SMOOTH PLS REGRESSION

The relation between the two smooth PLS regression algorithms is explored here from a theoretical viewpoint. The loading and regression vectors resulting from the two smooth PLS regression implementations are investigated. Results are given in the following propositions, while the proofs are provided separately in the Appendix.

Proposition 5.1. *The regression loadings $\widetilde{\boldsymbol{p}}_m$ and $\ddot{\boldsymbol{p}}_m$ are identical.*

Proposition 5.2. *The smooth PLS regression loadings $\boldsymbol{p}^{\star}_m$ computed in Algorithm 2 are orthogonal.*

Proposition 5.3. *The two smooth PLS regression algorithms generate the same sequence of approximate regression solutions, that is:*

$$\widehat{\boldsymbol{\beta}}^{\text{spls.1}}_{m,\ell} = \widehat{\boldsymbol{\beta}}^{\text{spls.2}}_{m,\ell} = \widehat{\boldsymbol{\beta}}^{\text{spls}}_{m,\ell}\,. \tag{5.1}$$

Proposition 5.4. *Both algorithms approximate the solution of the linear system of equations*

$$\mathsf{M}\,\mathsf{A}\,\boldsymbol{\beta}^{\star}_m = \mathsf{M}\,\mathsf{b}\,, \qquad \textit{with} \;\; \mathsf{M} = \mathcal{W}'_\ell\,\mathcal{W}_\ell \quad \textit{for} \;\; m \le p \;\; \textit{and} \;\; 2^\ell \le p\,, \tag{5.2}$$

iteratively through Krylov subspace approximations.

As a direct consequence of Proposition 5.4 we state the following proposition.

Proposition 5.5. *For $m \le p$ and increasing wavelet scale ℓ such that $2^\ell \to p$ the sequence of smooth PLS regression solutions generates the same subspaces and converges to the sequence of ordinary PLS regression solutions, that is:*

$$\widehat{\boldsymbol{\beta}}^{\text{spls}}_{m,\ell} \to \widehat{\boldsymbol{\beta}}^{\text{pls}}_{m}\,.$$

For both ordinary and smooth PLS regression the reduction of the dimension of the regression problem from large-p to small-m is almost identical. This is stated in the proposition below by employing the term of equivalence. The proof for Proposition 5.6 is given in the Appendix.

Proposition 5.6. *Ordinary and smooth PLS regression models are equivalent in reducing the dimension of the regression problem.*

Proper model selection is crucial for smooth PLS regression as it is for ordinary PLS regression. Prior to applying and assessing smooth PLS regression one needs to identify the dimension of the regression model, that is, the number of PLS regression components to be retained. This is done in the following section by means of cross validation prior to investigating smooth PLS regression on three well known NIR data sets.

6. EXPERIENCE WITH NIR DATA

Three well-known data from NIR spectroscopy are used here to assess smooth PLS regression. These are the diesel, the gasoline, and the biscuit data sets. All of them are available through the internet. The diesel data has been downloaded from the Eigenvector Research site at `http://www.eigenvector.com/data/SWRI/`, while the gasoline and the biscuit data have been downloaded from the `R` packages **`pls`** ([22]) and **`ppls`** ([23]) through the `R` website at `http://www.r-project.org/`. All three NIR data sets have been extensively used in the literature; see for instance [2], [24], [7], [8], and [9].

The diesel and the gasoline data sets quantify the cetane and the octane number of 381 diesel and 60 gasoline samples, respectively. The cetane number for diesel samples is the equivalent of the octane number for gasoline samples. The biscuit data measure fat concentration of 71 cookies. The data include information on 72 biscuit samples, yet, observation 23 is removed as a reported outlier. One can find more information on these three NIR data sets in the references given above. All three data sets use spectra for predictors. The NIR for the analyzed samples are registered over a broad range of wavelengths, measured in nanometers (nm). We retained in the analysis the appropriate wavelength ranges in order to build spectra of appropriate length (equal to a power of 2). For all three data sets the length of the spectra equals $256 = 2^8$.

The data have been centered prior to regression analysis by subtracting column means. They have been randomly split on 10 folds, and a 10-fold cross validation (see [25], Chapter 7) has been used in order to assess the number of PLS components. The NIR data ($\mathcal{D}$) have been split into 10 mutually exclusive groups, forming a training set $\mathcal{D}_{\text{train}}$ (used for model construction) and a test set $\mathcal{D}_{\text{test}} = \mathcal{D}^\star$ (used for model validation), where $\mathcal{D}_{\text{train}} \cap \mathcal{D}_{\text{test}} = \emptyset$ and $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}} = \mathcal{D}$. The cross validated mean squared prediction error $\mathsf{MSEP}^{\text{cv}}$ for a regression model based on m components, has been computed according to:

$$\mathsf{MSEP}_m^{\text{cv}} = \mathsf{E}_K\Big[\mathsf{E}_k\Big(\mathcal{L}\big(\boldsymbol{y}^\star,\, \widehat{\boldsymbol{y}}_m^{\star(-k)}\big)\Big)\Big], \tag{6.1}$$

where the superscript * is used to indicate the observations in $\mathcal{D}^\star$, and $k = 1, ..., K$ the part of the $K = 10$ groups of data which are left out. The notation E_K highlights average over the K different splits, while E_k indicates average over the number of observations inside the k^{th} test set. The suffix $^{(-k)}$ indicates that the fits are given by the investigated regression model on the data set excluding the k^{th} part. Using the same splits we did the same for the smooth PLS regression using wavelet approximation including wavelet scales up to $\ell = 6$ and $\ell = 7$. The results for the model selection study are reported in Table 1.

Table 1: NIR data: 10-fold cross-validation estimates for the prediction loss of the PLS and the smooth PLS regression models ($sPLS_\ell$) including 1 to 10 components for $\ell = 7$ and $\ell = 6$, respectively.

Data Set	Regression Model	Components									
		1	2	3	4	5	6	7	8	9	10
diesel	PLS	3.09	2.84	2.64	2.27	2.09	2.26	1.99	2.09	2.17	2.15
	$sPLS_7$	2.60	2.44	2.02	2.35	2.12	2.39	2.20	2.04	2.05	2.07
	$sPLS_6$	2.04	2.03	1.98	1.98	1.77	1.75	1.53	1.53	1.55	1.55
gasoline	PLS	0.79	0.29	0.23	0.25	0.25	0.26	0.30	0.28	0.27	0.24
	$sPLS_7$	0.83	0.23	0.11	0.15	0.14	0.13	0.15	0.19	0.15	0.13
	$sPLS_6$	0.79	0.21	0.11	0.11	0.12	0.10	0.13	0.17	0.18	0.17
biscuit	PLS	1.25	1.33	0.79	0.42	0.25	0.30	0.28	0.30	0.28	0.27
	$sPLS_7$	1.86	1.80	1.34	0.92	0.637	0.45	0.39	0.37	0.37	0.35
	$sPLS_6$	1.07	1.12	0.58	0.43	0.40	0.28	0.23	0.27	0.25	0.24

The PLS regression model selection results in Table 1 are similar to the ones already known from the existing literature. Furthermore, the model selection results for the smooth PLS regression are almost identical to the PLS regression results. As expected, the minimum prediction loss for smooth PLS regression is reached after retaining almost the same number of components as for ordinary PLS regression. The estimated out-of-sample prediction error for smooth PLS regression is sometimes even reduced compared to ordinary PLS regression prediction error. Notably for the gasoline data the prediction performance for smooth PLS improves substantially compared to ordinary PLS regression. Yet, this is not the case for the biscuit data.

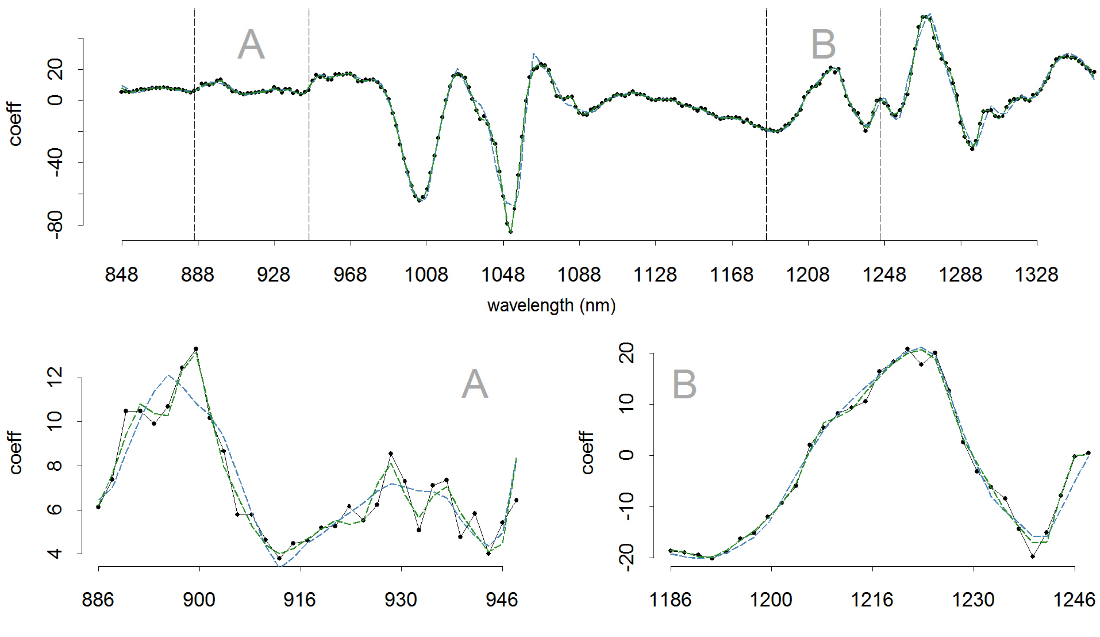


Figure 2: Diesel data. Regression coefficient for a regression model including 7 components. Response is the cetane number of the diesel samples and predictors are the NIR spectra over the wavelength region from 848 to 1358 nanometers (nm). Black points and black thin line correspond to the PLS regression coefficient. The smooth PLS regression coefficients with $\ell = 7$ and $\ell = 6$ are plotted in green and blue dashed lines, respectively. Selected wavelength regions (A and B) are magnified in the lower left and right panels.

Figures 2, 3, and 4 illustrate the regression solutions for PLS and smooth PLS regression. Black solid lines and points are used to depict the PLS regression solution, while dashed lines are used for smooth PLS regression results. For illustration purposes selected wavelength regions are magnified and plotted in the lower left and right panels. These allow better inspecting the smoothness induced by the use of the smooth PLS regression.

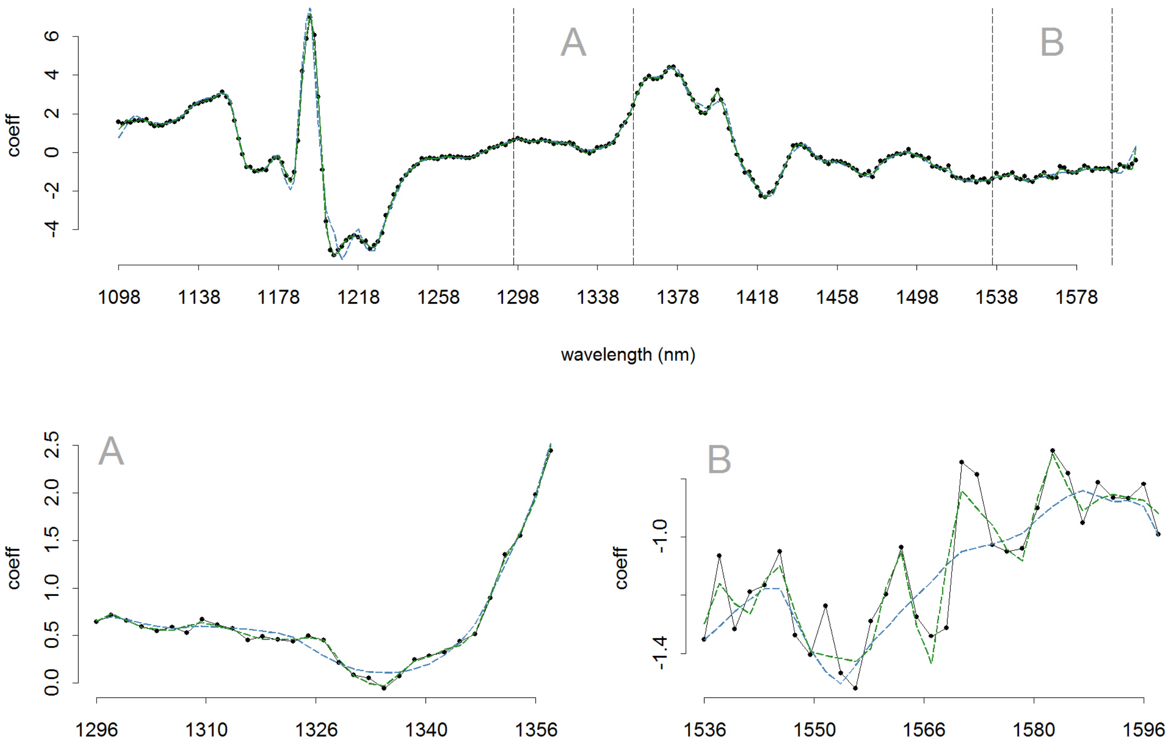


Figure 3: Gasoline data. Regression coefficient vector for a regression model including 3 components. Response is the octane number of 60 gasoline samples and predictors are the NIR spectra over the wavelength region from 1098 to 1608 nanometers (nm) in steps of two. Black points and black thin line correspond to the PLS regression coefficient. The smooth PLS regression coefficients with $\ell = 7$ and $\ell = 6$ are plotted in green and blue dashed lines, respectively. Selected wavelength regions (A and B) are magnified in the lower left and right panels.

For the diesel and the gasoline data set in Figures 2 and 3 the smooth PLS regression solution efficiently smooths the PLS regression coefficient vector especially for $\ell = 6$, see the light gray (blue) dashed line. The lower panel plots help discriminating between the three solutions. The smooth PLS regression coefficient is less efficient in smoothing the final solution for the biscuit data; see Figure 4. The ordinary PLS regression solution for this data set was already rather smooth.

Finally it is worth noting that smooth PLS regression may improve the prediction performance notably when the PLS regression solution is noisy. Smoothing reduces the prediction error in the diesel and the gasoline data. In contrast this is not the case in the biscuit data where PLS regression is already smooth.

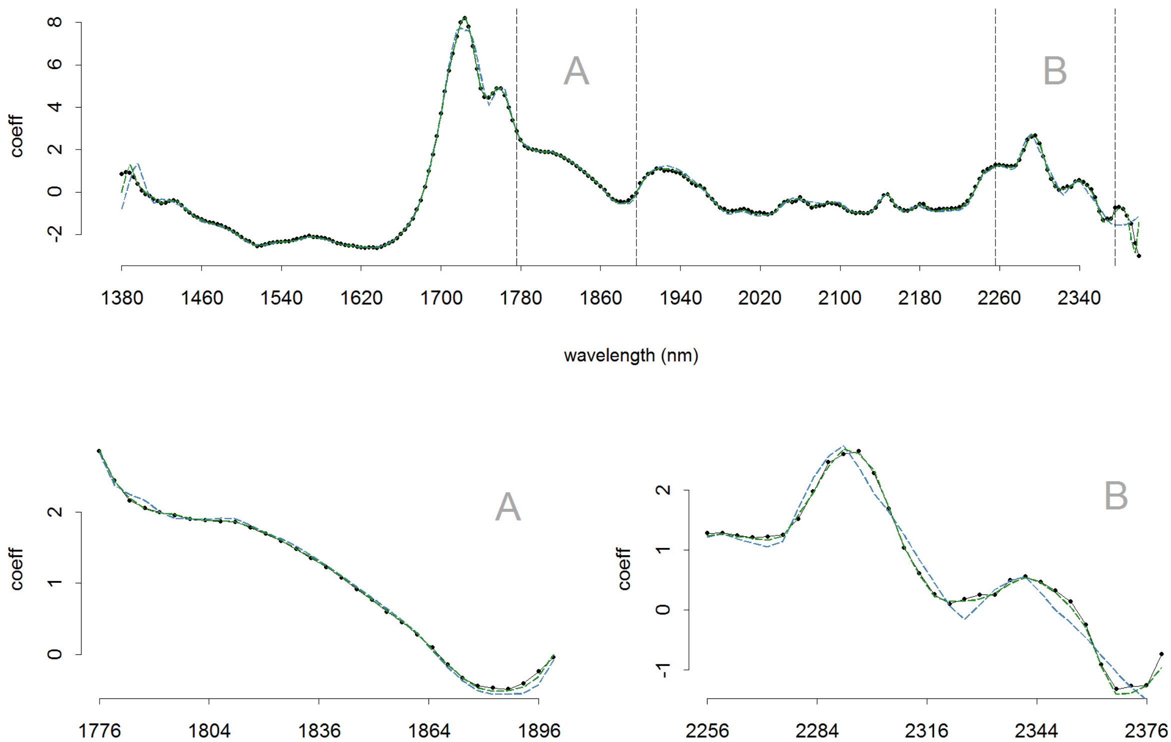


Figure 4: Biscuit data. Regression coefficient vector for a regression model including 5 components. Response is the fat concentration of biscuit samples and predictors are the NIR spectra over the wavelength region from 1100 to 2498 nanometers (nm). Black points and black thin line correspond to the PLS regression coefficient. The smooth PLS regression coefficients with $\ell = 7$ and $\ell = 6$ are plotted in green and blue dashed lines, respectively. Selected wavelength regions (A and B) are magnified in the lower left and right panels.

7. CONCLUSIONS

Most spectral data used in chemometrics are high dimensional and very often functional. PLS regression methods are well suited for high dimensional data. Wavelets are well suited for functional data. We explored the combination of these two in order to build smooth alternatives for PLS regression. The rationale behind smooth PLS regression stemmed from the fact that PLS regression coefficients are low dimensional approximations for the regression solution and should exhibit some degree of smoothness.

We showed that PLS regression can be effectively combined to wavelets for functional data analysis and provide smooth regression solutions to high dimensional regression problems. Wavelet expansion and truncation allowed us building two equivalent smooth PLS regression algorithms. The two algorithmic implementations for smooth PLS regression have been proven to be equivalent and to produce the same sequence of approximate solutions. These are regression solutions approximated through Krylov subspaces of dimension $m \leq p$. They are, therefore, PLS regression solutions. Working in the framework of spectral data we focused on near infra-red experiments which have been used to illustrate the potential of smooth PLS regression using wavelets. Three well known NIR data sets from the literature have been used to confirm that smooth PLS regression is a valuable alternative to ordinary PLS regression for smoothing the final regression solution while maintaining good prediction performance and dimension reduction.

The two presented smooth PLS regression algorithms have been implemented based on the PLS regression algorithm on orthogonal loadings. It is straightforward to implement both using the PLS regression algorithm on orthogonal scores; the results will be identical. The implementation of the proposed methods is straightforward. We used the `S-PLUS` wavelet package `S+WAVELETS` in our implementation; see [17]. Similar computer packages for wavelet analysis exist in `R`, as well; see for instance the `wavethresh` package in `R` (see [22]). Existing computational tools give all that is required for further smooth PLS regression developments.

A. APPENDIX

Prior to the proof of the propositions in Section 5 we state two lemmas required for the development of the proofs. The proof for Lemma A.1 is a direct consequence of wavelet properties and is omited; the interest reader can see [18], paragraph 4.3.1. The proof for Lemma A.2 is provided below using mathematical induction. Finally the notation $2^\ell \to p$ is used to denote the increasing order approximation of $\boldsymbol{X}$ by allowing finner scales to be included in the rows of matrix $\mathcal{W}_\ell$.

Lemma A.1. *For the truncated matrix $\mathcal{W}_\ell$ of dimension $2^\ell < p$ we have:*

1. *All cross-product matrices $\mathcal{W}_\ell' \mathcal{W}_\ell$ with $2^\ell < p$ are block-diagonal, with*
$$\mathcal{W}_\ell' \mathcal{W}_\ell \to I_p \qquad \text{as} \quad 2^\ell \to p\,,$$
where I_p is used to denote the identity matrix of order p.
2. *All cross-product matrices $\mathcal{W}_\ell \mathcal{W}_\ell'$ with $2^\ell \le p$ satisfy:*
$$\mathcal{W}_\ell \mathcal{W}_\ell' = I_p\,.$$

Lemma A.2. *For all $m \le p$, $\boldsymbol{E}_m \mathcal{W}_\ell' = \widetilde{\boldsymbol{E}}_m^{(\ell)}$.*

Proof of Lemma A.2: We use mathematical induction. For $m = 1$ the lemma holds given:
$$\boldsymbol{E}_0 \mathcal{W}_\ell' = \boldsymbol{X} \mathcal{W}_\ell' = \widetilde{\boldsymbol{X}}^{(\ell)} = \widetilde{\boldsymbol{E}}_0^{(\ell)}\,.$$
Let it be true for $m-1$, that is assume that:
$$\boldsymbol{E}_{m-1} \mathcal{W}_\ell' = \widetilde{\boldsymbol{E}}_{m-1}^{(\ell)}\,.$$
We will prove that this also holds for m, that is:

$$\boldsymbol{E}_m \mathcal{W}_\ell' = \widetilde{\boldsymbol{E}}_m^{(\ell)}\,. \tag{A.1}$$

We develop seperately both sides of Expression (A.1). For the left hand side of Expression (A.1) we have:
$$\begin{aligned}
\boldsymbol{E}_m \mathcal{W}_\ell' &= \big(\boldsymbol{E}_{m-1} - \boldsymbol{t}_m^\star \boldsymbol{p}_m^{\star\prime}\big)\,\mathcal{W}_\ell' \\
&= \big(\boldsymbol{E}_{m-1} - \boldsymbol{E}_{m-1}\,\boldsymbol{p}_m^\star \boldsymbol{p}_m^{\star\prime}\big)\,\mathcal{W}_\ell' \\
&= \big(\boldsymbol{E}_{m-1} - \boldsymbol{E}_{m-1}\mathcal{W}_\ell'\,\ddot{\boldsymbol{p}}_m \ddot{\boldsymbol{p}}_m' \mathcal{W}_\ell\big)\,\mathcal{W}_\ell' \\
&= \boldsymbol{E}_{m-1}\mathcal{W}_\ell' - \boldsymbol{E}_{m-1}\mathcal{W}_\ell'\,\ddot{\boldsymbol{p}}_m \ddot{\boldsymbol{p}}_m' \mathcal{W}_\ell \mathcal{W}_\ell' \\
&= \boldsymbol{E}_{m-1}\mathcal{W}_\ell' - \boldsymbol{E}_{m-1}\mathcal{W}_\ell'\,\ddot{\boldsymbol{p}}_m \ddot{\boldsymbol{p}}_m' \\
&= \boldsymbol{E}_{m-1}\mathcal{W}_\ell'\big(I - \ddot{\boldsymbol{p}}_m \ddot{\boldsymbol{p}}_m'\big)\,.
\end{aligned}$$

For the right hand side of Equation (A.1) we have:
$$\begin{aligned}
\widetilde{\boldsymbol{E}}_m^{(\ell)} &= \widetilde{\boldsymbol{E}}_{m-1}^{(\ell)} - \widetilde{\boldsymbol{t}}_m \widetilde{\boldsymbol{p}}_m' \\
&= \widetilde{\boldsymbol{E}}_{m-1}^{(\ell)} - \widetilde{\boldsymbol{E}}_{m-1}^{(\ell)} \widetilde{\boldsymbol{p}}_m \widetilde{\boldsymbol{p}}_m' \\
&= \widetilde{\boldsymbol{E}}_{m-1}^{(\ell)}\big(I - \widetilde{\boldsymbol{p}}_m \widetilde{\boldsymbol{p}}_m'\big)\,.
\end{aligned}$$

Furthermore, given Expression (4.1) we have:

$$\ddot{\boldsymbol{p}}_m \ddot{\boldsymbol{p}}'_m = \mathcal{W}_\ell\, \boldsymbol{p}_m \boldsymbol{p}'_m \mathcal{W}'_\ell = \widetilde{\boldsymbol{p}}_m \widetilde{\boldsymbol{p}}'_m ,$$

which completes the proof. □

Proof of Proposition 5.1: Recall that for univariate PLS regression there is no need to deflate the response vector $\boldsymbol{y}$. The loading vector $\ddot{\boldsymbol{p}}_m$ in Expression (4.1) can be written in matrix form as $\mathcal{W}_\ell\, \boldsymbol{p}_m$; it then follows:

$$\ddot{\boldsymbol{p}}_m = \mathcal{W}_\ell\, \boldsymbol{p}_m = \mathcal{W}_\ell\, \boldsymbol{E}'_{m-1} \boldsymbol{y} = \big(\boldsymbol{E}_{m-1} \mathcal{W}'_\ell\big)' \boldsymbol{y} = \widetilde{\boldsymbol{E}}^{(\ell)\,\prime}_{m-1} \boldsymbol{y} = \widetilde{\boldsymbol{p}}_m .$$ □

Proof of Proposition 5.2: Using Proposition 5.1 and noting that the loading vectors $\widetilde{\boldsymbol{p}}$ are orthogonal by construction (they are the ordinary PLS regression loadings in the wavelet domain), it follows that:

$$\boldsymbol{p}^{\star\,\prime}_i \boldsymbol{p}^\star_j = \ddot{\boldsymbol{p}}'_i \mathcal{W}_\ell \mathcal{W}'_\ell\, \ddot{\boldsymbol{p}}_j = \widetilde{\boldsymbol{p}}'_i \mathcal{W}_\ell \mathcal{W}'_\ell\, \widetilde{\boldsymbol{p}}_j = \widetilde{\boldsymbol{p}}'_i \widetilde{\boldsymbol{p}}_j = 0 , \qquad \text{for} \quad i \neq j \quad \text{and} \quad i, j \leq p .$$

Therefore the smooth PLS regression loadings $\boldsymbol{p}^\star$ are orthogonal. □

Proof of Proposition 5.3: The smooth regression coefficients $\widehat{\boldsymbol{\beta}}^{\text{spls.1}}_m$ and $\widehat{\boldsymbol{\beta}}^{\text{spls.2}}_m$ are identical, as:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\text{spls.2}}_m &= \boldsymbol{P}^\star_m \widehat{\boldsymbol{q}}^\star_m \\
&= \mathcal{W}'_\ell\, \ddot{\boldsymbol{P}}_m \widehat{\boldsymbol{q}}^\star_m \\
&= \mathcal{W}'_\ell\, \widetilde{\boldsymbol{P}}_m \widehat{\widetilde{\boldsymbol{q}}}_m \\
&= \mathcal{W}'_\ell\, \widehat{\widetilde{\boldsymbol{\beta}}}^{\text{pls}}_{m,\ell} = \widehat{\boldsymbol{\beta}}^{\text{spls.1}}_m .
\end{aligned}$$

Note that $\widehat{\widetilde{\boldsymbol{q}}}_m = \widehat{\boldsymbol{q}}^\star_m$. This is justified by the fact that both are implied by the loading's matrix $\ddot{\boldsymbol{P}}_m$ and $\widetilde{\boldsymbol{P}}_m$, respectively. These are, yet, identical as shown in Proposition 5.1. □

Proof of Proposition 5.4: The link between PLS regression and conjugate gradients for solving large linear system of equations is well-known; see for instance [26]. The solution to the system of equations is approximated through Krylov subspaces. The system in (5.2) is pre-multipled by a non-singular matrix M. This is sometimes referred to in numerical analysis as a preconditioned system. While preconditioning mainly focuses on improvement in the convergence of iterative solution methods, such as the Krylov methods, here it is used to induce smoothness. This is done by using $\mathsf{M} = \mathcal{W}'_\ell \mathcal{W}_\ell$. The two smooth PLS regression algorithms are two facets of preconditioning the conjugate gradients. While the former operates on transformed coordinates ($\widetilde{\mathsf{A}}$ and $\widetilde{\mathsf{b}}$), the latter (Algorithm 2) iterates starting from directions determined by matrix M. The equivalence between these two algorithms is sketched below:

$$\begin{aligned}
\mathsf{M}\,\mathsf{A}\,\boldsymbol{\beta}^\star_m &= \mathsf{M}\,\mathsf{b} , \\
\mathcal{W}'_\ell \mathcal{W}_\ell\, \mathsf{A}\,\boldsymbol{\beta}^\star_m &= \mathcal{W}'_\ell \mathcal{W}_\ell\, \mathsf{b} , \\
\mathcal{W}_\ell\, \mathsf{A}\, \mathcal{W}'_\ell\, \widetilde{\boldsymbol{\beta}}_m &= \mathcal{W}_\ell\, \mathsf{b} , \\
\widetilde{\mathsf{A}}\, \widetilde{\boldsymbol{\beta}}_m &= \widetilde{\mathsf{b}} , \qquad \text{for} \quad m \leq p .
\end{aligned}$$

The final solution $\widetilde{\boldsymbol{\beta}}$ can be transformed back in the original coordinates according to:

$$\boldsymbol{\beta}^{\star}_m = \mathcal{W}'_\ell \, \widetilde{\boldsymbol{\beta}}_m \,,$$

in exactly the same manner that the loading vectors $\widetilde{\boldsymbol{p}}$ can be also transformed back in original coordinates as:

$$\boldsymbol{p}^{\star}_m = \mathcal{W}'_\ell \, \widetilde{\boldsymbol{p}}_m \,. \qquad \square$$

Proof of Lemma 5.5: For $\mathsf{M} = I_p$ in the system of equations (5.2) the ordinary PLS regression solution is recovered. This happens for increasing ℓ as $2^\ell \to p$. The PLS regression solution is a Krylov solution, that is:

$$\widehat{\boldsymbol{\beta}}^{\text{pls}}_m \in \mathcal{K}_m(\mathsf{b}, \mathsf{A}) \,, \qquad \text{for} \quad m \le p \,.$$

The smooth PLS regression solution given in Expression (3.4) as:

$$\widehat{\boldsymbol{\beta}}^{\text{spls}}_m = \mathcal{W}'_\ell \, \widehat{\widetilde{\boldsymbol{\beta}}}^{\text{pls}}_m \,, \qquad \text{for} \quad m \le p \,,$$

is a Krylov solution. Combining Remark 2.1 and expression (2.5) to the orthogonality of the DWT matrix $\mathcal{W}$, as long as $2^\ell \to p$ one gets:

$$\widehat{\boldsymbol{\beta}}^{\text{spls}}_m \in \mathcal{W}'_\ell \, \mathcal{K}_m(\mathcal{W}_\ell \, \mathsf{b}, \mathcal{W}_\ell \, \mathsf{A} \, \mathcal{W}'_\ell) = \mathcal{W}'_\ell \, \mathcal{W}_\ell \, \mathcal{K}_m(\mathsf{b}, \mathsf{A}) \approxeq \mathcal{K}_m(\mathsf{b}, \mathsf{A}) \,, \qquad \text{for} \quad m \le p \,. \qquad \square$$

Proof of Proposition 5.6: The dimension reduction performance of both ordinary and smooth PLS regression is determined by the minimum number of iterations required to achieve the best approximate solution to the system of equations in (5.2). This is strongly dependent on the spectrum of A and $\mathcal{M}\mathsf{A}$ for ordinary and smooth PLS regression, respectively. Let $S(\mathsf{A})$ be the spectrum of a symmetric matrix A as given by its eigen decomposition $\mathsf{A} = V \Lambda \mathsf{V}'$ with $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ denoting the diagonal matrix of eigenvalues of A, and V its orthonormal set of eigenvectors. Similarly, let $S(\mathcal{M}\mathsf{A})$ be the spectrum of the symmetric matrix $\widetilde{\mathsf{A}}$. A sufficient condition for Proposition 5.6 to hold is given below:

Ordinary and smooth PLS regression are approximately equivalent in reducing the dimension of the regression problem whenever:

$$S(\mathcal{M}\mathsf{A}) \approx S(\mathsf{A}) \,.$$

Consider the eigen decomposition of matrix $\widetilde{\mathsf{A}}$ as follows:

$$\begin{pmatrix} \widetilde{V}_\ell & \widetilde{V}_{\bar{\ell}} \end{pmatrix} \times \begin{pmatrix} \widetilde{\Lambda}_\ell & 0 \\ 0 & \widetilde{\Lambda}_{\bar{\ell}} \end{pmatrix} \times \begin{pmatrix} \widetilde{V}'_\ell \\ \widetilde{V}'_{\bar{\ell}} \end{pmatrix} ,$$

for $\widetilde{V}_\ell = \mathcal{W}_\ell V_\ell$ and $\widetilde{V}_{\bar{\ell}} = \mathcal{W}_{\bar{\ell}} V_{\bar{\ell}}$, where the subscript ℓ is used to denote the ℓ-scales wavelet approximation and $\bar{\ell}$ used to denote the excluded wavelet scales. The expression above simplifies to:

$$\mathsf{W}_\ell V_\ell \widetilde{\Lambda}_\ell V'_\ell \, \mathsf{W}'_\ell + \mathsf{W}_{\bar{\ell}} V_{\bar{\ell}} \widetilde{\Lambda}_{\bar{\ell}} V'_{\bar{\ell}} \, \mathsf{W}'_{\bar{\ell}} \,, \tag{A.2}$$

We discuss the two following cases:

1. When $2^\ell = p$, the second term in Expression (A.2) disappears and $S(\widetilde{\mathsf{A}}) = S(\mathsf{A})$ since $\mathcal{W}_\ell$ is the identity matrix and $V_\ell = V$. The two regression methods are then identical in reducing the dimension of the regression problem.
2. When $2^\ell < p$ the second term in Expression (A.2) is generally much smaller than the first term, especially for collinear and functional data (such as the NIR data) where PLS regression is used. The diagonal entries in $\widetilde{\Lambda}_{\bar{\ell}}$ are close to zero and the second term in Expression (A.2) vanishes; hence the spectrum of A is approximated by the first term and $S(\mathcal{M}\mathsf{A}) \approx S(\mathsf{A})$. □

ACKNOWLEDGMENTS

The author would like to thank the two referees for their very constructive comments and their suggestions. The author would also like to express his gratitude to Prof. Joe Whittaker and Prof. Alina Matei for their support.

REFERENCES

[1] Höskuldsson, A. (1988). PLS regression methods, *Journal of Chemometrics*, **2**, 211–228.

[2] Goutis, C. and Fearn, T. (1996). Partial least squares regression on smooth factors, *Journal of the American Statistical Association*, **91**(434), 627–632.

[3] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[4] Durand, J.F. and Sabatier, R. (1997). Additive splines for partial least squares regression, *Journal of the American Statistical Association*, **92**(440), 1546–1554.

[5] Trygg, J. and Wold, S. (1998). PLS regression on wavelet compressed NIR spectra, *Chemometrics and Intelligent Laboratory Systems*, **42**(1), 209–220.

[6] Preda, C. and Saporta, G. (2005). PLS regression on a stochastic process, *Computational Statistics & Data Analysis*, **48**(1), 149–158.

[7] Reiss, P.T. and Ogden, R.T. (2007). Functional principal component regression and functional partial least squares, *Journal of the American Statistical Association*, **102**, 984–996.

[8] Krämer, N.; Boulesteix, A.L. and Tutz, G. (2008). Penalized partial least squares with applications to B-spline transformations and functional data, *Chemometrics and Intelligent Laboratory Systems*, **94**(1), 60–69.

[9] Kondylis, A. and Whittaker, J. (2013). Feature selection for functional PLS, *Chemometrics and Intelligent Laboratory Systems*, **121**, 82–89.

[10] Wold, S.; Martens, H. and Wold, H. (1983). *The multivariate calibration problem in chemistry solved by the PLS method.* In "Proc. Conf. Matrix Pencils" (A. Ruhe and B. Kagström, Eds.), Lecture Notes in Mathematics, Springer-Verlag, pp. 286–293.

[11] Martens, H. and Naes, T. (1989). *Multivariate Calibration*, John Wiley & Sons, New York.

[12] Helland, I.S. (1988). On the structure of partial least squares regression, *Communications in Statistics – Simulation and Computation*, **17**, 581–607.

[13] de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression, *Chemometrics and Intelligent Laboratory Systems*, **18**, 251–263.

[14] Goutis, C. (1997). A fast method to compute orthogonal loadings partial least squares, *Journal of Chemometrics*, **11**, 33–38.

[15] Jolliffe, I.T. (2002). *Principal Component Analysis*, Springer Verlag.

[16] Parlett, B. (1980). *The Symmetric Eigenvalue Problem*, Prentice-Hall Series in Applied Mathematics, New Jersey.

[17] Bruce, A. and Gao, H.Y. (1996). *Applied Wavelet Analysis with S-PLUS*, Springer-Verlag, New York.

[18] Vidakovic, B. (1999). *Statistical Modeling by Wavelets*, John Wiley & Sons, New York.

[19] Mallat, S.G. (1989). A theory for multiresolution signal decomposition: the wavelet representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

[20] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, **90**(432), 1200–1224.

[21] Jansen, M. (2000). *Noise Reduction by Wavelets Thresholding*, Springer, Lecture Notes in Statistics, New York.

[22] Nason, G. (2013). *Wavethresh: wavelets statistics and transforms*, R package version 4.6.6, `http://CRAN.R-project.org/package=wavethresh`.

[23] Krämer, N. and Boulesteix, A.L. (2014). *PPLS: Penalized Partial Least Squares*, R package version 1.6.1, `http://CRAN.R-project.org/package=ppls`.

[24] Kalivas, J.H. (1997). Two data sets of near infrared spectra, *Chemometrics and Intelligent Laboratory Systems*, **37**, 255–259.

[25] Hastie, T.; Tibshirani, R. and Friedman, J. (2009). *Elements of Statistical Learning*, 2nd Edition, Springer Series in Statistics, New York.

[26] Phatak, A. and de Hoog, F. (2001). *PLSR, Lanczos, and Conjugate Gradients*, Commonwealth Scientific and Industrial Research Organisation, CMIS 01/122.

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 481–500
https://doi.org/10.57805/revstat.v20i4.385

Rényi Entropy of k-Records: Properties and Applications

Authors: Jitto Jose
– Department of Statistics, Rajagiri College of Social Sciences,
Kalamassery, Cochin, Kerala, India – 683104
jittojosevkm@gmail.com

E.I. Abdul Sathar ✉
– Department of Statistics, University of Kerala,
Trivandrum, India – 695581
sathare@gmail.com

Received: October 2019 Revised: November 2020 Accepted: November 2020

Abstract:

- In this paper, we discuss on Rényi entropy of k-records arising from any continuous distribution in detail. The relevance of constructing k-records from random sample in the context of information contained in a random variable has been described in the study. Some important properties of Rényi entropy of k-records have been derived as well. Two relevant applications of Rényi entropy of k-records are discussed in this work. Finally, a simple estimator is proposed for Rényi entropy of k-records and a numerical illustration has been carried out using a real life data set.

Keywords:

- *characterization; k-records; monotone property; Rényi entropy ordering; Rényi divergence; uniform distribution.*

AMS Subject Classification:

- 62B10, 94A15, 94A17.

✉ Corresponding author.

1. INTRODUCTION

Shannon [33] defined the entropy of a system which measures uncertainty contained in a random variable. The Shannon entropy measure of uncertainty is inversely related to the occurrence probability of the event. For a non-negative and absolutely continuous random variable X with probability density function (pdf) $f(x)$, the Shannon entropy is defined by

$$H(X) = -\int_0^\infty f(x)\ln f(x)dx.$$

Moreover, Rényi [30] introduced one parameter extension of Shannon entropy by defining an entropy of order α called Rényi entropy. The Rényi entropy of X with pdf $f(x)$ is defined by

$$H_\alpha(X) = \frac{1}{1-\alpha}\ln\int_{-\infty}^{\infty} f^\alpha(x)dx, \quad \alpha > 0,\ (\alpha \neq 1). \tag{1.1}$$

It can be easily shown that $\lim_{\alpha\to 1} H_\alpha(X) = H(X)$. Some important properties of Rényi entropy are as follows: $H_\alpha(X)$ can be negative, $H_\alpha(X)$ is invariant under a location transformation, $H_\alpha(X)$ is not invariant under a scale transformation and for any $\alpha_1 < \alpha_2$, we have $H_{\alpha_1}(X) \geq H_{\alpha_2}(X)$, the equality occurs if and only if X is uniformly distributed. The Rényi divergence of order α between two random variables X and Y with density functions $f(x)$ and $g(y)$, respectively, given by

$$D_\alpha(f,g) = \frac{1}{\alpha-1}\int_{-\infty}^{\infty}\left[\frac{f(x)}{g(x)}\right]^{\alpha-1} f(x)dx. \tag{1.2}$$

For details, see Golshani and Pasha [19] and Contreras-Reyes [8]. The intriguing properties and applications of Rényi entropy have been extensively studied in literature.

Morales *et al.* [27] studied properties of Rényi entropy with respect to testing of hypothesis in parametric models. The connection of Rényi information with log-likelihood of the random variable derived from the gradient of the spectrum of Rényi information is discussed in Song [34]. Csiszár [10] gave Rényi's entropy and divergence of order α operational characterizations in terms of block coding and hypothesis testing. In the field of statistical mechanics, the ergodic diffusion processes in terms of Rényi entropy has been discussed in De Gregorio and Iacus [12]. Further, Kirchanov [24] uses Rényi entropy to describe quantum dissipative systems. For more details about the application of Rényi entropy, one may refer Nadarajah and Zografos [28], Asadi *et al.* [5], Contreras-Reyes [8] and Contreras-Reyes and Cortés [9].

This paper is structured as follows: Section 2 gives a brief introduction about k-records. Section 3 expresses Rényi Entropy of k-records arising from any continuous distribution. In Section 4, we discuss some important properties of Rényi entropy of upper and lower k-records. Section 5 presents two applications of Rényi entropy of k-records. The overall findings are stated in Section 6.

2. BACKGROUND OF k-RECORDS

Chandler [7] defined records as successive extremes occuring in a sequence of independent and identically distributed (iid) random variables. Records are of great importance in several real life problems involving weather, economic studies, sports, etc. Prediction of next record value is an interesting problem in many real life situations. For example, the prediction of next record level of water that a dam can capture is helpful in holding or discharge of the water. Similarly, prediction of lowest share value in stock markets is essential to plan for the investment strategies. More applications of record values are available in Arnold *et al.* [4] and Ahsanullah [3].

In many events associated with athletics, temperature, wind velocity, etc., one is compelled to depend upon the available record data to deal with statistical inference problems of the parent distribution. But, statistical inferences based on records are difficult to make since the records occurs rarely in real life situations. We can observe that the expected waiting time for every record after the first observation is infinite. One may overcome this difficulty by the use of k-records introduced by Dziubdziela and Kopociński [13] which occur more frequently than the classical records. For example, consider first 10 observations from the data given in David and Nagaraja [11]: 0.464, 0.060, 1.486, 1.022, 1.394, 0.906, 1.179, -1.501, -0.690, 1.372. The records observed from the data are: 0.464 and 1.486. We can construct upper k-records from the data as given below:

Table 1: Sequences of k-records for $k = 2, 3, 4$.

2-Records	0.060, 0.464, 1.022, 1.394.
3-Records	0.060, 0.464, 1.022, 1.179, 1.372
4-Records	0.060, 0.464, 0.906, 1.022, 1.179

It is well known that if the number of observations on the random variable increases the statistical inferences becomes more reliable. In other words, the uncertainty contained in the random variable reduces.

Many works are going on to detect outliers in a data so as to delete them for devising more reasonable statistical methods to the problem of interest. The integer parameter k involved in k-records can be chosen in such a manner that the record data generated will exclude the specified number of outliers which are feared to be crept into the data. For example, if some initial scrutiny of the data reveals that there is a possibility of occurrence of only one outlier in terms of its largeness in the data, then it is enough to consider upper 2-records as the desirable record data that may be used for further analysis. Hence, it is beneficial to construct k-records from a sequence of random variables than constructing classical record values in such situations.

Suppose $\{X_i,\ i \geq 1\}$ is a sequence of iid random variables. If for a positive integer k, we collect those observations in the sequence which occupy the k-th largest position but exceeds in value for the first time the just previously recorded k-th largest value.

Then, the resulting sequence is known as the sequence of k-th upper records or simply k-records. We denote the times at which upper k-record values occur as $T_{n(k)}$ for $n = 1, 2, ...$ and are defined by $T_{1(k)} = k$ and for $n > 1$, $T_{n+1(k)} = \min\{j : j > T_{n(k)}, X[j : j + k - 1] > X[T_{n(k)} - k + 1 : T_{n(k)}]\}$, where $X[p : q]$ is the p-th order statistic in a random sample of size q. Then we define the sequence of upper k-record values denoted by $U_{n(k)}$ as $U_{n(k)} = X[T_{n(k)} - k + 1 : T_{n(k)}]$. If the parent distribution is absolutely continuous with survival function $\bar{F}_X(x)$ and pdf $f_X(x)$, then, the pdf of the n-th upper k-record value $U_{n(k)}$ is given by (see Arnold *et al.* [4])

$$f_{n(k)}(x) = \frac{k^n}{\Gamma(n)}\left[-\ln \bar{F}_X(x)\right]^{n-1}[\bar{F}_X(x)]^{k-1} f_X(x),\ \ n = 1, 2, ... \tag{2.1}$$

Similarly, we can define the lower k-records. For a positive integer k, if we denote the times at which lower k-records occur as $T^L_{n(k)}$ for $n = 1, 2, ...$ and are defined by $T^L_{1(k)} = k$ and for $n > 1$, $T^L_{n+1(k)} = \min\{j : j > T^L_{n(k)}, X[j : j + k - 1] < X[T^L_{n(k)} - k + 1 : T^L_{n(k)}]\}$. Then we define the sequence of lower k-records denoted by $L_{n(k)}$ as $L_{n(k)} = X[T^L_{n(k)} - k + 1 : T^L_{n(k)}]$. If the parent distribution is absolutely continuous with cumulative distribution function (cdf) $F_X(x)$ and pdf $f_X(x)$, then, the pdf of the n-th lower k-record value $L_{n(k)}$ is given by (see Ahsanullah [3])

$$g_{n(k)}(x) = \frac{k^n}{\Gamma(n)}[-\ln F(x)]^{n-1}[F(x)]^{k-1} f(x), \quad n = 1, 2, ... \tag{2.2}$$

Several applications of k-records are available in the literature. In reliability, a k-out-of-n system breaks down at the time of the failure of $(n - k + 1)$-th component. The life time of a k-out-of-n system is the $(n - k + 1)$-th order statistic in a sample of size n. Consequently, the n-th upper k-record value can be regarded as the life length of a k-out-of-$T_{n(k)}$ system. In actuarial science, there arises situations where second or third largest set of values are of special interest when the insurance claim of some non-life insurance is considered. One may refer Kamps [23] for more details. Detailed description on the theoretical aspects as well as applications of k-records are available in Arnold *et al.* [4], Nevzorov [29] and Ahsanullah [3].

Many authors have discussed about the information measures of classical records and its generalized version (k-records) arising from probability distribution. Hofmann and Nagaraja [21] derived some general results on the Fisher information contained in the classical record values and Hofmann and Balakrishnan [20] derived some general results on the Fisher information contained in the k-record values generated from an iid sample of fixed size from a continuous distribution. Madadi and Tata [25] present results on the Shannon information contained in classical record values and Madadi and Tata [26] present results on the Shannon information contained in k-record values. They have established a relationship between the Shannon information content of a random sample of fixed size and the Shannon information in the data consisting of sequential maxima. Also, they have considered the information contained in the k-record data from an inverse sampling plan as well. Goel *et al.* [18] discussed the measure of entropy for past lifetime distributions based on k-records. Recently, Jose and Sathar [22] studied some important properties of residual extropy of k-record values as well. It is to be noted that, when $k = 1$, we can easily obtain classical record values from k-records. Hence, k-records can be also considered as a generalized version of classical records. Baratpour *et al.* [6] studied entropy properties of classical records. Abbasnejad and Arghami [2] have discussed about the information contained in classical record values in detail and have derived some important properties as well. But to the best of our knowledge, no attention has been paid to the study of Rényi information contained in k-records.

Through this paper, the Rényi entropy of k-records arising from any continuous distribution has been discussed in detail. We also explore some of its important properties and have presented two applications of Rényi entropy of k-records.

3. RÉNYI ENTROPY OF k-RECORDS

Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with parent distribution $f(x)$. Then, analogous to (1.1), the Rényi entropy of n-th upper k-record value $\big(U_{n(k)}\big)$ is defined by

$$H_\alpha(U_{n(k)}) = \frac{1}{1-\alpha} \ln \int_x f^\alpha_{n(k)}(x)dx, \quad \alpha > 0,\ (\alpha \neq 1). \tag{3.1}$$

In the following example, we illustrate that Rényi entropy measure of uncertainty contained in the original random variable is more when compared to that of k-records arising from the observations on the original random variable.

Example 3.1. Assume X is a random variable following $U(2,4)$ with pdf given by

$$f_X(x) = \begin{cases} \frac{1}{2}, & 2 \leq x \leq 4, \\ 0, & \text{otherwise.} \end{cases}$$

We use the Rényi entropy to measure the uncertainty involved in the random variable X. Let $H_\alpha(X)$ denote the Rényi entropy of X. Then from (1.1), we get $H_\alpha(X) = \ln 2$. Also, the Rényi entropy of n-th upper k-record value arising from $U(2,4)$ is obtained from (3.1) as

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha} \ln\left[\frac{k^{\alpha n}}{\Gamma^\alpha(n)2^{\alpha-1}} \frac{\Gamma(\alpha(n-1)+1)}{(\alpha(k-1)+1)^{\alpha(n-1)+1}}\right].$$

It is to be noted that $H_\alpha(X)$ is independent of α. Moreover, $H_\alpha(X) - H_\alpha\big(U_{n(k)}\big) \geq 0$ for $\alpha > 0$. This means that the uncertainty of X is more than $U_{n(k)}$. Thus, the predictability of X is smaller than the predictability of $U_{n(k)}$. The graphical representation of Rényi entropy of X and the Rényi entropy of $U_{n(k)}$ for varying α is given in Figure 1.

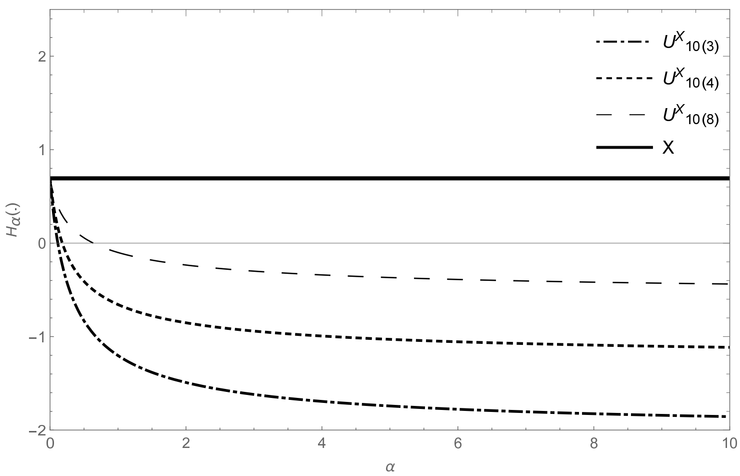


Figure 1: Rényi entropy of X and $U_{n(k)}$ for various values of α.

Fashandi and Ahmadi [15] have represented Rényi entropy of n-th upper k-record value in terms of Rényi entropy of n-th upper k-record value arising from $U(0,1)$. But they have not used that representation to study the properties of Rényi entropy of n-th upper k-record value arising from any continuous distribution. In this paper, we use the expression of Rényi entropy of n-th upper k-record value in terms of Rényi entropy of n-th upper k-record value arising from $U(0,1)$ to carry out investigation on properties and divergence of Rényi entropy of n-th upper k-record value. Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with a common distribution $U(0,1)$. Let $U^*_{n(k)}$ denote the n-th upper k-record value arising from the sequence $\{X_i,\ i \geq 1\}$. Using (2.1) in (3.1), we get

$$H_\alpha\left(U^*_{n(k)}\right) = \frac{1}{1-\alpha} \ln \int_{-\infty}^{\infty} \frac{k^{\alpha n}}{\Gamma^\alpha(n)} [\ln(1-x)]^{\alpha(n-1)} [1-x]^{\alpha(k-1)} dx.$$

Using the transformation $z = -\ln(1-x)$, we have

$$H_\alpha\left(U^*_{n(k)}\right) = \frac{1}{1-\alpha} \ln \int_0^{\infty} \frac{k^{\alpha n}}{\Gamma^\alpha(n)} z^{\alpha(n-1)} \mathrm{e}^{-z(\alpha(k-1)+1)} dz.$$

Then, the Rényi entropy of $U^*_{n(k)}$ is given by

$$H_\alpha\left(U^*_{n(k)}\right) = \frac{1}{1-\alpha} \ln \left[\frac{k^{\alpha n}}{\Gamma^\alpha(n)} \frac{\Gamma(\alpha(n-1)+1)}{(\alpha(k-1)+1)^{\alpha(n-1)+1}} \right]. \tag{3.2}$$

Then, for a sequence of iid random variables $\{X_i,\ i \geq 1\}$ with cdf $F(x)$ and pdf $f(x)$. If we denote $U_{n(k)}$ the n-th upper k-record value of the sequence $\{X_i\}$. Applying (2.1) in (3.1), we get

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha} \ln \frac{k^{\alpha n}}{\Gamma^\alpha(n)} \int_{-\infty}^{\infty} [-\ln(1-F(x))]^{\alpha(n-1)} [1-F(x)]^{\alpha(k-1)} f^\alpha(x) dx.$$

Using the transformation $u = -\ln(1-F(x))$ and on integrating, we get

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha} \ln \left\{ \frac{k^{\alpha n}}{\Gamma^\alpha(n)} \frac{\Gamma(\alpha(n-1)+1)}{(\alpha(k-1)+1)^{\alpha(n-1)+1}} E\big[f^{\alpha-1}\big(F^{-1}(1-\mathrm{e}^{-V})\big)\big] \right\},$$

where V follows gamma distribution with parameters $\alpha(n-1)+1$ and $\alpha(k-1)+1$ and we denote it by $V \sim \text{Gamma}(\alpha(n-1)+1, \alpha(k-1)+1)$. Then, from (3.2), the Rényi entropy of $U_{n(k)}$ is given by

$$H_\alpha\big(U_{n(k)}\big) = H_\alpha\left(U^*_{n(k)}\right) + \frac{1}{1-\alpha} \ln\big\{E\big[f^{\alpha-1}\big(F^{-1}(1-\mathrm{e}^{-V})\big)\big]\big\}. \tag{3.3}$$

Similarly, the Rényi entropy of n-th lower k-record value arising from any continuous distribution can be expressed in terms of Rényi entropy of n-th lower k-record value arising from $U(0,1)$. Let $L_{n(k)}$ denote the n-th lower k-record value of the sequence $\{X_i\}$. Then, the Rényi entropy of $L_{n(k)}$ is given by

$$H_\alpha\big(L_{n(k)}\big) = H_\alpha\left(L^*_{n(k)}\right) + \frac{1}{1-\alpha} \ln\big\{E\big[f^{\alpha-1}\big(F^{-1}(e^{-V})\big)\big]\big\}, \tag{3.4}$$

where $H_\alpha\left(L^*_{n(k)}\right)$ denote the Rényi entropy of n-th lower k-record value arising from $U(0,1)$ and $V \sim \text{Gamma}(\alpha(n-1)+1, \alpha(k-1)+1)$.

As an illustration, we obtain the Rényi entropy of k-records arising from exponential and Pareto distribution in the following examples.

Example 3.2. Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables having a common Pareto distribution with density function given by

$$f(x) = \frac{\beta}{\sigma}\left(\frac{x}{\sigma}\right)^{-\beta-1}, \quad x > \sigma.$$

Here,

$$F^{-1}(x) = \sigma[1-x]^{-\frac{1}{\beta}}.$$

Now, we have

$$E\big[f(F^{-1}(1-\mathrm{e}^{-V_n}))\big] = \frac{\beta^{\alpha n}}{\sigma^{\alpha-1}}\left[\frac{\alpha(k-1)+1}{\alpha(\beta k+1)-1}\right]^{\alpha(n-1)+1}.$$

Then, from (3.2) and (3.3), we get

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha}\ln\left[\frac{k^{\alpha n}}{\Gamma^\alpha(n)}\frac{\beta^{\alpha n}\Gamma(\alpha(n-1)+1)}{\sigma^{\alpha-1}[\alpha(\beta k+1)-1]^{\alpha(n-1)+1}}\right].$$

The graphical representation of the Rényi entropy of $U^X_{n(k)}$ arising from Pareto distribution with shape parameter $\beta = 3$ and scale parameter $\sigma = 2$ is given in Figure 2, for varying α.

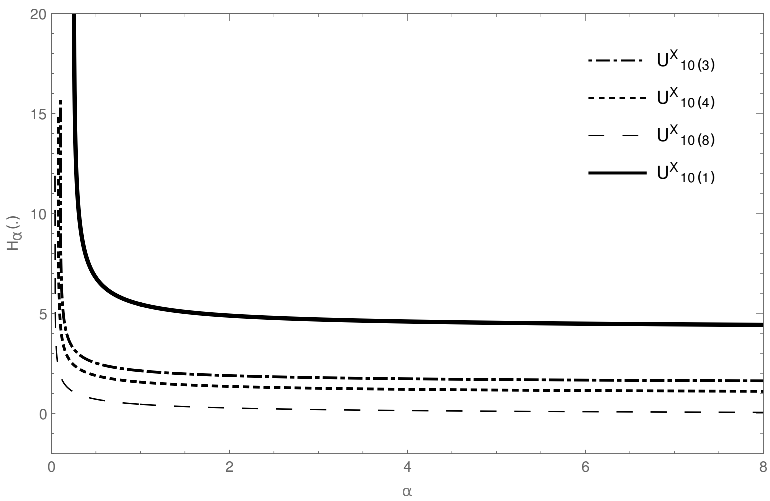


Figure 2: Rényi entropy of $U^X_{n(k)}$ for various values of α.

If we put $k = 1$, we can easily obtain the classical records from the sequence of k-records. From the figure, it can be observed that the Rényi entropy of classical upper record values (when $k = 1$) is greater than the Rényi entropy of upper k-records. This means that the uncertainty contained in classical records is more than that of k-records. Hence, one may observe certain situations where the predictability of classical records is less than the predictability of k-records when analyzed using Rényi entropy.

Example 3.3. Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables having a common exponential distribution with density function given by

$$f(x) = \theta \mathrm{e}^{-\theta x}, \quad x > 0,\ \theta > 0.$$

Here,

$$F^{-1}(x) = -\frac{1}{\theta} \ln(1 - x).$$

Now, we have

$$E\big[f^{\alpha-1}(F^{-1}(1 - \mathrm{e}^{-V}))\big] = \left[\frac{\alpha(k-1)+1}{\alpha k}\right]^{\alpha(n-1)+1} \theta^{\alpha-1}.$$

Then, from (3.2) and (3.3), we get

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha} \ln\left[\frac{k^{\alpha n}}{\Gamma^\alpha(n)} \frac{\theta^{\alpha-1}\Gamma(\alpha(n-1)+1)}{(\alpha k)^{\alpha(n-1)+1}}\right].$$

4. PROPERTIES OF RÉNYI ENTROPY OF k-RECORDS

In this section, we discuss some important properties of Rényi entropy of upper and lower k-records arising from any continuous distribution. To determine the monotonicity of Rényi entropy of upper and lower k-records arising from any continuous distribution we make use of the following definitions of stochastic and likelihood ratio orders given in Shaked and Shanthikumar [32].

Definition 4.1. Let X and Y be two non-negative random variables with cdfs F and G and with pdfs f and g respectively, then X is said to be smaller than Y:

(**1**) in the likelihood ratio order, denoted by $X \leq_{\text{lr}} Y$, if $\frac{f(x)}{g(x)}$ is decreasing in $x \geq 0$;

(**2**) in the usual stochastic order, denoted by $X \leq_{\text{st}} Y$, if $\bar{F}(x) \leq \bar{G}(x)$ for all $x \geq 0$, where $\bar{H}(\cdot)$ is the survival function.

It is well known that $X \leq_{\text{lr}} Y \Longrightarrow X \leq_{\text{st}} Y$ and $X \leq_{\text{st}} Y$ if and only if $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing functions ϕ.

Definition 4.2. The random variable X is said to be less than or equal to the random variable Y in Rényi entropy ordering, denoted by $X \leq_{\text{RE}} Y$, if $H_\alpha(X) \leq H_\alpha(Y)$ for all $\alpha > 0$.

The following theorem reveals the monotone behaviour of Rényi entropy of upper k-record values based on n.

Theorem 4.1. *Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with a common cdf $F(x)$ and pdf $f(x)$. Let $U_{n(k)}$ denote the n-th upper k-record value. If $f(x)$ is non-decreasing in x, then for $n > k$, $H_\alpha\big(U_{n(k)}\big)$ is non-increasing in n.*

Proof: The proof is straightforward as in Theorem 2.1 of Abbasnejad and Arghami [2]. □

In a similar way, we can state the monotone behaviour of Rényi entropy of lower k-records as given in the following theorem. The proof is not included since it easily follows as in Theorem 4.1.

Theorem 4.2. *Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with a common cdf $F(x)$ and pdf $f(x)$. Let $L_{n(k)}$ denote the n-th lower k-record value. If $f(x)$ is non-increasing in x, then for $n > k$, $H_\alpha(L_{n(k)})$ is non-increasing in n.*

We will now discuss about the Rényi entropy ordering of n-th upper k-record value of two random variables. Abbasnejad and Arghami [2] have used Rényi entropy ordering of the random variables to establish their Rényi entropy ordering of classical record values. In the following theorem, we make use of Rényi entropy ordering of the random variables to establish their Rényi entropy ordering of n-th upper k-record value.

Theorem 4.3. *Let X and Y be two continuous random variables with cdfs $F(x)$ and $G(y)$ and pdfs $f(x)$ and $g(y)$ respectively. Suppose that $U^X_{n(k)}$ and $U^y_{n(k)}$ represents the n-th upper k record value arising from X and Y respectively. Assume that*

$$\Lambda_1 = \left\{ v > 0 \,\middle|\, \frac{g(G^{-1}(1-\mathrm{e}^{-v}))}{f(F^{-1}(1-\mathrm{e}^{-v}))} \leq 1 \right\},$$

$$\Lambda_2 = \left\{ v > 0 \,\middle|\, \frac{g(G^{-1}(1-\mathrm{e}^{-v}))}{f(F^{-1}(1-\mathrm{e}^{-v}))} > 1 \right\}$$

and $X \leq_{\mathrm{RE}} Y$. If $\inf \Lambda_1 \geq \sup \Lambda_2$, then $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$, $\forall\ n \geq 1$ and $n > k$.

Proof: The proof is omitted since it is similar to that of Theorem 2.3 in Abbasnejad and Arghami [2]. □

In the following example, we apply Theorem 4.3 to obtain Rényi entropy ordering of two random variables following exponential distribution based on upper k-records.

Example 4.1. Let X and Y be two random variables having common exponential distribution with different scale parameters σ and λ respectively, where $\sigma > \lambda$. Then from (1.1), we get

$$H_\alpha(X) = \frac{1}{1-\alpha} \ln(\alpha) - \ln(\sigma).$$

It can be easily verified that $H_\alpha(X)$ is a decreasing function of σ. Thus, we have $H_\alpha(X) \leq H_\alpha(Y)$ and thereby $X \leq_{\mathrm{RE}} Y$. We have $f(F^{-1}(1-\mathrm{e}^{-x})) = \frac{1}{\sigma}\mathrm{e}^{-x}$ and $\inf \Lambda_1 = \sup \Lambda_2$. Hence, by Theorem 4.3 we get $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$.

Similar to Theorem 4.3, we establish the Rényi entropy ordering of two random variables based on lower k-records.

Theorem 4.4. *Let X and Y be two continuous random variables with cdfs $F(x)$ and $G(y)$ and pdfs $f(x)$ and $g(y)$ respectively. Suppose*

$$\Lambda_1 = \left\{ v > 0 \,\middle|\, \frac{g(G^{-1}(\mathrm{e}^{-v}))}{f(F^{-1}(\mathrm{e}^{-v}))} \leq 1 \right\},$$

$$\Lambda_2 = \left\{ v > 0 \,\middle|\, \frac{g(G^{-1}(\mathrm{e}^{-v}))}{f(F^{-1}(\mathrm{e}^{-v}))} > 1 \right\}$$

and $X \leq_{\mathrm{RE}} Y$. If $\inf \Lambda_1 \geq \sup \Lambda_2$, then $L^X_{n(k)} \leq_{\mathrm{RE}} L^Y_{n(k)}$, $\forall\, n \geq 1$ and $n > k$.

The following lemma explains the effect of location-scale transformation on random variable in the case of Rényi entropy of k-records. The proof is simple and hence omitted.

Lemma 4.1. *Consider a non-negative random variable X with pdf f and cdf F. Let $Z = aX + b$ be a transformation on X, where $a > 0$ and $b \geq 0$ are constants. Then*

$$H_\alpha\left(U^Z_{n(k)}\right) = H_\alpha\left(U^X_{n(k)}\right) + \ln a, \tag{4.1}$$

where $U^Z_{n(k)}$ and $U^X_{n(k)}$ are the n-th k-record corresponding to Z and X respectively.

Thus, the Rényi entropy of k-records changes due to the change in scale, but it does not change due to the change in location. The next theorem will discuss on the Rényi entropy ordering of k-records under location-scale transformation.

Theorem 4.5. *Consider two absolutely continuous random variables X and Y. Assume that $U^Z_{n(k)}$ and $U^X_{n(k)}$ are the n-th upper k-record corresponding to X and Y respectively. Let $U^{Z_1}_{n(k)} = a_1 U^X_{n(k)} + b_1$ and $U^{Z_2}_{n(k)} = a_2 U^Y_{n(k)} + b_2$, where $a_1, a_2 > 0$ and $b_1, b_2 \geq 0$ are constants. If $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$, then $U^{Z_1}_{n(k)} \leq_{\mathrm{RE}} U^{Z_2}_{n(k)}$ for $a_1 \leq a_2$.*

Proof: If $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$, then

$$H_\alpha\left(U^X_{n(k)}\right) \leq H_\alpha\left(U^Y_{n(k)}\right).$$

Since $a_1 \leq a_2$, $\ln a_1 \leq \ln a_2$. Hence,

$$\ln a_1 + H_\alpha\left(U^X_{n(k)}\right) \leq \ln a_2 + H_\alpha\left(U^Y_{n(k)}\right).$$

Thus, from (4.1), we get $U^{Z_1}_{n(k)} \leq_{\mathrm{RE}} U^{Z_2}_{n(k)}$. Hence the theorem. □

We will now deduce the following corollary which removes the restriction on the scale constants.

Corollary 4.1. *Consider two absolutely continuous random variables X and Y. Assume that $U^Z_{n(k)}$ and $U^X_{n(k)}$ are the n-th upper k-record corresponding to X and Y respectively. Let $U^{Z_1}_{n(k)} = aU^X_{n(k)} + b$ and $U^{Z_2}_{n(k)} = aU^Y_{n(k)} + b$, where $a > 0$ and $b \geq 0$ are constants. If $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$, then $U^{Z_1}_{n(k)} \leq_{\mathrm{RE}} U^{Z_2}_{n(k)}$.*

We will now discuss the effect of monotone transformation for Rényi entropy of k-records through the following theorem.

Theorem 4.6. *Assume a strictly convex function ψ having $\psi(0) = 0$ and $\psi(\infty) = \infty$. Consider, if $Y = \psi(X)$ then*

$$H_\alpha\left(U^{Y}_{n(k)}\right) = H_\alpha\left(U^{*}_{n(k)}\right) + \frac{1}{1-\alpha}\ln\left\{E\left[\frac{f(F^{-1}(1-\mathrm{e}^{-V_n}))}{\psi'(F^{-1}(1-\mathrm{e}^{-V_n}))}\right]^{\alpha-1}\right\}, \tag{4.2}$$

where $V_n \sim \mathrm{Gamma}(\alpha(n-1)+1, \alpha(k-1)+1)$. Here, $U^{Y}_{n(k)}$ are the n-th upper k-record value corresponding to Y.

Proof: Let $g_{n(k)}(y)$ and $\bar{G}_{n(k)}(y)$ be the pdf and survival function of n-th upper k-record value corresponding to Y. Then, from (2.1) we get

$$H_\alpha\left(U^{Y}_{n(k)}\right) = \frac{1}{1-\alpha}\ln\int_0^\infty \frac{k^{\alpha n}}{\Gamma^\alpha(n)}\left[-\ln\bar{G}(y)\right]^{\alpha(n-1)}[\bar{G}(y)]^{\alpha(k-1)}g^\alpha(y)dy.$$

Applying the transformation $Y = \psi(X)$, we have

$$H_\alpha\left(U^{Y}_{n(k)}\right) = \frac{1}{1-\alpha}\ln\frac{k^{\alpha n}}{\Gamma^\alpha(n)}\int\limits_0^\infty\left[-\ln\bar{F}(x)\right]^{\alpha(n-1)}[\bar{F}(x)]^{\alpha(k-1)}\left(\frac{f(x)}{\psi'(x)}\right)^\alpha\psi'(x)dx.$$

Using the substitution $u = -\ln\bar{F}(x)$ in the integral, the theorem follows. □

The following theorem, establishes the Rényi entropy ordering of strictly increasing convex functions of two n-th upper k-records based on the Rényi entropy ordering of their respective k-records.

Theorem 4.7. *Suppose X and Y are non-negative random variables such that $U^{X}_{n(k)} \leq_{\mathrm{RE}} U^{Y}_{n(k)}$ and ψ be a strictly increasing convex function with $\psi(0) = 0$, $\psi(\infty) = \infty$, $\psi'(x)$ exists and is continuous with $\psi'(0) \geq 1$. Then $\psi\left(U^{X}_{n(k)}\right) \leq_{\mathrm{RE}} \psi\left(U^{Y}_{n(k)}\right)$, where $U^{X}_{n(k)}$ and $U^{Y}_{n(k)}$ denote the n-th upper k-record value corresponding to X and Y respectively.*

Proof: Since $U^{X}_{n(k)} \leq_{\mathrm{RE}} U^{Y}_{n(k)}$, we have $H_\alpha\left(U^{X}_{n(k)}\right) \leq H_\alpha\left(U^{Y}_{n(k)}\right)$. This implies

$$H_\alpha\left(U^{X*}_{n(k)}\right)E\left[f^{\alpha-1}\left(F^{-1}(1-\mathrm{e}^{-V_n})\right)\right] \leq H_\alpha\left(U^{Y*}_{n(k)}\right)E\left[g^{\alpha-1}\left(G^{-1}(1-\mathrm{e}^{-V_n})\right)\right], \tag{4.3}$$

where $V_n \sim \mathrm{Gamma}(\alpha(n-1)+1, \alpha(k-1)+1)$. Then, from (4.2), we have

$$\begin{aligned} H_\alpha\left(\psi\left(U^{X}_{n(k)}\right)\right) - H_\alpha\left(\psi\left(U^{Y}_{n(k)}\right)\right) &= \\ &= H_\alpha\left(U^{X*}_{n(k)}\right) - H_\alpha\left(U^{Y*}_{n(k)}\right) + \frac{1}{1-\alpha}\ln\left\{\frac{E\left[\frac{f(F^{-1}(1-\mathrm{e}^{-V_n}))}{\psi'(F^{-1}(1-\mathrm{e}^{-V_n}))}\right]^{\alpha-1}}{E\left[\frac{g(G^{-1}(1-\mathrm{e}^{-V_n}))}{\psi'(G^{-1}(1-\mathrm{e}^{-V_n}))}\right]^{\alpha-1}}\right\}. \end{aligned}$$

Since $\psi'(0) \geq 1$ and from (4.3), we obtain $H_\alpha\left(\psi\left(U^{X}_{n(k)}\right)\right) - H_\alpha\left(\psi\left(U^{Y}_{n(k)}\right)\right) \leq 0$. Hence, $\psi\left(U^{X}_{n(k)}\right) \leq_{\mathrm{RE}} \psi\left(U^{Y}_{n(k)}\right)$. □

Therefore, we can observe that the Rényi entropy ordering of two random variables determine the Rényi entropy ordering of their respective k-records and the Rényi entropy ordering of the respective convex function of k-records are determined by the Rényi entropy ordering of their respective k-records. The following example discusses the same.

Example 4.2. Consider a convex function $\psi(x) = \beta x$, where $\beta \geq 1$. Hence ψ be a strictly increasing convex function with $\psi(0) = 0$, $\psi(\infty) = \infty$, $\psi'(x)$ exists and is continuous with $\psi'(0) \geq 1$. From Example 4.1, we have $U^X_{n(k)} \leq_{\mathrm{RE}} U^Y_{n(k)}$. Thus, the assumptions of Theorem 4.7 are satisfied and therefore, we can directly obtain $\psi\Big(U^X_{n(k)}\Big) \leq_{\mathrm{RE}} \psi\Big(U^Y_{n(k)}\Big)$ in which X and Y have common exponential distribution with different scale parameters σ and λ respectively, where $\sigma > \lambda$.

We will now study another property regarding the bound of Rényi entropy of k-records. Through the following theorem, we present a lower bound for the Rényi entropy of upper k-records arising from any continuous distribution.

Theorem 4.8. *Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with a common distribution function $F(x)$ and density function $f(x)$. Let $H_\alpha\big(U_{n(k)}\big)$ denote the Rényi entropy of n-th upper k-record value arising from the sequence and $H_\alpha\Big(U^*_{n(k)}\Big)$ denote the Rényi entropy of n-th upper k-record value arising from $U(0,1)$. Suppose that $M = f(m)$ exists, where M is the mode of X, then for $\alpha > 0$*

$$H_\alpha\big(U_{n(k)}\big) \geq H_\alpha\Big(U^*_{n(k)}\Big) - \ln M. \tag{4.4}$$

Proof: Since M is the mode of X, we have

$$f(F^{-1}(y)) \leq M.$$

Using the transformation $y = 1 - \mathrm{e}^{-V}$, we get

$$\begin{aligned} f(F^{-1}(1-\mathrm{e}^{-U})) &\leq M, \\ f^{\alpha-1}(F^{-1}(1-\mathrm{e}^{-U})) &\leq M^{\alpha-1}. \end{aligned}$$

Taking expectations on both sides, we obtain

$$E\big[f^{\alpha-1}(F^{-1}(1-\mathrm{e}^{-U}))\big] \leq M^{\alpha-1}. \tag{4.5}$$

Similarly, for $0 < \alpha < 1$

$$E\big[f^{\alpha-1}(F^{-1}(1-\mathrm{e}^{-U}))\big] \geq M^{\alpha-1}. \tag{4.6}$$

From (4.5) and (4.6), for $\alpha > 0$, we have

$$\frac{1}{1-\alpha} \ln E\big[f^{\alpha-1}(F^{-1}(1-\mathrm{e}^{-U}))\big] \geq -\ln M. \tag{4.7}$$

Using (3.3) in (4.7), we get

$$\begin{aligned} H_\alpha\big(U_{n(k)}\big) - H_\alpha\Big(U^*_{n(k)}\Big) &\geq -\ln M \\ H_\alpha\big(U_{n(k)}\big) &\geq H_\alpha\Big(U^*_{n(k)}\Big) - \ln M. \end{aligned}$$

Hence the theorem. □

In the following example, we make use of Theorem 4.8 to obtain bound for Rényi entropy of upper k-record value arising from Gompertz distribution.

Example 4.3. The pdf of Gompertz distribution with shape parameter λ and scale parameter β is given by

$$f(x) = \beta\lambda \mathrm{e}^{\beta x+\lambda(1-\mathrm{e}^{\beta x})}, \quad x > 0,\ \beta,\ \lambda > 0.$$

We know that mode of this distribution is $\frac{1}{\beta}\ln\frac{1}{\lambda}$. Thus, from (4.4) we have

$$H_\alpha\big(U_{n(k)}\big) \geq \frac{1}{1-\alpha}\ln\left[\frac{k^{\alpha n}\beta}{\ln\lambda\Gamma^\alpha(n)}\frac{\Gamma(\alpha(n-1)+1)}{(\alpha(k-1)+1)^{\alpha(n-1)+1}}\right].$$

In the following theorem, similar to Theorem 4.8, we obtain lower bound for Rényi entropy of lower k-records arising from any continuous distribution.

Theorem 4.9. *Let $\{X_i,\ i \geq 1\}$ be a sequence of iid random variables with a common distribution function $F(x)$ and density function $f(x)$. Let $H_\alpha\big(L_{n(k)}\big)$ denote the Rényi entropy of n-th lower k-record value arising from the sequence and $H_\alpha\Big(L^*_{n(k)}\Big)$ denote the Rényi entropy of n-th lower k-record value arising from $U(0,1)$. Suppose that $M = f(m)$ exists, where M is the mode of X, then for $\alpha > 0$*

$$H_\alpha\big(L_{n(k)}\big) \geq H_\alpha\Big(L^*_{n(k)}\Big) - \ln M. \tag{4.8}$$

In the following example, we make use of Theorem 4.9 to obtain lower bound for Rényi entropy of lower k-records arising from Fréchet distribution.

Example 4.4. The density function of Frechet distribution with shape parameter a and scale parameter s is given by

$$f(x) = \frac{a}{s}\left(\frac{x}{s}\right)^{-1-a}\mathrm{e}^{-\left(\frac{x}{s}\right)^{-a}}, \quad x > 0;\ a,\ s > 0.$$

We know that mode of this distribution is $s\left(\frac{a}{1+a}\right)^{\frac{1}{a}}$. Thus, from (4.8), we get

$$H_\alpha\big(U_{n(k)}\big) \geq \frac{1}{1-\alpha}\ln\left\{\left[\frac{a}{a+1}\right]^a\frac{k^{\alpha n}}{s\ \Gamma^\alpha(n)}\frac{\Gamma(\alpha(n-1)+1)}{(\alpha(k-1)+1)^{\alpha(n-1)+1}}\right\}.$$

5. APPLICATIONS OF RÉNYI ENTROPY OF k-RECORDS

This section deals with the applications of Rényi entropy of k-records. One application of Rényi entropy of k-records is that it can be used to characterize a class of distributions of non-negative random variables. Another application of Rényi entropy of k-records is that it determines Rényi divergence between the distribution of k-record values and its parent distribution.

5.1. Characterization of exponential distribution

Ebrahimi [14] suggested that maximum entropy paradigm can be used to produce a model for the data generating distribution. In the maximum entropy procedure, a model that best approximates the unknown distribution is derived based on the partial knowledge about this distribution in terms of a set of information constraints. Then, the inference is based on the model that maximizes the entropy of the random variables subject to the information constraints. In this subsection, we derive exponential distribution as the distribution that maximizes the Rényi entropy of k-records under some information constraints.

Let ξ be a class of distributions $F(x)$ of non-negative random variables X with $F(0) = 0$ such that

(i) $r(x, \theta) = a(\theta)b(x)$,

(ii) $b(x) \geq \beta, \quad \beta > 0$,

where $b(x) = B'(x)$ is a non-negative function of x and $a(\theta)$ is a non-negative function of θ.

Abbasnejad and Arghami [2] derived exponential distribution as the distribution that maximizes the Rényi entropy of classical record values under some information constraints. In the following theorem we characterize ξ using the Rényi entropy of n-th upper k-record value.

Theorem 5.1. *Let $U_{n(k)}$ be the n-th upper k-record value of $F(x;\theta)$, where $F(x;\theta)$ is in class ξ. Then, the n-th upper k-record value of the distribution $F(x;\theta)$ has maximum Rényi entropy in ξ if and only if $F(x;\theta) = 1 - \mathrm{e}^{-a(\theta)\beta x}$.*

Proof: The proof follows similar steps to that of Theorem 4.1 in Abbasnejad and Arghami [2]. □

5.2. Rényi divergence of k-records

Several applications of entropy divergence measures in formulating test statistics for testing of hypotheses and goodness-of-fit tests are available in literature. Gil *et al.* [16] presented closed form expressions of Rényi divergence for nineteen commonly used univariate continuous distributions as well as those for multivariate Gaussian and Dirichlet distributions. Salicrú *et al.* [31] suggested test statistics using some families of divergence like ϕ-divergence. Vasicek [35] used the sample Shannon entropy estimate to test normality. Abbasnejad [1] obtained a test statistic for exponentiality based on Rényi divergence. Abbasnejad and Arghami [2] studied Rényi divergence between parent distribution and distribution of classical record value as well. Through the following theorem, we study Rényi divergence between parent distribution and distribution of n-th upper k-record value.

Theorem 5.2. *The Rényi divergence between distribution of n-th upper k-record value and its parent distribution is given by*

$$D_\alpha(f_{n(k)}, f) = -H_\alpha\Big(U^*_{n(k)}\Big),$$

where $f_{n(k)}$ is the pdf of $U_{n(k)}$ and $U^*_{n(k)}$ is the n-th upper k-record value arising from $U(0,1)$. Moreover, $D_\alpha(f_{n(k)}, f)$ is increasing in n.

Proof: Using (2.1) in (1.2) and by the transformation $u = -\ln \bar{F}(x)$, we get

$$\begin{aligned} D_\alpha(f_{n(k)}, f) &= \frac{1}{\alpha-1}\ln\int_0^\infty \frac{k^{\alpha n}}{\Gamma^\alpha(n)} u^{\alpha(n-1)} \mathrm{e}^{-u(\alpha(k-1)+1)} du, \\ &= -H_\alpha\Big(U^*_{n(k)}\Big). \end{aligned}$$

Hence, the Rényi divergence between the distribution of the n-th upper k-record value and the parent distribution is distribution free. Moreover, taking the derivative of $H_\alpha\Big(U^*_{n(k)}\Big)$ with respect to n, we get

$$\frac{dH_\alpha\Big(U^*_{n(k)}\Big)}{dn} = \frac{\alpha}{\alpha-1}(1-\ln k) - \frac{1}{\alpha-1}\xi(\alpha(n-1)+1) + \frac{\alpha}{\alpha-1}\xi(n),$$

where $\xi(u) = \dfrac{d\ln\Gamma(u)}{du}$. For every u, the function $\xi(u)$ is non-decreasing and therefore $H_\alpha\Big(U^*_{n(k)}\Big)$ is non-increasing in n. Thus the result follows. □

Thus, by increasing n, we expect that the divergence between the distribution of the n-th upper k-record value and the parent distribution increases.

5.3. Numerical illustration

In this subsection, we propose a simple estimator for the Rényi entropy of the n-th upper k-record value and discuss the merit of k-records over classical records and parent random variable in terms of uncertainty. To estimate the Rényi entropy based on n-th upper k-record value, kernel density has been applied to estimate the density function and empirical distribution has been used as an estimator for the distribution function. The estimator is proposed for Rényi entropy obtained by replacing the density of the parent random variable by the density of n-th upper k-record value and hence much complexities arises while deriving the properties of the proposed estimator directly. Therefore, the proposed simple estimator for Rényi entropy based on n-th upper k-record value can be analysed numerically by evaluating the average bias and MSE for different sample sizes which examines the bias and consistency characteristics of the proposed estimator. A numerical illustration has been presented with an intention to describe the benefit of applying Rényi entropy based on n-th k-record in comparison to that of the parent random variable. Using (2.1) in (3.1), the Rényi entropy of the n-th upper k-record can be expressed as

$$H_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha}\ln\int\limits_0^\infty \frac{k^{\alpha n}}{\Gamma^\alpha(n)}\big[-\ln\bar{F}(x)\big]^{\alpha(n-1)}\big[\bar{F}(x)\big]^{\alpha(k-1)} f^\alpha(x)dx. \tag{5.1}$$

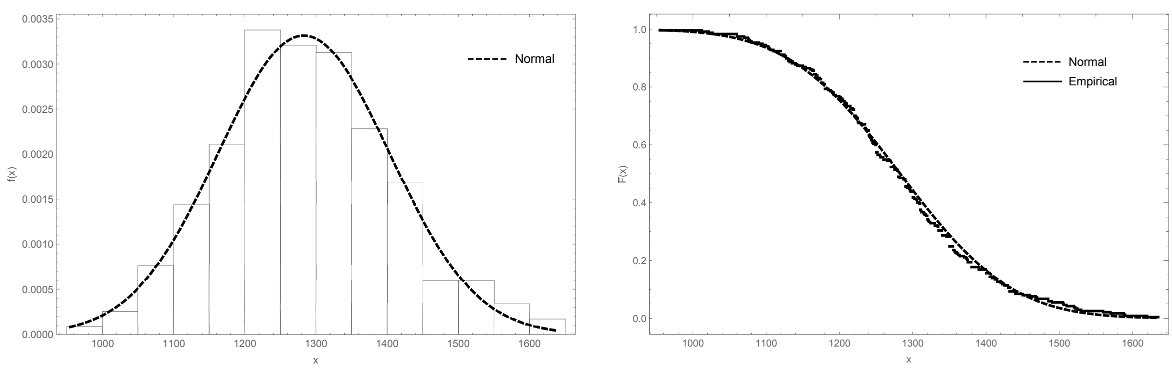
A simple estimator for the Rényi entropy of the n-th upper k-records value based on a random sample of size n is given by

$$\hat{H}_\alpha\big(U_{n(k)}\big) = \frac{1}{1-\alpha} \ln \int_0^\infty \frac{k^{\alpha n}}{\Gamma^\alpha(n)} \Big[-\ln \hat{\bar{F}}(x)\Big]^{\alpha(n-1)} \Big[\hat{\bar{F}}(x)\Big]^{\alpha(k-1)} \hat{f}^\alpha(x)dx, \tag{5.2}$$

where $\hat{f}(x) = \frac{1}{nb_n} \sum\limits_{j=1}^{n} K\big(\frac{x-X_j}{b_n}\big)$, denotes the kernel density estimator with the bandwidth b_n. Also $K(\cdot)$ is a kernel function satisfying the condition $\int_{-\infty}^{\infty} K(x)dx = 1$ and is usually a symmetric pdf. Also, $\hat{\bar{F}}(x) = \frac{1}{n} \sum\limits_{i=1}^{n} I(X_i \geq x)$ is the empirical survival function and $I(X_i \geq x)$ is the indicator function.

In the following illustration, we use a real life data set to compute Rényi entropy of the n-th upper k-record value and make a comparison with that of classical records and parent random variable.

Dataset 1: Let the random variable X represents the brain weight (in grams) of 237 adults discussed in Gladstone [17]. The brain weight of an adult is not so easy to obtain and hence for more reliable inferences on the random variable X, the distribution of X should possess less uncertainty. The study focus on the uncertainty contained in the distribution of the random variable X. Initial study on distribution of X suggests the normal distribution with location parameter $\mu = 1282.87$ and scale parameter $\sigma = 120.86$ is a good fit for the data set with Kolmogrove-Smirnov (K-S) statistic $= 0.03914$ and p-value $= 0.9755$. Since the normal distribution is a good fit for the proposed data, a Gaussian kernel can be chosen for estimation procedure using the given data set. The fit of normal distribution to data is depicted in Figure 3.



(a) Histogram and normal curve. **(b)** Normal and empirical CDF.

Figure 3: Modelling brain weight data using normal distribution.

To estimate Rényi entropy of the n-th upper k-record value the Gaussian kernel with $b_n = 120$ is applied in (5.2). The closeness of the estimators of Rényi entropy based on n-th upper k-record value and the parent random variable with the theoretical value of Rényi entropy which has been obtained by assuming normal distribution for the random variable with parameter values $\mu = 1282.87$ and $\sigma = 120.86$ (ML estimates) for different choices of α are presented in Table 2.

Table 2: Comparison of theoretical values and estimates of Rényi entropy based on X and $U_{n(k)}$ where $k = 1, 2, 5, 7, 9$ and 10.

α	$H_\alpha(X)$	$\hat{H}_\alpha(X)$	$\hat{H}_\alpha(U_{n(1)})$	$\hat{H}_\alpha(U_{n(2)})$	$\hat{H}_\alpha(U_{n(5)})$	$\hat{H}_\alpha(U_{n(7)})$	$\hat{H}_\alpha(U_{n(9)})$	$\hat{H}_\alpha(U_{n(10)})$
0.10	6.9885	8.9341	7.2943	6.4240	6.3835	6.2870	5.9981	5.8530
0.30	6.5692	9.5116	8.8650	8.7573	8.7103	8.6019	8.1354	7.5349
0.50	6.4024	11.1967	10.3061	10.2393	9.7591	9.6735	9.6341	9.5618
0.70	6.2846	20.6588	13.7849	13.7476	12.6784	12.6646	12.6296	12.5986
1.15	6.1556	17.3662	12.4395	12.3814	12.3616	12.2889	11.8381	11.7295
1.40	6.1147	10.8954	10.0227	9.2979	9.2531	9.1823	9.1673	9.1643
1.75	6.0823	3.7508	7.7072	7.5870	7.8013	7.8826	7.5907	7.6997
2.00	6.0558	3.1864	6.8005	6.7703	6.7178	6.7033	6.5973	6.5817
2.25	6.0336	2.2530	6.4741	6.1328	6.0768	6.0333	5.9477	5.8533
2.50	6.0147	1.6343	5.0568	4.9876	4.7800	4.6415	4.5031	4.4339
3.25	5.9839	0.8677	3.9306	3.8767	3.7152	3.6076	3.4999	3.4460
3.50	5.9598	0.4133	3.3674	3.3213	3.1828	3.0906	2.9983	2.9521

From Table 2, we can observe that the estimates of Rényi entropy based on n-th upper k-record value is closer to its theoretical value than the estimate of Rényi entropy based parent random variable. Also, when $k = 1$, k-records becomes classical records. In terms of uncertainty, we have compared three different estimates (based on parent random variable, classical records and k-records) for Rényi entropy which can be obtained from a random sample. Hence, from Table 2, one may conclude that there are situations where construction of k-records or classical records from random sample gives closer estimate than the estimate obtained based on random variable. Moreover, the k-records or classical records are ordered random variables which carry an additional information about their ranks when compared to the parent random variable.

Table 3: Average bias and MSE of the estimate of Rényi entropy of the n-th upper k-record value for different choices of α.

n	k	$\alpha = 0.25$		$\alpha = 0.75$		$\alpha = 1.50$		$\alpha = 3.0$	
		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
20	1	1.14072	1.09495	1.12052	1.06003	1.11085	1.03190	1.09435	1.03083
	3	1.07479	1.00485	1.07209	0.99914	1.06225	0.92091	1.05467	0.90367
	6	1.00195	0.89941	0.99823	0.85221	0.96188	0.81123	0.94057	0.80224
	8	0.92137	0.79354	0.91658	0.77677	0.91307	0.77391	0.89728	0.75497
	10	0.85722	0.74954	0.83132	0.73397	0.81957	0.71637	0.80547	0.68588
60	1	0.93818	0.84330	0.90040	0.84103	0.88101	0.82568	0.87493	0.82497
	3	0.85666	0.81577	0.83509	0.80718	0.82274	0.74541	0.79530	0.73895
	6	0.79185	0.73802	0.78635	0.73201	0.78544	0.71686	0.76749	0.70849
	8	0.76585	0.65507	0.75573	0.61861	0.72946	0.57439	0.70244	0.57347
	10	0.68611	0.53900	0.67284	0.50139	0.66842	0.49361	0.65933	0.44052
100	1	0.76797	0.69524	0.76709	0.68935	0.75349	0.68063	0.75052	0.68010
	3	0.74507	0.59512	0.73545	0.59152	0.72329	0.56915	0.71883	0.55927
	6	0.71429	0.53813	0.70983	0.52673	0.69858	0.51378	0.65525	0.48069
	8	0.64116	0.46515	0.63902	0.43912	0.62873	0.43155	0.61199	0.42260
	10	0.58717	0.41116	0.57277	0.40205	0.57109	0.38755	0.56469	0.37685

To study the effect of the estimator suggested for Rényi entropy of the n-th upper k-record value denoted as $H_\alpha\big(U_{n(k)}\big)$, we have obtained average bias and mean square error (MSE) of the estimator using bootstrapping procedure. The bias and MSE of the estimates are evaluated from value of Rényi entropy of the n-th upper k-record obtained using the parameter estimates $\mu = 1282.87$ and scale parameter $\sigma = 120.86$ in (5.1) which we have considered as the true value of $H_\alpha\big(U_{n(k)}\big)$. The average bias and MSE of $H_\alpha\big(U_{n(k)}\big)$ based on 100 bootstrap estimates from samples of sizes 20, 60 and 100 are presented in Table 3. It can be observed that the average bias and MSE of the estimator of Rényi entropy of the n-th upper k-record value diminishes as sample size becomes large.

6. CONCLUSION

The study explains the relevance of k-records in measuring uncertainty using Rényi entropy after comparing it with Rényi entropy of classical records as well as with Rényi entropy of original random variable. Fashandi and Ahmadi [15] have expressed Rényi entropy for k-records arising from any continuous distribution in terms of Rényi entropy of k-records arising from uniform distribution and we have used that representation to derive some important properties of Rényi entropy of k-records. The monotone behaviour of Rényi entropy of k-records have been derived. We have shown that the Rényi entropy ordering of random variables determines the Rényi entropy ordering of their respective k-record values. The Renyi entropy ordering of k-records determines the Renyi entropy ordering of their respective linear transformations of k-records as well as their convex function of k-records. A lower bound for the Rényi entropy of k-records have been obtained in this work. We have applied Rényi entropy of k-records to characterize exponential distribution by maximization of Rényi entropy based on certain information constraints. The study also establishes that the Rényi divergence between the distribution of k-records and its parent distribution is distribution free and the divergence increases with increase in n. A simple estimator for Rényi entropy of k-records has been proposed and compared estimates of Renyi entropy of k-records, classical records and parent random variable using a real life data set.

ACKNOWLEDGMENTS

We acknowledge the valuable suggestions from the editor, associate editor and the referees that helped to improve the article to a great extent.

REFERENCES

[1] Abbasnejad, M. (2011). Some goodness of fit tests based on renyi information, *Applied Mathematical Sciences*, **5**(39), 1921–1934.

[2] Abbasnejad, M. and Arghami, N.R. (2011). Rényi entropy properties of records, *Journal of Statistical Planning and Inference*, **141**(7), 2312–2320.

[3] Ahsanullah, M. (2004). *Record Values – Theory and Applications*, University Press of America, Maryland, United States.

[4] Arnold, B.C.; Balakrishnan, N. and Nagaraja, H.N. (1998). *Records*, volume 768, John Wiley & Sons, New York.

[5] Asadi, M.; Ebrahimi, N. and Soofi, E.S. (2005). Dynamic generalized information measures, *Statistics & Probability Letters*, **71**(1), 85–98.

[6] Baratpour, S.; Ahmadi, J. and Arghami, N.R. (2007). Entropy properties of record statistics, *Statistical Papers*, **48**(2), 197–213.

[7] Chandler, K. (1952). The distribution and frequency of record values, *Journal of the Royal Statistical Society. Series B (Methodological)*, **14**, 220–228.

[8] Contreras-Reyes, J.E. (2015). Rényi entropy and complexity measure for skew-Gaussian distributions and related families, *Physica A: Statistical Mechanics and its Applications*, **433**, 84–91.

[9] Contreras-Reyes, J.E. and Cortés, D.D. (2016). Bounds on Rényi and Shannon entropies for finite mixtures of multivariate skew-normal distributions: application to swordfish (xiphias gladius linnaeus), *Entropy*, **18**(11), 382.

[10] Csiszár, I. (1995). Generalized cutoff rates and Rényi's information measures, *IEEE Transactions on Information Theory*, **41**(1), 26–34.

[11] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, Wiley, New York.

[12] De Gregorio, A. and Iacus, S.M. (2009). On Rényi information for ergodic diffusion processes, *Information Sciences*, **179**(3), 279–291.

[13] Dziubdziela, W. and Kopociński, B. (1976). Limiting properties of the k-th record values, *Applicationes Mathematicae*, **2**(15), 187–190.

[14] Ebrahimi, N. (2000). The maximum entropy method for lifetime distributions, *Sankhyā: The Indian Journal of Statistics, Series A*, 236–243.

[15] Fashandi, M. and Ahmadi, J. (2012). Characterizations of symmetric distributions based on Rényi entropy, *Statistics & Probability Letters*, **82**(4), 798–804.

[16] Gil, M.; Alajaji, F. and Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions, *Information Sciences*, **249**, 124–131.

[17] Gladstone, R.J. (1905). A study of the relations of the brain to the size of the head, *Biometrika*, **4**(1/2), 105–123.

[18] Goel, R.; Taneja, H. and Kumar, V. (2018). Measure of entropy for past lifetime and k-record statistics, *Physica A: Statistical Mechanics and its Applications*, **503**, 623–631.

[19] Golshani, L. and Pasha, E. (2010). Rényi entropy rate for Gaussian processes, *Information Sciences*, **180**(8), 1486–1491.

[20] Hofmann, G. and Balakrishnan, N. (2004). Fisher information in k-records, *Annals of the Institute of Statistical Mathematics*, **56**(2), 383–396.

[21] Hofmann, G. and Nagaraja, H. (2003). Fisher information in record data, *Metrika*, **57**(2), 177–193.

[22] Jose, J. and Sathar, E.I.A. (2019). Residual extropy of k-record values, *Statistics & Probability Letters*, **146**(C), 1–6.

[23] Kamps, U. (1995). A concept of generalized order statistics, *Journal of Statistical Planning and Inference*, **48**(1), 1–23.

[24] Kirchanov, V.S. (2008). Using the renyi entropy to describe quantum dissipative systems in statistical mechanics, *Theoretical and Mathematical Physics*, **156**(3), 1347–1355.

[25] Madadi, M. and Tata, M. (2011). Shannon information in record data, *Metrika*, **74**(1), 11–31.

[26] Madadi, M. and Tata, M. (2014). Shannon information in k-records, *Communications in Statistics-Theory and Methods*, **43**(15), 3286–3301.

[27] Morales, D.; Pardo, L. and Vajda, I. (1997). Some new statistics for testing hypotheses in parametric models, *Journal of Multivariate Analysis*, **62**(1), 137–168.

[28] Nadarajah, S. and Zografos, K. (2003). Formulas for Rényi information and related measures for univariate distributions, *Information Sciences*, **155**(1–2), 119–138.

[29] Nevzorov, V.B. (2001). *Records: Mathematical Theory*, American Mathematical Society, Providence, Rhode Island.

[30] Rényi, A. (1961). *On measures of entropy and information.* In "Proceedings of Fourth Berkely Symposium on Mathematics", Statistics and Probability 1960, volume 1, pp. 547–561, University of California Press, Berkely, California.

[31] Salicrú, M.; Morales, D.; Menéndez, M. and Pardo, L. (1994). On the applications of divergence type measures in testing statistical hypotheses, *Journal of Multivariate Analysis*, **51**(2), 372–391.

[32] Shaked, M. and Shanthikumar, J.G. (2007). *Stochastic Orders*, Springer Science & Business Media, Berlin, Germany.

[33] Shannon, C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal*, **27**(3), 379–423.

[34] Song, K.S. (2001). Rényi information, loglikelihood and an intrinsic distribution measure, *Journal of Statistical Planning and Inference*, **93**(1–2), 51–69.

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

REVSTAT – Statistical Journal
Volume 20, Number 4, July 2022, 501–528
https://doi.org/10.57805/revstat.v20i4.384

A Study on Discrete Bilal Distribution with Properties and Applications on Integer-Valued Autoregressive Process

Authors: Emrah Altun ✉
– Department of Mathematics, Bartin University,
Bartin 74100, Turkey
emrahaltun@bartin.edu.tr

M. El-Morshedy
– Department of Mathematics, College of Science and Humanities,
in Al-Kharj Prince Sattam bin Abdulaziz University,
Al-Kharj 11942, Saudi Arabia
m.elmorshedy@psau.edu.sa
– Department of Mathematics, Faculty of Science, Mansoura University,
Mansoura 35516, Egypt
mah_elmorshedy@mans.edu.eg

M. S. Eliwa
– Department of Mathematics, Faculty of Science, Mansoura University,
Mansoura 35516, Egypt
mseliwa@mans.edu.eg
– Department of Statistics and Operation Research, College of Science, Qassim University,
P.O. Box 6644, Buraydah 51482, Saudi Arabia
m.eliwa@qu.edu.sa

Received: December 2019 Revised: November 2020 Accepted: November 2020

Abstract:

- This study proposes a new one-parameter discrete distribution, called a discrete Bilal distribution. The structural properties of the proposed distribution, including the shape of the probability mass function, mode, moments, skewness, kurtosis, index of dispersion, mean deviation, stress-strength reliability, and order statistics, are derived. These properties are expressed in closed-forms. The maximum likelihood and method of moments estimation methods are considered to estimate the unknown model parameter. An extensive simulation study is carried out to examine the finite sample performance of estimation methods. The usefulness of the proposed model is illustrated in the first-order integer-valued autoregressive process. The empirical importance of the proposed models is proved through three real data applications.

Keywords:

- *Bilal distribution; INAR(1) process; method of moments; maximum likelihood; simulation.*

AMS Subject Classification:

- 60E05, 62E10, 62E15, 62F10, 62N05.

✉ Corresponding author.

1. INTRODUCTION

The count data sets arise in different fields such as yearly number destructive earthquakes, monthly traffic accidents and hourly bacterial growth and among others. These kind of data sets are modeled with discrete probability distributions. Poisson and negative-binomial distributions are the most popular distributions and are widely used to model these kind data sets. In recent years, researchers have shown great interest to introduce new discrete distributions by discretizing a continuous failure time model. Let the continuous random variable X has the survival function (sf) $S(x) = \Pr(X > x)$. The probability mass function (pmf) dealing with the continuous random variable X is given by

$$\Pr(X = x) = S(x) - S(x + 1), \quad x = 0, 1, 2, ...$$

Many researchers have introduced sophisticated discrete distributions by applying the discretization method to the continuous failure time models. For instance, discrete Lindley distribution by Gómez-Déniz and Calderín-Ojeda (2011) [12], discrete Rayleigh distribution by Roy (2004) [28], discrete inverse Rayleigh distribution by Hussain and Ahmad (2014) [13], discrete Pareto distribution by Buddana and Kozubowski (2014) [6], discrete Weibull distribution by Nakagawa and Osaki (1975) [21], discrete Lomax distribution by Para and Jan (2016a) [24], discrete generalized Weibull distribution by Para and Jan (2017) [26] and exponentiated discrete Lindley by El-Morshedy *et al.* (2019) [10], discrete flexible one parameter distribution by Eliwa and El-Morshedy (2020) [7] and discrete gompertz-G by Eliwa *et al.* (2020a) [8] and among others. The discrete analogue of the Burr-Hatke distribution was introduced by El-Morshedy *et al.* (2020) [11] with its regression model and residual analysis. More recently, Eliwa *et al.* (2020b) [9] introduced the discrete analogue of the three-parameter Lindley distribution and demonstrated its performance in modeling the time series of counts.

In this paper, we introduce a new one-parameter discrete distribution by applying the discretization method to the Bilal distribution, proposed by Abd-Elrahman (2013) [4]. The arising distribution is called as the discrete Bilal (DBL) distribution. The DBL distribution has simple probability mass and cumulative distribution functions and statistical properties such as mean, mode, skewness, kurtosis measures, mean deviation and also stress-strength reliability are obtained in explicit forms. The DBL distribution provides an opportunity to model different types of the count data sets such over and under-dispersed. We illustrate the importance of DBL distribution in first-order integer-valued autoregressive (INAR(1)) process by applying the DBL distribution as an innovation process of INAR(1) process, introduced by McKenzie (1985) [20] and Al-Osh and Alzaid (1987) [1]. INAR(1) process is widely used to model time series of counts. Several researchers have done important studies on the INAR(1) processes with more flexible innovation distributions. For instance, Jazi *et al.* (2012) [14] introduced the INAR(1) process with geometric innovations (INAR(1)G) to model the over-dispersed time series of counts. Similarly, Lívio *et al.* (2018) [19] introduced the INAR(1) process with Poisson-Lindley innovations (INAR(1)PL) for over-dispersed time series of counts. More recently, Altun (2020a) [2] introduced a new generalization of the geometric and demonstrated its performance in INAR(1) process. More recently, Altun (2020b) [3] introduced a mixed Poisson distribution and defined a new INAR(1) process for over-dispersed time series of counts.

The remaining parts of the presented study is organized as follows. The statistical properties of the DBL distribution are obtained in Section 2. The parameter estimation of the DBL distribution is discussed in Section 3. The INAR(1) process with DBL innovations is introduced in Section 4 with its parameter estimation. In Section 5, we discuss the finite-sample performance of the parameter estimation methods via two simulation studies. In Section 6, three data sets are analyzed with DBL and other competitive models to prove the importance of the DBL distribution practically. Section 7 deals with the concluding remarks of the study.

2. THE DISCRETE-BILAL DISTRIBUTION

Recently, Abd-Elrahman (2013) [4] proposed a new flexible model, called Bilal (BL) distribution. The cumulative distribution function (cdf) of the BL distribution is

$$\Pi(x;\beta) = 1 - \left(3 - 2e^{-\frac{x}{\beta}}\right) e^{-\frac{2x}{\beta}}, \ x \geq 0, \beta > 0. \tag{2.1}$$

The sf and probability density function (pdf) of (2.1) are given, respectively, by

$$S(x;\beta) = \left(3 - 2e^{-\frac{x}{\beta}}\right) e^{-\frac{2x}{\beta}}, \ x \geq 0, \beta > 0, \tag{2.2}$$

$$\pi(x;\beta) = \frac{6}{\beta}\left(1 - e^{-\frac{x}{\beta}}\right) e^{-\frac{2x}{\beta}}, \ x \geq 0, \beta > 0. \tag{2.3}$$

Now, we introduce a DBL distribution by discretizing the sf of the BL distribution. Let the parameter $p = e^{-\frac{1}{\beta}}$, the cdf of DBL distribution is given by

$$F(x;p) := F(X \leq x) = 1 - \left(3 - 2p^{x+1}\right) p^{2(x+1)}, \ x = 0, 1, 2, 3, ... \tag{2.4}$$

The corresponding sf and pmf to (2.4) are given, respectively, by

$$S(x;p) = \left(3 - 2p^{x+1}\right) p^{2(x+1)}, \tag{2.5}$$

and

$$f(x;p) := P(X = x) = 2(p^3 - 1)p^{3x} - 3(p^2 - 1)p^{2x}, \ \ x = 0, 1, 2, 3, ... \tag{2.6}$$

The pmf in (2.6) is log-concave for all values of p, where

$$\frac{f(x+1;p)}{f(x;p)} = \frac{2p^{x+6} - 2p^{x+3} - 3p^4 + 3p^2}{2p^{x+3} - 3p^2 - 2p^x + 3} \tag{2.7}$$

is a decreasing function in x for all value of p. The possible pmf shapes of the DBL distribution are displayed in Figure 1. These figures show that the DBL distribution has right-skewed shapes and it has long right-tails.

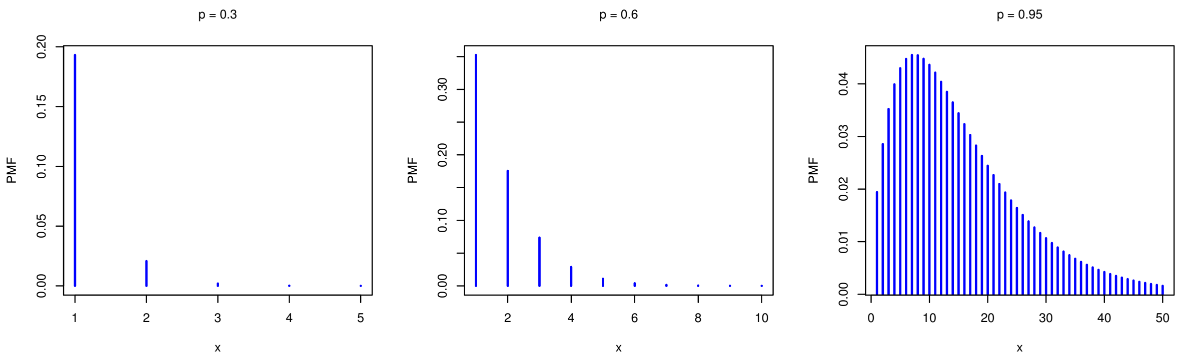


Figure 1: The pmf plots of the DBL distribution.

The hazard rate function (hrf) is

$$h(x;p) = \frac{2(p^3-1)p^x - 3(p^2-1)}{3-2p^x}, \quad x \in \mathbb{N}_0, \tag{2.8}$$

where $h(x;p) = \frac{f_x(x;p)}{R(x-1;p)}$. The reversed hazard rate function (rhrf) is

$$r(x;p) = \frac{2(p^3-1)p^{3x} - 3(p^2-1)p^{2x}}{1-(3-2p^{x+1})\,p^{2(x+1)}}, \quad x \in \mathbb{N}_0, \tag{2.9}$$

where $r(x;p) = \frac{f_x(x;p)}{F(x;p)}$. Figure 2 shows the hrf and rhrf plots for different values of the parameter p.

It is clear that the hrf of the DBL distribution increases up to time t where $0<t<x<\infty$, whereas the hrf is constant after time t. Regarding to the rhrf, it is seen that it always decreases for all x.

Suppose X_1 and X_2 are two independent random variables following the DBL distribution with the parameters p_1 and p_2, respectively. Let $W=\min(X_1,X_2)$ be a random variable which has a hrf

$$\begin{aligned} h_W(x;p_1,p_2) &= \frac{P(\min(X_1,X_2)\geq x) - P(\min(X_1,X_2)\geq x+1)}{P(\min(X_1,X_2)\geq x)} \\ &= \frac{2(p_1^3-1)p_1^x - 3(p_1^2-1)}{3-2p_1^x} + \frac{2(p_2^3-1)p_2^x - 3(p_2^2-1)}{3-2p_2^x} \\ &\quad - \frac{\left\{2(p_1^3-1)p_1^x - 3(p_1^2-1)\right\}\left\{2(p_2^3-1)p_2^x - 3(p_2^2-1)\right\}}{(3-2p_1^x)(3-2p_2^x)}. \end{aligned} \tag{2.10}$$

The extra term $h_1(x;p_1)h_2(x;p_2)$ arises because in the discrete case $P(X_1=x, X_2=x)\neq 0$, where $h_1(x;p_1)$ and $h_2(x;p_2)$ are the hrfs of X_1 and X_2, respectively. The rest of this section contains the statistical properties of the DBL distribution.

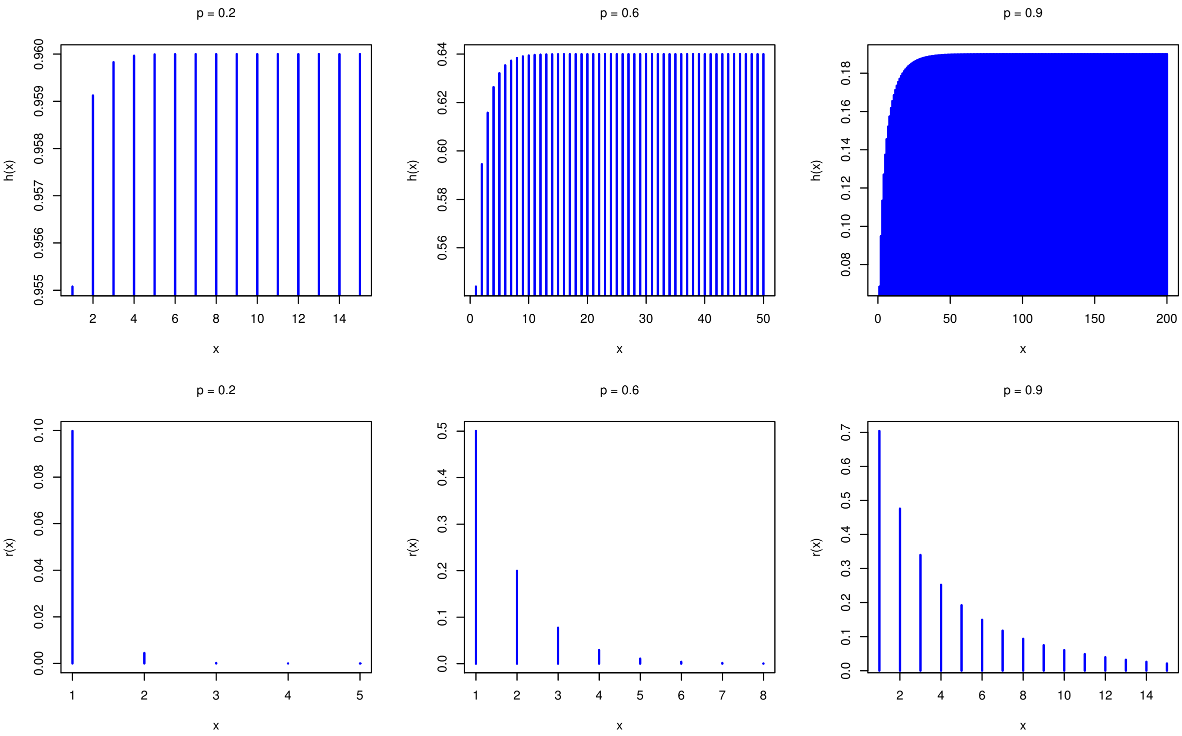


Figure 2: The hrf and rhrf of the DBL distribution.

2.1. Mode

The mode of the DBL distribution is obtained by solving (2.11):

$$6(p^3-1)p^{3x}\ln(p)-6(p^2-1)p^{2x}\ln(p)=0. \tag{2.11}$$

By solving (2.11), we have

$$\text{Mode}(X)=\frac{\ln(p+1)-\ln(p^2+p+1)}{\ln(p)}. \tag{2.12}$$

As seen from (2.12), mode of the DBL distribution is an increasing function of the parameter p.

2.2. Moments, skewness and kurtosis

The probability generating function (pgf) of the DBL distribution is obtained as follows:

$$\begin{aligned} G_X(s) &= \sum_{x=0}^{\infty} s^x f_x(x;p) \\ &= 2\sum_{x=0}^{\infty}(p^3-1)\left(p^3 s\right)^x - 3\sum_{x=0}^{\infty}(p^2-1)\left(p^2 s\right)^x \\ &= \frac{2(p^3-1)}{1-p^3 s}-\frac{3(p^2-1)}{1-p^2 s}, \end{aligned} \tag{2.13}$$

where $\sum_{x=0}^{\infty} aq^x = \frac{a}{1-q}$. Replacing s with e^s, the moment generating function (mgf) of the DBL distribution is

$$M_X(s) = \frac{2(p^3-1)}{1-p^3e^s} - \frac{3(p^2-1)}{1-p^2e^s}. \tag{2.14}$$

Using the mgf, given in (2.14), we obtain the mean, variance, skewness and kurtosis of the DBL distribution, given, respectively, by

$$\mathrm{E}(X) = \frac{p^2(p^2+p+3)}{(p^2+p+1)(1-p^2)}, \tag{2.15}$$

$$\mathrm{Var}(X) = \frac{p^2(3p^4+4p^3-p^2+4p+3)}{(p^2+p+1)^2(p^2-1)^2}, \tag{2.16}$$

$$\mathrm{Sk}(X) = -\frac{3p^8+7p^7-3p^6+6p^5+44p^4+6p^3-3p^2+7p+3}{p(3p^4+4p^3-p^2+4p+3)^{3/2}}, \tag{2.17}$$

and

$$\mathrm{Ku}(X) = \frac{\begin{array}{r}3p^{12}+10p^{11}+19p^{10}+72p^9+224p^8+206p^7+ \\ +21p^6+206p^5+224p^4+72p^3+19p^2+10p+3\end{array}}{[p(3p^4+4p^3-p^2+4p+3)]^2}. \tag{2.18}$$

The behavior of the mean, variance, skewness and kurtosis are displayed in Figures 3.

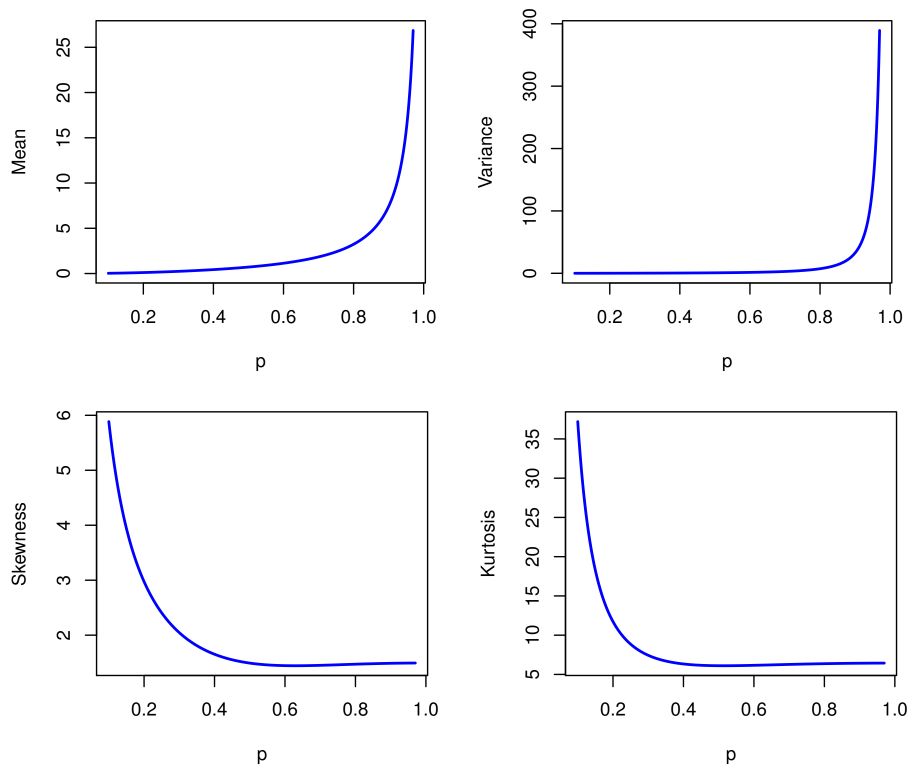


Figure 3: The mean, variance, skewness and kurtosis values of the DBL distribution.

According to results in Figure 3, the following observations are obtained:

1. The mean and variance increase as $p \to 1$;

2. The skewness and kurtosis decrease as $p \to 1$;

3. The proposed distribution is suitable model for the positively skewed count data sets;

4. The proposed distribution is leptokurtic since its kurtosis is always greater than 3.

2.3. Dispersion index and coefficient of variation

The dispersion index (DI) is calculated as variance to mean ratio. When DI is greater than 1, the distribution is over-dispersed, opposite case shows the under-dispersion. When DI is equal to 1, the distribution is equi-dispersed. The coefficient of variation (CV) is also very similar measure to DI. It is calculated as a ratio of the standard deviation to the mean. The DI and CV measures of the DBL distribution are given, respectively, by

$$\mathrm{DI}(X) = \frac{(3p^4 + 4p^3 - p^2 + 4p + 3)}{(p^2 + p + 1)(p^2 + p + 3)(1 - p^2)}, \tag{2.19}$$

$$\mathrm{CV}(X) = \frac{\sqrt{3p^4 + 4p^3 - p^2 + 4p + 3}}{p(p^2 + p + 3)}. \tag{2.20}$$

Figure 4 shows the DI and CV plots of the DBL distribution for various values of the model parameter. It is observed that DI can be either smaller or larger than one.

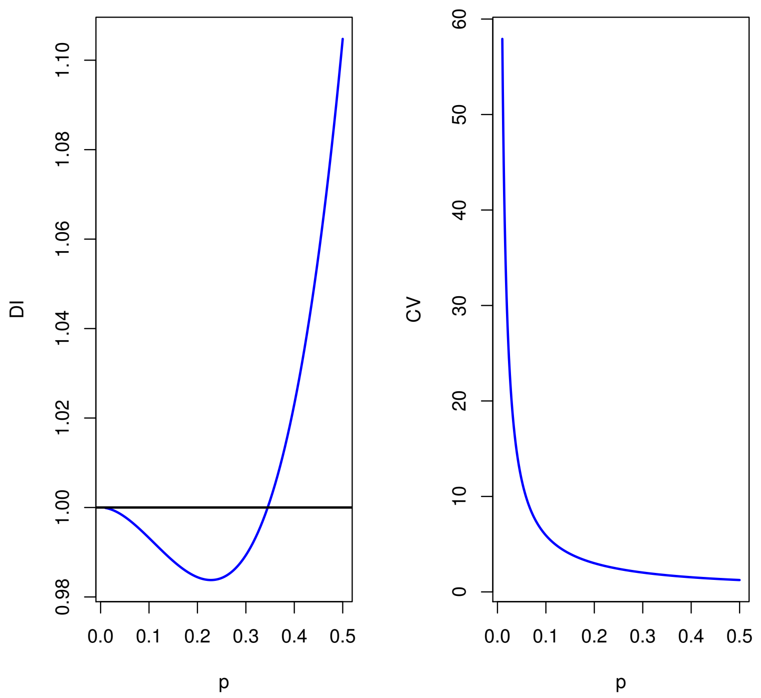


Figure 4: The DI and CV plots of the DBL distribution.

2.4. Mean deviation

The mean deviation (MD) about the mean measures the amount of scatter in a population. For a random variable X having a DBL distribution, the MD is defined as

$$\begin{aligned}
\mathrm{MD}(X) &= \sum_{x=0}^{\infty} |x-\mathrm{E}(X)|\, f(x;p) \\
&= \sum_{x=0}^{\mathrm{E}(X)} (\mathrm{E}(X)-x) f(x;p) \;+ \sum_{x=\mathrm{E}(X)+1}^{\infty} (x-\mathrm{E}(X)) f(x;p) \\
&= 2\mathrm{E}(X) F(\mathrm{E}(X);p) \;-\; 2\sum_{x=0}^{\mathrm{E}(X)} x f(x;p) \\
&= \frac{-2}{p^8+2p^7+p^6-2p^5-4p^4-2p^3+p^2+2p+1}
\left\{\begin{array}{l}
-6p^{\frac{p^4+p^3-6p^2-3p-3}{(p^2+p+1)(p^2-1)}} + 4p^{\frac{p^4+p^3-9p^2-4p-4}{(p^2+p+1)(p^2-1)}} \\
+2p^{\frac{2p^4+2p^3-9p^2-5p-5}{(p^2+p+1)(p^2-1)}} + 6p^{\frac{2(2p^4+2p^3-3p^2-3p-3)}{(p^2+p+1)(p^2-1)}} \\
-4p^{\frac{4p^4+4p^3-9p^2-7p-7}{(p^2+p+1)(p^2-1)}} + 6p^{\frac{5p^4+5p^3-6p^2-7p-7}{(p^2+p+1)(p^2-1)}} \\
-2p^{\frac{5p^4+5p^3-9p^2-8p-8}{(p^2+p+1)(p^2-1)}} + 3p^{\frac{2(3p^4+3p^3-3p^2-4p-4)}{(p^2+p+1)(p^2-1)}} \\
-6p^{\frac{2(p^4+p^3-3p^2-2p-2)}{(p^2+p+1)(p^2-1)}} - 2p^{\frac{3(p^4+p^3-3p^2-2p-2)}{(p^2+p+1)(p^2-1)}} \\
-3p^{\frac{-2(3p^2+p+1)}{(p^2+p+1)(p^2-1)}} + 2p^{\frac{-3(3p^2+p+1)}{(p^2+p+1)(p^2-1)}}
\end{array}\right\}.
\end{aligned}$$

The MD increases with $p \to 1$.

2.5. Stress-strength reliability

Stress-strength reliability (SSR) analysis is widely used in reliability engineering. Assume that both stress and strength are in the positive domain. Let $X_{\text{stress}} \sim \mathrm{DBL}(p)$ and $X_{\text{strength}} \sim \mathrm{DBL}(q)$. Then, the expected SSR can be expressed in a closed form as

$$\mathrm{SSR} := P\left[X_{\text{stress}} \le X_{\text{strength}}\right] = \sum_{x=0}^{\infty} f_{X_{\text{stress}}}(x;p) R_{X_{\text{strength}}}(x;q). \tag{2.21}$$

Using (2.5) and (2.6), we get

$$\mathrm{SSR} = \frac{4q^3(p^3-1)}{p^3q^3-1} + \frac{6q^2(1-p^3)}{p^3q^2-1} + \frac{6q^3(1-p^2)}{p^2q^3-1} + \frac{9q^2(p^2-1)}{p^2q^2-1}. \tag{2.22}$$

Figure 5 shows the SSR for various values of the parameters p and q. According to Figure 5, we concluded that:

- **(i)** The SSR increases for $q \to 1$ with fixed value of p;
- **(ii)** The SSR decreases for $p \to 1$ with fixed value of q.

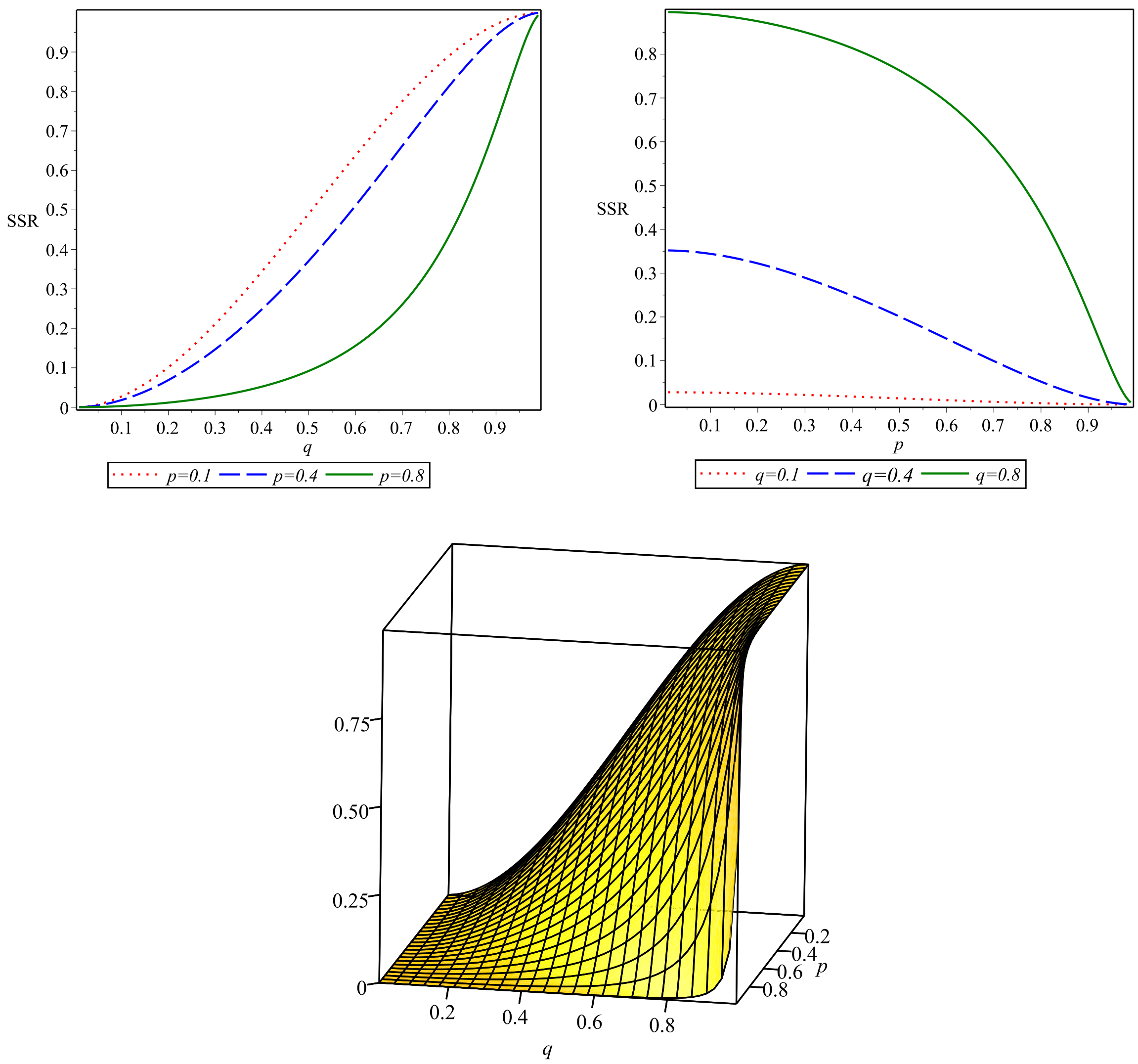


Figure 5: The SSR utilizing the DBL distribution.

2.6. Order statistics

Let $x_{1:n}, x_{2:n}, ..., x_{n:n}$ be the order statistics of a random sample from the DBL distribution. The cdf of i-th order statistics for an integer value of x is given by

$$
\begin{aligned}
F_{i:n}(x;p) &= \sum_{k=i}^{n} \binom{n}{k} \left[F_i(x;p)\right]^k \left[1 - F_i(x;p)\right]^{n-k} \\
&= \sum_{k=i}^{n} \sum_{j=0}^{n-k} \Upsilon_{(m)}^{(n,k)} \left[F_i(x;p)\right]^{k+j} \\
&= \sum_{k=i}^{n} \sum_{j=0}^{n-k} \Upsilon_{(m)}^{(n,k)} F_i(x;p,k+j), \qquad (2.23)
\end{aligned}
$$

where $\Upsilon_{(m)}^{(n,k)} := (-1)^j \binom{n}{k} \binom{n-k}{j}$ and $F_i(x;p,k+j) = \left[1 - \left(3 - 2p^{x+1}\right) p^{2(x+1)}\right]^{k+j}$ represents the cdf of the exponentiated DBL distribution with power parameter $k+j$.

The corresponding pmf to (2.23) is given by

$$
\begin{aligned}
f_{i:n}(x;p) &= F_{i:n}(x;p) - F_{i:n}(x-1;p) \\
&= \sum_{k=i}^{n}\sum_{j=0}^{n-k} \Upsilon_{(m)}^{(n,k)} f_i(x;p,k+j), \qquad (2.24)
\end{aligned}
$$

where $f_i(x;p,k+j)$ represents the pmf of the exponentiated DBL distribution with power parameter $k+j$. Thus, the b-th moments of $X_{i:n}$ can be written as

$$
\mathrm{E}(X_{i:n}^b) = \sum_{x=0}^{\infty}\sum_{k=i}^{n}\sum_{j=0}^{n-k} \Upsilon_{(m)}^{(n,k)} x^b f_i(x;p,k+j). \qquad (2.25)
$$

3. ESTIMATION METHODS

We use two estimation methods to estimate the unknown parameter of the DBL distribution. These methods are maximum likelihood estimation (MLE) and method of moments (MM).

3.1. Maximum likelihood estimation

Let $X_1, X_2, ..., X_n$ be random variables from the DBL distribution. The log-likelihood function (L) of the DBL distribution is

$$
L(x;p) = n\ln(p-1) + 2\ln p \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \ln\left[2p^{x_i}\left(p^2+p+1\right) - 3p - 3\right]. \qquad (3.1)
$$

By differentiating (3.1) with respect to the parameter p, we have the following equation:

$$
\frac{n}{p-1} + \frac{2}{p}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \frac{2p^{x_i}(2p+1) + 2x_i p^{x_i-1}\left(p^2+p+1\right) - 3}{2p^{x_i}\left(p^2+p+1\right) - 3p - 3} = 0. \qquad (3.2)
$$

The solution of the above equation gives MLE of the parameter p. However, it is not possible to obtain the exact form of the MLE of the parameter p since the equation has non-linear functions. For this reason, it has to be solved numerically. The other possible way to obtain the MLE of the parameter p is to direct minimization of the negative log-likelihood function. To do this, we use the `constrOptim` function of `R` software.

3.2. Moment estimation

The MM estimator of the parameter p is obtained by solving

$$
\frac{p^2(p^2+p+3)}{(p^2+p+1)(p^2-1)} - \bar{x} = 0, \qquad (3.3)
$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$. We use `nleqslv` to solve (3.3).

4. INAR(1) PROCESS WITH DBL INNOVATIONS

Time series of counts arise in different fields such as econometrics, actuarial and medical sciences. For instance, yearly incidents of terrorism, daily number of doctor visits, yearly number of traffic accidents and among others. McKenzie (1985) [20] and Al-Osh and Alzaid (1987) [1] introduced the INAR(1) process with Poisson innovations to analyze these kind of data sets. It is said that $\{X_t\}_{t\in\mathbb{Z}}$ follows a stable INAR(1) process if

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{4.1}$$

where $0 \leq \alpha < 1$. The innovation process, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, constitutes a sequence of the independent and identically distributed (iid) discrete random variables. The mean and variance of the innovation process are $E(\varepsilon_t) = \mu_\varepsilon$ and $\operatorname{Var}(\varepsilon_t) = \sigma_\varepsilon^2$, respectively. This model was shortly denoted as INAR(1)P process. Note that the innovations, $\{\varepsilon_t\}_{t\in\mathbb{Z}}$, are independent from X_{t-k}, $k \geq 1$. The binomial thinning operator, $\circ$, is defined by

$$\alpha \circ X_{t-1} := \sum_{j=1}^{X_{t-1}} W_j, \tag{4.2}$$

where $\{W_j\}_{j\geq 1}$ is a sequence of iid Bernoulli random variables with probabilities $\Pr(W_j = 1) = 1 - \Pr(W_j = 0) = \alpha$. The one-step transition probability of the INAR(1) process is

$$\Pr(X_t = k | X_{t-1} = l) = \sum_{i=1}^{\min(k,l)} \Pr(B_l^\alpha = i) \Pr(\varepsilon_t = k - i), k, l \geq 0, \tag{4.3}$$

where $B_n^\alpha \sim \text{Binomial}(\alpha, n)$ and $\alpha \in [0, 1)$. According to the works of Al-Osh and Alzaid (1987) [1] and McKenzie (1985) [20], we introduce a new INAR(1) model with a more flexible innovation distribution. We assume that the innovations follow a DBL distribution with parameter p. We call this process as INAR(1)DBL. Since the dispersion of the DBL can be under or over the value 1, the INAR(1)DBL can be used to model both under-dispersed and over-dispersed time series of counts. Using (4.3), the one-step transition probability of INAR(1)DBL process is given by

$$\begin{aligned}\gamma_{i,j} &= \Pr(X_t = k | X_{t-1} = l) \\ &= \sum_{i=1}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \left[2\left(p^3 - 1\right) p^{3(k-i)} - 3\left(p^2 - 1\right) p^{2(k-i)}\right].\end{aligned} \tag{4.4}$$

The equation in (4.4) represents the one-step transition probability of the process from state l to state k. The marginal probability function of the INAR(1)DBL process is

$$\begin{aligned}\gamma_j &= \Pr(X_t = k) \\ &= \sum_{l=0}^{\infty} \gamma_{ij} \Pr(X_{t-1} = l) \\ &= \sum_{l=0}^{\infty} \sum_{i=1}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \left[2\left(p^3 - 1\right) p^{3(k-i)} - 3\left(p^2 - 1\right) p^{2(k-i)}\right] \gamma_i,\end{aligned} \tag{4.5}$$

where $k = 0, 1, 2, \ldots$, (see Jazi *et al.*, 2012 [14]). Following the results given in Al-Osh and Alzaid (1987) [1], we obtain the mean, variance and DI of the INAR(1)DBL process and given, respectively, by

$$\mu_X = \frac{p^2\left(p^2+p+3\right)}{\left(p^2+p+1\right)\left(1-p^2\right)(1-\alpha)}, \tag{4.6}$$

$$\sigma_X^2 = \frac{\alpha}{\alpha^2-1}\left(\frac{3p^2\left(p^2-1\right)}{\left(p^2-1\right)^2} - \frac{2p^2\left(p^2-1\right)}{\left(p^3-1\right)^2}\right) - \frac{p^2\left(3p^4+4p^3-p^2+4p+3\right)}{\left(\alpha^2-1\right)\left(p^4+p^3-p-1\right)^2}, \tag{4.7}$$

$$\mathrm{DI}_X = \left(\alpha - \frac{3\,p^4+4\,p^3-p^2+4\,p+3}{p^6+2\,p^5+4\,p^4+2\,p^3-2\,p^2-4\,p-3}\right)(\alpha+1)^{-1}. \tag{4.8}$$

According to Al-Osh and Alzaid (1987) [1], the conditional expectation and variance of INAR(1)DBL process are given, respectively, by

$$\mathrm{E}\left(X_t \middle| X_{t-1}\right) = \alpha X_{t-1} + \frac{p^2(p^2+p+3)}{(p^2+p+1)(1-p^2)}, \tag{4.9}$$

$$\mathrm{Var}\left(X_t \middle| X_{t-1}\right) = \alpha(1-\alpha)X_{t-1} + \frac{p^2(3p^4+4p^3-p^2+4p+3)}{(p^2+p+1)^2(p^2-1)^2}. \tag{4.10}$$

4.1. Estimation of INAR(1)DBL process

Bourguignon *et al.* (2019) [5] and Lívio *et al.* (2018) [19] used three estimation methods to obtain the parameters of INAR(1) process defined under different innovation distributions. These methods are conditional least squares (CLS), Yule-Walker (YW) and the conditional maximum likelihood (CML) estimation methods. They compared the finite sample performance of these estimation methods for different sample sizes and parameter settings and concluded that CML estimation method provides better results than CLS and YW estimation methods. Here, we use these three estimation methods to obtain the unknown parameters of the INAR(1)DBL process. However, there are no explicit forms for the CLS and YW estimators of the INAR(1)DBL process because of the non-linearity of the equations.

Conditional maximum likelihood

The conditional log-likelihood function of the INAR(1)DBL process is

$$\begin{aligned}\ell(\boldsymbol{\Theta}) &= \sum_{t=2}^{T} \ln\left[\Pr\left(X_t = k \middle| X_{t-1} = l\right)\right] \\ &= \sum_{t=2}^{T} \ln\left[\begin{array}{l}\displaystyle\sum_{i=0}^{\min(x_t,x_{t-1})} \binom{x_{t-1}}{i} \alpha^i (1-\alpha)^{x_{t-1}-i} \\ \times\left\{2\left(p^3-1\right)p^{3(x_t-i)} - 3\left(p^2-1\right)p^{2(x_t-i)}\right\}\end{array}\right],\end{aligned} \tag{4.11}$$

where $\boldsymbol{\Theta} = (\alpha_{cml}, p_{cml})$ is the unknown parameter vector. The CML estimator of $\boldsymbol{\Theta}$, say $\hat{\boldsymbol{\Theta}}$ can be obtained by maximizing the equation (4.11). It is well-known that the maximization of (4.11) is equivalent to minimization of the negative of (4.11). Minimization of the negative of (4.11) could be done by using different software such as `R`, `MATLAB`, `C++` or `S-Plus`. Here, we prefer `constrOptim` function of `R` software to minimize the negative of (4.11). Note that the CML estimators are asymptotically normal and consistent under the regularity conditions (Bourguignon *et al.*, 2019 [5]).

Yule-Walker

The YW estimators are obtained by simultaneous solution of the equations for the theoretical and empirical moments of the INAR(1)DBL process. The autocorrelation function (ACF) of the INAR(1) process at lag h is $\rho_X(h) = \alpha^h$, and $\rho_X(1) = \alpha$ for $h = 1$. Therefore, the YW estimator of the parameter α is

$$\hat{\alpha}_{YW} = \frac{\sum_{t=2}^{T}\left(X_t - \bar{X}\right)\left(X_{t-1} - \bar{X}\right)}{\sum_{t=1}^{T}\left(X_t - \bar{X}\right)^2}. \tag{4.12}$$

The YW estimator of the parameter p, say $\hat{p}_{YW}$, can be obtained by solving

$$\frac{p^2\left(p^2 + p + 3\right)}{\left(p^2 + p + 1\right)\left(1 - p^2\right)\left(1 - \hat{\alpha}_{YW}\right)} = \bar{X}, \tag{4.13}$$

where $\bar{X} = \sum_{t=1}^{T} X_t \Big/ T$. However, it is not possible to obtain the explicit forms of the YW estimators of the parameter p. Therefore, (4.13) has to solved numerically by using the software such as `R` or `MATLAB`. We use the `uniroot` function of the `R` software to obtain $\hat{p}_{YW}$.

Conditional least squares

The CLS estimators of the parameters α and p can be obtained by minimizing

$$S(\boldsymbol{\eta}) = \sum_{t=2}^{T}\left(X_t - E\left(X_t | X_{t-1}\right)\right)^2, \tag{4.14}$$

where $\boldsymbol{\eta} = (\alpha_{cls}, p_{cls})$ and $E\left(X_t | X_{t-1}\right)$ is given in (4.9). Replacing $E\left(X_t | X_{t-1}\right)$ with (4.9) in (4.14), we have

$$S(\boldsymbol{\eta}) = \sum_{t=2}^{T}\left(X_t - \alpha X_{t-1} - \frac{p^2\left(p^2 + p + 3\right)}{\left(p^2 + p + 1\right)\left(1 - p^2\right)}\right)^2. \tag{4.15}$$

The derivatives of (4.15) with respect to the parameters α and p and equating them to zero, we have

$$\text{(4.16)} \qquad \frac{\partial S(\boldsymbol{\eta})}{\partial p} = \sum_{t=2}^{T} - \frac{12\,p\left(X_t - \alpha\,X_{t-1} + \frac{p^2\,(p^2+p+3)}{(p^2-1)\,(p^2+p+1)}\right)\left(p^4+p^3+p^2+p+1\right)}{\left(-p^4-p^3+p+1\right)^2} = 0,$$

$$\text{(4.17)} \qquad \frac{\partial S(\boldsymbol{\eta})}{\partial \alpha} = \sum_{t=2}^{T} -2\,X_{t-1}\left(X_t - \alpha\,X_{t-1} + \frac{p^2\,\left(p^2+p+3\right)}{\left(p^2-1\right)\,\left(p^2+p+1\right)}\right) = 0.$$

The simultaneous solutions of (4.16) and (4.17) give the CLS estimators of the parameter α and p. However, since the mean of the DBL distribution has non-linear functions, it is not possible to obtain the p_{cls} in explicit form. However, when the parameter p is known, the CLS estimator of the parameter α is

$$\text{(4.18)} \qquad \hat{\alpha}_{cls} = \sum_{t=2}^{T} \frac{(X_t+1)\left(p^4+p^3\right) - X_t\,(p+1) + 3p^2}{\left(p^4+p^3-p-1\right)X_{t-1}},$$

where p can be replaced with $\hat{p}_{cml}$ (see, Bourguignon *et al.*, 2019 [5]).

5. SIMULATION STUDIES

Here, two simulation studies are given to evaluate the parameter estimation performance of proposed models.

5.1. Simulation of DBL model

The finite-sample performances of the MLE and MM methods are compared for small and large sample sizes based on the simulation study. The below simulation steps are used for this goal:

1. Generate $N = 10{,}000$ samples of size $n = 20, 50, 100, 200$ and 500 from DBL(0.1), DBL(0.5) and DBL(0.7), respectively.
2. Using each generated sample, compute the MLE and MM estimator of the parameter p, say $\widehat{p}_j$ where $j = 1, 2, ..., 10{,}000$.
3. Compute the biases, mean-squared errors (MSEs) and mean relative errors (MREs) using the following equations:

$$\text{Bias}(p) = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{p}_j - p\right), \quad \text{MSE}(p) = \frac{1}{N}\sum_{j=1}^{N}\left(\widehat{p}_j - p\right)^2 \ \text{ and } \ \text{MRE} = \frac{1}{N}\sum_{j=1}^{N}\frac{\hat{p_j}}{p_j}.$$

The simulation results are reported in Table 1. The following remarks are obtained according to the results in Table 1:

1. The estimated biases always decrease and near the zero when $n \to \infty$.

2. The estimated MSEs decrease and near the zero when $n \to \infty$.

3. The estimated MREs are near the desired value, 1, especially for large sample sizes.

4. Both estimation methods work well for estimating the parameter p and produce similar results.

Similar results can be obtained for different values of the parameter p.

Table 1: The simulation results of DBL distribution.

Parameter	Sample size	Bias		MSE		MRE	
		MLE	MM	MLE	MM	MLE	MM
$p = 0.1$	20	−0.036585	−0.036506	0.006924	0.006926	0.634153	0.634937
	50	−0.016756	−0.016717	0.003237	0.003238	0.832445	0.832833
	100	−0.006808	−0.006800	0.001424	0.001424	0.931918	0.932002
	200	−0.002833	−0.002825	0.000523	0.000523	0.971665	0.971747
	500	−0.002338	−0.002341	0.000196	0.000196	0.976622	0.976590
$p = 0.5$	20	−0.008975	−0.008716	0.003892	0.003873	0.982050	0.982567
	50	−0.002803	−0.002843	0.001605	0.001610	0.994394	0.994314
	100	−0.001900	−0.001884	0.000682	0.000682	0.996201	0.996231
	200	−0.000803	−0.000765	0.000317	0.000317	0.998394	0.998470
	500	−0.000145	−0.000146	0.000150	0.000151	0.999101	0.999089
$p = 0.7$	20	−0.004901	−0.004959	0.001647	0.001647	0.992999	0.992915
	50	−0.001908	−0.001971	0.000700	0.000702	0.997275	0.997184
	100	−0.000833	−0.000854	0.000330	0.000329	0.998810	0.998780
	200	−0.000734	−0.000764	0.000170	0.000170	0.998952	0.998909
	500	−0.000856	−0.000859	0.000075	0.000075	0.998777	0.998773

5.2. Simulation of INAR(1)DBL process

We carry out a simulation study to evaluate the asymptotic behaviours of the CML, YW and CLS estimators of INAR(1)DBL process for small and sufficiently large sample sizes. The number of simulation replications is $N = 10,000$ and three sample sizes are used: $n = 25$, 50 and 100. Four parameter vectors are also used. These are ($\alpha = 0.3$, $p = 0.9$), ($\alpha = 0.5$, $p = 0.5$), ($\alpha = 0.2$, $p = 0.3$) and ($\alpha = 0.7$, $p = 0.6$). The biases, MSEs and MREs are used to evaluate the simulation results.

We expect that when the sample size is sufficiently large, the biases and MSEs near the zero and MREs are near the one. The simulation results are summarized in Table 2. As seen from the simulation results, the results of the CML and YW estimation methods are very near each other. However, the CML estimation method approaches to the desired values of the biases, MSEs and MREs more faster than those of the CLS and YW estimation methods.

The performance of the CML method is better than the CLS and YW estimation methods for both small and sufficiently large sample sizes. Therefore, we suggest to use the CML estimation to obtain the unknown parameters of the INAR(1)DBL process.

Table 2: Simulation results of INAR(1)DBL process.

Sample size	Parameters	CML			YW			CLS		
		Bias	MSE	MRE	Bias	MSE	MRE	Bias	MSE	MRE
$\alpha = 0.3,\ p = 0.9$										
$n = 25$	α	−0.0020	0.0092	0.9959	−0.1239	0.0469	0.7620	−0.1198	0.0494	0.7768
	p	−0.0055	0.0016	0.9931	0.0209	0.0028	1.0261	0.0187	0.0030	1.0234
$n = 50$	α	−0.0032	0.0042	0.9936	−0.0686	0.0195	0.8629	−0.0685	0.0203	0.8631
	p	−0.0010	0.0007	0.9987	0.0140	0.0014	1.0175	0.0132	0.0016	1.0165
$n = 100$	α	−0.0003	0.0023	0.9993	−0.0270	0.0088	0.9461	−0.0257	0.0090	0.9487
	p	−0.0016	0.0004	0.9980	0.0038	0.0008	1.0048	0.0035	0.0009	1.0043
$\alpha = 0.5,\ p = 0.5$										
$n = 25$	α	−0.0295	0.0257	0.9409	−0.1326	0.0533	0.7444	−0.1286	0.0579	0.7524
	p	0.0002	0.0049	1.0003	0.0356	0.0079	1.0713	0.0333	0.0087	1.0665
$n = 50$	α	−0.0122	0.0112	0.9756	−0.0623	0.0207	0.8754	−0.0616	0.0218	0.8768
	p	0.0014	0.0024	1.0029	0.0196	0.0035	1.0392	0.0193	0.0039	1.0386
$n = 100$	α	−0.0025	0.0054	0.9950	−0.0310	0.0095	0.9380	−0.0310	0.0100	0.9381
	p	−0.0010	0.0013	0.9979	0.0096	0.0020	1.0192	0.0098	0.0021	1.0197
$\alpha = 0.2,\ p = 0.3$										
$n = 25$	α	−0.0285	0.0304	0.9661	−0.0910	0.0513	0.9550	−0.0814	0.0599	1.0285
	p	−0.0076	0.0046	0.9924	0.0033	0.0047	1.0111	−0.0007	0.0064	1.0053
$n = 50$	α	−0.0276	0.0222	0.9762	−0.0502	0.0290	0.8860	−0.0493	0.0298	0.8980
	p	−0.0049	0.0023	0.9838	−0.0009	0.0023	0.9969	−0.0014	0.0024	0.9954
$n = 100$	α	−0.0141	0.0116	0.9896	−0.0206	0.0135	0.9221	−0.0198	0.0134	0.9253
	p	−0.0017	0.0011	0.9944	−0.0006	0.0012	0.9980	−0.0007	0.0011	0.9976
$\alpha = 0.7,\ p = 0.6$										
$n = 25$	α	−0.0134	0.0082	0.9808	−0.1689	0.0591	0.7590	−0.1657	0.0637	0.7649
	p	−0.0073	0.0047	0.9878	0.0667	0.0119	1.1112	0.0610	0.0141	1.1017
$n = 50$	α	−0.0058	0.0036	0.9917	−0.0855	0.0216	0.8779	−0.0856	0.0227	0.8777
	p	−0.0014	0.0021	0.9977	0.0401	0.0063	1.0669	0.0396	0.0069	1.0660
$n = 100$	α	−0.0051	0.0019	0.9928	−0.0433	0.0079	0.9381	−0.0434	0.0083	0.9380
	p	−0.0006	0.0011	0.9990	0.0208	0.0029	1.0346	0.0207	0.0031	1.0345

6. EMPIRICAL STUDIES

This section is devoted to illustrate the importance of the DBL distribution by analyzing the three real data sets with proposed and competitive models. The performance of fitted models are compared using goodness-of-fit criteria, Kolmogorov-Smirnov (K-S) test with its corresponding p-value.

6.1. Number of fires in Greece

The first data set deals with the number of fires in Greece for the period from 1 July 1998 to 31 August 1998. This data set was reported by Karlis and Xekalaki (2001) [16] and also is given in the Appendix. The performance of the DBL distribution is compared with competitive models listed in Table 3.

Table 3: The competitive models of the DBL distribution.

Distribution	Abbreviation	Author(s)
Geometric	Geo	—
Discrete Lindley	DLi	Gómez-Déniz and Calderín-Ojeda (2011) [12]
Discrete Rayleigh	DR	Roy (2004) [28]
Discrete inverse Rayleigh	DIR	Hussain and Ahmad (2014) [13]
Discrete Pareto	DPa	Krishna and Pundir (2009) [17]
Poisson	Poi	Poisson (1837) [27]
Discrete generalized exponential type II	DGE-II	Nekoukhou *et al.* (2013) [22]
Discrete Weibull	DW	Nakagawa and Osaki (1975) [21]
Discrete inverse Weibull	DIW	Jazi *et al.* (2010) [15]
Discrete Burr type II	DB-XII	Para and Jan (2016a) [24]
Exponentiated discrete Lindley	EDLi	El-morshedy *et al.* (2019) [10]
Discrete log-logistic	DLog-L	Para and Jan (2016b) [25]
Exponentiated discrete Weibull	EDW	Nekoukhou and Bidram (2015) [23]

Tables 4 and 5 contain the MLEs of the parameters for each fitted distribution with their standard errors (std-er). The asymptotic confidence intervals (CI) and the results of the goodness-of-fit test are also reported in these tables.

Table 4: The MLEs, CIs, K-S and p-values of fitted models with one-parameter for the number of fires in Greece.

Statistic		Model						
		DBL	Geo	DLi	DR	DIR	DPa	Poi
MLE_p		0.867	0.844	0.741	0.980	0.018	0.546	5.398
Std-er_p		0.008	0.013	0.014	0.023	0.007	0.029	0.209
95% CI	Lower_p	0.852	0.818	0.712	0.935	0.004	0.488	4.988
	Upper_p	0.883	0.869	0.769	1.00	0.033	0.605	5.809
K-S		0.096	0.164	0.097	0.183	0.429	0.355	0.854
p-value		0.202	0.003	0.198	< 0.001	0	< 0.001	0

According to Tables 4 and 5, two model provide the sufficient results for analyzing the number of fires in Greece since the p-values of these models are greater than 0.05. These are DBL and DLi distributions. However, DBL distribution has the smallest value of K-S statistic and the largest p-value among all competitive models as well as DLi distribution.

Table 5: The MLEs, CIs, K-S and p-values of fitted models with two and more parameters for the number of fires in Greece.

Statistic		Model						
		DGE-II	DW	DIW	DB-XII	EDLi	DLog-L	EDW
MLE_p		0.822	0.879	0.079	0.761	0.766	4.226	0.860
Std-er_p		0.019	0.023	0.022	0.043	0.021	0.389	0.099
95% CI	Lower_p	0.785	0.835	0.035	0.677	0.725	3.462	0.665
	Upper_p	0.859	0.924	0.123	0.845	0.808	4.989	1.055
MLE_α		1.255	1.131	1.035	2.503	0.797	1.717	1.081
Std-er_α		0.175	0.082	0.079	0.487	0.113	0.138	0.238
95% CI	Lower_α	0.912	0.969	0.881	1.548	0.575	1.446	0.615
	Upper_α	1.598	1.292	1.189	3.457	1.018	1.988	1.549
MLE_θ		—	—	—	—	—	—	1.092
Std-er_θ		—	—	—	—	—	—	0.448
95% CI	Lower_θ	—	—	—	—	—	—	0.214
	Upper_θ	—	—	—	—	—	—	1.969
K-S		0.130	0.123	0.208	0.299	0.124	0.149	0.125
p-value		0.031	0.047	< 0.001	< 0.001	0.046	0.009	0.042

Figures 6 and 7 show the estimated cdfs and probability-probability (PP) plots. These figures support the results reported in Tables 4 and 5.

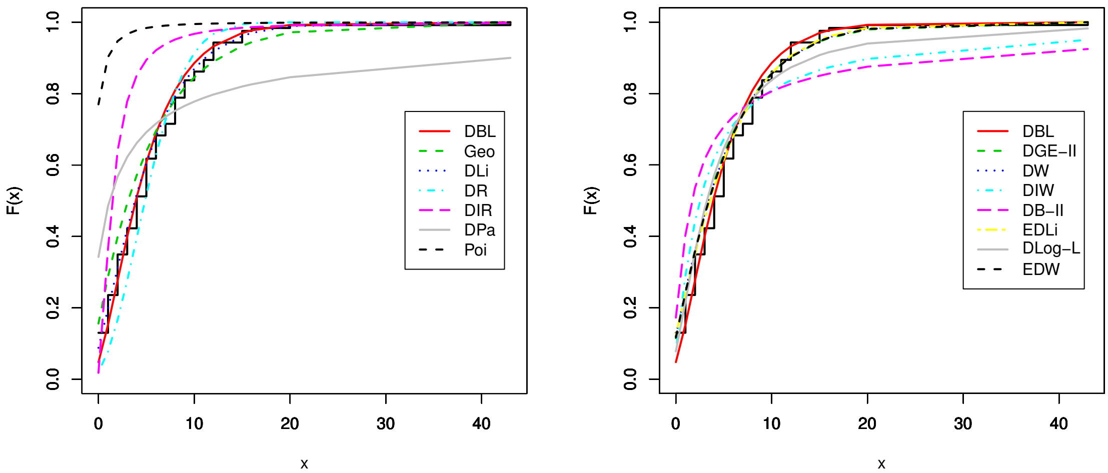


Figure 6: The estimated CDFs of fitted models.

Figure 8 shows the log-likelihood profile of $\widehat{p}$ where $L = -346.902$. It is found that the log-likelihood profile of $\widehat{p}$ is unimodal-shaped. Thus, this estimator is a unique and considered the best for the used data set.

Table 6 shows the results of MM method for the DBL parameter. It is clear that MM method works well for estimating the parameter p.

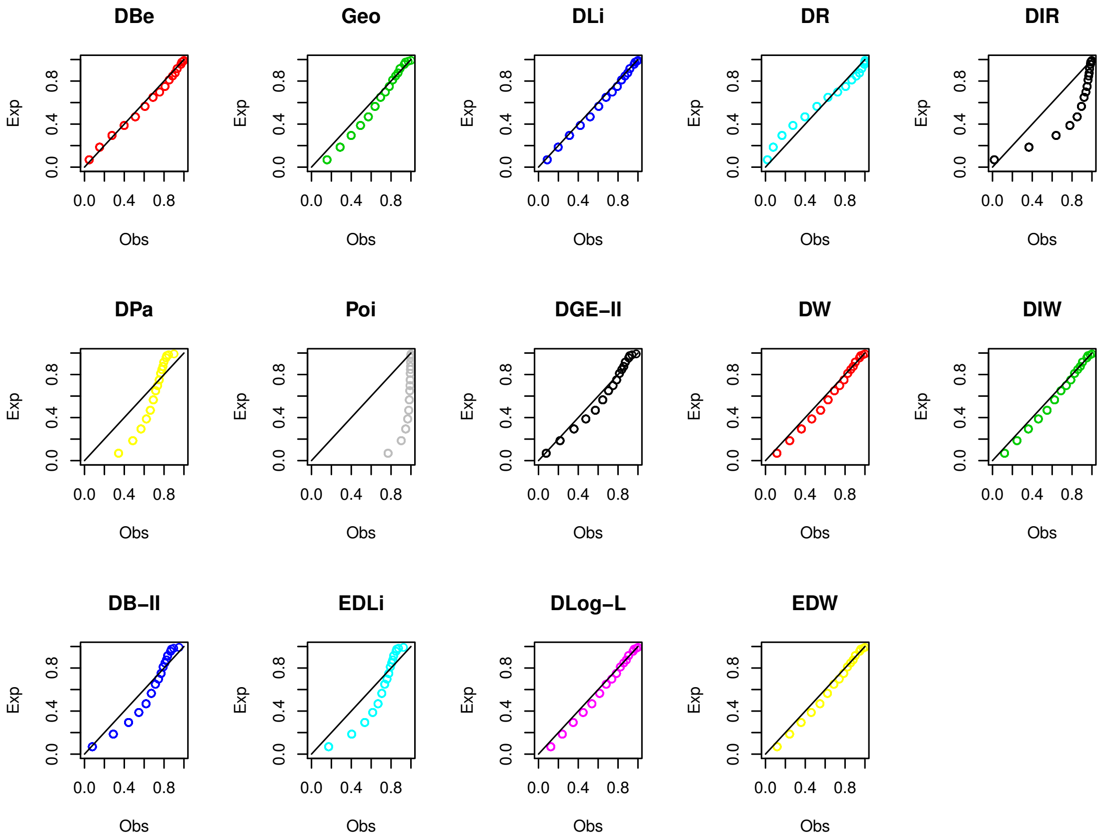


Figure 7: The PP plots of fitted models.

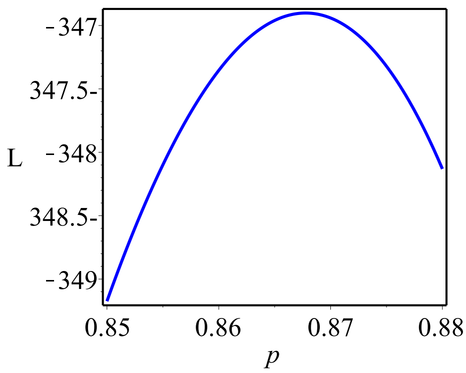


Figure 8: The log-likelihood profile of $\widehat{p}$ for the number of fires in Greece data set.

Table 6: The estimated parameter of DBL distribution with MM method.

Method	Measure		
	$\widehat{p}$	K-S	p-value
MM	0.868	0.095	0.220

Using the MM estimator of the parameter of p, the statistical properties of DBL distribution such as mean, mode, variance, DI, MD, CV, skewness and kurtosis values are listed in Table 7.

Table 7: The statistical properties of DBL distribution for the number of fires in Greece.

Method	Measure							
	Mean	Mode	Variance	DI	MD	CV	Skewness	Kurtosis
MM	5.3867	2.3936	18.1002	3.3601	3.2218	0.7897	1.4837	6.4127

6.2. Failure times

The data used represents the failure times for a sample of 15 electronic components in an acceleration life test (see Lawless, 2003 [18]). The performance of the DBL distribution is compared with discrete flexible model with one parameter (DFx-I), Geo, DR, DIR, DPa, DGE-II, DIW, DLog-L, DB-XII and discrete Lomax (DLo) distributions. The results of the fitted models with goodness-of-fit test are given in Tables 8 and 9.

Table 8: The MLEs, CIs, K-S and p-values of fitted models with one-parameter for the failure times data.

Statistic		Model					
		DBL	DFx-I	Geo	DR	DIR	DPa
MLE_p		0.971	0.973	0.965	0.999	1.8×10^{-7}	0.720
Std-er_p		0.005	0.006	0.009	2.58×10^{-4}	0.055	0.061
95% CI	Lower_p	0.960	0.961	0.948	0.998	0	0.600
	Upper_p	0.981	0.985	0.982	0.999	0.107	0.839
K-S		0.114	0.146	0.177	0.216	0.698	0.405
p-value		0.978	0.864	0.673	0.433	9.1×10^{-7}	0.009

Table 9: The MLEs, CIs, K-S and p-values of fitted models with two-parameters for the failure times data.

Statistic		Model				
		DGE-II	DIW	DLog-L	DB-XII	DLo
MLE_p		0.956	2.2×10^{-4}	21.463	0.975	0.012
Std-er_p		0.013	7.75×10^{-4}	5.387	0.051	0.039
95% CI	Lower_p	0.930	0	10.904	0.874	0
	Upper_p	0.981	0.001	32.021	1	0.088
MLE_α		1.491	0.875	1.791	13.367	104.506
Std-er_α		0.535	0.164	0.388	27.785	84.409
95% CI	Lower_α	0.441	0.554	1.031	0	0
	Upper_α	2.540	1.196	2.551	67.824	269.947
K-S		0.129	0.209	0.136	0.388	0.205
p-value		0.937	0.482	0.913	0.015	0.491

It is found that the DFx-I, Geo, DR, DGE-II, DIW, DLog-L and DLo distributions work quite well besides the DBL distribution. But the DBL distribution is the best among all tested models because it has the smallest value of K-S as well as it has the highest p-value. Figures 9 and 10 show the estimated cdfs and PP plots for the failure times data.

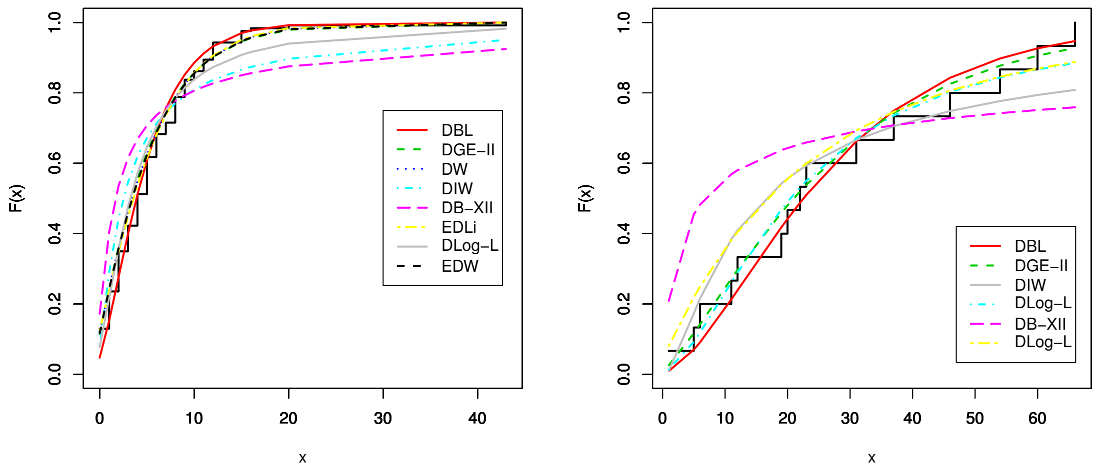


Figure 9: The estimated cdfs for the failure times data.

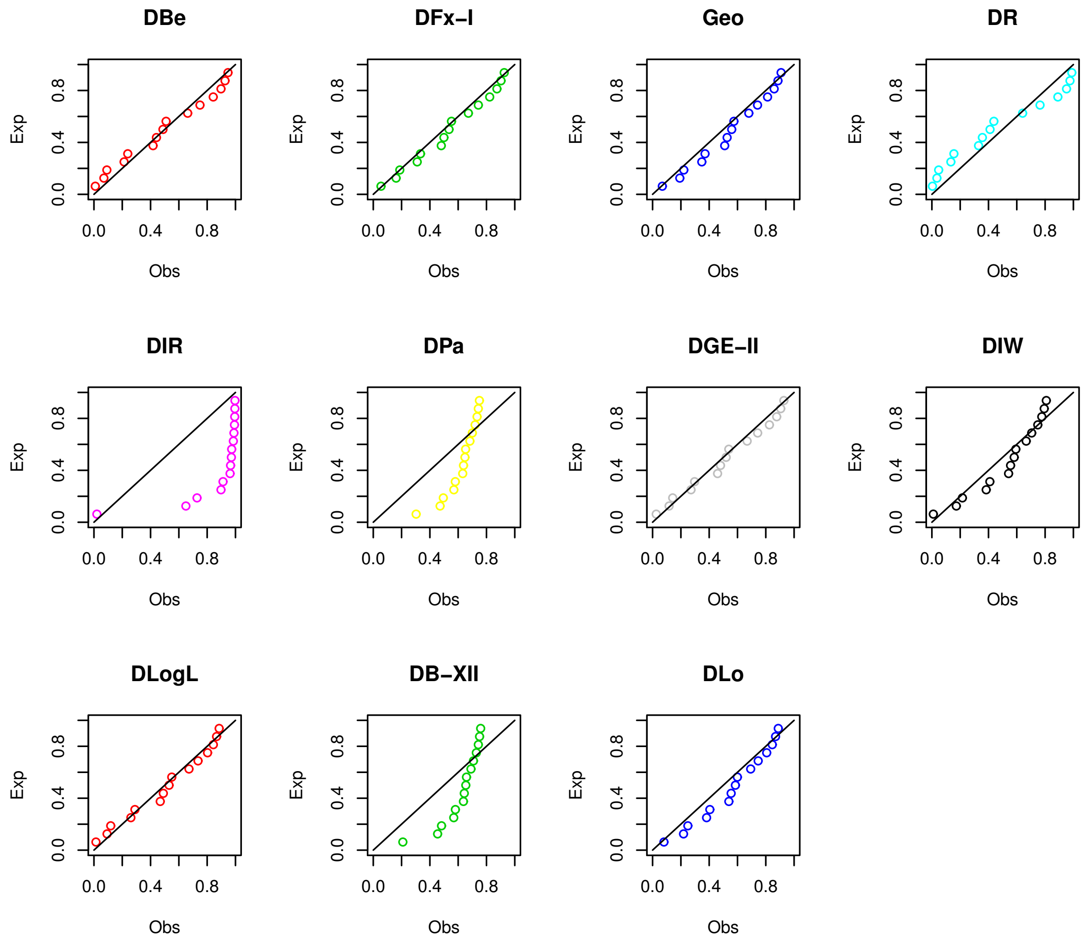


Figure 10: The PP plots for the failure times data.

It is clear that the DBL, DFx-I, Geo, DR, DGE-II, DIW, DLog-L and DLo distributions are suitable choices for this data set. However, the DBL distribution is the best choice since it has lowest value of the K-S test statistic. Figure 11 shows the TTT plot and log-likelihood profile of $\widehat{p}$, where $L = -64.784$.

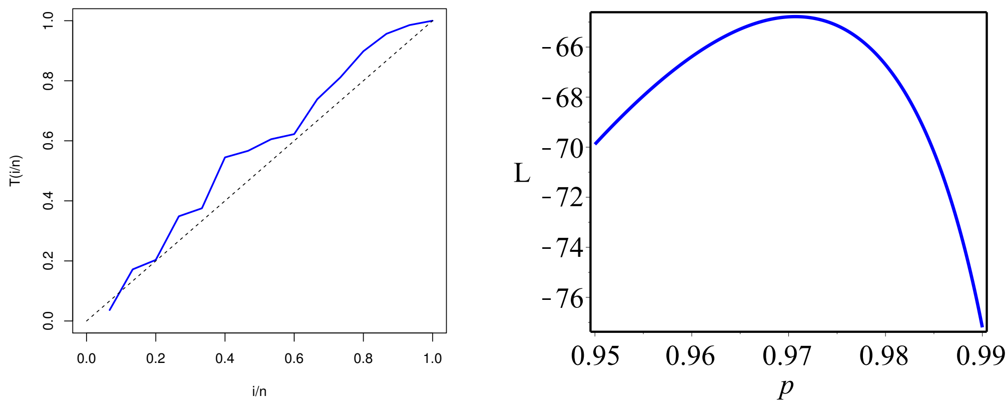


Figure 11: The TTT plot (left panel) and log-likelihood profile of $\widehat{p}$ (right panel) for the failure times data.

Regarding Figure 11, it is clear that the shape of the hrf can be increasing and the log-likelihood profile of $\widehat{p}$ is unimodal-shaped. Table 10 shows the estimation of the proposed model using the MM for the failure times data.

Table 10: Estimation and goodness of fit test for the failure times data.

Method	Statistic		
	p	K-S	p-value
MM	0.971	0.109	0.994

According to the p-value of the K-S test, MM method works quite well besides the MLE method for estimating the unknown parameter. But the MM is the best. Using the MM estimator of the parameter p, some statistics of the DBL distribution are reported in Table 11.

Table 11: Some descriptive statistics for data set II.

Method	Statistic							
	Mean	Mode	Variance	DI	MD	CV	Skewness	Kurtosis
MM	27.816	13.284	417.044	14.992	15.533	0.734	1.493	6.442

The data herein is suffering from over dispersion phenomena as DI > 1. Furthermore, it is moderately skewed right with leptokurtic.

6.3. Burglary crimes

The performance of the INAR(1)DBL process is compared with the INAR(1)P, INAR(1)PL and INAR(1)G processes. The one-step translation probabilities of the competitive INAR(1) models are given below:

1. INAR(1)P

$$\Pr\left(X_t = k \middle| X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \frac{\exp(-\lambda)\,\lambda^{k-i}}{(k-i)!}, \quad \lambda > 0.$$

2. INAR(1)PL

$$\Pr\left(X_t = k \middle| X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \frac{\theta^2\,(k-i+\theta+2)}{(\theta+1)^{k-i+3}}, \quad \theta > 0,$$

3. INAR(1)G

$$\Pr\left(X_t = k \middle| X_{t-1} = l\right) = \sum_{i=1}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \left[p(1-p)^{k-i}\right], \quad 0 < p < 1.$$

The CML estimation method is used to obtain unknown parameters of INAR(1)DBL, INAR(1)PL, INAR(1)G and INAR(1)P models. To decide the best model, two information criteria are used: Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC). The smallest values of AIC and BIC and indicate the best fitted model on the data set.

The series of monthly counts of burglary crimes in the 22th police car beat in Pittsburgh is used to compare the performance of INAR(1)DBL, INAR(1)PL, INAR(1)G and INAR(1)P processes. The data set consists of 144 monthly observations between the date of January 1990 and December 2001 and is given in the Appendix. The data set can be also found in `http://www.forecastingprinciples.com/index.php/crimedata`. The mean, variance and DI values of the used data set are 6.111, 13.372 and 2.188, respectively. It is clear that monthly counts of burglary crimes exhibit over-dispersion. So, the innovation distribution of INAR(1) process should be able to model over-dispersion. Therefore, INAR(1) process with DBL innovations could be a good choice to model these data set.

The autocorrelation function (ACF) and partial ACF plots of the used data set are displayed in Figure 12. As seen from these plots, ACF has clear cut-off after the first lag. Therefore, AR(1) process could be a good choice for analyzing these data set.

The estimated parameters of the fitted INAR(1) process and model selection criteria are listed in Table 12. Since the INAR(1)DBL model has the smaller values of AIC and BIC statistics than those of INAR(1)P, INAR(1)PL and INAR(1)G processes, the INAR(1)DBL process provides better fits than other competitive INAR(1) processes. More importantly, the obtained DI value of INAR(1)DBL process is very near the empirical one. It is obvious that INAR(1)DBL astoundingly explains the characteristics of the data set.

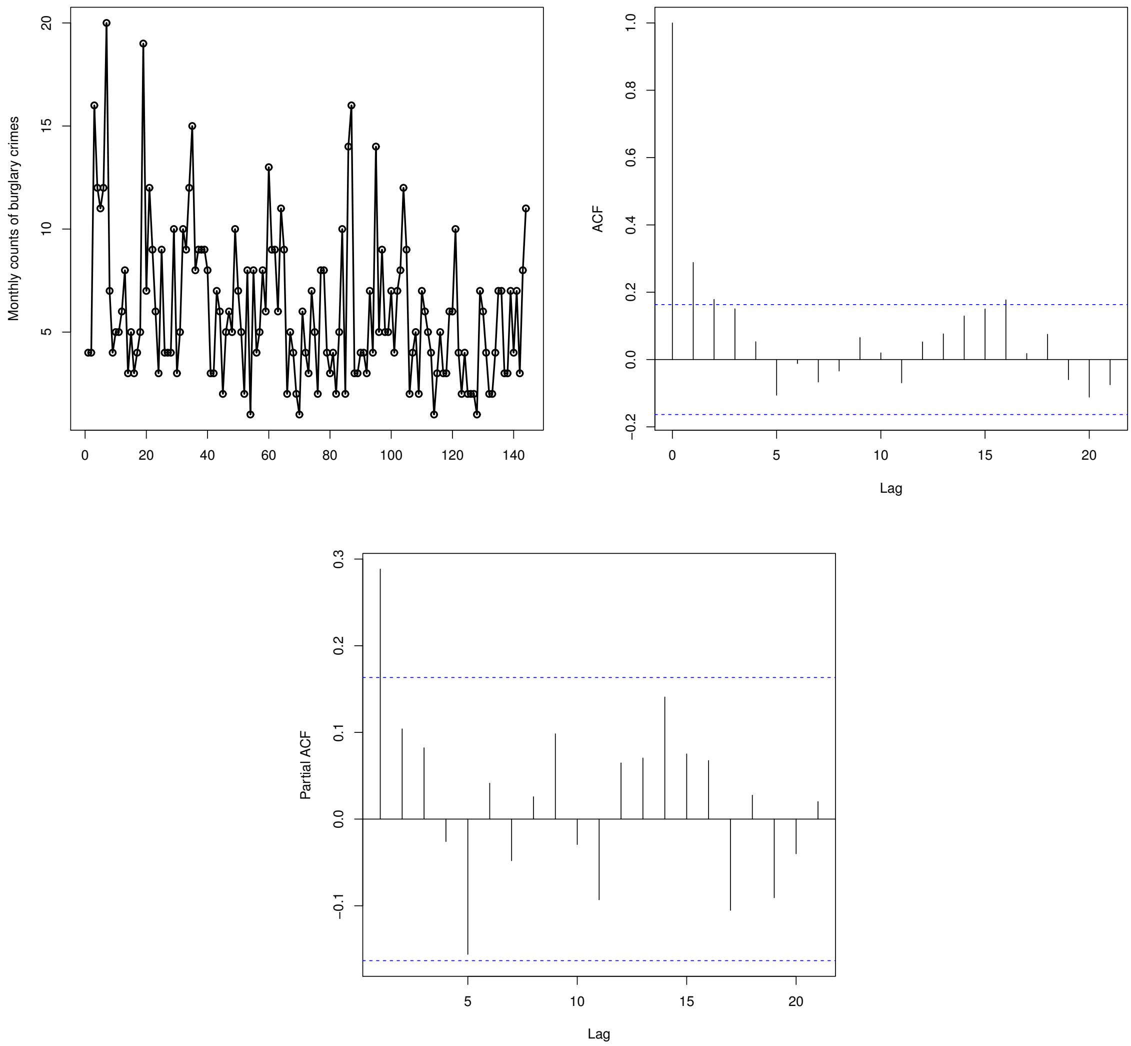


Figure 12: The plots of monthly counts of burglary crimes and its corresponding ACF and PACF plots.

Table 12: The CML estimates of INAR(1)DBL and INAR(1)P process and goodness-of-fit statistics.

Model	Parameters	Estimate	Std-er	AIC	BIC	μ_X	σ^2_X	DI
INAR(1)DBL	α p	0.3032 0.8402	0.0467 0.0121	733.1232	739.0628	6.1505	14.6336	2.3792
INAR(1)PL	α θ	0.3842 0.4451	0.0365 0.0147	739.8960	745.8356	6.1731	17.4559	2.8277
INAR(1)G	α p	0.4319 0.2221	0.0376 0.0192	747.7226	753.6622	6.1649	21.2445	3.4460
INAR(1)P	α λ	0.1952 4.9402	0.0194 0.0537	778.3730	784.3126	6.1381	6.1381	1
Empirical						6.1111	13.3722	2.1882

Additionally, the residual analysis is conducted to evaluate the accuracy of the fitted INAR(1)DBL model for the data used. The Pearson residuals of the INAR(1)DBL process are given by

$$r_t = \frac{X_t - \mathrm{E}\,(X_t|\,X_{t-1})}{\mathrm{Var}(X_t|\,X_{t-1})^{1/2}} \tag{6.1}$$

where $\mathrm{E}\,(X_t|\,X_{t-1})$ and $\mathrm{Var}\,(X_t|\,X_{t-1})$ are defined in (4.9) and (4.10), respectively. When the fitted INAR(1) process is valid for the modeled data, the Pearson residuals should have zero mean and unit variance as well as uncorrelated. The Pearson residuals of the INAR(1)DBL process are calculated by using (6.1). The mean and variance of these residuals are obtained as 0.0005 and 0.9917, respectively. The obtained values of the mean and variance of the Pearson residuals are very closed to the desired values. Moreover, the predicted values of the burglary crimes and the ACF plot of the Pearson residuals are displayed in Figure 13 which ensures that the residuals are uncorrelated.

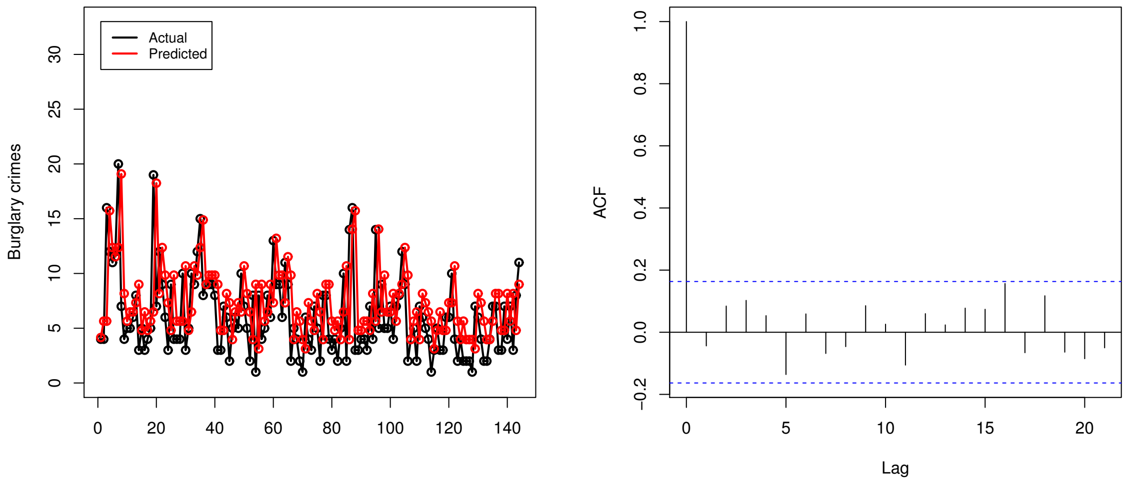


Figure 13: The predicted values of the burglary crimes (left) and the ACF plot of the Pearson residuals (right).

7. CONCLUSIONS

A new one-parameter discrete model is introduced. The statistical properties of proposed model are studied extensively. Two parameter estimation method are used. These are the maximum likelihood and method of moments estimation methods. The relative efficiency of parameter estimation methods are discussed via simulation study. Three applications to three real data sets are given to convince the readers in favour of DBL model. Empirical findings show that the DBL model is an attractive model and produce more reliable results than other its counterparts. More importantly, INAR(1) process with DBL innovations produce better results than INAR(1)P model in case of over-dispersion. We hope that DBL distribution gains much more attention and is applied to wider range of application fields.

A. APPENDIX

The data set used in Section 6.1:

Number of fires:	0	1	2	3	4	5	6	7	8	9	10	11	12	15	16	20	43
Observed values:	16	13	14	9	11	13	8	4	9	6	3	4	6	4	1	1	1

The data set used in Section 6.2:

1.0, 5.0, 6.0, 11.0, 12.0, 19.0, 20.0, 22.0, 23.0, 31.0, 37.0, 46.0, 54.0, 60.0, 66.0

The data set used in Section 6.3:

4	4	16	12	11	12	20	7	4	5	5	6	8	3	5	3	4	5	19	7
12	9	6	3	9	4	4	4	10	3	5	10	9	12	15	8	9	9	9	8
3	3	7	6	2	5	6	5	10	7	5	2	8	1	8	4	5	8	6	13
9	9	6	11	9	2	5	4	2	1	6	4	3	7	5	2	8	8	4	3
4	2	5	10	2	14	16	3	3	4	4	3	7	4	14	5	9	5	5	7
4	7	8	12	9	2	4	5	2	7	6	5	4	1	3	5	3	3	6	6
10	4	2	4	2	2	2	1	7	6	4	2	2	4	7	7	3	3	7	4
7	3	8	11																

ACKNOWLEDGMENTS

The author El-Morshedy would like to acknowledge that this publication was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia.

REFERENCES

[1] Al-Osh, M.A. and Alzaid, A.A. (1987). First-order integer-valued autoregressive (INAR(1)) process, *Journal of Time Series Analysis*, **8**(3), 261–275.

[2] Altun, E. (2020a). A new generalization of geometric distribution with properties and applications, *Communications in Statistics – Simulation and Computation*, **49**(3), 793–807.

[3] Altun, E. (2020b). A new one-parameter discrete distribution with associated regression and integer-valued autoregressive models, *Mathematica Slovaca*, **70**(4), 979–994.

[4] Abd-Elrahman, A.M. (2013). Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications, *Journal of Probability and Statistical Science*, **11**(2), 153–164.

[5] Bourguignon, M.; Rodrigues, J. and Santos-Neto, M. (2019). Extended Poisson INAR(1) processes with equidispersion, underdispersion and overdispersion, *Journal of Applied Statistics*, **46**, 101–118.

[6] Buddana, A. and Kozubowski, T.J. (2014). Discrete Pareto distributions, *Economic Quality Control*, **29**(2), 143–156.

[7] Eliwa, M.S. and El-Morshedy, M. (2020). A one-parameter discrete distribution for over-dispersed data: statistical and reliability properties with applications, *Journal of Applied Statistics*, forthcoming.

[8] Eliwa, M.S.; Alhussain, Z.A. and El-Morshedy, M. (2020a). Discrete Gompertz-G family of distributions for over-and under-dispersed data with properties, estimation, and applications, *Mathematics*, **8**(3), 358.

[9] Eliwa, M.S.; Altun, E.; El-Dawoody, M. and El-Morshedy, M. (2020b). A new three-parameter discrete distribution with associated INAR(1) process and applications, *IEEE Access*, **8**, 91150–91162.

[10] El-Morshedy, M.; Eliwa, M.S. and Nagy, H. (2019). A new two-parameter exponentiated discrete Lindley distribution: properties, estimation and applications, *Journal of Applied Statistics*, `https://doi.org/10.1080/02664763.2019.1638893`.

[11] El-Morshedy, M.; Eliwa, M.S. and Altun, E. (2020). Discrete Burr-Hatke distribution with properties, estimation methods and regression model, *IEEE Access*, **8**, 74359–74370.

[12] Gómez-Déniz, E. and Calderín-Ojeda, E. (2011). The discrete Lindley distribution: properties and applications, *Journal of Statistical Computation and Simulation*, **81**(11), 1405–1416.

[13] Hussain, T. and Ahmad, M. (2014). Discrete inverse Rayleigh distribution, *Pakistan Journal of Statistics*, **30**(2), 203–222.

[14] Jazi, A.M.; Jones, G. and Lai, C.D. (2012). Integer valued AR(1) with geometric innovations, *Journal of the Iranian Statistical Society*, **11**(2), 173–190.

[15] Jazi, A.M.; Lai, D.C. and Alamatsaz, H.M. (2010). Inverse Weibull distribution and estimation of its parameters, *Statistical Methodology*, **7**(2), 121–132.

[16] Karlis, D. and Xekalaki, E. (2001). *On some discrete valued time series models based on mixtures and thinning.* In "Proceedings of the Fifth Hellenic-European Conference on Computer Mathematics and its Applications", pp. 872–877.

[17] Krishna, H. and Pundir, P.S. (2009). Discrete Burr and discrete Pareto distributions, *Statistical Methodology*, **6**, 177–188.

[18] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

[19] L'ivio, T.; Khan, N.M.; Bourguignon, M. and Bakouch, H.S. (2018). An INAR(1) model with Poisson-Lindley innovations, *Economics Bulletin*, **38**(3), 1505–1513.

[20] McKenzie, E. (1985). Some simple models for discrete variate time series, *Journal of the American Water Resources Association*, **21**, 645–650.

[21] Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution, *IEEE Transactions on Reliability*, **24**(5), 300–301.

[22] Nekoukhou, N.; Alamatsaz, M.H. and Bidram, H. (2013). Discrete generalized exponential distribution of a second type, *Statistics*, **47**(4), 876–887.

[23] Nekoukhou, V. and Bidram, H. (2015). The exponentiated discrete Weibull distribution, *SORT*, **39**(1), 127–146.

[24] Para, B.A. and Jan, T.R. (2016a). On discrete three-parameter Burr type XII and discrete Lomax distributions and their applications to model count data from medical science, *Biometrics and Biostatistics International Journal*, **4**(2), 1–15.

[25] Para, B.A. and Jan, T.R. (2016b). Discrete version of log-logistic distribution and its applications in genetics, *International Journal of Modern Mathematical Sciences*, **14**(4), 407–422.

[26] Para, B.A. and Jan, T.R. (2017). Discrete generalized Weibull distribution: properties and applications in medical sciences, *Pakistan Journal of Statistics*, **33**, 337–354.

[27] Poisson, S.D. (1837). *Recherches sur la Probabilité des Jugements en Matiére Criminelle et en Matiére Civile, Précédées des Régles Générales du Calcul des Probabilitiés*, Paris, France: Bachelier, pp. 206–207.

[28] Roy, D. (2004). Discrete Rayleigh distribution, *IEEE Transactions on Reliability*, **53**(2), 255–260.

REVSTAT-Statistical journal

Aims and Scope

The aim of REVSTAT-Statistical Journal is to publish articles of high scientific content, developing Statistical Science focused on innovative theory, methods, and applications in different areas of knowledge. Important survey/review contributing to Probability and Statistics advancement is also welcome.

Background

Statistics Portugal started in 1996 the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, a quarterly publication whose goal was the publication of papers containing original research results, and application studies, namely in the economic, social and demographic fields. Statistics Portugal was aware of how vital statistical culture is in understanding most phenomena in the present-day world, and of its responsibilities in disseminating statistical knowledge.

In 1998 it was decided to publish papers in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work. At the time, the editorial board was mainly composed by Portuguese university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal. In 2001, the *Revista de Estatística* published a three volumes special issue containing extended abstracts of the invited and contributed papers presented at the 23rd European Meeting of Statisticians (EMS). During the EMS 2001, its editor-in-chief invited several international participants to join the editorial staff.

In 2003 the name changed to REVSTAT-Statistical Journal, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

Editorial policy

REVSTAT-Statistical Journal is an open access peer-reviewed journal published quarterly, in English, by Statistics Portugal.

The editorial policy of REVSTAT is mainly placed on the originality and importance of the research. The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage revstat.ine.pt based in Open Journal System (OJS). The only working language allowed is English. Authors intending to submit any work must register, login and follow the guidelines.

There are no fees for publishing accepted manuscripts that will be made available in open access.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external. Authors can suggest an editor or reviewer who is expert on the paper subject providing her/his complete information, namely: name, affiliation, email and, if possible, personal URL or ORCID number.

All published works are Open Access (CC BY 4.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Also, in the context of archiving policy, REVSTAT is a *blue* journal welcoming authors to deposit their works in other scientific repositories regarding the use of the published edition and providing its source.

Journal prints may be ordered at expenses of the author(s), and prior to publication.

Abstract and Indexing services

REVSTAT-Statistical Journal is covered *by Journal Citation Reports - JCR (Clarivate); Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; Scimago Journal & Country Rank; Scopus*

Author guidelines

The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage https://revstat.ine.pt/ based in Open Journal System (OJS). Authors intending to submit any work must ***register, login*** and follow the indications choosing ***Submissions***.

REVSTAT - Statistical Journal adopts the COPE guidelines on publication ethics.

Work presentation

- the only working language is English;
- the first page should include the name, ORCID iD (optional), Institution, country, and mail-address of the author(s);
- a summary of fewer than one hundred words, followed by a maximum of six keywords and the MSC 2020 subject classification should be included also in the first page;
- manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and a maximum of 25 pages;
- the title should be with no more than 120 characters (with spaces);
- figures must be a minimum of 300dpi and will be reproduced online as in the original work, however, authors should take into account that the printed version is always in black and grey tones;
- authors are encouraged to submit articles using LaTeX which macros are available at *REVSTAT style*;
- citations in text should be included in the text by name and year in parentheses, as in the following examples: § article title in lowercase (Author 1980); § This theorem was proved later by AuthorB and AuthorC (1990); § This

subject has been widely addressed (AuthorA 1990; AuthorB et al. 1995; AuthorA and AuthorB 1998).

- references should be listed in alphabetical order of the author's scientific surname at the end of the article;
- acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text;
- authors are welcome to suggest one of the Editors or Associate Editors or yet other reviewer expert on the subject providing a complete information, namely: name, affiliation, email and personal URL or ORCID number in the Comments for the Editor (submission form).

Accepted papers

After final revision and acceptance of an article for publication, authors are requested to provide the corresponding LaTeX file, as in REVSTAT style.

Supplementary files may be included and submitted separately in .tiff, .gif, .jpg, .png .eps, .ps or .pdf format. These supplementary files may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Copyright Notice

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, in order to ensure the widest possible dissemination of information.

According to REVSTAT's *archiving policy*, after assigning the copyright form, authors may cite and use limited excerpts (figures, tables, etc.) of their works accepted/published in REVSTAT in other publications and may deposit only the published edition in scientific repositories providing its source as REVSTAT while the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.