Journal of Statistical Distributions and Applications

RESEARCH Open Access

A new trivariate model for stochastic episodes

Francesco Zuniga, Tomasz J. Kozubowski and Anna K. Panorska*

*Correspondence: ania@unr.edu
Department of Mathematics & Statistics, University of Nevada, Reno 89557, NV, USA

Abstract

We study the joint distribution of stochastic events described by (X, Y, N), where N has a 1-inflated (or deflated) geometric distribution and X, Y are the sum and the maximum of N exponential random variables. Models with similar structure have been used in several areas of applications, including actuarial science, finance, and weather and climate, where such events naturally arise. We provide basic properties of this class of multivariate distributions of mixed type, and discuss their applications. Our results include marginal and conditional distributions, joint integral transforms, moments and related parameters, stochastic representations, estimation and testing. An example from finance illustrates the modeling potential of this new model.

Keywords: BEG model, BGGE distribution, BTLG distribution, Extremes, Financial data, Geometric distribution, Maximum likelihood estimation, Random sum, Stochastic representation, TETLEG model, Zero-altered distribution, Zero-inflated distribution

Classification codes: 60E05; 60G50; 60G70; 62E15; 62H05; 62H12; 62H15; 62P05; 62P15

1 Introduction

This paper introduces a new model for stochastic events such as growth/decline periods of a financial return, flood, drought, or a heat wave, among others. The model describes the *duration* N, the *magnitude* X and the *peak value* Y of such events through the random structure

$$(X, Y, N) \stackrel{d}{=} \left(\sum_{i=1}^{N} X_i, \bigvee_{i=1}^{N} X_i, N \right), \tag{1}$$

where the $\{X_i\}_{i\geq 1}$ are independent and identically distributed (IID) exponential random variables given by the probability density function (PDF)

$$f(x) = \beta e^{-\beta x}, \ \ x \in \mathbb{R}_+, \ \ \beta > 0, \tag{2}$$

N is an integer-valued random variable, independent of the $\{X_i\}$, and $\bigvee_{i=1}^{N} X_i$ denotes the maximum of $\{X_i\}_{i=1,\ldots,N}$. Events like this arise in many applications. A common process generating such observations is Peaks-over-Threshold process where we are interested in observations exceeding (or below) a threshold. For example, in finance, time periods with positive/negative log-returns give growth/decline periods for an asset. In climate/hydrology, a flood may be described as stage of a stream exceeding a levy, heat wave may be described as consecutive days when the maximum daily temperature exceeds a high threshold, deluge can be thought of as consecutive observations (daily, hourly, etc.) of precipitation exceeding a high threshold. In energy research, heating degree days are those days when the maximum daily temperature is below 64 degrees Fahrenheit (the threshold temperature varies by country and region). Drought may be considered to be a time period when the maximum precipitation (daily, annual, etc.) is below a seasonal low for a given region. Various bivariate and trivariate models for stochastic episodes/events with the structure of (1) were developed and applied in different fields in Arendarczyk et al. (2018a, b), Barreto-Souza (2012), Barreto-Souza and Silva (2019), Biondi et al. (2002, 2005, 2008), Kozubowski and Panorska (2005, 2008), and Kozubowski et al. (2008a, b, 2010, 2011). Most of the existing models assume that the underlying observations $\{X_i\}$ are IID exponential or (dependent) Pareto variables, thus cover the cases of light and heavy tailed processes. In the aforementioned work the duration N was modeled with geometric distribution.

Our interest in extending the models developed earlier was motivated by the observation that the duration of (exceedingly) many events in processes such as financial asset returns is one time period. While geometric distribution allows for values of one, in many processes the events of duration 1 (above/below threshold) are either more or less common than in the geometric model. For example, heat waves are often a "hot day" (so 1-inflated), financial returns often switch daily from positive to negative (1 inflated), large precipitation is often lasting one day (1 -inflated). Summarizing, while the geometric durations work well for some applications, they do not work well for every applications. Thus, we introduced the generalization allowing 1-inflation or 1-deflation.

Thus, we developed an extension of the models with geometric duration, to allow more flexibility when accounting for the frequency of data with the duration of one. Models for count process, such as duration, are typically discrete, positive or non-negative, integer valued random variables such as geometric, negative binomial or Poisson. There are many works in the literature (medical, ecology, social science, actuarial science, etc.) describing counts data with a very large number of zeros. The models used to account for the "excess" zeros fall to two general types: zero inflated (ZI) or hurdle (H) models. We discuss and provide examples of applications for these models in Section 2. We also provide a representation for the ZI and H models via waiting times for the first success in independent Bernoulli trials with different probabilities of success.

This paper is organized as follows. Section 2 is devoted to the discussion of ZI nad H models and introduces our model for 1-inflated geometric distribution. Section 3 introduces our trivariate model and presents its basic properties. Section 4 provides information about marginal and conditional distributions for the trivariate vector. Section 5 is devoted to estimation and testing connected with the new model. An illustrative data example is given in Section 6. Selective proofs and auxiliary results are collected in the Supplementary Material.

2 Mixture models for duration

In this section we briefly discuss zero inflated and hurdle models and introduce our (shifted hurdle) model for duration. As noted in the introduction, the two common ways of dealing with extensive zeros in the literature are *zero -inflated* (ZI) (or zero-adjusted, zero-altered) and *hurdle* (H) models (see, e.g., Cameron and Trivedi 1998, 2005; Lambert 1992; Mullahy 1986; 1997; Panicha 2018; Zuur et al. 2009; Alshkaki 2016; and references therein). These two approaches to account for large number of zeros involve mixture distributions with two components, but they differ in the way that zeros can occur. The models are mixtures of a point mass at zero and a counting distribution. In the ZI models, zero can occur as an outcome of the point mass variable or the counting variable. On the other hand, in H models zero can only occur as an outcome of the point mass while the counting variable is truncated at zero.

Examples of zero-inflated or hurdle models used in the literature include applications in econometrics (see, e.g., Cameron and Trivedi 1998, 2005; Zeileis et al. 2008), ecology (Panicha 2018; Zuur et al. 2009), public health, epidemiology and bioinformatics (Hu et al. 2011; Zelterman 2004; Chipeta et al. 2014). There are also interesting applications in social science, criminology and actuarial science (Aryal 2011; Constantinescu et al. 2019; Iwunor 1995; Pandey and Tiwari 2011; Sharma and Landge 2013; Tüzen and Erbaş 2018). In ecology, the use of ZI nad H models is connected with estimation population sizes using various capture-recapture type methods. In public health and epidemiology, these models are used to estimate the number of sick with a given disease, in bioinformatics ZI and H models serve for estimation of the size of the population of drug users, and in criminology and social science to estimate the size of rural-urban migration, the size of homeless populations or violators of a certain law, or the number of highway crashes (see, e.g., Famoye and Singh 2006; Iwunor 1995; Pandey and Tiwari 2011; Sharma and Landge 2013). In actuarial science, zero-inflated discrete and dependent by mixture Pareto distribution was used in modeling probability of ruin in the compound binomial risk model in Constantinescu et al. (2019).

We now turn to the definitions of the ZI and H models. We start with the notation used in the rest of this paper: the set of non-negative integers (including zero) shall be denoted by $\mathbb{N}_0$, while $\mathbb{N}$ shall stand for the set of natural numbers (excluding zero).

2.1 The zero-inflated model

The ZI model is a mixture of point mass at zero and a counting random variable N. In practice, the latter is often chosen to follow a standard discrete distribution such as Poisson, geometric or negative binomial (see, e.g., Mullahy 1986; Lambert 1992). It is important to note that in the ZI model, the zeros may come from two different sources: the point mass or the count variable. The probability mass function (PMF) f_{ZI} of a zero-inflated random variable N_{ZI}, derived from a "base" discrete random variable N with the PMF f, is of the form

$$f_{ZI}(n) = qI_{\{0\}}(n) + (1-q)f(n), \quad n \in \mathbb{N}_0, \ 0 \leq q \leq 1,$$

where I_A is the indicator function of the set A.

Remark 1 *The corresponding mixture representation, connecting the relevant random variables, is as follows:*

$$N_{ZI} \overset{d}{=} JN,$$

where J is a Bernoulli random variable with parameter $1-q$, independent of N.

2.2 The hurdle model

The hurdle model is also a mixture, where the components are a point mass at zero and a counting "base" random variable N with the PMF f. However, the base random variable is *truncated* below at zero before mixing. Due to the truncation, the PMF f of N is converted to f_T, where the latter is the PMF of the conditional distribution of N given that $N \geq 1$,

$$f_T(n) = \frac{f(n)}{1-f(0)}, \quad n \in \mathbb{N}.$$

Mixing this distribution with a point mass at zero leads to the hurdle distribution based on N, with the PMF of the form

$$f_H(n) = qI_{\{0\}}(n) + (1-q)f_T(n), \quad n \in \mathbb{N}_0. \tag{3}$$

Remark 2 *Similarly to the ZI case, a random variable N_H with the PMF (3) admits the mixture representation of the form*

$$N_H \stackrel{d}{=} JN_T,$$

where J is a Bernoulli random variable with parameter $1-q$, independent of N_T. Note that in the hurdle model the value of zero can only come from the Bernoulli trail J.

Both the IZ and H models have appeared in the literature. The practical convenience of the hurdle model comes from the ease of estimation procedures compared with those for the IZ model. Since in this work the count random variable N represents the duration of an event, our N is always greater than or equal to one. Further, our data may show an unusual frequency of ones (not zeros). Thus, we use a hurdle-type model (shifted up by one) for the duration N. We discuss it in more details below.

2.3 A hurdle-type geometric distribution

We start with the definition of geometric random variable we use in this work: a random variable with the PMF

$$f(n) = p(1-p)^{n-1}, \quad n \in \mathbb{N}, \ 0 \leq p \leq 1, \tag{4}$$

will be referred to as a geometric random variable with parameter (probability of success) p, and denoted by $N_p \sim \mathcal{GEO}(p)$. Note that this variable "starts" at one, and accounts for *the number of trials* until the first success in a series of IID Bernoulli trials with parameter p. Using this variable, we can define a hurdle-type model with the PMF of the form

$$f(n) = \begin{cases} q & \text{for } n = 0 \\ (1-q)p(1-p)^{n-1} & \text{for } n \in \mathbb{N}. \end{cases} \tag{5}$$

Next we define our counting variable N that would represent the duration in the trivariate model (1). Namely, this will be the distribution given by (5) shifted up by one, with the PMF of the form

$$f(n) = \begin{cases} q & \text{for } n = 1 \\ (1-q)p(1-p)^{n-2} & \text{for } n \in \{2, 3, \ldots, \}. \end{cases} \tag{6}$$

We shall denote this distribution by $\mathcal{HGEO}(p, q)$, which stands for **h**urdle - **g**eometric distribution, and write $N \sim \mathcal{HGEO}(p, q)$ when the random variable N follows this distribution. Note, that depending on the relation between p and q, the $\mathcal{HGEO}(p, q)$ model may over-inflate the number of ones ($p < q$) or under-inflate the number of ones ($p > q$)

compared to geometric distribution with probability of success p. The following result provides a useful stochastic representation of this distribution.

Proposition 1 *If* $N \sim \mathcal{HGEO}(p, q)$ *then*

$$N \overset{d}{=} 1 + IN_p, \tag{7}$$

where N_p *is geometric with the PMF (4) and* I *is Bernoulli with parameter* $1 - q$*, independent of* N_p*.*

It is easy to see that when $p = q$ then the HGEO model (7) and its PMF (6) reduce to geometric distribution with parameter p, and the PMF given by (4). It is interesting to compare the hurdle-type HGEO model above with one analogous to zero-inflation, and also built upon the geometric distribution. The PMF of the latter will be of the form

$$g(n) = \begin{cases} q + (1-q)p & \text{for } \ n = 1 \\ (1-q)p(1-p)^{n-1} & \text{for } \ n \in \{2, 3, \ldots\}. \end{cases} \tag{8}$$

Remark 3 *Both of the models introduced above have interpretations as waiting times for the first success in a sequence of independent Bernoulli trials* $\{I_j\}$*. Namely, if the probabilities of success are given by* $\mathbb{P}(I_1 = 1) = q$ *and* $\mathbb{P}(I_j = 1) = p$ *for* $j \geq 2$ *then the number of trials till the first success will have the HGEO distribution given by the PMF (6). On the other hand, if the probabilities of success are the same as above for* $n \geq 2$ *while for* $n = 1$ *we have* $\mathbb{P}(I_1 = 1) = q + (1-q)p$*, then the corresponding waiting time will have a distribution with the PMF (8). Because of this, it is clear that the first model is more flexible than the second: for the hurdle type model, we have* $\mathbb{P}(I_1 = 0) = 1 - q$*, which can fall anywhere in the unit interval, while an analogous probability for the second model is equal to* $(1-p)(1-q)$*, which does not cover the entire unit interval as* q *changes in* $(0, 1)$*.*

3 Definition and basic properties of the new trivariate model

In this section we formally define the new distribution of (1) and derive its basic properties. Here, and elsewhere in the paper, the notation $\mathcal{EXP}(\beta)$ stands for the exponential distribution with the PDF (2).

Definition 1 *The random vector* (X, Y, N) *with the stochastic representation given in (1), where the* $\{X_i\}$ *are IID exponential random variables with the PDF (2) and* $N \sim \mathcal{HGEO}(p, q)$ *with the PMF (6), independent of the* $\{X_i\}$*, is said to have a generalized TETLG (GT) distribution, denoted by* $\mathcal{GT}(p, q, \beta)$*.*

We note that when $p = q$, then the variable N has a geometric distribution with parameter p, and the random vector (X, Y, N) above has the TETLG distribution studied in Kozubowski et al. (2011), where the name stands for **T**rivariate distribution with **E**xponential, **T**runcated **L**ogistic and **G**eometric marginal distributions. Our construction provides a flexible generalization of the TETLG model that accounts for the excess of ones in the data.

We now derive basic characteristics of the GT model, starting with its PDF. For this, we use the bivariate distribution of $\left(\sum_{i=1}^{n} X_i, \bigvee_{i=1}^{n} X_i\right)$, developed in Qeadan et al. (2012), which is the conditional distribution of (X, Y) given $N = n$ in the GT model. This distribution, referred to as the BGGE model in Qeadan et al. (2012), has the PDF of the form

$$f(x, y|n) = \beta^n e^{-\beta x} H(x, y, n), \tag{9}$$

where, for all $n \in \mathbb{N}$,

$$H(x,y,n) = \begin{cases} \sum_{s=1}^{k} \frac{n(n-1)}{(s-1)!(n-s)!}(x-sy)^{n-2}(-1)^{s+1} & \text{for } n \geq 2,\ (x,y) \in S_k,\ k = 1, \ldots, n-1, \\ 1 & \text{for } n = 1,\ (x,y) \in S_0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

The values of the joint PDF in (9) depend on which of the *sectors* S_k the input (x, y) belongs to, where

$$S_0 = \{(x, y) : x = y > 0\},$$

and

$$S_k = \{(x,y) \in \mathbb{R}^2 : 0 \leq \frac{1}{k+1}x \leq y < \frac{1}{k}x\}, \quad k = 1, 2, \ldots, n-1.$$

The support of this distribution is the set

$$A_n = \cup_{k=0}^{n-1} S_k = \{(x, y) : \frac{x}{n} \leq y \leq x\}, \quad n \in \mathbb{N}.$$

The joint distribution of $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ can now be derived via a standard hierarchical approach, where $N \sim \mathcal{HGEO}(p, q)$ with the PMF f given by (6) and $(X, Y)|N = n$ has the PDF $f(x, y|n)$ given by (9), so that $f(x, y, n)$ of (X, Y, N) is $f(x, y, n) = f(x, y|n)f(n)$. This leads to the following result.

Proposition 2 *The PDF of* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *is given by*

$$f(x, y, n) = \begin{cases} q\beta e^{-\beta x} & \text{for } n = 1 \\ \beta^n e^{-\beta x} H(x, y, n)(1-q)p(1-p)^{n-2} & \text{for } n > 1, \end{cases} \tag{11}$$

where the function H is defined in (10).

Remark 4 *We note that the support of the GT distribution is the same as that of its special case of TETLG distribution, which consists of the set* $\{(x, y, n) : n \in \mathbb{N}, (x, y) \in A_n\}$. *In analogy with the TETLG model, if* $n = 1$ *and* $(x, y) \in S_0$, *so that the point* (x, y, n) *is in the set* $A_1 \times \{1\}$, *the joint PDF reduces to*

$$f(x, y, n) = q\beta e^{-\beta x} I_{A_1 \times \{1\}}(x, y, n) = qf_1(x, y, n),$$

where $f_1(x, y, n) = \beta e^{-\beta x} I_{A_1 \times \{1\}}(x, y, n)$. *In other words, with probability q, the distribution is concentrated on the set* $A_1 \times \{1\}$, *and represents the random vector* $(E_0, E_0, 1)$, *where* $E_0 \sim \mathcal{EXP}(\beta)$. *However, when* $n \geq 2$, *the conditional distribution of* (X, Y) *given* $N = n$ *is absolutely continuous with the PDF* $f(x, y|n)$ *given in (9), in which case the joint PDF in (11) becomes*

$$f(x, y, n) = (1-q)f(x, y|n)p(1-p)^{n-2} = (1-q)f_2(x, y, n), \quad n \geq 2, (x, y) \in A_n,$$

where $f_2(x, y, n) = f(x, y|n)p(1-p)^{n-2}$. *Altogether, we have*

$$f(x, y, n) = qf_1(x, y, n) + (1-q)f_2(x, y, n), \tag{12}$$

showing that the GT distribution is a mixture of a degenerate distribution of $(E_0, E_0, 1)$ *and a proper, trivariate distribution (with the PDF* f_2*), where the mixing probabilities correspond to the events* $N = 1$ *and* $N \geq 2$*, respectively. This absolutely continuous component is the same as its analogue in the TETLG model.*

The mixture representation (12) of the GT distribution can also be seen from the stochastic representation of this model, stated below.

Proposition 3 *If* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *then*

$$(X, Y, N) \overset{d}{=} \sum_{i=1}^{N} \left(E_i, \frac{E_i}{i}, 1 \right), \tag{13}$$

where $N \sim \mathcal{HGEO}(p, q)$ *and the* $\{E_i\}$ *are independent* $\mathcal{EXP}(\beta)$ *variables, independent of* N*.*

Next, we provide an alternative stochastic representation, involving a geometric variable N_p rather than the mixed geometric variable N, which is the part of the random vector (X, Y, N). Both of these representations are useful for deriving further properties of the GT distribution.

Proposition 4 *If* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *then*

$$(X, Y, N) \overset{d}{=} (E_0, E_0, 1) + I \left(\sum_{i=1}^{N_p} E_i, \sum_{i=1}^{N_p} \frac{E_i}{i+1}, N_p \right), \tag{14}$$

where all the variables on the right-hand-side of (14) are mutually independent, $N_p \sim \mathcal{GEO}(p)$*,* I *is a Bernoulli random variable with parameter* $(1 - q)$*, and the* $\{E_i\}$ *are independent* $\mathcal{EXP}(\beta)$ *random variables.*

Remark 5 *We note that in the special case* $p = q$*, both of the above stochastic representations result in the TETLG distribution studied in Kozubowski et. al. (2011), describing the random vector*

$$(\tilde{X}, \tilde{Y}, \tilde{N}) \overset{d}{=} \left(\sum_{i=1}^{N_p} E_i, \bigvee_{i=1}^{N_p} E_i, N_p \right),$$

where N_p *and the* $\{E_i\}$ *are as above. Further, it can be shown that in general the random vector* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *can be directly related to a TETLG random vector* $(\tilde{X}, \tilde{Y}, \tilde{N})$ *viz. another stochastic representation, which involves the operation* $\oplus_j$ *defined below. This operation acts component-wise on two vectors,* $\mathbf{x} = (x_1, \ldots, x_n)$ *and* $\mathbf{y} = (y_1, \ldots, y_n)$*, and returns another vector, denoted by* $\mathbf{x} \oplus_j \mathbf{y}$*. The latter is obtained by adding the corresponding coordinates of* $\mathbf{x}$ *and* $\mathbf{y}$*, with the exception of the j-th coordinate, where we take* $x_j \vee y_j$ *(the maximum of* x_j *and* y_j*). Thus, we have*

$$\mathbf{x} \oplus_j \mathbf{y} = (x_1 + y_1, \ldots, x_{j-1} + y_{j-1}, x_j \vee y_j, x_{j+1} + y_{j+1}, \ldots, x_n + y_n), \quad j \in \{1, 2, \ldots, n\}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n. \tag{15}$$

With this notation we have

$$(X, Y, N) \overset{d}{=} (E_0, E_0, 1) \oplus_2 I(\tilde{X}, \tilde{Y}, \tilde{N}), \tag{16}$$

where E_0 and I are as above and the operation $\oplus_2$ acts as the maximum of the second coordinates and the sum of the first and the third coordinates, so that

$$(x_1, x_2, x_3) \oplus_2 (y_1, y_2, y_3) = (x_1 + y_1, x_2 \vee y_2, x_3 + y_3).$$

Using the above representations, we obtain the characteristic function (ChF) of the GT model, presented below.

Proposition 5 *The characteristic function of $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ is given by*

$$\begin{aligned}\phi(t_1, t_2, t_3) &= \mathbb{E}e^{t_1X+t_2Y+t_3N}\\ &= \frac{q\beta e^{it_3}}{\beta - i(t_1 + t_2)} + \sum_{n=2}^{\infty} e^{it_3 n} \prod_{j=1}^{n} \frac{(1-q)\beta p(1-p)^{n-2}}{\beta - i(t_1 + t_2/j)}, \quad t_1, t_2, t_3 \in \mathbb{R}.\end{aligned} \tag{17}$$

When $p = q$, Eq. (17) yields the ChF of the TETLG model of Kozubowski et al. (2011). The joint moments of the GT model are presented below.

Proposition 6 *If $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ then*

$$\begin{aligned}\mu_{k,l,m} &= \mathbb{E}\left\{X^k Y^l N^m\right\}\\ &= \frac{(k+l)!}{\beta^{k+l}} q + \frac{1}{\beta^{k+l}} \sum_{n=2}^{\infty} n^m (1-q) p (1-p)^{n-2}\\ &\quad \sum_{k_1,\ldots,k_n} \sum_{l_1,\ldots,l_n} \binom{k}{k_1 \cdots k_n} \binom{l}{l_1 \cdots l_n} \prod_{j=1}^{n} \frac{(l_j + k_j)!}{j^{l_j}},\end{aligned} \tag{18}$$

where the double summation is taken over all sets of non-negative integers $k_1, \ldots, k_n$ and $l_1, \ldots, l_n$ that add up to k and l, respectively.

One can use the above result to obtain the mean vector and the covariance matrix of the GT distribution, which are presented below.

Proposition 7 *If $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$, then*

$$\mathbb{E}X = \frac{1}{\beta}\left[1 + \frac{1-q}{p}\right], \quad \mathbb{E}Y = \frac{1}{\beta}\left[1 + \frac{(1-q)[p - 1 - \log(p)]}{(1-p)^2}\right], \quad \mathbb{E}N_p = 1 + \frac{1-q}{p},$$

and the elements of the covariance matrix $\mathbf{\Sigma} = [\sigma_{i,j}]$ are as follows:

$$\begin{aligned}
\sigma_{1,1} &= \mathbf{V}ar(X) = \frac{1 + p^2 - q^2}{\beta^2 p^2},\\
\sigma_{1,2} &= \mathbf{C}ov(X, Y) = \frac{p(1-p)^2 + (1-p)(1-q)^2 + (1-q)(p-q)\log(p)}{\beta^2 p(1-p)^2},\\
\sigma_{1,3} &= \mathbf{C}ov(X, N) = \frac{(1-q)(2-p) - (1-q)^2}{\beta p^2},\\
\sigma_{2,2} &= \mathbf{V}ar(Y) = \frac{2(q-p)}{(1-p)\beta^2} + \frac{(1-q)pc_p}{(1-p)^3\beta^2} - \left[\frac{q-p}{\beta(1-p)} - \frac{(1-q)\log(p)}{\beta(1-p)^2}\right]^2,\\
\sigma_{2,3} &= \mathbf{C}ov(Y, N) = \frac{(1-q)[(2p-q)\log(p) + (1-p)(p-q+1)]}{\beta(1-p)^2 p},\\
\sigma_{3,3} &= \mathbf{V}ar(N) = \frac{(1-q)[1-p+q]}{p^2},
\end{aligned}$$

where the constant c_p in the expression for $\sigma_{2,2}$ is given by

$$c_p = \int_0^\infty \frac{u^2 e^{-u} du}{\left[e^{-u} + p/(1-p)\right]^2}. \tag{19}$$

4 Marginal and conditional distributions

In this section we present the marginal and (selected) conditional distributions of the new trivariate GT distribution.

4.1 Bivariate margins

Here we discuss the three bivariate marginal distributions of (X, N), (Y, N), and (X, Y), starting with the joint distribution of (X, N).

4.1.1 The marginal distribution of X, N

In view of the stochastic representation in (1), the joint PDF of (X, N) can be derived through a standard conditioning argument using the fact that $N \sim \mathcal{HGEO}(p, q)$ and, given $N = n$, the variable X is the sum of n IID exponential variables with parameter $\beta > 0$. Thus, $X|N = n$ has a gamma distribution $\mathcal{GAM}(n, \beta)$, with the PDF given by:

$$f_{X|N=n}(x) = \frac{\beta^n}{(n-1)!} x^{n-1} e^{-\beta x}, \quad x \in \mathbb{R}_+, \ \beta > 0. \tag{20}$$

We obtain the joint PDF of (X, N) by multiplying the conditional PDF (20) by the marginal PMF of N given by (6), leading to the following result.

Proposition 8 *If $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ then the joint PDF of (X, N) is given by*

$$f_{X,N}(x, n) = \begin{cases} q\beta e^{-\beta x} & \textit{for} \ \ n = 1 \\ \frac{\beta^n}{\Gamma(n)} x^{n-1} e^{-\beta x}(1-q)p(1-p)^{n-2} & \textit{for} \ \ n \geq 2. \end{cases} \tag{21}$$

Remark 6 *Clearly, in the special case $p = q$ we obtain the bivariate BEG distribution (see Kozubowski and Panorska 2005), describing the random vector*

$$(\tilde{X}, \tilde{N}) \stackrel{d}{=} \left(\sum_{i=1}^{N_p} E_i, N_p \right),$$

where $N_p \sim \mathcal{GEO}(p)$ and the $\{E_i\}$ are IID $\mathcal{EXP}(\beta)$. Further, it can be seen from Proposition 4 or the relation (16) that in general the random vector (X, N) with the PDF (21) can be related to a BEG random vector $(\tilde{X}, \tilde{N})$ viz.

$$(X, N) \stackrel{d}{=} (E_0, 1) + I(\tilde{X}, \tilde{N}), \tag{22}$$

where $E_0 \sim \mathcal{EXP}(\beta)$, I is Bernoulli with parameter $1 - q$, and all the variables on the right-hand-side of (22) are mutually independent.

4.1.2 The marginal distribution of Y, N

We now turn to the joint distribution of (Y, N). Here, given $N = n$, the variable Y is the maximum of n IID exponential random variables. Thus, it has a *generalized exponential* distribution with the PDF

$$f_{Y|N=n}(y) = n\beta e^{-\beta y}(1 - e^{-\beta y})^{n-1}, \ \ y \in \mathbb{R}_+. \tag{23}$$

By proceeding as above, we obtain the joint PDF of (Y, N) as follows.

Proposition 9 *If* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *then the joint PDF of* (Y, N) *is given by*

$$f_{Y,N}(y, n) = \begin{cases} q\beta e^{-\beta y} & \text{for } n = 1, \\ n\beta e^{-\beta y}(1 - e^{-\beta y})^{n-1}(1 - q)p(1 - p)^{n-2} & \text{for } n \geq 2. \end{cases} \tag{24}$$

Remark 7 *We note that in the special case* $p = q$ *we obtain the bivariate BTLG distribution (see Kozubowski and Panorska 2008), describing the random vector*

$$(\tilde{Y}, \tilde{N}) \stackrel{d}{=} \left(\bigvee_{i=1}^{N_p} E_i, N_p \right),$$

where N_p *and the* $\{E_i\}$ *are as above. Further, it can be seen from the relation (16) that in general the random vector* (Y, N) *with the PDF (24) can be related to a BTLG random vector* $(\tilde{Y}, \tilde{N})$ *viz.*

$$(Y, N) \stackrel{d}{=} (E_0, 1) \oplus_1 I(\tilde{Y}, \tilde{N}), \tag{25}$$

where E_0 *and* I *are as above and the operation* $\oplus_j$ *is given by (15), so the* $\oplus_1$ *above acts as the maximum of the first coordinates and the sum of the second coordinates,*

$$(x_1, x_2) \oplus_1 (y_1, y_2) = (x_1 \vee y_1, x_2 + y_2).$$

4.1.3 The marginal distribution of *X*, *Y*

Finally, we turn to the last of the three bivariate distributions, the distribution of (X, Y). Its PDF can be obtained in a standard way by adding up the trivariate PDF of the GT model (given in Theorem 2) across all the values of $n \in \mathbb{N}$. The support of this new bivariate distribution is the set

$$A = \cup_{k=0}^{\infty} S_k = \cup_{k=1}^{\infty} A_k = \{(x, y) : 0 < y \leq x\}.$$

Lengthy algebra produces the result below, which can be proven in the same way as Theorem 3.1 in Kozubowski et al. (2011).

Proposition 10 *If* $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ *then the marginal PDF of* (X, Y) *is given by*

$$f_{X,Y}(x, y) = qg_1(x, y) + (1 - q)g_2(x, y), \tag{26}$$

where $g_1(x, y) = \beta e^{-\beta x} I_{A_1}(x, y)$ *and*

$$g_2(x, y) = p\beta^2 e^{-\beta x} \sum_{s=1}^{k} \frac{(-1)^{s+1}}{(s-1)!} W_s(\beta[\,1 - p\,]\,[\,x - sy\,]\,), \quad (x, y) \in S_k,\ k \in \mathbb{N}, \tag{27}$$

with

$$W_s(u) = e^u(s(s-1)u^{s-2} + 2su^{s-1} + u^s) - \sum_{i=0}^{k-s} \frac{(i+s)(i+s-1)u^{i+s-2}}{i!}, \quad s \in \mathbb{N}, u > 0. \tag{28}$$

Remark 8 *The result shows that the distribution of* (X, Y) *is a mixture of a degenerate distribution of the vector* (E_0, E_0) *with* $E_0 \sim \mathcal{EXP}(\beta)$*, which is a singular part of the distribution (corresponding to the event* $N = 1$*, which occurs with probability* q*) and an absolutely continuous component with the PDF* g_2 *given by (27), supported on the set* A *(corresponding to the event* $N \geq 2$*, which occurs with probability* $1 - q$*). Similar interpretation applies to the special case of the BETL distribution discussed in Kozubowski et al. (), obtained here when* $p = q$*. In that case the above proposition yields the PDF of*

$$(\tilde{X}, \tilde{Y}) \stackrel{d}{=} \left(\sum_{i=1}^{N_p} E_i, \bigvee_{i=1}^{N_p} E_i \right),$$

where N_p and the $\{E_i\}$ are as above. Further, it can be seen from the relation (16) that in general the random vector (X, Y) with the PDF (26) can be related to a BETL random vector $(\tilde{X}, \tilde{Y})$ viz.

$$(X, Y) \stackrel{d}{=} (E_0, E_0) \oplus_2 I(\tilde{X}, \tilde{Y}), \tag{29}$$

where E_0 and I are as above and the operation $\oplus_j$ is given by (15), so that

$$(x_1, x_2) \oplus_2 (y_1, y_2) = (x_1 + y_1, x_2 \vee y_2).$$

4.2 Univariate margins

We now discuss the univariate margins. Since N has a hurdle-type generalized geometric distribution given by (6), we shall focus on the marginal distributions of X and Y, starting with X.

4.2.1 The marginal distribution of X

The PDF of X can be calculated in a straightforward way by summing up the joint PDF of X and N given by (21) across all the values of $n \in \mathbb{N}$, leading to the result below.

Proposition 11 *If $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ then the PDF of X is*

$$f_X(x) = \left(1 - \frac{1-q}{1-p}\right) \beta e^{-\beta x} + \left(\frac{1-q}{1-p}\right) p\beta e^{-p\beta x}, \quad x \in \mathbb{R}_+, \tag{30}$$

with the corresponding CDF of the form

$$F_X(x) = \begin{cases} 0 & \textit{for } x < 0 \\ \left(1 - \frac{1-q}{1-p}\right)\left(1 - e^{-\beta x}\right) + \frac{1-q}{1-p}\left(1 - e^{-p\beta x}\right) & \textit{for } x \geq 0. \end{cases}$$

Remark 9 *The results shows that X is a generalized mixture of two exponential distributions, one with parameter β and another with parameter $p\beta$. The term "generalized" signifies the fact that the two weights in (30), although add up to one, are not necessarily restricted to the unit interval. It is worth noting that the distribution of X is also a proper mixture of two distributions, of which one is again exponential with parameter β while the other has a hypoexponential distribution, also known as generalized Erlang distribution (see, e.g., Johnson et al. 1994), given by the PDF*

$$g(x) = \frac{1}{1-p} p\beta e^{-p\beta x} - \frac{p}{1-p} \beta e^{-\beta x}, \quad x \in \mathbb{R}_+. \tag{31}$$

The above is the PDF of $X_1 + X_2$, where X_1 is exponential with parameter $p\beta$ and X_2 is exponential with parameter β, independent of X_1 (the term "hypoexponential" describes convolutions of exponential variables with different parameters). While the above hypoexponential variable is also a generalized mixture of exponential distributions, the distribution of X is a proper mixture as its PDF can be expressed as

$$f_X(x) = q\beta e^{-\beta x} + (1-q) g(x), \quad x \in \mathbb{R}_+,$$

with $g(\cdot)$ given by (31). This representation can be obtained directly from Proposition 4, which shows that X is either equal to X_1 (representing the exponential part E_0, with probability q) or $X_1 + X_2$, where X_2 is exponential with parameter $p\beta$. This X_2 corresponds to *the first sum on the right-hand-side of the representation (14), and the hypoexponential component $X_1 + X_2$ occurs with probability $1 - q$.*

Remark 10 *The above mixture representations of X provide a convenient tool in deriving basic characteristics of the distribution of X. For example, the moment generating function (MGF) of X is of the form*

$$M_X(t) = \mathbb{E}\left(e^{tX}\right) = \left(1 - \frac{1-q}{1-p}\right)\frac{\beta}{\beta - t} + \left(\frac{1-q}{1-p}\right)\frac{p\beta}{p\beta - t}, \quad t < p\beta,$$

while the moments of X are given by

$$\mathbb{E}X^\eta = \left(1 - \frac{1-q}{1-p}\right)\frac{\Gamma(\eta+1)}{\beta^\eta} + \left(\frac{1-q}{1-p}\right)\frac{\Gamma(\eta+1)}{(p\beta)^\eta}, \quad \eta > -1.$$

4.2.2 The marginal distribution of Y

The PDF of Y shown in the result below can be calculated as that of X, by summing up the joint PDF of Y and N given by (24) across all the values of $n \in \mathbb{N}$.

Proposition 12 *If $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$ then the PDF of Y is*

$$f_Y(y) = \beta e^{-\beta y}\left[q + \frac{(1-q)p}{1-p}\left(\frac{1}{[1-(1-p)(1-e^{-\beta y})]^2} - 1\right)\right], \quad y \in \mathbb{R}_+,$$

while the CDF of Y is

$$F_Y(y) = \begin{cases} 0 & \textit{for } y < 0 \\ \left[1 - e^{-\beta y}\right]\left[q + (1-q)\frac{p(1-e^{-\beta y})}{1-(1-p)(1-e^{-\beta y})}\right] & \textit{for } y \geq 0. \end{cases}$$

Remark 11 *The structure of the above CDF reveals that the distribution of Y is also a two component mixture. Indeed, Y can be thought of as either an exponential variable E_0 with parameter β (with probability q) or the maximum of two independent variables, $E_0 \vee \tilde{Y}$, where*

$$\tilde{Y} \stackrel{d}{=} \bigvee_{i=1}^{N_p} E_i$$

with N_p and $\{E_i\}$ as above. The variable $\tilde{Y}$ has a truncated logistic distribution on $\mathbb{R}_+$, with the CDF

$$F_{\tilde{Y}}(y) = \frac{p(1-e^{-\beta y})}{1-(1-p)(1-e^{-\beta y})}, \quad y \in \mathbb{R}_+,$$

studied by Marshall and Olkin (1997). This can also be seen from the representation (29), showing that $Y \stackrel{d}{=} E_0 \vee I\tilde{Y}$.

4.3 Conditional distributions

Here we summarize basic facts concerning bivariate and univariate conditional distributions connected with the GT distribution. Since the results below are established by routine derivations involving ratios of the relevant PDFs, their elementary proofs are omited.

4.3.1 Bivariate conditional distributions

Here we consider the three bivariate conditional distributions of $(X, Y)|N = n$, $(X, N)|Y = y$, and $(Y, N)|X = x$.

4.3.2 The distribution of X and Y given N = n

The conditional distribution of (X, Y) given $N = n \in \mathbb{N}$ was studied in Qeadan et al. (2012), and is known as the $BGGE(\beta, n)$ model. Its PDF, given by (9) - (10), provided the basis for our derivation of the GT PDF. In particular, the conditional distribution of X given $N = n$ is Gamma with the PDF given in (20) while the conditional distribution of Y given $N = n$ is generalized exponential (see, e.g., Gupta and Kundu 2007) with the PDF given in (23).

4.3.3 The distribution of X and N given Y = y

Next, we consider the conditional PDF of (X, N) given $Y = y > 0$, which turns out to be of the form

$$f(x, n|y) = (\beta(1-p))^{n-1} e^{-\beta(x-y)} H(x, y, n) \cdot \begin{cases} \frac{q}{v(y)} & \text{for } n = 1 \\ \frac{p}{1-p}\frac{1-q}{v(y)} & \text{for } n \geq 2, \end{cases}$$

where the function $H(x, y, n)$ is given by (10) and

$$v(y) = q + \frac{(1-q)p}{1-p}\left(\frac{1}{[1-(1-p)(1-e^{-\beta y})]^2} - 1\right), \quad y \in \mathbb{R}_+.$$

We note that when $n = 1$, this conditional PDF is non-zero only if $x = y$, in which case it takes on the value of $q/v(y)$. We also note that when $n \geq 2$, the function $H(x, y, n)$ will be non-zero only if x satisfies $ky < x \leq (k+1)y$, $k = 1, \ldots, n-1$, in which case

$$H(x, y, n) = \sum_{s=1}^{k} \frac{n(n-1)}{(s-1)!\,(n-s)!}(x - sy)^{n-2}(-1)^{s+1}. \tag{32}$$

4.3.4 The distribution of Y and N given X = x

Finally, we have the following expression for the PDF of (Y, N) given $X = x > 0$:

$$f(y, n|x) = (\beta(1-p))^{n-1} H(x, y, n) \cdot \begin{cases} \frac{q(1-p)}{u(x)} & \text{for } n = 1 \\ \frac{p(1-q)}{u(x)} & \text{for } n \geq 2, \end{cases}$$

where the function $H(x, y, n)$ is given by (10) and

$$u(x) = q - p + (1-q)pe^{(1-p)\beta x}, \quad x \in \mathbb{R}_+.$$

Again, when $n = 1$ the PDF above is non-zero only if $y = x$, in which case it takes on the value of $q(1-p)/u(x)$, and when $n \geq 2$ the function $H(x, y, n)$ will be non-zero only if x satisfies $x/(k+1) \leq y < x/k$, $k = 1, \ldots, n-1$, in which case we have (32).

4.3.5 Univariate conditional distributions

It turns out that all three univariate conditional distributions of X, Y, and N given the other two variables are the same as their counterparts in the special TETLG case $(p = q)$ studied by Kozubowski et al. (2011). We present their formulas below for the convenience of the reader.

4.3.6 The distribution of X given Y = y, n = n

The PDF of the conditional distribution of X given $Y = y > 0, N = n \in \mathbb{N}$ is given by

$$f(x|y, n) = \left(\frac{\beta}{1-e^{-\beta y}}\right)^{n-1} \frac{e^{-\beta(x-y)}}{n} H(x, y, n).$$

Similarly to the cases discussed above, when $n = 1$ the PDF is non-zero only if $y = x$, in which case it takes on the value of 1. In turn, when $n \geq 2$ the function $H(x, y, n)$ will be non-zero only if x satisfies $ky < x \leq (1+k)y$, $k = 1, \ldots, n-1$, in which case we have (32).

4.3.7 The distribution of Y given X = x, n = n

The conditional PDF of Y given $X = x > 0, N = n \in \mathbb{N}$ is of the form

$$f(y|x, n) = \frac{(n-1)!}{x^{n-1}} H(x, y, n).$$

Again, for $n = 1$ we have $f(y|x, n) = 1$ if $y = x$ (and zero otherwise), while for $n \geq 2$ the function $H(x, y, n)$ will be non-zero only if y satisfies $x/(k+1) \leq y < x/k$, $k = 1, \ldots, n-1$, in which case we have (32). We also note that this particular distribution is parameter-free.

4.3.8 The distribution of N given X = x, y = y

As in the TETLG case, the conditional distribution of N given $X = x > 0$ and $Y = y > 0$ reduces to a point mass at 1 when $x = y$. On the other hand, for $(x, y) \in S_k$, $k \in \mathbb{N}$, the PMF of this distribution is of the form

$$f(n|x, y) = \frac{(\beta(1-p))^{n-2}}{\sum_{s=1}^{k} \frac{(-1)^{s+1}}{(s-1)!} W_s(\beta[\,1-p\,]\,[\,x-sy\,]\,)} \cdot \begin{cases} H(x, y, n) & \text{for } n \geq k+1 \\ 0 & \text{otherwise,} \end{cases}$$

where W_s is defined in (28).

5 Estimation and testing

In this section we consider the problems of estimating the parameters of the GT model and testing the hypothesis that $p = q$ (so that the GT model reduces to the TETLG model) based on a random sample $(X_1, Y_1, N_1), \ldots, (X_k, Y_k, N_k)$ from the $\mathcal{GT}(p, q, \beta)$ distribution.

5.1 Maximum likelihood estimation

We start with the Fisher information matrix $I(p, q, \beta)$ corresponding to the distribution of $(X, Y, N) \sim \mathcal{GT}(p, q, \beta)$. Routine calculations lead to

$$I(p, q, \beta) = \begin{bmatrix} \frac{1-q}{p^2(1-p)} & 0 & 0 \\ 0 & \frac{1}{q(1-q)} & 0 \\ 0 & 0 & \frac{1+p-q}{p\beta^2} \end{bmatrix}. \tag{33}$$

Next, we turn to the parameter estimation viz. maximum likelihood. While the PDF of the GT model is rather complicated, fortunately the function $H(x, y, n)$ is parameter-free and the derivation of the maximum likelihood estimators (MLEs) is straightforward. Indeed, the likelihood function can be written as

$$L(p, q, \beta) = C\beta^{k\overline{N}_k} e^{-\beta k\overline{X}_k} q^{k_1} \left[(1-q)p\right]^{k-k_1} (1-p)^{k\overline{N}_k - 2k + k_1}, \tag{34}$$

where C is parameter-free, k_1 is the number of data points with $N_i = 1$, and $\overline{X}_k, \overline{N}_k$ are the sample means of the $\{X_i\}$ and the $\{N_i\}$, respectively. Thus, the statistics k_1/k, $\overline{X}_k$ and $\overline{N}_k$ are jointly sufficient. We note that these statistics do not involve the values of $\{Y_i\}_{i=1,\ldots,k}$. This is due to the fact that the conditional distribution of Y given X and N is parameter free (see Section 4.3.7). Thus the values of N and X carry all the information necessary to estimate all the parameters. We also note that the likelihood function can be maximized with respect to each parameter separately from the other parameters, leading to explicit MLEs provided below.

Proposition 13 *Let* $(X_1, Y_1, N_1), \ldots, (X_k, Y_k, N_k)$ *be IID observations from* $\mathcal{GT}(p, q, \beta)$ *distribution such that* $\overline{X}_k > 0$, $\overline{N}_k > 1$. *Then, there exist unique MLEs of the three parameters, given by*

$$\hat{p}_k = \left(1 - \frac{k_1}{k}\right)\frac{1}{\overline{N}_k - 1}, \quad \hat{q}_k = \frac{k_1}{k}, \quad and \hat{\beta}_k = \frac{\overline{N}_k}{\overline{X}_k}. \tag{35}$$

We note that since the distribution of X is absolutely continuous, we have $\mathbb{P}(\overline{X}_k > 0) = 1$. Additionally, while it is possible to have $\overline{N}_k = 1$, which occurs only if all sample values are equal to 1, the probability of this event converges to zero as the sample size k goes to infinity. However, if such an event does occur, the MLEs of q and β still exist and are unique (with values of 1 and $1/\overline{X}_k$, respectively) while the MLE of p is undefined.

Further, the estimators do not involve the values $\{Y_i\}$, and the estimator of β is exactly the same as its counterpart in the TETLG model of Kozubowski et al. (2011). Moreover, if only the data on the $\{X_i, N_i\}$ were available, and the values of the third variable were missing, we would still obtain exactly the same set of three estimators. In fact, the estimators of p and q are only dependent on the univariate observations of the $\{N_i\}$, and would be exactly the same if the rest of the data was missing, or if we only have information on the $\{Y_i, N_i\}$ while the $\{X_i\}$ were missing. However, in the latter case, the estimator of β is no longer the same as above, and a numerical search is needed to find it. The estimators are also different from those above if we only worked with bivariate data on $\{X_i, Y_i\}$ or only univariate data involving either the $\{X_i\}$ or the $\{Y_i\}$. In these three cases the underlying models are mixtures, and finding the estimators is not straightforward. We now turn to the asymptotic properties of the MLEs, where we have the following result.

Proposition 14 *The vector MLE* $(\hat{p}_k, \hat{q}_k, \hat{\beta}_k)^\top$ *given in Proposition 13 is*

1. *Consistent;*
2. *Asymptotically normal, that is* $\sqrt{n}[\,(\hat{p}_k, \hat{q}_k, \hat{\beta}_k)^\top - (p, q, \beta)^\top]$ *converges in distribution to a trivariate normal distribution with the (vector) mean zero and the covariance matrix*

$$\mathbf{\Sigma}_{MLE} = \begin{bmatrix} \frac{p^2(1-p)}{1-q} & 0 & 0 \\ 0 & q(1-q) & 0 \\ 0 & 0 & \frac{p\beta^2}{1+p-q} \end{bmatrix}; \tag{36}$$

3. *Asymptotically efficient, that is the asymptotic covariance matrix (36) coincides with the inverse of the Fisher information matrix (33).*

Remark 12 *The above result allows to derive approximate* $(1 - \alpha) \times 100\%$ *confidence intervals for the parameters in large sample setting, leading to*

$$\hat{p}_k \pm z_{\alpha/2}\hat{\sigma}_{ASY}(p)/k, \quad \hat{q}_k \pm z_{\alpha/2}\hat{\sigma}_{ASY}(q)/k, \quad \hat{\beta}_k \pm z_{\alpha/2}\hat{\sigma}_{ASY}(\beta)/k,$$

where the quantities $\hat{\sigma}_{ASY}(p)$, $\hat{\sigma}_{ASY}(q)$, *and* $\hat{\sigma}_{ASY}(\beta)$ *are the square roots of the diagonal entries of the asymptotic covariance matrix (36) with the parameters replaced by their MLEs.*

5.2 Testing for $p = q$ under the GT model

The objective of this section is to develop a likelihood ratio (LR) test for the null hypothesis $H_0 : p = q$ under the assumption that the data follow the $\mathcal{GT}(p, q, \beta)$ distribution. Before setting up the test, let us consider the parameter space of this model, and its subspace corresponding to the null hypothesis. Clearly, we must have $\beta > 0$ and p, q must belong to the unit interval. However, care is needed in regard to the boundary values of p and q in order to assure the parameterization is identifiable and the possible values of p and q are consistent with the results of estimation. With this in mind, we denote the vector-parameter by $\theta = (\theta_1, \theta_2, \theta_3)^\top$, where $\theta_1 = p$, $\theta_2 = q$, and $\theta_3 = \beta$, and propose to set the general parameter space Θ as follows:

$$\Theta = \{(\theta_1, \theta_2, \theta_3) : (\theta_1, \theta_2) \in \Theta^{1,2}, \theta_3 > 0\},$$

where

$$\Theta^{1,2} = \{(\theta_1, \theta_2) : \theta_1, \theta_2 \in (0, 1)\} \cup \{(\theta_1, \theta_2) : 0 < \theta_1 \leq 1, \theta_2 = 0\} \cup \{(\theta_1, \theta_2) : \theta_1 = 1, \theta_2 \in [0, 1]\}.$$

With this definition of the parameter subspace for p and q, all the boundary values with $p = 0$ are excluded, regardless of q, and so are all the boundary values with $q = 1$, with the exception of the "corner" of the unit square where we have $p = q = 1$. Clearly, the null subset of Θ where $p = q$ corresponds to the set

$$\Theta_0 = \{(\theta_1, \theta_2, \theta_3) : 0 < \theta_1 = \theta_2 \leq 1, \theta_3 > 0\}.$$

With the above set-up, we wish to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1, \tag{37}$$

where $\Theta_1 = \Theta - \Theta_0$. The classical LR test rejects the null hypothesis in (37) in favor of the alternative for large values of the LR test statistic

$$\Lambda = \frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}, \tag{38}$$

where $L(\cdot)$ is the likelihood function (of the full GT model) given by (34). The evaluation of the likelihood ratio test statistic (38) is straightforward. Indeed, the numerator in (38) is simply the value of the likelihood evaluated at the three MLEs given by (35), resulting in $L(\hat{p}_k, \hat{q}_k, \hat{\beta}_k)$, given explicitly. The denominator is given by $L(\hat{p}_k^0, \hat{q}_k^0, \hat{\beta}_k^0)$, where the triple $(\hat{p}_k^0, \hat{q}_k^0, \hat{\beta}_k^0)$ are the values of the parameter θ that maximizes the likelihood over Θ_0. What this means in our case is that we set $q = p$ in the likelihood function (34), resulting in

$$L_0(p, \beta) = L(p, p, \beta) = C\beta^{k\overline{N}_k} e^{-\beta k \overline{X}_k} p^k (1 - p)^{k\overline{N}_k - k}, \tag{39}$$

and subsequently maximize the function $L_0(p, \beta)$ with respect to $p \in (0, 1]$ and $\beta > 0$. We recognize the function in (39) as the likelihood based on the TETLG model of Kozubowski et al. (2011), which is known to be maximized by

$$\hat{p}_k^0 = 1/\overline{N}_k, \quad \hat{\beta}_k^0 = \overline{N}_k/\overline{X}_k. \tag{40}$$

Thus, the denominator in (38) becomes $L_0(\hat{p}_k^0, \hat{\beta}_k^0) = L(\hat{p}_k^0, \hat{p}_k^0, \hat{\beta}_k^0)$, and is also given explicitly. By putting these facts together, we arrive at the following result.

Proposition 15 *The LR statistic (38) for testing the hypotheses in (37) based on a random sample of size k from a $\mathcal{GT}(p,q,\beta)$ distribution is given by*

$$\Lambda = \left(\frac{\hat{q}_k}{\hat{p}_k^0}\right)^{k_1} \left(\frac{(1-\hat{q}_k)\hat{p}_k}{(1-\hat{p}_k^0)\,\hat{p}_k^0}\right)^{k-k_1} \left(\frac{1-\hat{p}_k}{1-\hat{p}_k^0}\right)^{k\overline{N}_k-2k+k_1}, \tag{41}$$

where k_1 is the number of sample values with $N_i = 1$, $\hat{p}_k$, $\hat{q}_k$, $\hat{\beta}_k$ are given by (35), and $\hat{p}_k^0$, $\hat{\beta}_k^0$ are given by (40).

Remark 13 *We note that the LR statistic does not involve the values of $\{X_i\}$ and $\{Y_i\}$. In fact, the exact same statistic comes up in connection with testing the hypotheses*

$$H_0: p = q \;\; versus \;\; H_1: p \neq q \tag{42}$$

in the context of univariate $\mathcal{HGEO}(p,q)$ distribution, based on a random sample $N_1,\ldots,N_k$. The PMF of this distribution is given in (6). Under the null hypothesis in (42) this 1- inflated geometric distribution reduces to the classical geometric distribution $\mathcal{GEO}(p)$, given by the PMF (4).

By the standard large sample theory, the quantity $2\log\Lambda$ has approximately chi-square distribution when the sample size k is large, which helps to set-up the critical region in practice.

Proposition 16 *Let Λ be the LR test statistic (38), based on a random sample of size k from $\mathcal{GT}(p,q,\beta)$ distribution. Then, as $k \to \infty$, the quantity $2\log\Lambda$ converges in distribution to a chi-square random variable with 1 degree of freedom.*

Remark 14 *While the calculation of the LR test statistic or the quantity $2log\Lambda$ in practice is straightforward, some care is required when dealing with certain exceptional cases, where the ratios in (41) may seem to be undefined. Careful examination of the likelihood function, the relevant MLEs, and the LR statistic reveals five different cases, which can be described as follows:*

1. *If all the values of N_i are 1 (so that $k_1 = k$) then $2\log\Lambda = 0$,*
2. *If all the values of N_i are 2 (so that $k_1 = 0$) then $2\log\Lambda = 2k\log 4$,*
3. *If all the values of N_i are either 1 or 2, but they are not all the same, then*

$$2\log\Lambda = 2k\left[\overline{N}_k\log(\overline{N}_k) + \frac{k_1}{k}\log\left(\frac{k_1}{k}\right)\right],$$

4. *If $N_i \geq 2$ for all $i = 1,\ldots,k$ and at least one value is greater than 2 then*

$$2\log\Lambda = 2k\left[(\overline{N}_k-2)\log(\overline{N}_k-2) - 2(\overline{N}_k-1)\log(\overline{N}_k-1) + \overline{N}_k\log(\overline{N}_k)\right],$$

5. *If at least one $N_i = 1$ and at least one $N_i > 2$ then*

$$\begin{aligned}2\log\Lambda \;=\; 2k\Bigg[&\frac{k_1}{k}\log\left(\frac{k_1}{k}\right) + 2\left(1-\frac{k_1}{k}\right)\log\left(1-\frac{k_1}{k}\right) - 2(\overline{N}_k-1)\log(\overline{N}_k-1)\\ &+(\overline{N}_k-1-\left(1-\frac{k_1}{k}\right)\log\left(\overline{N}_k-1-\left(1-\frac{k_1}{k}\right)\right) + \overline{N}_k\log(\overline{N}_k)\Bigg].\end{aligned} \tag{43}$$

In order to have a practical guide as to when one can use the limiting distribution as a reasonable approximation to the distribution of $2\log\Lambda$ we performed a Monte Carlo

Table 1 Sample sizes for the limiting distribution to work as an approximation for the $2 \log \Lambda$

p	0.01	0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.98	0.99
k	510	160	60	25	15	12	10	20	35	60	85	240	1,000	8,000	27,000

The top row lists a selection of true values of the parameter *p*. The bottom row lists the corresponding minimal samples sizes for which the χ_1^2 to be a reasonable approximation to the distribution of $2 \log \Lambda$

study. Noting that the speed of convergence may depend on the true value of p, we simulated 10,000 samples of (varying) size k from $\mathcal{GEO}(p)$ distribution for selected values of p. We then found the smallest k for which the (empirical) distribution of $2 \log \Lambda$ can be assumed to be χ_1^2. We used Kolmogorov-Smirnov goodness-of-fit test with significance level of 0.05 to assess whether the distribution of $2 \log \Lambda$ can be reasonably considered to be χ_1^2. We summarized the results of this simulation study in Table 1.

The simulation shows that when the true value of p is between 0.1 and 0.9, the sample size needed for the limiting distribution to be a good approximation for the distribution of $2 \log \Lambda$ is below 100. However, once the value of p becomes closer to 0 or 1, the sample sizes required for a reasonable approximation are growing. In particular, note that the sample size required for large values of p (close to 1) are much larger than those for the small values of p (close to 0).

6 An illustrative data example

In this section, we illustrate potential applications of the new GT model using $S\&P500$ index return data.

The data was downloaded from 'Yahoo! Finance' historical data archive. The initial data were the daily closing prices for the $S\&P500$ index, covering the period from Dec 30, 1927 to April 17, 2020. These were converted to (n = 23,183) daily log-returns, i.e. natural logarithms of the ratios of the closing prices for two consecutive days. Finally, we converted these data to the *growth periods*, where the daily closing prices increase from one day to the next one, so that the log-returns stay positive (5,540 growth periods). In this case, the N_i are the durations (in days) of the growth periods, while the X_i and the Y_i are the magnitude and the maximum daily return for the i-th growth period. We call this data set $S\&P500$.

We estimated the parameters p, q, and β of the underlying $\mathcal{GT}(p, q, \beta)$ model viz. maximum likelihood, using the results given in Proposition 13. The resulting estimates are shown in Table 2, along with estimated margins of error (ME) of the 95% (asymptotic) confidence intervals described in the remark following Proposition 14.

Note, that in this case, $p < q$ and so we have an example of under-inflated ones data. This may be reflection of the fact that $S\&P500$ is a composite index, which is more stable, that is less prone to changes from growth to decline, than an individual stock return. Further, we tested the hypotheses $H_0 : p = q$ versus $H_1 : p \neq q$ on significance level 0.05 using the likelihood ratio test described in Section 5, and obtained the test statistic

Table 2 Maximum likelihood estimates of the parameters for the *S&P*500 data with the corresponding (approximate) 95% margins of error in parentheses

	$\hat{p}$ (ME)	$\hat{q}$ (ME)	$\hat{\beta}$ (ME)
*S&P*500	0.47 (0.00007)	0.441 (0.00009)	133.575 (2.8832)

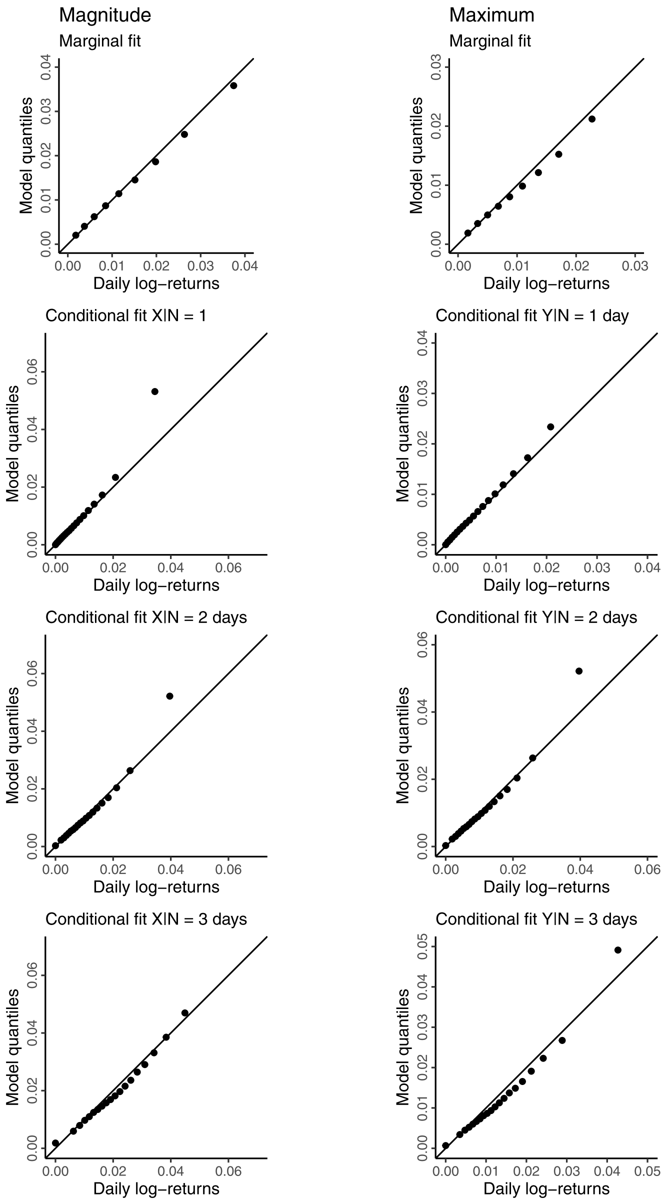

Fig. 1 Fit of the GT model to *S&P*500 data. The graphs in the left/right column show the fit of magnitude/maximum. The top row has QQ plots of the marginal distributions of magnitude (left panel) and maximum (right panel). Rows 2, 3, and 4 show QQ plots of the conditional distributions of the magnitude (left panels) and maximum (right panels) given duration $N = 1, 2$, and 3

$2 \log \Lambda = 10.06$, with (approximate) p-value of 0.003. Thus, we rejected the null hypothesis and conclude that durations are coming from the shifted hurdle model. Thus, the use of GT model for the growth episodes arising from our data is better than the standard TETLG model (connected with $p = q$), which has been used before in similar settings (see, Kozubowski et al. 2011).

Although we did not aim at a formal goodness of fit analysis, we wanted to present visual evidence of the reasonable fit of our model. We fitted the marginal distributions of

Table 3 Duration Fit for *S&P*500 data

N in days	1	2	3	4	5	6	7	8	9	10 +
Freq.	2444	1406	825	418	205	121	58	33	13	17
Rel. freq.	0.441	0.254	0.149	0.075	0.037	0.022	0.010	0.006	0.002	0.003
Model prob.	0.441	0.263	0.139	0.074	0.039	0.021	0.011	0.006	0.003	0.003

The table contains observed frequency (second row), relative frequency (third row), and model probabilities (last row) for the growth periods of given durations (first row)

X and Y, and the conditional distributions for X given N and Y given N, when $N = 1, 2, 3$. The fit is illustrated in Fig. 1 below with QQ plots.

All QQ plots show reasonable to very good fit. Note that the fit of the marginal of maximum is particularly impressive, given that the values of the maxima of the episodes did not play any role in the estimation.

Next, we present the fit of the $\mathcal{HGEO}(p, q)$ model for duration. Table 3 contains the observed frequencies/relative frequencies along with estimated model probabilities for the duration of $S\&P500$ growth events.

The relative frequencies and model probabilities for our data are reasonably close, and we conclude that the fit of the $\mathcal{HGEO}(p, q)$ model is quite good for this data set as well. We believe that the fit of the GT model to the $S\&P500$ data will start a common use of this model for data sets with an excessive number of ones.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40488-021-00114-3.

Additional file 1: Supplementary material for "A new trivariate model for stochastic episodes".

Abbreviations

The following abbreviations are used in this manuscript (in alphabetical order):
BGGE: Bivariate distribution with geometric and generalized exponential margins; CDF: Cumulative distribution function; ChF: Characteristic function; EXP: Exponential distribution; GEO: Geometric; GT: Generalized TETLG; H: Hurdle model; HGEO: Hurdle geometric distribution; IID: Independent and identically distributed; LR: Likelihood ratio; MLE: Maximum likelihood estimator; PDF: Probability density function; PMF: Probability mass function; TETLG: Trivariate distribution with exponential, truncated logistic and geometric margins; ZI: Zero inflated distribution

Acknowledgements

The authors thank Ms. Ilaria Vinci for helpful discussions and help with checking of tedious calculations. We also thank the two reviewers for their thoughtful and constructive comments, that helped us improve this manuscript.

Authors' contributions

FZ is a PHD student, who derived majority of the results, performed simulations and data analysis, and produced all figures. AKP is his adviser, who reviewed all this work from inception to the end of the manuscript writing, and contributed to some of the results. TJK is a research collaborator who contributed to some of the results and revision of the paper. The author(s) read and approved the final manuscript.

Funding

NA.

Availability of data and materials

All data is publicly available. Here is the link to 'Yahoo! Finance' historical archive: https://finance.yahoo.com/quote/%5EGSPC/history?p=%5EGSPC

Competing interests

The authors declare that they have no competing interests.

Received: 10 August 2020 Accepted: 12 January 2021
Published online: 26 February 2021

References

Alshkaki, R. S. A.: On zero-one inflated geometric distribution. Internat. Res. J. Math. Eng. IT. **3**(8), 10–21 (2016)

Arendarczyk, M., Kozubowski, T. J., Panorska, A. K.: A bivariate distribution with Lomax and geometric margins. J. Korean Statist. Soc. **47**, 405–422 (2018a)

Arendarczyk, M., Kozubowski, T. J., Panorska, A. K.: The joint distribution of the sum and the maximum of dependent Pareto risks. J. Multivar. Anal. **167**, 136–156 (2018b)

Aryal, T.: Inflated geometric distribution to study the distribution of rural outmigrants. J. Instit. Eng. **8**(1), 266–268 (2011)

Barreto-Souza, W.: Bivariate gamma-geometric law and its induced Lévy process. J. Multivar. Anal. **109**, 130–145 (2012)

Barreto-Souza, W., Silva, R. B.: A bivariate infinitely divisible law for modeling the magnitude and duration of monotone periods of log-returns. Statist. Neerlandica. **73**, 211–233 (2019)

Biondi, F., Kozubowski, T. J., Panorska, A. K.: Stochastic modeling of regime shifts. Clim. Res. **23**, 23–30 (2002)

Biondi, F., Kozubowski, T. J., Panorska, A. K.: A new model for quantifying climate episodes. Internat. J. Climatol. **25**, 1253–1264 (2005)

Biondi, F., Kozubowski, T. J., Panorska, A. K., Saito, L: A new stochastic model of episode peak and duration for eco-hydro-climatic applications. Ecol. Modell. **211**, 383–395 (2008)

Cameron, A. C., Trivedi, P. K.: Regression Analysis of Count Data. Cambridge University Press, Cambridge (1998)

Cameron A.C., Trivedi. P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, Cambridge (2005)

Chipeta, M. G., Ngwira, B. M., Simoonga, C., Kazembe, L. N.: Zero adjusted models with applications to analyzing helminths count data. BMC Res Notes. **7**, 7–856 (2014)

Constantinescu, C. D., Kozubowski, T. J., Qian, H. H.: Probability of ruin in discrete insurance risk model with dependent Pareto claims. Depend. Model. **7**(1), 215–233 (2019)

Famoye, F., Singh, K. P.: Zero-inflated generalized Poisson regression model with an application to domestic violence data. J. Data Sci. **4**, 117–130 (2006)

Gupta, D., Kundu, D.: Generalized exponential distribution: Existing results and some recent developments. J. Statist. Plann. Infer. **137**(11), 3537–3547 (2007)

Hu, M.-C, Pavlicova, M, Nunes, E. V: Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. Am. J. Drug Alcohol Abuse. **37**(5), 367–375 (2011)

Iwunor, C. C. O.: Estimation of parameters of the inflated geometric distribution for rural out-migration. Genus. **51**(3/4), 253–260 (1995)

Johnson, N. L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, vol. 1, 2nd ed. Wiley, New York (1994)

Kozubowski, T. J., Panorska, A. K.: A mixed bivariate distribution with exponential and geometric marginals. J. Statist. Plann. Infer. **134**, 501–520 (2005)

Kozubowski, T. J., Panorska, A. K.: A mixed bivariate distribution connected with geometric maxima of exponential variables. Comm. Statist. Theory Methods. **37**, 2903–2923 (2008)

Kozubowski, T. J., Panorska, A. K., Biondi, F.: Mixed multivariate models for random sums and maxima. In: SenGupta, A. (ed.) Advances in Multivariate Statistical Methods, Statistical Science and Interdisciplinary Research - Vol. 4, pp. 145–171. World Scientific, Singapore, (2008a)

Kozubowski, T. J., Panorska, A. K., Podgórski, K.: A bivariate Lévy process with negative binomial and gamma marginals. J. Multivar. Anal. **99**, 1418–1437 (2008b)

Kozubowski, T. J., Panorska, A. K., Qeadan, F.: The distributions of the peak to average and peak to sum ratios under exponentiality. In: Wells, M. T., SenGupta, A. (eds.) Advances in Directional and Linear Statistics, Festschrift Volume for J.S. Rao, Physica-Verlag, Heidelberg, pp. 131–141, (2010)

Kozubowski, T. J, Panorska, A. K, Qeadan, F: A new multivariate model involving geometric sums and maxima of exponentials. J. Statist. Plann. Infer. **141**(7), 2353–2367 (2011)

Lambert, D.: Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. **34**, 1–14 (1992)

Marshall, A. W., Olkin, I.: A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika. **84**(3), 641–652 (1997)

Mullahy, J.: Specification and testing of some modified count data models. J. Econ. **33**, 341–365 (1986)

Mullahy, J.: Heterogeneity, excess zeros, and the structure of count data models. J. Appl. Econ. **12**(3), 337–350 (1997)

Pandey, H., Tiwari, R.: An inflated probability model for the rural out-migration. Recent Res. Sci. Tech. **3**(7), 100–103 (2011)

Panicha, K.: Capture-recapture estimation and modelling for one-inflated count data. Dissertation, University of Southampton (2018)

Qeadan, F., Kozubowski, T. J., Panorska, A. K.: The joint distribution of the sum and the maximum of n i.i.d. exponential random variables. Comm. Statist. Theory Methods. **41**(3), 544–569 (2012)

Sharma, A. K., Landge, V. S.: Zero inflated negative binomial for modeling heavy vehicle crash rate on Indian rural highway. Internat. J. Adv. Eng. Tech. **5**(2), 292–301 (2013)

Tüzen, M. F., Erbaş, S.: A comparison of count data models with an application to daily cigarette consumption of young persons. Comm. Statist. Theory Methods. **47**(23), 5825–5844 (2018)

Zeileis, A., Kleiber, C., Jackman, S.: Regression models for count data in R. J. Statist. Softw. **27**(8), 1–25 (2008)

Zelterman, D.: Discrete Distributions, Applications in the Health Sciences. Wiley, New Jersey (2004)

Zuur, A. F., Leno, E. N., Walker, N. J., Saveliev A.A., Smith G.M.: Mixed effects models and extensions in ecology with R. Statistics for Biology and Health. Springer Science and Business Media (2009). https://doi.org/10.1007/978-0-387-87458-611

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.