An Acad Bras Cienc (2022) 94(2): e20201972 DOI 10.1590/0001-3765202220201972
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

MATHEMATICAL SCIENCES

A competitive family to the Beta and Kumaraswamy generators: Properties, Regressions and Applications

GAUSS M. CORDEIRO, JULIO CEZAR S. VASCONCELOS, EDWIN M.M. ORTEGA & PEDRO RAFAEL D. MARINHO

Abstract: We define two new flexible families of continuous distributions to fit real data by compounding the Marshall–Olkin class and the power series distribution. These families are very competitive to the popular beta and Kumaraswamy generators. Their densities have linear representations of exponentiated densities. In fact, as the main properties of thirty five exponentiated distributions are well-known, we can easily obtain several properties of about three hundred fifty distributions using the references of this article and five special cases of the power series distribution. We provide a package implemented in R software that shows numerically the precision of one of the linear representations. This package is useful to calculate numerical values for some statistical measurements of the generated distributions. We estimate the parameters by maximum likelihood. We define a regression based on one of the two families. The usefulness of a generated distribution and the associated regression is proved empirically.

Key words: generating function, Marshall–Olkin family, maximum likelihood, moment, power series distribution.

INTRODUCTION

The *Marshall–Olkin* ("MO") family (Marshall & Olkin 1997) adds one parameter to a parent distribution. Let $G(z) = G(z;\boldsymbol{\tau})$ be the parent cumulative distribution function (cdf) of a random variable Z with parameter vector $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_q)^\top$. The survival function and probability density function (pdf) of Z are $\bar{G}(z) = \bar{G}(z;\boldsymbol{\tau})$ and $g(z) = g(z;\boldsymbol{\tau})$, respectively.

The cdf $H(z)$ and survival function $\bar{H}(z) = 1 - H(z)$ of the MO class with baseline $G(z;\boldsymbol{\tau})$ are

$$H(z) = H(z;\alpha,\boldsymbol{\tau}) = \frac{G(z;\boldsymbol{\tau})}{1-\bar{\alpha}\bar{G}(z;\boldsymbol{\tau})}, \quad z \in \mathbb{R}, \quad \alpha > 0, \tag{1}$$

and

$$\bar{H}(z) = \bar{H}(z;\alpha,\boldsymbol{\tau}) = \frac{\alpha\bar{G}(z;\boldsymbol{\tau})}{1-\bar{\alpha}\bar{G}(z;\boldsymbol{\tau})}, \tag{2}$$

respectively, where $\bar{\alpha} = 1 - \alpha$.

Equation (1) can generate many continuous distributions from popular ones. The MO-G density function can be expressed as

$$h(z) = h(z;\alpha,\boldsymbol{\tau}) = \frac{\alpha\, g(z;\boldsymbol{\tau})}{[1-\bar{\alpha}\bar{G}(z;\boldsymbol{\tau})]^2}. \tag{3}$$

For $\alpha = 1$, $h(z) = g(z;\boldsymbol{\tau})$ is the simplest case of (3). Marshall & Olkin (1997) pioneered the MO-Weibull (MOW) distribution which is a useful extension of the Weibull.

Consider N random variables $Z_1, \ldots, Z_N$ independent and identically distributed (i.i.d.) with cdf $H(z)$ and pdf $h(z)$ given by (1) and (3), respectively. Here, N is a discrete random variable with support $\{1, 2, \ldots\}$. Henceforth, let $X = \max\{Z_1, \ldots, Z_N\}$ and $Y = \min\{Z_1, \ldots, Z_N\}$ be two random variables assuming that N has the zero-truncated power series (PS) distribution with probability mass function (pmf)

$$p_n = \mathbb{P}(N = n; \vartheta) = \frac{a_n \vartheta^n}{C(\vartheta)}, n = 1, 2, \ldots, \tag{4}$$

where $a_n > 0$ (for $n \geq 1$), ϑ is called the power parameter and $C(\vartheta) = \sum_{n=1}^{\infty} a_n \vartheta^n > 0$. The probability generating function (pgf) of N is $P(z) = E(z^N) = C(z\vartheta)/C(\vartheta)$.

Five important distributions are special cases of (4): the zero-truncated Poisson, logarithmic, negative binomial, geometric and zero-truncated binomial distributions.

The cdf of $X = \max\{Z_1, \ldots, Z_N\}$ conditional given $N = n$ is

$$F_X(x \mid N = n) = \mathbb{P}[X \leq x | N = n] = H(x; \alpha, \boldsymbol{\tau})^n,$$

and then the unconditional cdf of X follows from (4)

$$F_X(x) = \sum_{n=1}^{\infty} H(x; \alpha, \boldsymbol{\tau})^n \frac{a_n \vartheta^n}{C(\vartheta)} = \frac{C(\vartheta H(x; \alpha, \boldsymbol{\tau}))}{C(\vartheta)}. \tag{5}$$

The conditional cdf of $Y = \min\{Z_1, \ldots, Z_N\}$ under $N = n$ is

$$F_Y(y \mid N = n) = \mathbb{P}[Y \leq y | N = n] = 1 - \bar{H}(y; \alpha, \boldsymbol{\tau})^n,$$

and then the unconditional cdf of Y follows from (4) as

$$F_Y(y) = 1 - \sum_{n=1}^{\infty} \bar{H}(y; \alpha, \boldsymbol{\tau})^n \frac{a_n \vartheta^n}{C(\vartheta)} = 1 - \frac{C(\vartheta \bar{H}(y; \alpha, \boldsymbol{\tau}))}{C(\vartheta)}. \tag{6}$$

Equations (5) and (6) define two *Marshall–Olkin Power Series-G* (MOPS-G) families under baseline G. They provide a strong motivation for explaining the failure time of any mechanism formed by an unknown number N of identical and independent (parallel or serial) components. The densities of X and Y are obtained by differentiating (5) and (6). We emphasize that these equations can generate many MOPS models. For each baseline G, we can generate ten (2×5) associated models from the five discrete distributions in Equation (4). For $\alpha = 1$, we have the *Power Series-G* (PS-G) classes under baseline G.

The minimum (Y) and maximum (X) statistics can be applied in several series and parallel systems with identical components and have many industrial and biological applications. In parallel systems, the random variable Y models the time of the first component to fail, while X models the time for the breakout system. A dual interpretation can be given for systems with serial components. These random variables are also very useful in oncology. For example, suppose we are studying a recurrence of a certain type of cancerous tumor of an individual after undergoing any kind of treatment. So, the time for the first cell to activate to produce cancer cells can be modeled by the generated distribution

of Y, while the disease manifestation (if it occurs only after an unknown number of factors have been active) can be modeled by the generated distribution of X.

Four new distributions based on the MOPS construction are introduced for illustrative purposes in Section Four special models. We derive linear representations for the densities of X and Y in Section Expansions. A package in R is presented in Section Numerical evaluation to calculate numerically several mathematical properties for the generated distributions based on the linear representations. General structural properties for the two families are addressed in Section Properties. In Section Estimation, we estimate the parameters for one of the families. We introduce in Section Regression the *Marshall–Olkin Truncated Poisson Weibull* regression defined from one of the families. In Section Two simulation studies, some simulations examine the accuracy of the maximum likelihood estimates (MLEs) and the quantile residuals (qrs). Two applications prove the utility of our finding in Section Applications. Finally, we offer concluding remarks in Section Conclusions.

FOUR SPECIAL MODELS

First, consider the zero-truncated Poisson in (4). The cdfs of the *Marshall–Olkin Zero-Truncated Poisson-G* (MOTP-G) distributions are determined from Equations (5) and (6) as

$$F_X(x) = (e^{\vartheta} - 1)^{-1} \left[\exp\{\vartheta\, H(x;\alpha,\boldsymbol{\tau})\} - 1\right] \tag{7}$$

and

$$F_Y(y) = 1 - (e^{\vartheta} - 1)^{-1} \left[\exp\left\{\vartheta\, \bar{H}(y;\alpha,\boldsymbol{\tau})\right\} - 1\right]. \tag{8}$$

The Weibull cdf with scale parameter $\lambda > 0$ and shape parameter $\beta > 0$ is (for $x \geq 0$)

$$G(z;\lambda,\beta) = 1 - \exp\left[-(\lambda z)^{\beta}\right].$$

Then, the cdf and survival function of the MO-Weibull (MOW) distribution are

$$H(z) = H(z;\alpha,\lambda,\beta) = \frac{1 - \exp\left[-(\lambda z)^{\beta}\right]}{1 - \bar{\alpha}\exp\left[-(\lambda z)^{\beta}\right]}$$

and

$$\bar{H}(z) = \bar{H}(z;\alpha,\lambda,\beta) = \frac{\alpha\exp\left[-(\lambda z)^{\beta}\right]}{1 - \bar{\alpha}\exp\left[-(\lambda z)^{\beta}\right]},$$

respectively.

By inserting the last two formulae in Equations (7) and (8) and differentiating the resulting expressions, we obtain the MOTP-Weibull (MOTPW) densities

$$f_X(x) = \frac{\alpha\,\vartheta\,\beta\,\lambda^{\beta}\,x^{\beta-1}\,e^{-u}}{(e^{j} - 1)\,(1 - ae^{-u})^2}\, \exp\left[\frac{\vartheta\,(1 - e^{-u})}{(1 - \bar{\alpha}e^{-u})}\right] \tag{9}$$

and

$$f_Y(y) = \frac{\alpha\,\vartheta\,\beta\,\lambda^{\beta}\,x^{\beta-1}\,e^{-u}}{(e^{j} - 1)\,(1 - ae^{-u})^2}\, \exp\left[\frac{\alpha\,\vartheta\,e^{-u}}{(1 - \bar{\alpha}e^{-u})}\right], \tag{10}$$

respectively, where $u = u(x) = (\lambda x)^{\beta}$ in $f_X(x)$ and $u = u(y) = (\lambda y)^{\beta}$ in $f_Y(y)$.

Second, consider the geometric distribution in (4). The cdfs of the *Marshall–Olkin Geometric-G* (MOG-G) classes follow from Equations (5) and (6)

$$F_X(x) = \frac{(1-\vartheta)}{\vartheta}\left[\frac{\vartheta\, H(x;\alpha,\boldsymbol{\tau})}{1-\vartheta\, H(x;\alpha,\boldsymbol{\tau})}\right] \tag{11}$$

and

$$F_Y(y) = 1 - \frac{(1-\vartheta)}{\vartheta}\left[\frac{\vartheta\, \bar{H}(y;\alpha,\boldsymbol{\tau})}{1-\vartheta\, \bar{H}(y;\alpha,\boldsymbol{\tau})}\right]. \tag{12}$$

The Burr XII (BXII) cdf is (for $x > 0$)

$$G(z;\beta,\lambda) = 1 - \left(1+z^{\beta}\right)^{-\lambda}, \tag{13}$$

where $\beta > 0$ and $\lambda > 0$ are shape parameters. For $\lambda = 1$ and $\beta = 1$ in Equation (13), we have the log-logistic (LL) and Lomax distributions, respectively.

Hence, the cdf and survival function of the *Marshall–Olkin Burr XII* (MOBXII) distribution are

$$H(z) = H(z;\alpha,\lambda,\beta) = \frac{1-(1+z^{\beta})^{-\lambda}}{1-\bar{\alpha}(1+z^{\beta})^{-\lambda}}$$

and

$$\bar{H}(z) = \bar{H}(z;\alpha,\lambda,\beta) = \frac{\alpha(1+z^{\beta})^{-\lambda}}{1-\bar{\alpha}(1+z^{\beta})^{-\lambda}},$$

respectively.

By inserting the last two formulae in Equations (11) and (12) and differentiating the resulting expressions with respect to x and y, respectively, we obtain the MOG-Burr XII (MOGBXII) densities

$$f_X(x) = \frac{\alpha\beta\lambda\,(1-\vartheta)\,x^{\beta-1}\,(1+x^{\beta})^{-\lambda-1}}{\left[1-\vartheta-(1-\alpha-\vartheta)(1+x^{\beta})^{-\lambda}\right]^2} \tag{14}$$

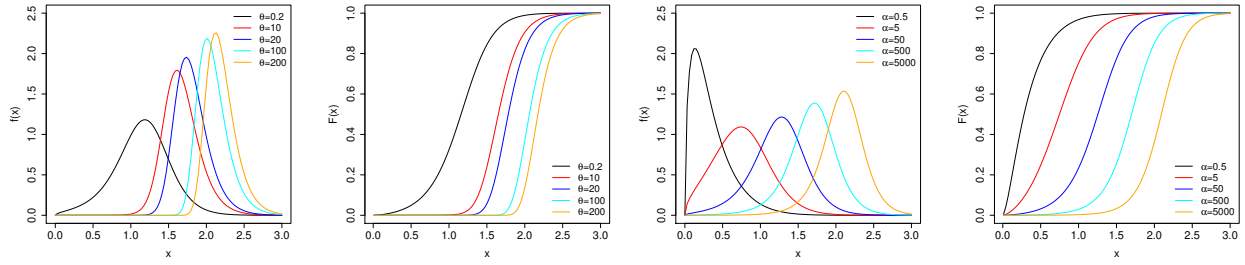
and

$$f_Y(y) = \frac{\alpha\beta\lambda\,(1-\vartheta)\,x^{\beta-1}\,(1+x^{\beta})^{-\lambda-1}}{\left\{1-\left[1-(1-\vartheta)\alpha\right](1+x^{\beta})^{-\lambda}\right\}^2}. \tag{15}$$

For the MOTPW and MOGBXII distributions (to the maximum X) referred to (9) and (14), some plots of the densities and cumulative functions are displayed in Figures 1 and 2, respectively. The various forms of the densities indicate more flexibility than the parent distributions.

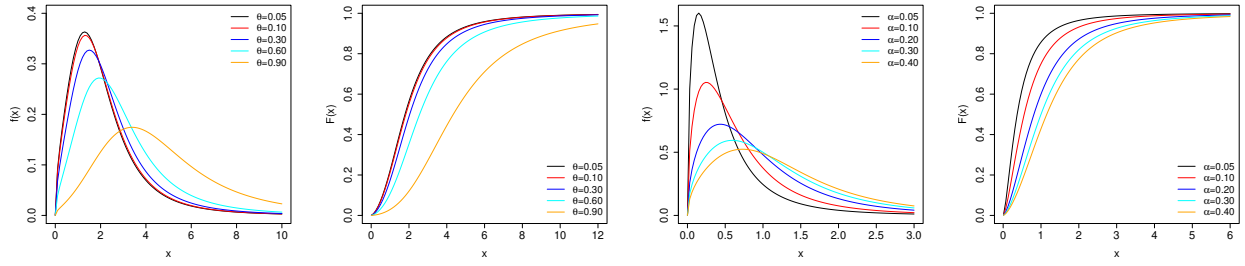
We can note increasing, decreasing, and unimodal shapes for the hrf of the MOTPW distribution in Figure 3. Also, we see a slightly different hrf with increasing, decreasing and increasing shape.

Graphics comparing the histograms from two simulated data sets and the MOTPW and MOGBXII densities of X under specified parameters are reported in Figure 4. They show good agreement between the simulated values and these densities.



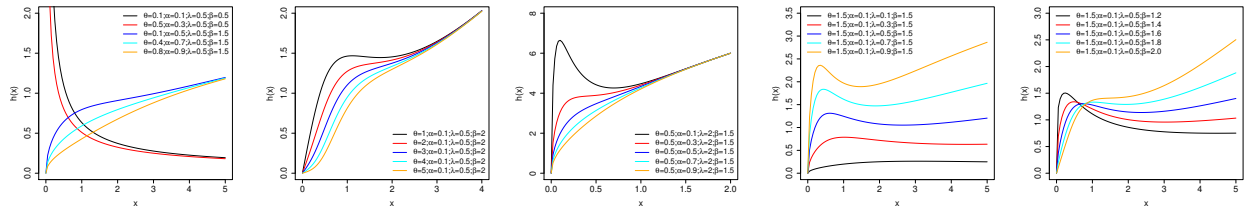
(a) (b) (c) (d)

Figure 1. Plots of the density and cumulative functions of the MOTPW distribution under four scenarios. (a) $\alpha = 30$, $\lambda = 2, \beta = 1.5$, and varying ϑ. (b) $\alpha = 30, \lambda = 2, \beta = 1.5$, and varying ϑ. (c) $\vartheta = 0.09, \lambda = 2, \beta = 1.5$, and varying α. (d) $\vartheta = 0.09, \lambda = 2, \beta = 1.5$, and varying α.



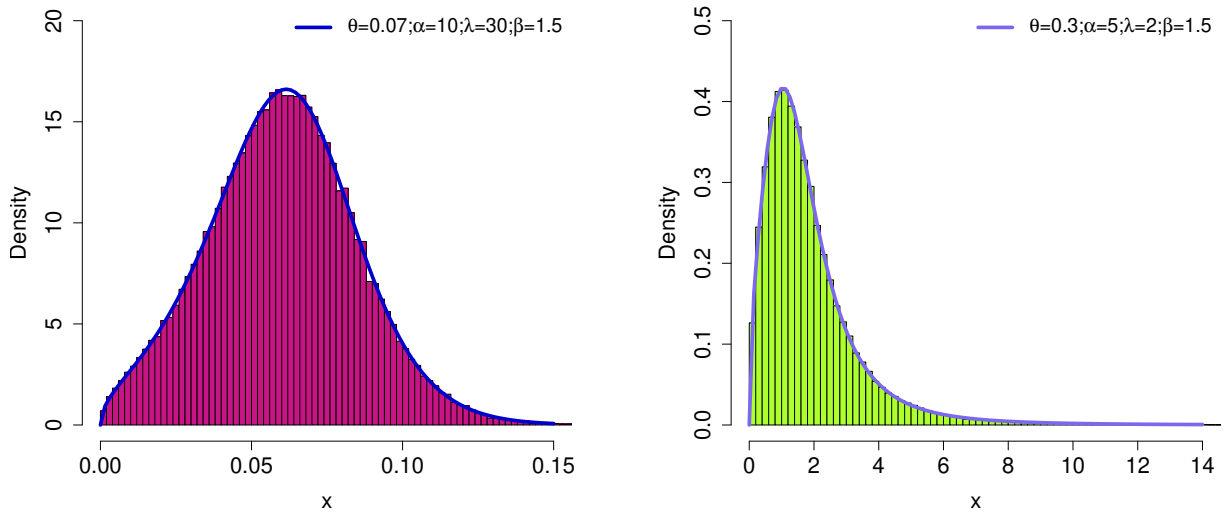
(a) (b) (c) (d)

Figure 2. Plots of the density and cumulative functions of the MOGBXII distribution under four scenarios. (a) $\alpha = 10$, $\lambda = 2, \beta = 1.5$, and varying ϑ. (b) $\alpha = 10, \lambda = 2, \beta = 1.5$, and varying ϑ. (c) $\vartheta = 0.9, \lambda = 2, \beta = 1.5$, and varying α. (d) $\vartheta = 0.9, \lambda = 2, \beta = 1.5$, and varying α.



(a) (b) (c) (d) (e)

Figure 3. Plots of the hrf of the MOTPW model.



(a) **(b)**

Figure 4. Plots of the MOTPW (a) and MOGBXII (b) densities and histograms of simulated data.

EXPANSIONS

We obtain useful linear representations for the density functions of X and Y for two separated cases $\alpha \in (0,1)$ and $\alpha > 1$. For $\alpha = 1$, we have $H(z;1,\boldsymbol{\tau}) = G(z;\boldsymbol{\tau})$.

By inserting (1) in Equation (5) and letting $\bar{G}(x) = \bar{G}(x;\boldsymbol{\tau})$, we can write

$$F_X(x) = \sum_{n=1}^{\infty} \frac{p_n\, G(x)^n}{[1-\bar{\alpha}\bar{G}(x)]^n}. \tag{16}$$

First, we consider the density of the maximum X when $\alpha \in (0,1)$. For $|z| < 1$ and $n = 1, 2, \ldots$, the negative binomial expansion holds

$$(1-z)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} (-z)^k. \tag{17}$$

Expanding $[1-\bar{\alpha}\bar{G}(z)]^{-n}$ as in Equation (17) since $\alpha \in (0,1)$, we have

$$F_X(x) = \sum_{n=1}^{\infty}\sum_{k=0}^{\infty} \binom{-n}{k} (-\bar{\alpha})^k\, p_n\; G(x)^n\, [1-G(x)]^k.$$

Henceforth, let $T_S \sim$ exp-G(s) be the exponentiated-G (exp-G) random variable with power parameter $s > 0$. Its cdf and pdf are $\Pi_S(x) = \Pi_S(x;\boldsymbol{\tau}) = G(x;\boldsymbol{\tau})^S$ and $\pi_S(x) = \pi_S(x;\boldsymbol{\tau}) = s\, G(x;\boldsymbol{\tau})^{s-1}\, g(x;\boldsymbol{\tau})$, respectively. Many exp-G properties have been studied exhaustively by several authors (Tahir & Nadarajah 2015). We can write

$$F_X(x) = \sum_{n=1}^{\infty} w_{n,0}\, \Pi_n(x) + \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} w_{n,k}\, \Pi_n(x)\, [1-G(x)]^k,$$

where $w_{n,k} = w_{n,k}(\alpha, \vartheta) = \binom{-n}{k} (-\bar{\alpha})^k p_n$ for $n = 1, 2, \ldots$ and $k = 0, 1, \ldots$ Further, using the binomial theorem, we obtain

$$F_X(x) = \sum_{n=1}^{\infty} w_{n,0}\, \Pi_n(x) + \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \sum_{i=0}^{k} w_{n,k,i}\, \Pi_{n+i}(x),$$

where $w_{n,k,i} = (-1)^i \binom{k}{i} w_{n,k}$ for $i = 0, 1, \ldots, k$.

By differentiating the last equation, we obtain

$$f_X(x) = \sum_{n=1}^{\infty} w_{n,0}\, \pi_n(x) + \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \sum_{i=0}^{k} w_{n,k,i}\, \pi_{n+i}(x). \tag{18}$$

We now move to the density of the maximum X when $\alpha > 1$. We modify the denominator in (16)

$$F_X(x) = \sum_{n=1}^{\infty} \frac{p_n\, G(x)^n}{\alpha^n \left[1 - (1 - \alpha^{-1}) G(x)\right]^n}$$

and then apply Equation (17) to find

$$F_X(x) = \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} v_{n,k}\, \Pi_{n+k}(x),$$

where $v_{n,k} = v_{n,k}(\alpha, \vartheta) = (-1)^k \binom{-n}{k} \alpha^{-n} (1-\alpha^{-1})^k p_n$ (for $n = 1, 2, \ldots$ and $k = 0, 1, \ldots$). By differentiating $F_X(x)$, the density of X follows as

$$f_X(x) = \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} v_{n,k}\, \pi_{n+k}(x). \tag{19}$$

Next, we consider the density of the minimum Y. By inserting (2) in Equation (6), we have

$$F_Y(y) = 1 - \sum_{n=1}^{\infty} \frac{\alpha^n\, p_n\, \bar{G}(y)^n}{\left[1 - \bar{\alpha} \bar{G}(y)\right]^n}. \tag{20}$$

For $\alpha \in (0, 1)$, we apply expansion (17) in the last equation to

$$F_Y(y) = 1 - \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} q_{n,k}\, \bar{G}(y)^{n+k},$$

where $q_{n,k} = q_{n,k}(\alpha, \vartheta) = (-1)^k \binom{-n}{k} \bar{\alpha}^k \alpha^n p_n$ for $n = 1, 2, \ldots$ and $k = 0, 1, \ldots$

By using the binomial theorem in $\bar{G}(y)^{n+k}$, we have

$$F_Y(y) = 1 + \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} \sum_{i=0}^{n+k} q_{n,k,i}\, \Pi_i(y),$$

where $q_{n,k,i} = (-1)^{i+1} \binom{n+k}{i} q_{n,k}$ for $i = 0, 1, \ldots, n + k$.

By differentiating $F_Y(y)$, the density of Y can be expressed as

$$f_Y(y) = \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} \sum_{i=1}^{n+k} q_{n,k,i}\, \pi_i(y). \tag{21}$$

We now obtain the density of Y when $\alpha > 1$. By changing the denominator in Equation (20), we have

$$F_Y(y) = 1 - \sum_{n=1}^{\infty} \frac{p_n\, \bar{G}(y)^n}{\left[1 - (1-\alpha^{-1})G(y)\right]^n}.$$

Applying expansion (17) in the last equation

$$F_Y(y) = 1 - \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} t_{n,k}\, \bar{G}(y)^n\, G(y)^k,$$

where $t_{n,k} = t_{n,k}(\alpha, \vartheta) = (-1)^k\, (1-\alpha^{-1})^k \binom{-n}{k} p_n$ (for $n = 1, 2, \ldots$ and $k = 0, 1, \ldots$).

Using the binomial theorem, we can rewrite $F_Y(y)$ as

$$F_Y(y) = 1 + \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} \sum_{i=0}^{n} t_{n,k,i}\, \Pi_{i+k}(y),$$

where $t_{n,k,i} = (-1)^{i+1} \binom{n}{i} t_{n,k}$ for $i = 0, 1, \ldots$ By simple differentiation

$$f_Y(y) = \sum_{n=1}^{\infty} \sum_{k=0}^{\infty} \sum_{i=0}^{n} t_{n,k,i}\, \pi_{i+k\geq 1}(y), \tag{22}$$

where $\pi_{i+k\geq 1}(y)$ is the exp-G density with power parameter $i + k \geq 1$.

Equations (18), (19), (21) and (22) are the main results of this section. These linear representations have great utility for deriving structural properties of the maximum X and minimum Y from well-known exp-G properties. More than thirty five exp-G models have been studied so far and then it is possible to construct at least three hundred fifty (70×5) MOPS-G models with properties determined from those exp-G properties. We can use statistical platforms with ten terms to have precise results.

NUMERICAL EVALUATION

In order to evaluate the analytical results presented in the previous sections, a package was implemented using the R programming language (R Core Team 2022). The **MarshallOlkinPSG** package was constructed in a generic way, that is, its most important functions allow generalizations for any baseline G distribution or even inform a zero-truncated PS distribution.

The library code can be obtained from GitHub at https://github.com/prdm0/MarshallOlkinPSG. On the library's website (see https://prdm0.github.io/MarshallOlkinPSG) it is possible to have more information on the functions implemented through the documentation and usage examples.

To install the package hosted and maintained on GitHub, it is necessary to previously install the **remotes** library. With the prerequisite met, the package **MarshallOlkinPSG** can be installed as:

```
# Install the remotes package:
# install.packages("remotes")
remotes::install_github("prdm0/MarshallOlkinPSG", force = TRUE)
```

The function `eq_19()` implements Equation (19) and compares, for example, with the exact MOTPW density in Equation (9). To facilitate comparison, the function `pdf_theorical()` implements this density function. By doing `help(eq_19)` it is possible to access an example of comparison of the two equations. Note that Equation (19) approximates (9) very well when finite sums are taken in applied problems. In other words, the results achieved by the function `eq_19()` approximates very well those from `pdf_theorical()`. The function `eq_19()` will also allow any baseline cdf $G(x)$ as an argument of `eq_19()`.

The function `eval_plot_moptw()` allows to validating numerically Equation (19) by means of plots. The true parameters for the MOTPW density are: $\alpha = 1.20$, $\vartheta = 1.50$, $\beta = 1.33$, and $\lambda = 2$. In addition, we require just a few terms in the sums to obtain a reasonable level of precision as shown in the plots in Figure 5, where six or eight terms provide very accurate approximations.

PROPERTIES

We now provide some mathematical properties of T_S that can be easily utilized in the linear representations of the previous section to find the corresponding properties of X and Y.

The nth ordinary moment of T_S has the form

$$\mu'_n = E(T_S^n) = s \int_{-\infty}^{\infty} t^n \, G(t;\boldsymbol{\tau})^{s-1} \, g(t;\boldsymbol{\tau}) dt = s \int_0^1 Q_G(u)^n \, u^{s-1} du, \tag{23}$$

where $Q_G(u;\boldsymbol{\tau}) = G^{-1}(u;\boldsymbol{\tau})$ is the qf of G.

Explicit expressions for several exp-G moments can be determined from (23).

The nth incomplete moment of T_S follows the previous algebra

$$m_n(y) = E(T_S^n | T_S < y) = s \int_0^{G(y;\boldsymbol{\tau})} Q_G(u)^n u^{s-1} du, \tag{24}$$

where the integral can be calculated for the great majority of G distributions. The first incomplete moment $m_1(y)$ is the most important case of (24) to find mean deviations and Lorenz and Bonferroni curves.

The moment generating function (mgf) of T_S follows as

$$M(w) = E(\mathrm{e}^{\mathrm{w T_s}}) = \mathrm{s} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{wt}} \, \mathrm{G(t;\boldsymbol{\tau})^{s-1}} \, \mathrm{g(t;\boldsymbol{\tau})dt} = \mathrm{s} \int_0^1 \exp\left[\mathrm{w\, Q_G(u)}\right] \, \mathrm{u^{s-1} du}. \tag{25}$$

The mgfs of exp-G distributions con be determined from Equation (25).

ESTIMATION

The MLEs are appropriate at least in large samples to determine confidence intervals for the parameters. We consider the random variable X defined from Equations (3) and (5) for any baseline

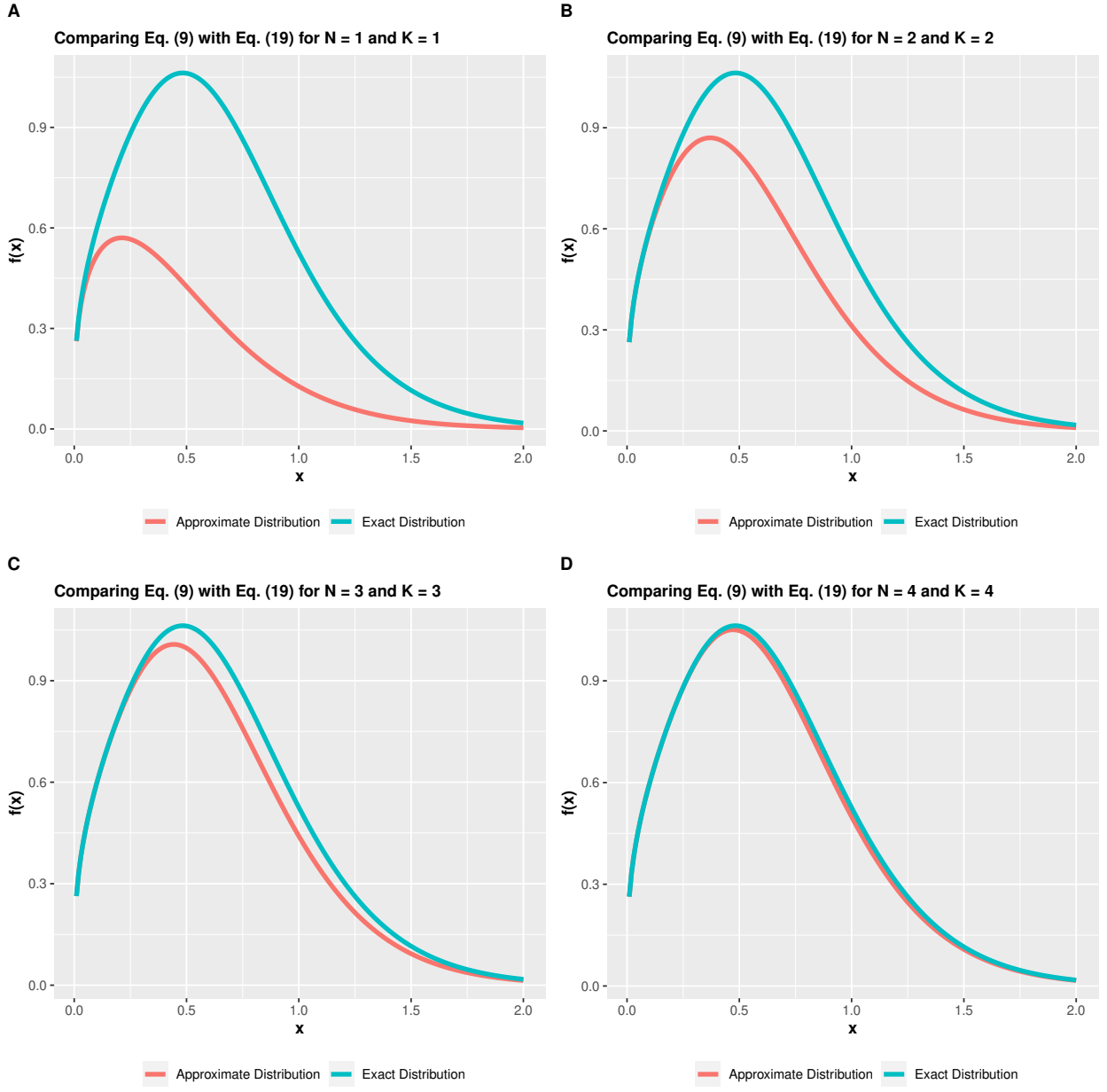


Figure 5. Numerical evaluation of (19) with finite sums, where *N* and *K* denote the upper limits of terms in the related sums with the running indices *n* and *k*, respectively.

G with any unknown parameter vector $\psi = (\alpha, \vartheta, \boldsymbol{\tau})^T$. By simple differentiation of (5), the density of X takes the form

$$f_X(x;\alpha,\vartheta,\boldsymbol{\tau}) = \frac{\alpha\,\vartheta\,g(x;\boldsymbol{\tau})\,C'\left(\vartheta\,H(x;\alpha,\boldsymbol{\tau})\right)}{C(\vartheta)\left[1-\bar{\alpha}\bar{G}(x;\boldsymbol{\tau})\right]^2}, \tag{26}$$

where $C'(\cdot)$ follows from (4) and $H(x;\alpha,\boldsymbol{\tau}) = G(x;\boldsymbol{\tau})/\left[1-\bar{\alpha}\,\bar{G}(x;\boldsymbol{\tau})\right]$.

The log-likelihood function for ψ from a random sample $x_1, \ldots, x_n$ of X is

$$\begin{aligned} \ell &= \ell(\psi) = \log\left[\frac{\alpha\vartheta}{C(\vartheta)}\right] + \sum_{i=1}^{n}\log\left[g(x_i;\boldsymbol{\tau})\right] + \sum_{i=1}^{n}\log\left[C'\left(\vartheta\, H(x_i;\alpha,\boldsymbol{\tau})\right)\right] \\ &- 2\sum_{i=1}^{n}\log\left[1-\bar{\alpha}\bar{G}(x;\boldsymbol{\tau})\right]. \end{aligned} \tag{27}$$

A similar development can be conducted for the random variable Y defined from Equation (6) for any baseline G.

We can find the MLE $\widehat{\psi}$ by maximizing Equation (27) using the `MaxBFGS` sub-routine (`Ox` program), `optim` function (`R`), and `PROC NLMIXED` (`SAS`). The *AdequacyModel* package can also maximize (27) using the PSO (particle swarm optimization) approach from the quasi-Newton BFGS, Nelder-Mead and simulated-annealing methods to maximize the log-likelihood function and it does not require initial values. Details are available at Marinho et al. (2019) and https://github.com/prdm0/AdequacyModel.

These scripts can be executed for a wide range of initial values and may lead to more than one maximum. However, in these cases, we consider the MLEs corresponding to the largest value of the maximum log-likelihood. There are sufficient conditions for the existence of these estimates such as compactness of the parameter space and the concavity of the log-likelihood function, but they can exist even when the conditions are not satisfied. In general, there is no explicit solution for the estimates from maximizing (27), but we can establish theoretical conditions on their existence and uniqueness for very special models by examining the ranges of the score components.

REGRESSION

Consider that $X_1, \ldots, X_n$ are independent random variables from any distribution in (9) assuming that the parameters λ and λ vary through them. We propose a new regression based on the response variable in (9) with the systematic components

$$\lambda_i = \exp(\mathbf{v}_i^T\eta_1) \qquad \text{and} \qquad \beta_i = \exp(\mathbf{v}_i^T\eta_2), \quad i = 1, \ldots, n, \tag{28}$$

respectively, where $\mathbf{v}_i^T = (v_{i1}, \ldots, v_{ip})$, $\eta_1 = (\eta_{11}, \ldots, \eta_{1p})^T$ and $\eta_2 = (\eta_{21}, \ldots, \eta_{2p})^T$. Equations (9) and (28) define the MOTPW regression. For $\alpha = 1$, it follows the *truncated Poisson Weibull* (TPW) regression.

In a similar manner, we can construct many other regressions based on other MOPS-G distributions defined from Equations (5) and (6).

The log-likelihood function for the vector $\psi = (\alpha, \vartheta, \eta_1^T, \eta_2^T)^T$ from the MOTPW regression can be reduced to

$$\begin{aligned} l(\psi) &= n\log\left[\frac{\alpha\vartheta}{\exp(\vartheta)-1}\right] + \sum_{i=1}^{n}\log(\beta_i) + \sum_{i=1}^{n}\beta_i\log(\lambda_i) + \sum_{i=1}^{n}(\beta_i-1)\log(x_i) - \\ &\sum_{i=1}^{n}(\lambda_i x_i)^{\beta_i} - \sum_{i=1}^{n}\log\left\{1-\bar{\alpha}\exp[-(\lambda_i x_i)^{\beta_i}]\right\} + \vartheta\sum_{i=1}^{n}\frac{\{1-\exp[-(\lambda_i x_i)^{\beta_i}]\}}{\{1-\bar{\alpha}\exp[-(\lambda_i x_i)^{\beta_i}]\}}. \end{aligned} \tag{29}$$

We obtain the MOTPW distribution for $\lambda_i = \lambda$ and $\beta_i = \beta$.

Let $\widehat{\psi}$ be the MLE of ψ. Equation (29) can also be maximized using the `gamlss` regression framework (Stasinopoulos & Rigby 2008) in `R`.

TWO SIMULATION STUDIES

We perform two simulation studies. The first one examines the accuracy of the MLEs of the parameter estimates in the MOTPW distribution. The second one does the same for the MOTPW regression.

The MOTPW distribution

First, we evaluate the precision of the estimates in the MOTPW distribution based on 1,000 Monte Carlo simulations using the R software. The simulation procedure follows as:

- The inverse function $Q(u) = F^{-1}(u)$ comes from (7)

$$Q(u) = \lambda^{-1}\left\{-\log\left[\frac{\vartheta - \log[u\exp(\vartheta) - u + 1]}{\vartheta + \alpha\log[u\exp(\vartheta) - u + 1] - \log[u\exp(\vartheta) - u + 1]}\right]\right\}^{\frac{1}{\beta}}. \quad (30)$$

- Generate $u \sim U(0,1)$ and obtain the values $x = Q(u)$ of the MOTPW distribution.

The true parameters are $\lambda = 3$, $\beta = 1$, $\vartheta = 1.5$ and $\alpha = 0.7$. The average estimates (AEs), biases, and mean squared errors (MSEs) are listed in Table I. The three measures decrease steadily when n becomes large.

Table I. Simulation results for the MOTPW distribution.

	$n = 100$			$n = 250$		
Parameter	AE	Bias	MSE	AE	Bias	MSE
λ	3.001	0.001	0.005	2.996	-0.00	0.005
β	0.998	-0.002	0.013	1.004	0.004	0.012
ϑ	1.585	0.085	0.081	1.567	0.0667	0.064
α	0.569	-0.130	0.090	0.563	-0.137	0.089

	$n = 500$			$n = 1,000$		
Parameter	AE	Bias	MSE	AE	Bias	MSE
λ	2.994	-0.006	0.004	2.995	-0.006	0.003
β	1.006	0.006	0.008	1.007	0.007	0.005
ϑ	1.546	0.046	0.035	1.526	0.026	0.016
α	0.597	-0.103	0.077	0.632	-0.068	0.063

The MOTPW regression

We perform some Monte Carlo simulations for some values of n to investigate the accuracy of the MLEs in the MOTPW regression under four scenarios: Scenario 1: $\vartheta = 0.6$ and $\alpha = 0.4$; Scenario 2: $\vartheta = 0.6$ and $\alpha = 1.4$; Scenario 3: $\vartheta = 1.7$ and $\alpha = 0.4$; Scenario 4: $\vartheta = 1.7$ and $\alpha = 1.4$. We take values greater than and less than one for ϑ and α.

The explanatory variables $v_1, \ldots, v_n$ are generated in the regression by taking $\lambda_i = 0.5 + 0.8\, v_i$, $\beta_i = 0.3 + 0.1\, v_i$, and $v_i \sim \textit{Bernoulli}(0.5)$.

For each scenario and value of n, one thousand samples are generated from the MOTPW regression fitted to each generated data set. The quantities reported in Table II are in good agreement with the asymptotic results for the MLEs.

Residual analysis

We investigate the quantile residuals (qrs) to verity the adequacy of the response distribution to determine outliers in the MOTPW regression. The same approach can be adopted to many other regressions defined from the distributions in (5) and (6). The qrs are given by (Dunn & Smyth 1996)

$$qr_i = \Phi^{-1}\left\{[\exp(\vartheta) - 1]^{-1}\exp\left\{\vartheta\frac{1 - \exp[-(\lambda_i x_i)^{\beta_i}]}{1 - \bar{\alpha}\exp[-(\lambda_i x_i)^{\beta_i}]}\right\} - 1\right\}, \tag{31}$$

where $\Phi(\cdot)$ is the normal cdf and λ_i and β_i are defined in Equation (28).

We consider the same scenarios for the simulations in Section Two Simulation Studies. For each fitted regression, the qrs are calculated from Equation (31). Figures 6, 7, 8, and 9 display QQ plots which show that the empirical distribution of these residuals is close to the standard normal distribution.

APPLICATIONS

The beta Weibull (BW) and Kumaraswamy Weibull (KwW) distributions have been widely used to fit real data in the last ten years or so. We compare the MOTPW distribution with the BW and KwW distributions since all of them have four parameters. The BW density pioneered by Lee et al. (2007) is

$$f(x) = \frac{c\lambda^c}{B(a,b)}x^{c-1}\exp\{-b(\lambda x)^c\}[1 - \exp\{-(\lambda x)^c\}]^{a-1},\ x > 0,$$

where all parameters are positive.

The KwW density introduced by Cordeiro & de Castro (2011) has the form

$$f(x) = a\, b\, c\, \lambda^c x^{c-1}\exp\{-(\lambda x)^c\}[1 - \exp\{-(\lambda x)^c\}]^{a-1}\{1 - [1 - \exp\{-(\lambda x)^c\}]^a\}^{b-1},\ x > 0,$$

where all parameters are positive.

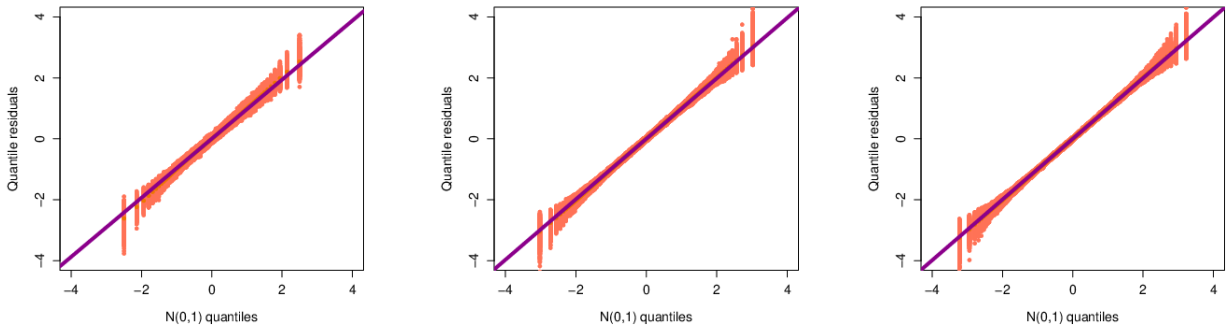
Application 1: Hourly dollar wage data

The first application refers to hourly dollar wages for $n = 534$ US workers. These data are obtained from the *SemiPar* package (Wand et al. 2005). Table III lists the estimates, standard errors (SEs) in parentheses, and three classical statistics. The lowest values of these measures reveal that the MOTPW is the best model. Next, the likelihood ratio (LR) statistic for comparing the MOTPW and TPW models is 6.159 (p-value < 0.013) which supports the wider distribution.

Figure 10a shows the histogram and the estimated MOTPW density. Figure 10b provides the empirical function and estimated MOTPW cdf, thus revealing that this distribution is appropriate for these data.

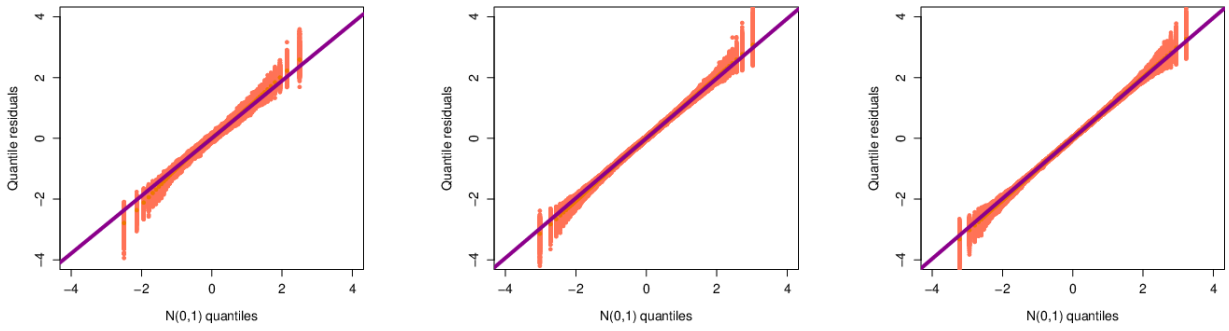
Table II. Simulation results for the MOTPW regression.

scenario 1									
	$n = 100$			$n = 500$			$n = 1,000$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
γ_{10}	0.614	0.114	0.074	0.557	0.057	0.029	0.527	0.027	0.017
γ_{11}	0.785	-0.015	0.031	0.792	-0.008	0.006	0.798	-0.002	0.002
γ_{20}	0.256	-0.044	0.031	0.271	-0.029	0.012	0.285	-0.015	0.007
γ_{21}	0.101	0.001	0.031	0.101	0.001	0.006	0.102	0.002	0.003
ϑ	0.734	0.134	0.164	0.651	0.051	0.093	0.621	0.021	0.070
α	0.477	0.077	0.089	0.440	0.040	0.072	0.413	0.013	0.067
scenario 2									
	$n = 100$			$n = 500$			$n = 1,000$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
γ_{10}	0.684	0.184	0.149	0.567	0.067	0.045	0.528	0.028	0.025
γ_{11}	0.779	-0.021	0.044	0.790	-0.009	0.007	0.798	-0.002	0.004
γ_{20}	0.235	-0.065	0.045	0.272	-0.028	0.015	0.289	-0.011	0.009
γ_{21}	0.096	-0.004	0.035	0.103	0.003	0.007	0.100	0.000	0.003
ϑ	0.637	0.037	0.157	0.578	-0.023	0.086	0.558	-0.042	0.077
α	1.722	0.322	0.382	1.530	0.130	0.180	1.467	0.067	0.123
scenario 3									
	$n = 100$			$n = 500$			$n = 1,000$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
γ_{10}	0.337	-0.161	0.079	0.483	-0.017	0.028	0.494	-0.007	0.019
γ_{11}	0.819	0.019	0.019	0.802	0.002	0.005	0.799	-0.001	0.002
γ_{20}	0.465	0.165	0.069	0.323	0.023	0.014	0.311	0.012	0.009
γ_{21}	0.094	-0.006	0.033	0.101	0.001	0.005	0.101	0.001	0.003
ϑ	1.349	-0.350	0.258	1.643	-0.057	0.035	1.679	-0.022	0.015
α	0.460	0.060	0.083	0.429	0.029	0.079	0.407	0.007	0.069
scenario 4									
	$n = 100$			$n = 500$			$n = 1,000$		
Parameter	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
γ_{10}	0.549	0.049	0.132	0.551	0.051	0.036	0.495	-0.005	0.015
γ_{11}	0.796	-0.004	0.038	0.795	-0.006	0.006	0.798	-0.002	0.003
γ_{20}	0.332	0.032	0.054	0.286	-0.014	0.012	0.307	0.007	0.006
γ_{21}	0.096	-0.004	0.032	0.100	0.000	0.005	0.103	0.003	0.003
ϑ	1.406	-0.294	0.222	1.643	-0.057	0.029	1.684	-0.016	0.013
α	1.913	0.513	0.739	1.604	0.204	0.240	1.408	0.008	0.090



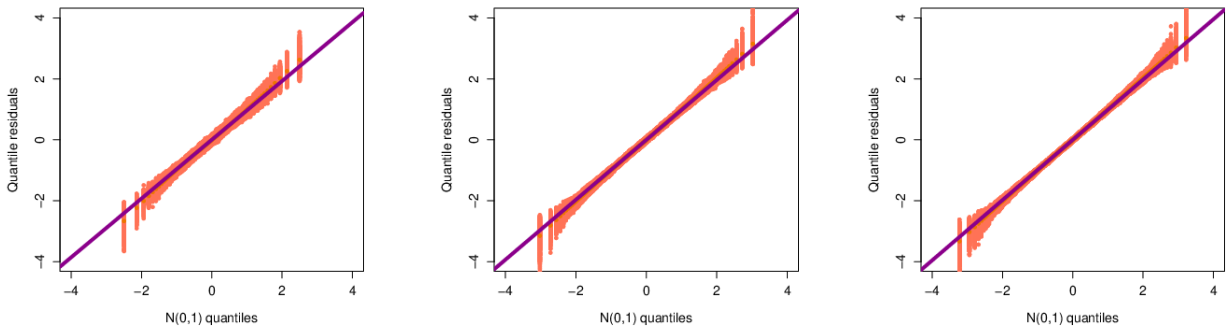
(a) **(b)** **(c)**

Figure 6. QQ plots for scenario 1 ($\vartheta = 0.6$ and $\alpha = 0.4$). (a) $n = 100$. (b) $n = 500$. (c) $n = 1,000$.



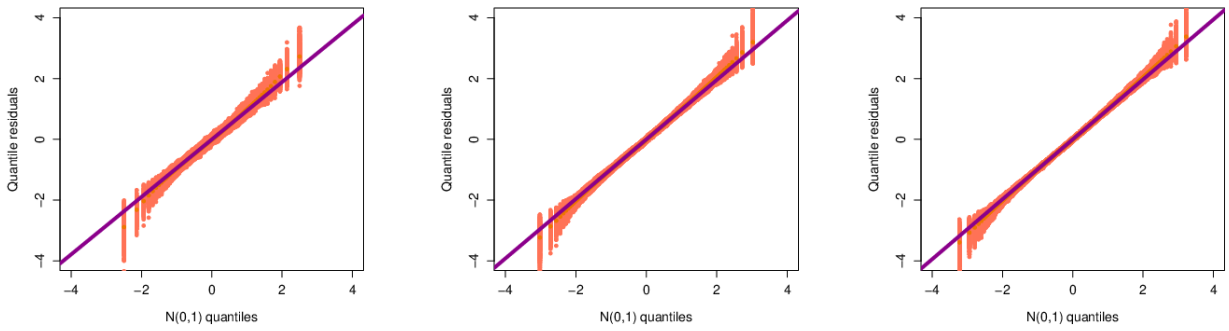
(a) **(b)** **(c)**

Figure 7. QQ plots for scenario 2 ($\vartheta = 0.6$ and $\alpha = 1.4$). (a) $n = 100$. (b) $n = 500$. (c) $n = 1,000$.



(a) **(b)** **(c)**

Figure 8. QQ plots for scenario 3 ($\vartheta = 1.7$ and $\alpha = 0.4$). (a) $n = 100$. (b) $n = 500$. (c) $n = 1,000$.



(a) **(b)** **(c)**

Figure 9. QQ plots for scenario 4 ($\vartheta = 1.7$ and $\alpha = 1.4$). (a) $n = 100$. (b) $n = 500$. (c) $n = 1,000$.

Table III. Results for hourly dollar wage data.

Model	log(λ)	log(β)	ϑ	α	AIC	BIC	GD
MOTPW	-2.720	0.694	11.210	0.019	**3031.288**	**3048.410**	**3023.288**
	(0.141)	(0.085)	(3.020)	(0.007)			
TPW	0.248	-0.541	31.100	(-)	3035.448	3048.289	3029.448
	(0.444)	(0.112)	(12.540)	(-)			
Model	log(λ)	*a*	*b*	log(*c*)	AIC	BIC	GD
KwW	-0.601	12.124	0.317	0.060	3034.039	3051.160	3026.039
	(0.023)	(0.802)	(0.013)	(0.014)			
BW	-5.453	2.327	126.000	0.216	3084.086	3101.208	3076.086
	(0.020)	(0.067)	(0.009)	(0.007)			

Application 2: Diabetes data

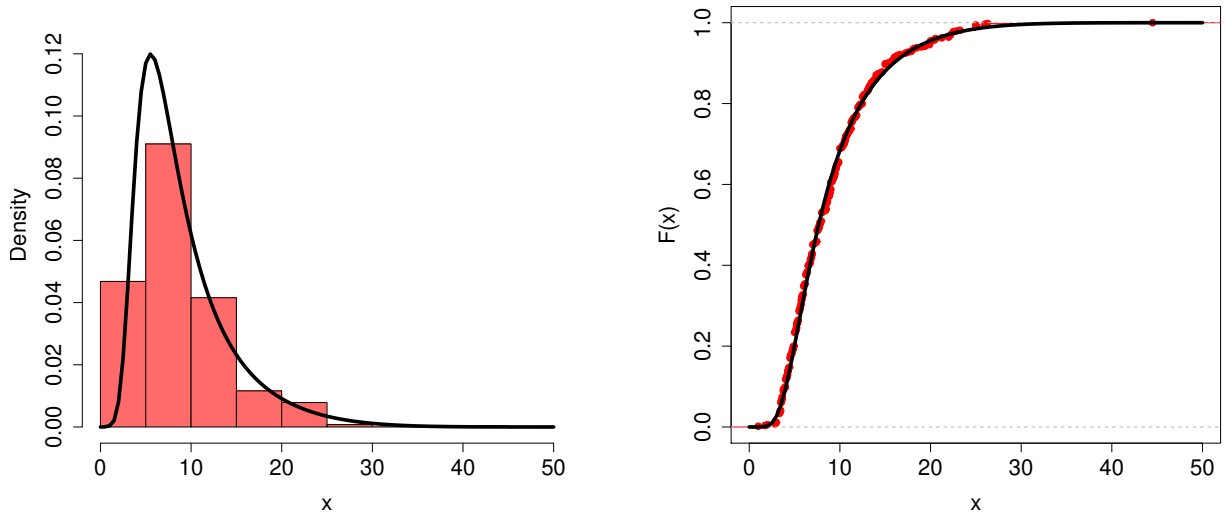
We consider two variables from the data reported by Reaven & Miller (1979): the response x_i is the relative weight defined by the ratio between the actual weight and the expected weight (given the person's height), and the explanatory variable v_{i1} indicates the diagnostic group (0 =normal, 1= chemical diabetes, 2 = overt diabetes). The diagnostic group has three levels and then we have two dummy variables (d_{ij}) (for $i = 1, \ldots, 145$ and $j = 1, 2$). The objective is to know what are the relations among the relative weight and the levels of the diagnostic group.

The systematic components for the MOTPW regression are

$$\lambda_i = \exp(\eta_{10} + \eta_{11} d_{i1} + \eta_{12} d_{i2}) \quad \text{and} \quad \beta_i = \exp(\eta_{20} + \eta_{21} d_{i1} + \eta_{22} d_{i2}), \quad i = 1, \ldots, 145.$$

The measures for the fitted regressions are reported in Table IV. Clearly, the MOTPW is the best regression for these data.

Table V provides the estimates, SEs and p-values for the best regression.



(a) **(b)**

Figure 10. (a) Estimated MOTPW pdf. (b) Estimated MOTPW cdf and the empirical cdf.

Table IV. Measures for diabetes data.

Model	AIC	BIC	GD
MOTPW	**-194.316**	**-170.502**	**-210.316**
TPW	-191.726	-170.889	-205.726
KwW	-188.769	-164.955	-204.769
BW	-185.607	-161.793	-201.607

Table V. Results for diabetes data.

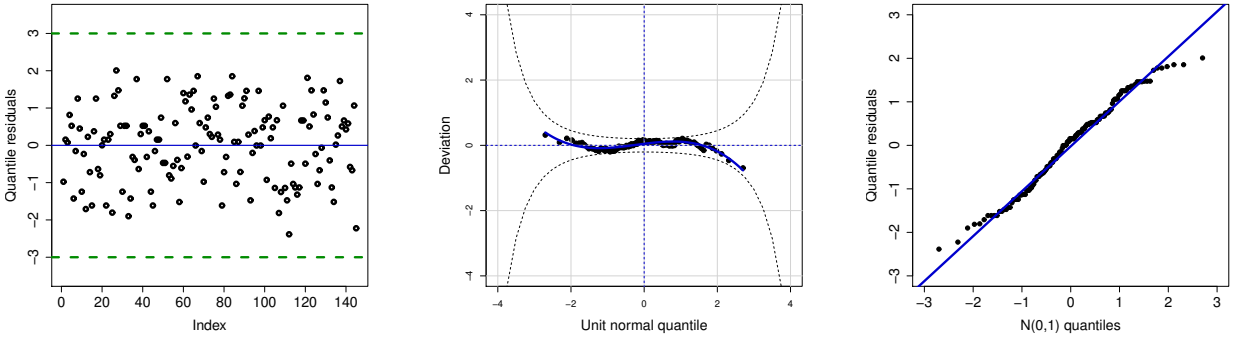
Parameter	Estimate	SE	p-Value
η_{10}	0.065	0.093	0.489
η_{11}	-0.119	0.036	0.001
η_{12}	-0.049	0.028	0.082
η_{20}	1.719	0.245	<0.001
η_{21}	0.373	0.140	0.009
η_{22}	0.131	0.141	0.355
ϑ	12.401	8.866	
α	0.095	0.079	

We note that the co-variable d_{i1} is significant and d_{i2} is not. So, there is a real difference between normal and chemical diabetes groups in relation to relative weight and no difference between normal and overt diabetes groups to relative weight. The same findings can be seen in Figure 12.

The LR statistic to compare the MOTPW and TPW regressions is $w = 4.590$ (p-value=0.032) that indicates that the fist regression is superior to the second regression to these data in terms of model fitting.

The plot of the residuals reported in Figure 11a does not detect outliers and departures from the general assumptions. The worm plot (Buuren & Fredriks 2001) of the residuals in Figure 11b and the QQ plot displayed in Figure 11c show the adequacy of the MOTPW regression for the current data.

A graphical comparison from the estimated cdfs in Figure 12 also supports the regression analysis.



(a) **(b)** **(c)**

Figure 11. **(a) Residual plot. (b) Worm plots. (c) QQ plot.**

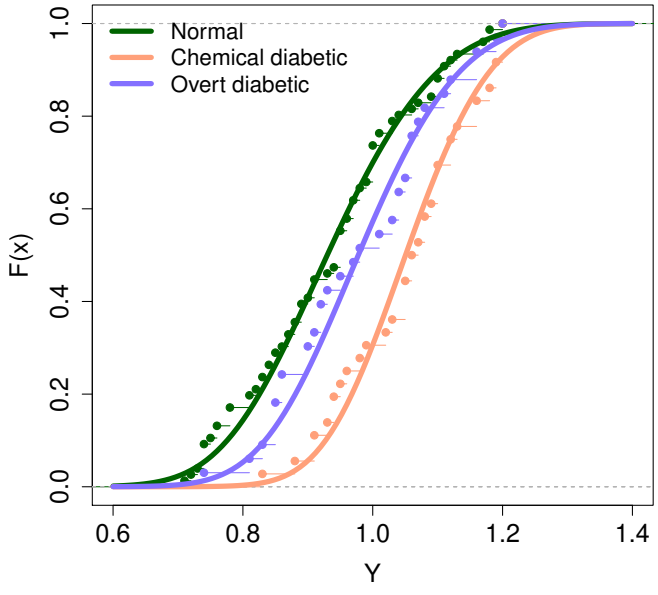


Figure 12. **Estimated cdf and the empirical cdf.**

CONCLUSIONS

We define two flexible Marshall–Olkin–Power-Series (MOPS) families of continuous distributions which can be very useful to fit real data. They are obtained by combining the Marshall–Olkin class (Marshall & Olkin 1997) and the power series distribution. Hundreds of continuous distributions can be easily formulated from the two families. We discuss some special distributions and maximum likelihood estimation. We introduce the *Marshall–Olkin Truncated Poisson Weibull* regression associated with one of the families. Some mathematical properties of these families are presented. We provide a package implemented in R software which can be used to determine numerically some mathematical properties for any distribution in the new families. The utility of the proposed models is proved empirically in two applications.

Acknowledgments

We gratefully acknowledge from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil.

REFERENCES

BUUREN SV & FREDRIKS M. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. Stat Med 20(8): 1259-1277.

CORDEIRO GM & DE CASTRO M. 2011. A new family of generalized distributions. J Stat Comput Simul 81(7): 883-898.

DUNN PK & SMYTH GK. 1996. Randomized quantile residuals. J Comput Graph Stat 5(3): 236-244.

LEE C, FAMOYE F & OLUMOLADE O. 2007. Beta-Weibull distribution: some properties and applications to censored data. J Mod Appl Stat Meth 6(1): 173-186.

MARINHO PRD, SILVA RB, BOURGUIGNON M, CORDEIRO GM & NADARAJAH S. 2019. AdequacyModel: An R package for probability distributions and general purpose optimization. PloS one 14(8): e0221487.

MARSHALL AW & OLKIN I. 1997. A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika 84(3): 641-652.

R CORE TEAM. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

REAVEN G & MILLER R. 1979. An attempt to define the nature of chemical diabetes using a multidimensional analysis. Diabetologia 16(1): 17-24.

STASINOPOULOS DM & RIGBY RA. 2008. Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Soft 23: 1-46.

TAHIR MH & NADARAJAH S. 2015. Parameter induction in continuous univariate distributions: Well-established G families. An Acad Bras Cienc 87: 539-568.

WAND M, COULL B, FRENCH J, GANGULI B, KAMMANN E, STAUDENMAYER J & ZANOBETTI A. 2005. SemiPar 1.0. R package. URL http://cran.r-project.org.

How to cite

CORDEIRO GM, VASCONCELOS JCS, ORTEGA EMM & MARINHO PRD. 2022. A competitive family to the Beta and Kumaraswamy generators: Properties, Regressions and Applications. An Acad Bras Cienc 94: e20201972. DOI 10.1590/0001-3765202220201972.

Manuscript received on December 28, 2020;
accepted for publication on June 7, 2021

GAUSS M. CORDEIRO[1]
http://orcid.org/0000-0002-3052-6551

JULIO CEZAR S. VASCONCELOS[2]
https://orcid.org/0000-0001-6794-3175

EDWIN M.M. ORTEGA[2]
https://orcid.org/0000-0003-3999-7402

PEDRO RAFAEL D. MARINHO[3]
http://orcid.org/0000-0003-1591-8300

[1]Universidade Federal de Pernambuco, Departamento de Estatística, Avenida Professor Moraes Rego, s/n, Bairro Iputinga, 50670-901 Recife, PE, Brazil
[2]Universidade de São Paulo, Departamento de Ciências Exatas, Avenida Pádua Dias, 11, Bairro São Dimas, 13418-900 Piracicaba, SP, Brazil
[3]Universidade Federal da Paraíba, Departamento de Estatística, Cidade Universitária, s/n, Bairro Castelo Branco, 58054-900 João Pessoa, PB, Brazil

Correspondence to: **Julio Cezar Souza Vasconcelos**
E-mail: juliocezarvasconcelos@hotmail.com

Author contributions

Gauss M. Cordeiro has contributed part of conceptualization, methodology, writing review and editing. Julio C. S. Vasconcelos has contributed part of Software, methodology and applications. Edwin M. M. Ortega has contributed part of formal analysis, methodology, writing original draft preparation. Pedro Rafael D. Marinho made contributions to the Section Numerical Evaluation, where he implemented the Marshall Olkin PSG computational library that numerically evaluates several mathematical properties presented in the article, allowing readers interested in the article to numerically verify the mathematical properties or even adapt the code to their interests. Also worked in Section Properties.