Journal of Data Science **10**(2012), 195-224

The Kumaraswamy Generalized Half-Normal Distribution for Skewed Positive Data

Gauss M. Cordeiro[1], Rodrigo R. Pescim[2] and Edwin M. M. Ortega[2*]
[1] *UFPE and* [2] *ESALQ - USP*

Abstract: For the first time, we propose and study the Kumaraswamy generalized half-normal distribution for modeling skewed positive data. The half-normal and generalized half-normal (Cooray and Ananda, 2008) distributions are special cases of the new model. Various of its structural properties are derived, including explicit expressions for the density function, moments, generating and quantile functions, mean deviations and moments of the order statistics. We investigate maximum likelihood estimation of the parameters and derive the expected information matrix. The proposed model is modified to open the possibility that long-term survivors may be presented in the data. Its applicability is illustrated by means of four real data sets.

Key words: Expected information, generalized half-normal distribution, half-normal distribution, hazard rate function, Kumaraswamy distribution, maximum likelihood estimation, mean deviation.

1. Introduction

For an arbitrary baseline cumulative distribution function (cdf) $G(x)$, Cordeiro and de Castro (2011) defined the probability density function (pdf) $f(x)$ and the cdf $F(x)$ of the Kumaraswamy-G ("Kw-G" for short) distribution by

$$f(x) = a\,b\,g(x)\,G^{a-1}(x)\,\{1 - G^a(x)\}^{b-1}, \tag{1}$$

and

$$F(x) = 1 - \{1 - G^a(x)\}^b, \tag{2}$$

respectively, where $g(x) = dG(x)/dx$ and $a > 0$ and $b > 0$ are additional shape parameters to the distribution of G. If X is a random variable with density (1),

[*]Corresponding author.

we write $X \sim$ Kw-G(a, b). Except for some special choices of the functions $g(x)$ and $G(x)$, the density function $f(x)$ will be difficult to deal with some generality. One major benefit of the Kw-G distribution is its ability of fitting skewed data that can not be properly fitted by existing distributions. This fact was demonstrated recently by Cordeiro *et al.* (2010) who apply the Kumaraswamy Weibull distribution to failure data.

A physical interpretation of (2) (for a and b positive integers) is as follows. Consider a system formed by b independent series components and that each component is made up of a parallel independent subcomponents. The system fails if any of the b components fails and that each component fails if all of the a subcomponents fail. The time to failure distribution of the entire system has precisely the Kw-G distribution.

The most popular models used to describe the lifetime process under fatigue are the half-normal (HN) and Birnbaum–Saunders (BS) distributions. When modeling monotone hazard rates, the HN and BS distributions may be initial choices because of their negatively and positively skewed density shapes. However, they do not provide a reasonable parametric fit for modeling phenomenon with non-monotone failure rates such as the bathtub shaped and the unimodal failure rates, which are common in reliability and biological studies. Such bathtub hazard curves have nearly flat middle portions and the corresponding densities have a positive anti-mode. The distributions which allow a bathtub fit are sufficiently complex (Nelson, 2004) and usually require five or more parameters. However, more recently, Díaz-García and Leiva (2005) introduced a new family of generalized BS distributions based on contoured elliptically distributions and Cooray and Ananda (2008) defined the generalized half-normal (GHN) distribution derived from a model for static fatigue. The distribution studied here extends the last distribution.

The GHN density function (Cooray and Ananda, 2008) with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$ has the form (for $x > 0$)

$$g(x) = \sqrt{\frac{2}{\pi}}\,\left(\frac{\alpha}{x}\right)\,\left(\frac{x}{\theta}\right)^{\alpha}\,\exp\Big[-\frac{1}{2}\left(\frac{x}{\theta}\right)^{2\alpha}\Big]. \tag{3}$$

Its cdf depends on the error function

$$G(x) = 2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right] - 1 = \operatorname{erf}\left(\frac{\left(\frac{x}{\theta}\right)^{\alpha}}{\sqrt{2}}\right), \tag{4}$$

where

$$\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \quad \text{and} \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\mathrm{d}t.$$

Its nth moment is given by (Cooray and Ananda, 2008) $E(X^n) = \sqrt{\frac{2^{\frac{n}{\alpha}}}{\pi}}\Gamma\left(\frac{n+\alpha}{2\alpha}\right)\theta^n$, where $\Gamma(\cdot)$ is the gamma function. The HN distribution is a special sub-model when $\alpha = 1$.

In this article, we study a new four-parameter distribution refereed to as the Kumaraswamy generalized half-normal (Kw-GHN) distribution, which contains as sub-models the HN and GHN distributions. The new distribution due to its flexibility in accommodating bathtub-shape form of the hazard function could be an important model in a variety of problems in survival analysis. It is also suitable for testing goodness of fit of the particular cases. By inserting (3) and (4) in (1), the Kw-GHN density function (for $x > 0$), with four parameters $\alpha > 0$, $\theta > 0$, $a > 0$ and $b > 0$, is given by

$$f(x) = ab\sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{x}\right)\left(\frac{x}{\theta}\right)^{\alpha}\exp\left[-\frac{1}{2}\left(\frac{x}{\theta}\right)^{2\alpha}\right]\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a-1} \times\left[1-\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\right]^{b-1}. \tag{5}$$

If X is a random variable with density (5), we write $X \sim$ Kw-GHN(α, θ, a, b). The new model contains some important sub-models. For $a = b = 1$, it gives the GHN distribution. If $\alpha = 1$, it yields the Kumaraswamy half-normal (Kw-HN) distribution. If $b = 1$, it leads to the exponentiated generalized half-normal (EGHN) distribution. Further, if $a = b = 1$, in addition to $\alpha = 1$, it reduces to the HN distribution. The cdf and hazard rate function corresponding to (5) are

$$F(x) = 1-\left\{1-\operatorname{erf}\left(\frac{\left(\frac{x}{\theta}\right)^{\alpha}}{\sqrt{2}}\right)^{\mathrm{a}}\right\}^{b}, \tag{6}$$

and

$$h(x) = \frac{ab\sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{x}\right)\left(\frac{x}{\theta}\right)^{\alpha}e^{-\left(\frac{x}{\theta}\right)^{2\alpha}/2}\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a-1}\left[1-\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\right]^{b-1}}{1-\left[1-\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\right]^{b}}, \tag{7}$$

respectively. Plots of these functions for selected parameter values, including the GHN and HN sub-models, are given in Figures 1 and 2, respectively.

Quantile functions are in widespread use in general statistics and often find representations in terms of lookup tables for key percentiles. Let $x = Q(u)$ be the Kw-GHN quantile function derived by inverting (6). If $x = Q_N(u) = \Phi^{-1}(u)$ denotes the standard normal quantile function, we can obtain

$$Q(u) = \theta\,[Q_N(\{1+[1-(1-u)^{1/b}]^{1/a}\}/2)]^{1/\alpha}. \tag{8}$$

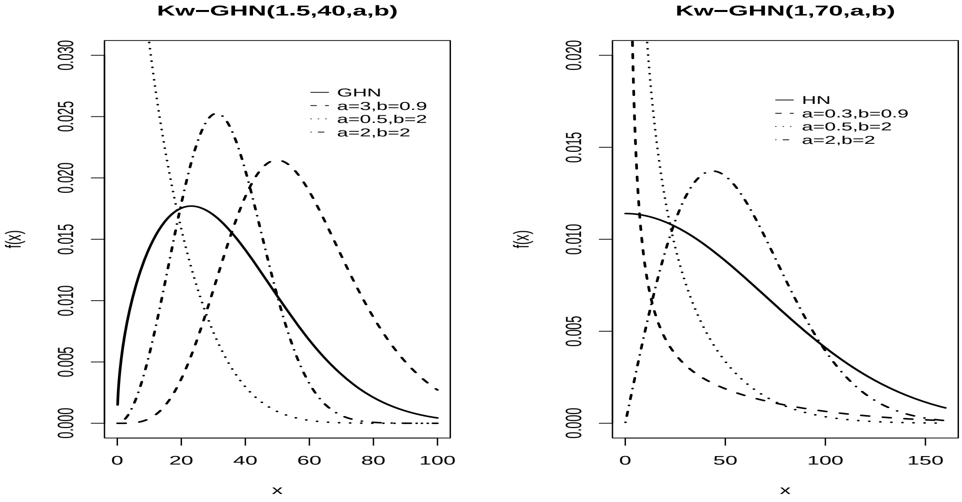

Figure 1: Plots of the density function (5) for some parameter values

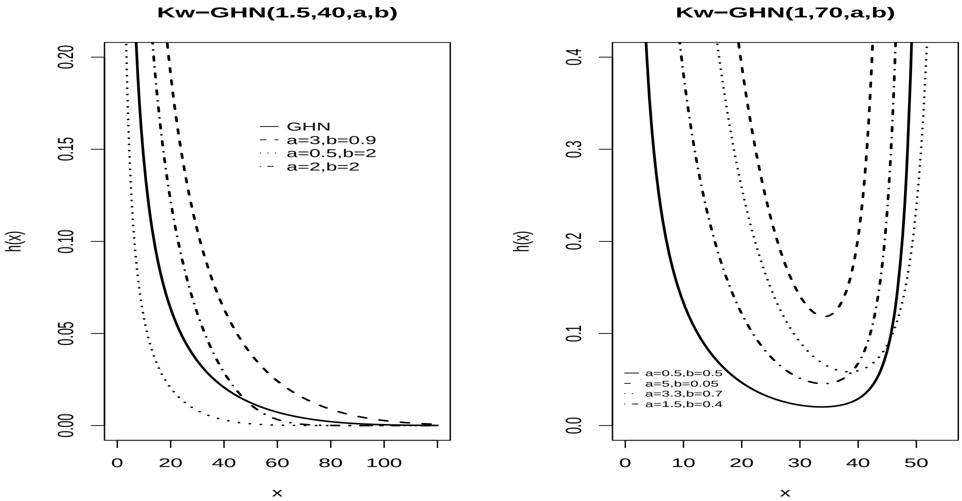

Figure 2: Plots of the hazard rate function (7) for some parameter values

Clearly, the Kw-GHN distribution is easily simulated by $X = Q(U)$, where U is an uniform random variable on the unit interval $(0, 1)$.

The article is organized as follows. In Section 2, we derive an expansion for the Kw-GHN density function. In Section 3, we study the behavior of the Bowley skewness and Moors kurtosis. In Section 4, we provide a general expansion for the moments. We derive power series expansions for the quantile and generating functions in Sections 5 and 6, respectively. Mean deviations are explored in Section 7. Expansions for the density of the order statistics and their moments are given in Section 8. The probability weighted moments (PWMs) are determined in Section 9. An alternative formula for moments of order statistics is given in Section 10. Maximum likelihood estimation is discussed in Section 11. In Section 12, we apply the proposed model for survival data with long-term survivors. Section 13 illustrates the importance of the new distribution applied to four real data sets. Finally, concluding remarks are given in Section 14.

2. Expansion for the Density Function

If $b > 0$ is a real non-integer, we can expand the binomial term in (1) to obtain

$$f(x) = \sum_{j=0}^{\infty} w_j \, h_{(j+1)a}(x), \tag{9}$$

where $w_j = (-1)^j \, (j+1)^{-1} \, b \binom{b-1}{j}$ and $h_a(x) = a \, G^{a-1}(x) \, g(x)$ denotes the exponentiated-G density with positive power a. We note that for $a > 1$ and $a < 1$ and for larger values of x, the multiplicative factor $a \, G^{a-1}(x)$ is greater and smaller than one, respectively. The reverse assertion is also true for smaller values of x. The latter immediately implies that the ordinary moments associated with the density function $f(x)$ are strictly larger (smaller) than those associated with the density $g(x)$ when $a > 1$ ($a < 1$).

(9) reveals that the Kw-G density function is a linear combination of exponentiated-G density functions. This result is useful to derive some mathematical properties of the Kw-G distribution from those of the exponentiated-G distributions. If $b > 0$ is an integer, the index j in the above sum stops at $b - 1$.

Some properties of the exponentiated distributions have been studied by several authors in recent years, see Mudholkar *et al.* (1995) for exponentiated Weibull, Gupta *et al.* (1998) for exponentiated Pareto, Gupta and Kundu (2001) for exponentiated exponential and Nadarajah and Gupta (2007) for exponentiated gamma distributions. We can apply (9) to the the EGHN density function with power $(j+1)a$ given by

$$h_{(j+1)a}(x) = (j+1)a\sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{x}\right)\left(\frac{x}{\theta}\right)^{\alpha}\exp\left[-\frac{1}{2}\left(\frac{x}{\theta}\right)^{2\alpha}\right]\left\{2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right]-1\right\}^{(j+1)a-1}. \tag{10}$$

We can obtain an expansion for $G(x)^{\beta}$ ($\beta > 0$ real non-integer) given by

$$G(x)^{\beta} = \sum_{r=0}^{\infty} s_r(\beta) \, G(x)^r, \tag{11}$$

where

$$s_r(\beta) = \sum_{j=r}^{\infty} (-1)^{r+j} \binom{\beta}{j} \binom{j}{r}.$$

Hence,

$$f(x) = g(x) \sum_{r=0}^{\infty} t_r \left\{ \operatorname{erf}\left(\frac{\left(\frac{x}{\theta}\right)^{\alpha}}{\sqrt{2}} \right) \right\}^{r}, \tag{12}$$

whose coefficients are

$$t_r = t_r(a,b) = a\, b \sum_{j=0}^{\infty} (-1)^j \binom{b-1}{j} s_r((j+1)a-1). \tag{13}$$

(12) is the main expansion for the Kw-GHN density function. This equation and other expansions derived here can be evaluated in symbolic computation software such as Mathematica and Maple. These symbolic software have currently the ability to deal with analytic expressions of formidable size and complexity.

3. Quantile Measures

The effect of the shape parameters a and b on the skewness and kurtosis of the new distribution can be considered based on quantile measures determined from (8). The Bowley skewness (Kenney and Keeping, 1962) is one of the earliest skewness measures defined by

$$B = \frac{Q(3/4) + Q(1/4) - 2Q(1/2)}{Q(3/4) - Q(1/4)}.$$

The Moors kurtosis (see Moors, 1988) based on octiles is defined by

$$M = \frac{Q(7/8) - Q(5/8) + Q(3/8) - Q(1/8)}{Q(6/8) - Q(2/8)}.$$

The measures B and M are less sensitive to outliers and they exist even for distributions without moments. For symmetric unimodal distributions, positive kurtosis indicates heavy tails and peakedness relative to the normal distribution, whereas negative kurtosis indicates light tails and flatness. For the normal distribution, $B = M = 0$.

In Figures 3 and 4, we plot the measures B and M for the Kw-GHN$(1.5, 40, a, b)$ distribution, as functions of a (for fixed b) and as functions of b (for fixed a), respectively. These plots indicate that the Bowley skewness always decreases when b increases (for fixed a) and first decreases steadily to a minimum value and then increases when a increases (for fixed b). On the other hand, the Moors kurtosis first decreases steadily to a minimum value and then increases when b increases (for fixed a) and always increases when a increases (for fixed b). So, these plots indicate that both measures can be very sensitive on these shape parameters, thus indicating the importance of the proposed distribution.

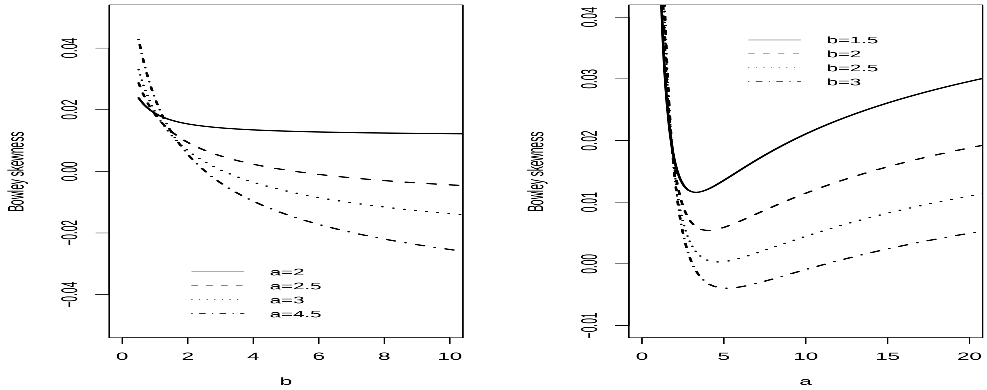

Figure 3: The Bowley skewness of the Kw-GHN distribution as function of b for some values of a and as function of a for some values of b

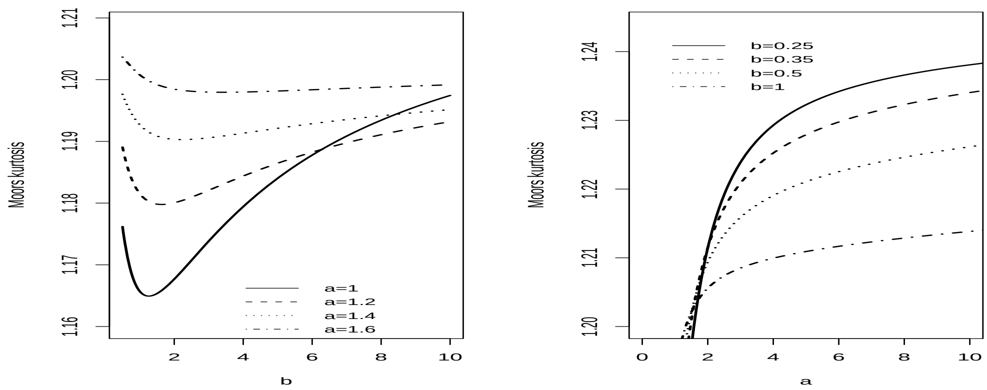

Figure 4: The Moors kurtosis of the Kw-GHN distribution as function of b for some values of a and as function of a for some values of b

The Bowley skewness can take positive and negative values. Tables 1 and 2 give the intervals for the parameter b (for fixed a) and for the parameter a (for fixed b) when the Bowley skewness is positive and negative. On the other hand, the Moors kurtosis always takes positive values. Tables 3 and 4 give the parameter b (for fixed a) and the parameter a (for fixed b) that yield the minimum Moors kurtosis.

Table 1: Intervals for the parameter b (for fixed a) when B is positive and negative

Some values of a	$B > 0$	$B < 0$
$a = 2$	$b > 0$	$\emptyset$
$a = 2.5$	$b \in (0, 5.280257)$	$b \in (5.280257, \infty)$
$a = 3$	$b \in (0, 3.231569)$	$b \in (3.231569, \infty)$
$a = 4.5$	$b \in (0, 2.584617)$	$b \in (2.584617, \infty)$

Table 2: Intervals for the parameter a (for fixed b) when B is positive and negative

Some values of a	$B > 0$	$B < 0$
$b = 1.5$	$a > 0$	$\emptyset$
$b = 2$	$a > 0$	$\emptyset$
$b = 2.5$	$a > 0$	$\emptyset$
$b = 3$	$a \in (0, 3.119547)$	$a \in (3.119547, \infty)$

Table 3: Values of the parameter b (for fixed a) when M has the minimum value

Some values of a	Minimum value for M
$a = 1$	$b = 1.275354$
$a = 1.2$	$b = 1.617821$
$a = 1.4$	$b = 2.198956$
$a = 1.6$	$b = 3.311047$

Table 4: Values of the parameter a (for fixed b) when M has the minimum value

Some values of b	Minimum value for M
$b = 0.25$	$a = 0.500965$
$b = 0.35$	$a = 0.500167$
$b = 0.5$	$a = 0.500001$
$b = 1$	$a = 0.500003$

4. Moments

Here and henceforth, let $X \sim \text{Kw-GHN}(\alpha, \theta, a, b)$. By setting $u = \left(\frac{x}{\theta}\right)^{\alpha}$, the nth moment of X comes from (12) as

$$E(X^n) = \theta^n \sqrt{\frac{2}{\pi}} \sum_{r=0}^{\infty} t_r \, I\left(\frac{n}{\alpha}, r\right),$$

where

$$I\left(\frac{n}{\alpha},r\right)=\int_0^\infty u^{\frac{n}{\alpha}}\exp\left(-\frac{u^2}{2}\right)\left[\operatorname{erf}\left(\frac{u}{\sqrt{2}}\right)\right]^r du.$$

From the power series expansion for the error function

$$\operatorname{erf}(x)=\frac{2}{\sqrt{\pi}}\sum_{m=0}^{\infty}\frac{(-1)^m\,x^{2m+1}}{(2m+1)\,m!},$$

and by calculating the resulting integral for any real number $r+n/\alpha$, we obtain

$$E(X^n)=\theta^n\sqrt{\frac{2}{\pi}}\sum_{r=0}^{\infty}t_r\,I\left(\frac{n}{\alpha},r\right), \tag{14}$$

where

$$I\left(\frac{n}{\alpha},r\right)=\pi^{-\frac{r}{2}}2^{r+\frac{n}{2\alpha}-\frac{1}{2}}\sum_{m_1,\cdots,m_r=0}^{\infty}\frac{(-1)^{m_1+\cdots+m_r}\Gamma\left(m_1+\cdots+m_r+\frac{r+\frac{n}{\alpha}+1}{2}\right)}{(m_1+1/2)\ldots(m_r+1/2)m_1!\cdots m_r!}. \tag{15}$$

Further, in the very special case when $r+n/\alpha$ is even, the integral $I\left(n/\alpha,r\right)$ can be expressed in terms of the Lauricella function of type A (Exton, 1978; Aarts, 2000) defined by

$$\begin{aligned}&F_A^{(n)}(a;b_1,\cdots,b_n;c_1,\cdots,c_n;x_1,\cdots,x_n)\\&=\textstyle\sum_{m_1=0}^{\infty}\cdots\sum_{m_n=0}^{\infty}\frac{(a)_{m_1+\cdots+m_n}(b_1)_{m_1}\cdots(b_n)_{m_n}}{(c_1)_{m_1}\cdots(c_n)_{m_n}}\frac{x_1^{m_1}\cdots x_n^{m_n}}{m_1!\cdots m_{n!}},\end{aligned}$$

where $(a)_k=a(a+1)\cdots(a+k-1)$ is the ascending factorial (with the convention that $(a)_0=1$). Numerical routines for the direct computation of the Lauricella function of type A are available, see Exton (1978) and Mathematica (Trott, 2006). Hence, $E(X^n)$ can be expressed in terms of the Lauricella functions of type A

$$E(X^n)=\theta^n\sqrt{\frac{2}{\pi}}\sum_{r=0}^{\infty}C_r\,F_A^{(r)}\left(\frac{r+\frac{n}{\alpha}+1}{2};\frac{1}{2},\cdots,\frac{1}{2};\frac{3}{2},\cdots,\frac{3}{2};-1,\cdots,-1\right),$$

where

$$C_r=\pi^{-\frac{r}{2}}\,2^{2r+\frac{n}{2\alpha}-\frac{1}{2}}\,\Gamma\left(\frac{r+\frac{n}{\alpha}+1}{2}\right)t_r.$$

An alternative expression for $E(X^n)$ follows from (9) as

$$E(X^n)=\sum_{j=0}^{\infty}w_j\,E(Y_{(j+1)a}^n), \tag{16}$$

where $Y_{(j+1)a} \sim$ EGHN$((j+1)a)$, and w_j and $\mathrm{h}_{(j+1)a}(x)$ are defined in (9) and (10), respectively. (16) gives the moments of the Kw-GHN distribution as an infinite linear combination of the corresponding EGHN moments.

The skewness and kurtosis measures can be calculated from the ordinary moments using well-known relationships. Plots of the skewness and kurtosis for some choices of b as function of a, and for some choices of a as function of b, for $\alpha = 1$ and $\theta = 45$, are given in Figures 5 and 6, respectively. These measures show a different behavior in terms of variation of the parameters a and b from the quantile measures studied in Section 3. In fact, these plots indicate that the skewness and kurtosis curves increase (decrease) with $b\,(a)$ for fixed $a\,(b)$.

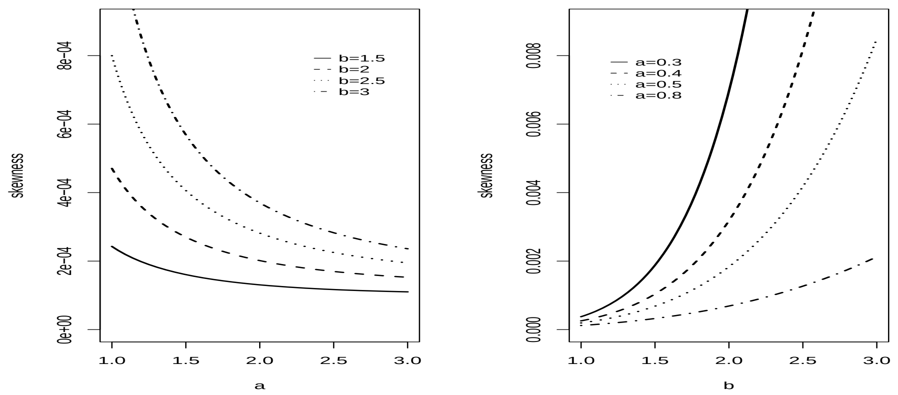

Figure 5: Skewness of the Kw-GHN distribution as function of a for some values of b and as function of b for some values of a

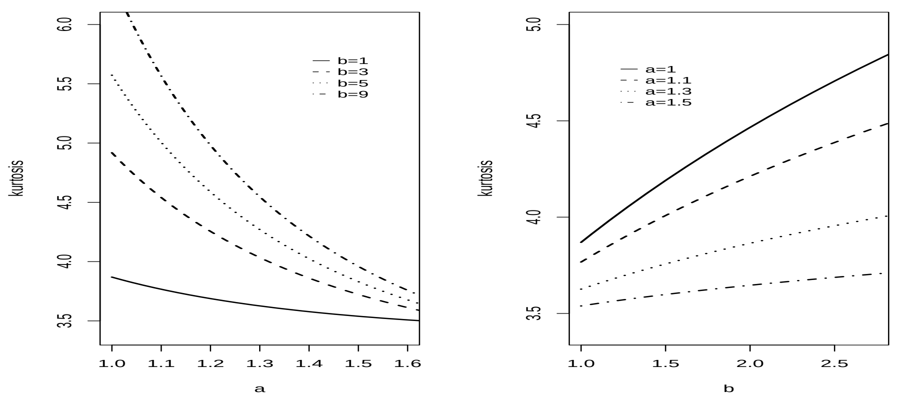

Figure 6: Kurtosis of the Kw-GHN distribution as function of a for some values of b and as function of b for some values of a

The pth descending factorial moment of X is

$$E(X^{(p)}) = E[X(X-1)\times\cdots\times(X-p+1)] = \sum_{n=0}^{p} s(p,n)\,E(X^n),$$

where $s(r,n) = (n!)^{-1}[d^n n^{(r)}/dx^n]_{x=0}$ is the Stirling number of the first kind. Other kinds of moments may also be obtained in closed-form, but we consider only the previous moments for reasons of space.

5. Quantile Expansion

Here, we obtain a power series expansion for the quantile function of X. By expanding the binomial terms in (8), we have

$$\frac{1}{2}\left\{1+[1-(1-u)^{1/b}]^{1/a}\right\} = \sum_{k=0}^{\infty} m_k\,u^k, \tag{17}$$

where $m_0 = [1+\sum_{j=0}^{\infty}(-1)^j\binom{1/a}{j}\binom{j/b}{j}]/2$ and $m_k = [\sum_{j=0}^{\infty}(-1)^{j+k}\binom{1/a}{j}\binom{j/b}{j}]/2$ for $k \geq 1$.

Following Steinbrecher (2002), the standard normal quantile function can be expanded as

$$Q_N(u) = \sum_{k=0}^{\infty} b_k\,w^{2k+1}, \tag{18}$$

where $w = \sqrt{2\pi}\,(u-1/2)$ and the quantities b_k can be calculated recursively from

$$b_{k+1} = \frac{1}{2(2k+3)}\sum_{r=0}^{k}\frac{(2r+1)\,(2k-2r+1)\,b_r\,b_{k-r}}{(r+1)\,(2r+1)}.$$

Here, $b_0 = 1$, $b_1 = 1/6$, $b_2 = 7/120$, $b_3 = 127/7560, \cdots$. The function $Q_N(u)$ can be expressed as a power series given by

$$Q_N(u) = \sum_{r=0}^{\infty} d_r\,u^r \tag{19}$$

where

$$d_r = \sum_{k=r}^{\infty}(-1/2)^{k-r}\binom{k}{r}\,e_k$$

and the quantities e_k are defined from the coefficients in (18) by $e_k = 0$ for $k = 0, 2, 4, \cdots$, and $e_k = (2\,\pi)^{k/2}\,b_{(k-1)/2}$ for $k = 1, 3, 5, \cdots$.

Combining (17) and (19), we obtain

$$Q_N\big(\{1+[1-(1-u)^{1/b}]^{1/a}\}/2\big)=\sum_{r=0}^{\infty}d_r\left(\sum_{k=0}^{\infty}m_k\,u^k\right)^r. \tag{20}$$

We use throughout an equation of Gradshteyn and Ryzhik (2000) for a power series raised to a positive integer j

$$\left(\sum_{i=0}^{\infty}a_i\,x^i\right)^j=\sum_{i=0}^{\infty}c_{j,i}\,x^i, \tag{21}$$

where the coefficients $c_{j,i}$ (for $i=1,2,\cdots$) are easily obtained from the recurrence equation

$$c_{j,i}=(ia_0)^{-1}\sum_{m=1}^{i}[m(j+1)-i]\,a_m\,c_{j,i-m}, \tag{22}$$

and $c_{j,0}=a_0^j$. From (20)-(22), we obtain

$$Q_N\big(\{1+[1-(1-u)^{1/b}]^{1/a}\}/2\big)=\sum_{k=0}^{\infty}h_k^\star\,u^k,$$

where $h_k^\star=\sum_{k=0}^{\infty}d_r\,g_{r,k}$ and the quantities $g_{r,k}$ are given by $g_{r,0}=m_0^r$ and $g_{r,k}=(k\,m_0)^{-1}\sum_{s=1}^{k}[s(r+1)-k]\,m_s\,g_{r,k-s}$ for $k\geq 1$. The argument of the standard normal quantile function guarantees that the sum $\sum_{k=0}^{\infty}h_k^\star\,u^k$ belongs to the interval $(0,5)$. Setting $h_k=h_k^\star/5$, the quantile function of X can be expressed as

$$Q(u)=\theta\,5^{1/\alpha}\left(\sum_{k=0}^{\infty}h_k\,u^k\right)^{1/\alpha},$$

and then it does involve a power series in the interval $(0,1)$. Using (11), we can write

$$\left(\sum_{k=0}^{\infty}h_k\,u^k\right)^{1/\alpha}=\sum_{r=0}^{\infty}s_r(\alpha^{-1})\left(\sum_{k=0}^{\infty}h_k\,u^k\right)^r,$$

where $s_r(\alpha^{-1})=\sum_{j=r}^{\infty}(-1)^{r+j}\binom{\alpha^{-1}}{j}\binom{j}{r}$. The above equation and (21) lead to

$$Q(u)=\sum_{k=0}^{\infty}v_k\,u^k, \tag{23}$$

where $v_k = \theta\, 5^{1/\alpha} \sum_{r=0}^{\infty} s_r(\alpha^{-1})\, q_{r,k}$ for $k \geq 0$, $q_{r,k} = (k\, h_0)^{-1} \sum_{m=1}^{k} [m(r+1) - k]\, h_m\, q_{r,k-m}$ for $k \geq 1$ and $q_{r,0} = h_0^r$. (23) is the main result of this section.

6. Generating Function

By setting $u = \left(\frac{x}{\theta}\right)^{\alpha}$, the moment generating function (mgf) of X can be obtained from (12) as

$$M(s) = \sqrt{\frac{2}{\pi}} \sum_{r=0}^{\infty} t_r \sum_{m=0}^{\infty} \frac{\theta^m\, s^m}{m!} \int_0^{\infty} u^{\frac{m}{\alpha}} \exp\left(-\frac{u^2}{2}\right) \left[\mathrm{erf}\left(\frac{\mathrm{u}}{\sqrt{2}}\right)\right]^r du.$$

Following similar lines of Section 4, $M(s)$ can be reduced to

$$M(s) = \sum_{m=0}^{\infty} \frac{A_m\, s^m}{m!}, \tag{24}$$

where

$$A_m = \sqrt{\frac{2}{\pi}}\, \theta^m \sum_{r=0}^{\infty} t_r\, \pi^{-\frac{r}{2}}\, 2^{r+\frac{m}{2\alpha}-\frac{1}{2}}\, I\left(\frac{m}{\alpha}, r\right),$$

and $I\left(\frac{m}{\alpha}, r\right)$ is defined by (15). Evidently, A_m gives a second representation for the mth moment of X.

An alternative equation for $M(s)$ can be derived from the quantile expansion (23). Using (21), we have

$$M(s) = \int_0^1 \exp\{s\, Q(u)\} du = \int_0^1 \sum_{k=0}^{\infty} \frac{s^k\, \left(\sum_{n=0}^{\infty} v_n\, u^n\right)^k}{k!} du = \sum_{k=0}^{\infty} \frac{B_k\, s^k}{k!}, \tag{25}$$

where $B_k = \sum_{n=0}^{\infty} d_{k,n}/(n+1)$ for $k = 0, 1, \cdots$, and $d_{k,n}$ can be calculated recursively by $d_{k,0} = v_0^k$ and $d_{k,n} = (i\, v_0)^{-1} \sum_{m=1}^{n} [m(k+1) - n]\, v_m\, d_{k,n-m}$ for $k \geq 1$. Clearly, B_k gives a third representation for the kth moment of X. (24) and (25) are the main results of this section.

7. Mean Deviations

We can derive the mean deviations about the mean $\mu = E(X)$ and about the median M of X from

$$\delta_1 = 2\big[\mu\, F(\mu) - J(\mu)\big] \qquad \text{and} \qquad \delta_2 = \mu - 2J(M), \tag{26}$$

where $J(q) = \int_0^q x\, f(x) dx$. The median M comes from (8) by $M = Q(1/2)$. By setting $u = (x/\theta)^{\alpha}$, this integral can be obtained from (12) as

$$J(q) = \theta \sqrt{\frac{2}{\pi}} \sum_{r=0}^{\infty} t_r \int_0^{(q/\theta)^{\alpha}} u^{\frac{1}{\alpha}} \exp\left(-\frac{u^2}{2}\right) \left[\mathrm{erf}\left(\frac{\mathrm{u}}{\sqrt{2}}\right)\right]^r du.$$

The error function erf $\left(\frac{\mathrm{u}}{\sqrt{2}}\right)$ admits the expansion $\left(\sum_{k=0}^{\infty} a_k u^k\right)$, where $a_{2k+1} = \frac{(-1)^k 2^{1-k}}{\sqrt{2\pi}(2k+1)k!}$ and $a_{2k} = 0$ for $k \in \mathbb{N}$. Thus,

$$J(q) = \theta \sqrt{\frac{2}{\pi}} \sum_{r,k=0}^{\infty} t_r \, c_{r,k} \int_0^{(q\,\theta^{-1})^{\alpha}} u^{k+\frac{1}{\alpha}} \exp\left(-\frac{u^2}{2}\right) du,$$

where the quantities $c_{r,k}$ are immediately calculated from the a_k's above by (22). Setting $v = u^2/2$, we can write

$$J(q) = \frac{2\theta}{\sqrt{\pi}} \sum_{r,k=0}^{\infty} t_r \, c_{r,k} \int_0^{(q\theta^{-1})^{2\alpha}/2} v^{(k+\alpha^{-1}-1)/2} \, \mathrm{e}^{-\mathrm{v}} \mathrm{dv}.$$

For $\lambda > 0$,

$$\int_0^y v^{\lambda-1} \, \mathrm{e}^{-\alpha\,\mathrm{v}} \mathrm{dv} = \alpha^{-\lambda} \, \gamma(\lambda, \alpha\,\mathrm{y}),$$

where $\gamma(a, y) = \int_0^y t^{a-1} \, \mathrm{e}^{-\mathrm{y}} \mathrm{dt}$ is the incomplete gamma function. Then,

$$J(q) = \frac{2\theta}{\sqrt{\pi}} \sum_{r,k=0}^{\infty} t_r c_{r,k} \gamma \left[(k + \alpha^{-1} + 1)/2, (q\theta^{-1})^{2\alpha}/2\right]. \tag{27}$$

From (26) and (27), we can obtain the mean deviations. These quantities immediately yield the Bonferroni and Lorenz curves that have applications in economics, reliability, demography, insurance and medicine. They are defined for a given probability π by

$$B(\pi) = \frac{J(q)}{\pi\, E(X)} \quad \text{and} \quad L(\pi) = \frac{J(q)}{E(X)},$$

respectively, where $q = Q(\pi)$ comes from the quantile function (8). Using (27), we can easily obtain $B(\pi)$ and $L(\pi)$. It is easy to verify that $L(\pi) \geq \pi$ and $L(0) = 0$ and $L(1) = 1$. In economics, if $\pi = F(q)$ is the proportion of units whose income is lower than or equal to q, $L(\pi)$ gives the proportion of total income volume accumulated by the set of units with an income lower than or equal to q.

8. Order Statistics

Order statistics make their appearance in many areas of statistical theory and practice. The density function $f_{i:n}(x)$ of the ith order statistic, for $i = 1, \cdots, n$, from i.i.d. random variables $X_1, \cdots, X_n$ following any Kw-G distribution, is simply given by

$$f_{i:n}(x) = \frac{f(x)}{B(i, n-i+1)} \sum_{j=0}^{n-i} (-1)^j \binom{n-i}{j} F(x)^{i+j-1},$$

where $B(\cdot,\cdot)$ denotes the beta function.

Cordeiro and de Castro (2011) derived an expression for the density function of the Kw-G order statistics as a function of the baseline density multiplied by infinite weighted sums of powers of $G(x)$. This result enables us to obtain the ordinary moments of the Kw-G order statistics as infinite weighted sums of PWMs of the G distribution. They demonstrated that

$$F(x)^{i+j-1} = \sum_{r=0}^{\infty} p_{i+j-1,r}\, G(x)^r, \tag{28}$$

where the coefficients $p_{u,r}(a,b)$ can be expressed as

$$p_{u,r} = p_{u,r}(a,b) = \sum_{k=0}^{u}(-1)^k \binom{u}{k} \sum_{m=0}^{\infty}\sum_{l=r}^{\infty}(-1)^{mr+l}\binom{kb}{m}\binom{ma}{l}\binom{l}{r} \tag{29}$$

for $r, u = 0, 1, \cdots$. Let

$$w_i = (-1)^i ab\binom{b-1}{i} \quad \text{and} \quad w_{i,j,r} = (-1)^{i+j+r}\, a\, b\binom{a(i+1)-1}{j}\binom{b-1}{i}\binom{j}{r}.$$

If a is an integer, Cordeiro and de Castro (2011) derived $f_{i:n}(x)$ in the form

$$f_{i:n}(x) = \frac{g(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{r,u=0}^{\infty} w_u p_{i+j-1,r} G(x)^{a(u+1)+r-1}, \tag{30}$$

whereas if a is a real non-integer, they obtained

$$f_{i:n}(x) = \frac{g(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{r,u,v=0}^{\infty}\sum_{t=0}^{v} w_{u,v,t}\, p_{i+j-1,r}\, G(x)^{r+t}. \tag{31}$$

(30) and (31) show that the density of the order statistics is a function of the baseline density multiplied by infinite weighted sums of powers of $G(x)$. These generalized moments for some baseline distributions can be accurate computationally by numerical integration as mentioned before. Let Y be a GHN random variable and $\tau_{s,r} = E[Y^s G(Y)^r]$ be the (s,r)th PWM of Y (for $r = 0, 1, \cdots$) as defined by Greenwood *et al.* (1979). We can calculate $\tau_{s,r}$ using the power series expansion given before as

$$\tau_{s,r} = \int_0^\infty x^s \mathrm{erf}\left(\frac{\left(\frac{x}{\theta}\right)^\alpha}{\sqrt{2}}\right)^r g(x)\,dx = \sum_{k=0}^{\infty} c_{r,k}\theta^{-k\alpha}\int_0^\infty x^{k\alpha+s} g(x)\,dx.$$

The integral in the last equation is the $k\alpha+s$ generalized moment of Y and then

$$\tau_{s,r}=\frac{\theta^s}{\sqrt{\pi}}\sum_{k=0}^{\infty}c_{r,k}\,2^{(k\alpha+s)/(2\alpha)}\,\Gamma\left(\frac{k\alpha+s+\alpha}{2\alpha}\right). \tag{32}$$

(32) is a new result for the PWMs of order (s,r) of the GHN distribution. Hence, the sth moment of the order statistics, say $E(X_{i:n}^s)$, can be obtained from (30), (31) and (32) as

$$E(X_{i:n}^s)=\frac{1}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{r,u=0}^{\infty}w_u\,p_{i+j-1,r}\,\tau_{s,a(u+1)+r-1},$$

and

$$E(X_{i:n}^s)=\frac{1}{B(i,n-i+1)}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}\sum_{r,u,v=0}^{\infty}\sum_{t=0}^{v}w_{u,v,t}\,p_{i+j-1,r}\,\tau_{s,r+t},$$

if a is an integer and a is a real non-integer, respectively. Clearly, these two equations should be used numerically with a large number in place of infinity.

9. Probability Weighted Moments

The (s,r)th PWM of X following the Kw-G distribution, say $\rho_{s,r}$, is formally defined by

$$\rho_{s,r}=E[X^sF(X)^r]=\int_{-\infty}^{\infty}x^sF^r(x)f(x)dx.$$

From (12) and (28), we can write

$$\rho_{s,r}=\sum_{r,m=0}^{\infty}t_rp_{m,r}\tau_{s,m+r}.$$

Using (32), we obtain

$$\rho_{s,r}=\frac{\theta^s}{\sqrt{\pi}}\sum_{r,m,k=0}^{\infty}2^{(k\alpha+s)/(2\alpha)}\Gamma\left(\frac{k\,\alpha+s+\alpha}{2\alpha}\right)t_rp_{m,r}c_{r+m,k}, \tag{33}$$

where the quantities t_r and $p_{m,r}$ can be calculated from (13) and (29), respectively. (33) is the main result of this section.

10. Alternative Formula for Moments of Order Statistics

We now offer an alternative formula for the moments of the order statistics based on the PWMs of the GHN distribution. We use the formula for the sth moment due to Barakat and Abdelkader (2004) applied to the independent and identically distributed case, subject to existence,

$$E(X_{i:n}^s) = s \sum_{j=n-i+1}^{n} (-1)^{j-n+i-1} \binom{j-1}{n-i} \binom{n}{j} I_j(s), \tag{34}$$

where $I_j(s)$ denotes the integral

$$I_j(s) = \int_0^\infty x^{s-1} \{1 - F(x)\}^j dx.$$

Using the binomial expansion and interchanging terms, the last integral becomes

$$I_j(s) = \sum_{p=0}^{j} (-1)^p \binom{j}{p} \rho_{s-1,p},$$

where $\rho_{s-1,p}$ is immediately obtained from (33).

Inserting the expression for $I_j(s)$ in (34) yields

$$E(X_{i:n}^s) = s \sum_{j=n-i+1}^{n} \sum_{p=0}^{j} (-1)^{j-n+i+p-1} \binom{j-1}{n-i} \binom{n}{j} \binom{j}{p} \rho_{s-1,p}. \tag{35}$$

Formula (35) is the main result of this section.

11. Estimation and Inference

The estimation of the model parameters is addressed by the method of maximum likelihood. If Y follows a Kw-GHN distribution with vector of parameters $\boldsymbol{\lambda} = (\alpha, \theta, a, b)^T$, the log-likelihood for $\boldsymbol{\lambda}$ from a single observation y of Y is given by

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \log(a) + \log(b) + \log\left(\sqrt{\frac{2}{\pi}}\right) + \log(\alpha) - \log(y) + \alpha \log\left(\frac{y}{\theta}\right) - \frac{1}{2}\left(\frac{y}{\theta}\right)^{2\alpha} \\ &+ (a-1)\log\left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right] - 1\right\} + (b-1)\log\left[1 - \left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right] - 1\right\}^{a}\right], \quad y > 0. \end{aligned}$$

The components of the unit score vector $\mathbf{U} = \left(\frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \theta}, \frac{\partial \ell}{\partial a}, \frac{\partial \ell}{\partial b}\right)^T$ are

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= \frac{1}{\alpha} + \log\left(\frac{y}{\theta}\right) - \log\left(\frac{y}{\theta}\right)\left(\frac{y}{\theta}\right)^{2\alpha} + \frac{2(a-1)}{\sqrt{2\pi}} \left\{ \frac{v \log\left(\frac{y}{\theta}\right)}{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right] - 1} \right\} \\ &+ \frac{2a(1-b)}{\sqrt{2\pi}} \left\{ \frac{v \log\left(\frac{y}{\theta}\right) \left[2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right] - 1\right]^{a-1}}{1 - \left[2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right] - 1\right]^{a}} \right\}, \end{aligned}$$

$$\frac{\partial \ell}{\partial \theta} = \frac{\alpha}{\theta}\left(\frac{y}{\theta}\right)^{2\alpha} - \left(\frac{\alpha}{\theta}\right) + \frac{2(1-a)}{\sqrt{2\pi}}\left\{\frac{v\left(\frac{\alpha}{\theta}\right)}{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1}\right\}$$
$$+ \frac{2a(1-b)}{\sqrt{2\pi}}\left\{\frac{v\left(\frac{\alpha}{\theta}\right)\left[2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right]^{a-1}}{1-\left[2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right]^{a}}\right\},$$

$$\frac{\partial \ell}{\partial a} = \frac{1}{a} + \log\left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right\}$$
$$+ (1-b)\left\{\frac{\left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\log\left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right\}}{1-\left[2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right]^{a}}\right\},$$

$$\frac{\partial \ell}{\partial b} = \frac{1}{b} + \log\left[1-\left\{2\Phi\left[\left(\frac{y}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\right],$$

where $v = \exp[-(y/\theta)^{2\alpha}/2](y/\theta)^{\alpha}$.

For a random sample $y = (y_1, \cdots, y_n)^T$ of size n from Y, the total log-likelihood is $\ell_n = \ell_n(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \ell^{(i)}(\boldsymbol{\lambda})$, where $\ell^{(i)}(\boldsymbol{\lambda})$ is the log-likelihood for the ith observation ($i = 1, \cdots, n$). The total score function is $\mathbf{U}_n = \sum_{i=1}^{n} \mathbf{U}^{(i)}$, where $\mathbf{U}^{(i)}$ has the form given before for $i = 1, \cdots, n$. The maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ is the solution of the system of nonlinear equations $\mathbf{U}_n = \mathbf{0}$. For interval estimation and tests of hypotheses on the parameters in $\boldsymbol{\lambda}$, we require the 4×4 unit expected information matrix

$$\mathbf{K} = \mathbf{K}(\boldsymbol{\lambda}) = \begin{pmatrix} \kappa_{\alpha,\alpha} & \kappa_{\alpha,\theta} & \kappa_{\alpha,a} & \kappa_{\alpha,b} \\ \kappa_{\theta,\alpha} & \kappa_{\theta,\theta} & \kappa_{\theta,a} & \kappa_{\theta,b} \\ \kappa_{a,\alpha} & \kappa_{a,\theta} & \kappa_{a,a} & \kappa_{a,b} \\ \kappa_{b,\alpha} & \kappa_{b,\theta} & \kappa_{b,a} & \kappa_{b,b} \end{pmatrix},$$

whose elements are given in the Appendix.

Under conditions that are fulfilled for parameters in the interior of the parameter space but not on the boundary, the asymptotic distribution of $\sqrt{n}\,(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is $\mathrm{N}_4(0, \mathbf{K}(\boldsymbol{\lambda})^{-1})$. The estimated asymptotic multivariate normal $\mathrm{N}_4(0, n^{-1}\mathbf{K}(\widehat{\boldsymbol{\lambda}})^{-1})$ distribution of $\widehat{\boldsymbol{\lambda}}$ can be used to construct approximate confidence intervals for the parameters and for the hazard rate and survival functions. An asymptotic confidence interval with significance level γ for each parameter λ_r is given by

$$\mathrm{ACI}(\lambda_r, 100(1-\gamma)\%) = (\hat{\lambda}_r - z_{\gamma/2}\sqrt{\hat{\kappa}^{\lambda_r,\lambda_r}}, \hat{\lambda}_r + z_{\gamma/2}\sqrt{\hat{\kappa}^{\lambda_r,\lambda_r}}),$$

where $\hat{\kappa}^{\lambda_r,\lambda_r}$ is the rth diagonal element of $n^{-1}\mathbf{K}(\boldsymbol{\lambda})^{-1}$ estimated at $\widehat{\boldsymbol{\lambda}}$, for $r = 1, \cdots, 4$, and $z_{\gamma/2}$ is the quantile $1 - \gamma/2$ of the standard normal distribution.

The likelihood ratio (LR) statistic is useful for comparing the new distribution with some of its special sub-models. For example, we may use the LR statistic to check if the fit using the Kw-GHN distribution is statistically "superior" to a fit using the GHN distribution for a given data set. In any case, considering the partition $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \boldsymbol{\lambda}_2^T)^T$, tests of hypotheses of the type $H_0 : \boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_1^{(0)}$ versus $H_A : \boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_1^{(0)}$ can be performed via the LR statistic $w = 2\{\ell(\widehat{\boldsymbol{\lambda}}) - \ell(\widetilde{\boldsymbol{\lambda}})\}$, where $\widehat{\boldsymbol{\lambda}}$ and $\widetilde{\boldsymbol{\lambda}}$ are the estimates of $\boldsymbol{\lambda}$ under H_A and H_0, respectively. Under the null hypothesis H_0, $w \xrightarrow{d} \chi_q^2$, where q is the dimension of the vector $\boldsymbol{\lambda}_1$ of interest. The LR test rejects H_0 if $w > \xi_\gamma$, where ξ_γ denotes the upper $100\gamma\%$ point of the χ_q^2 distribution.

12. A Kw-GHN Mixture Model for Survival Data with Cure Fraction

In population based cancer studies, cure is said to occur when the mortality in the group of cancer patients returns to the same level as that expected in the general population. The cure fraction is of interest to patients and also a useful measure when analyzing trends in cancer patient survival. Models for survival analysis typically assume that every subject in the study population is susceptible to the event under study and will eventually experience such event if the follow-up is sufficiently long. However, there are situations when a fraction of individuals are not expected to experience the event of interest, that is, those individuals are cured or not susceptible. For example, researchers may be interested in analyzing the recurrence of a disease. Many individuals may never experience a recurrence; therefore, a cured fraction of the population exists. Cure rate models have been used to estimate the cured fraction. These models are survival models which allow for a cured fraction of individuals. These models extend the understanding of time-to-event data by allowing for the formulation of more accurate and informative conclusions. These conclusions are otherwise unobtainable from an analysis which fails to account for a cured or insusceptible fraction of the population. If a cured component is not present, the analysis reduces to standard approaches of survival analysis. Cure rate models have been used for modeling time-to-event data for various types of cancers, including breast cancer, non-Hodgkins lymphoma, leukemia, prostate cancer and melanoma. Perhaps the most popular type of cure rate models is the mixture model (Berkson and Gage, 1952; Maller and Zhou, 1996). In this model, the population is divided into two sub-populations so that an individual either is cured with probability π, or has a proper survival function $S(x)$ with probability $1-\pi$. This formulation leads to an improper population survivor function $S^*(x)$ expressed in the mixture form

$$S^*(x) = \pi + (1-\pi)S(x), \quad S(\infty) = 0, \quad S^*(\infty) = \pi. \tag{36}$$

Common choices for $S(x)$ in (36) are the exponential and Weibull distributions. Here, we adopt the Kw-GHN distribution. Mixture models involving these distributions have been studied by several authors, including Farewell (1982), Sy and Taylor (2000) and Ortega *et al.* (2009). The book by Maller and Zhou (1996) provides a wide range of applications of the long-term survivor mixture model. The use of survival models with a cure fraction has become more and more frequent because traditional survival analysis do not allow for modeling data in which nonhomogeneous parts of the population do not represent the event of interest even after a long follow-up. Now, we propose an application of the new distribution to compose a mixture model for cure rate estimation.

Suppose that X_i's are independent and identically distributed random variables having the density function (5). Consider a sample $x_1, \cdots, x_n$, where x_i is either the observed lifetime or censoring time for the ith individual. Let a binary random variable z_i (for $i = 1, \cdots, n$) indicate that the ith individual in a population is at risk or not with respect to a certain type of failure, i.e. $z_i = 1$ indicates that the ith individual will eventually experience a failure event (uncured) and $z_i = 0$ indicates that the individual will never experience such event (cured). The proportion of uncured $1 - \pi$ individuals can be expressed such that the conditional distribution of z_i is $\Pr(z_i = 1) = 1 - \pi$. The probability that the individual i is cured is modeled by π and this proportion does not vary over the individuals.

The maximum likelihood method is used to estimate the parameters. So, the contribution of an individual that failed at x_i to the likelihood function reduces to

$$\begin{aligned}&(1-\pi)a\,b\sqrt{\frac{2}{\pi}}\left(\frac{\alpha}{x}\right)\left(\frac{x}{\theta}\right)^{\alpha}\exp\Big[-\frac{1}{2}\left(\frac{x}{\theta}\right)^{2\alpha}\Big]\Big\{2\Phi\Big[\left(\frac{x}{\theta}\right)^{\alpha}\Big]-1\Big\}^{a-1}\\&\times\Big[1-\Big\{2\Phi\Big[\left(\frac{x}{\theta}\right)^{\alpha}\Big]-1\Big\}^{a}\Big]^{b-1},\end{aligned}\tag{37}$$

whereas the contribution of an individual that is at risk at time x_i becomes

$$\pi+(1-\pi)\left\{1-\operatorname{erf}\left(\frac{\left(\frac{\mathrm{x}}{\theta}\right)^{\alpha}}{\sqrt{2}}\right)^{\mathrm{a}}\right\}^{b},\tag{38}$$

where the $\operatorname{erf}(\cdot)$ function is defined in Section 1. We refer to the new model (37) and (38) as the Kw-GHN mixture model with long-term survivors. For $a = b = 1$, we obtain a new model called the GHN mixture model with long-term survivors.

Thus, the log-likelihood function for the parameter vector $\boldsymbol{\lambda} = (a, b, \alpha, \theta, \pi)^T$

can be obtained from (37) and (38) as

$$l(\boldsymbol{\lambda}) = r\log\left[(1-\pi)a\,b\sqrt{2/\pi}\right] + \sum_{i\in F}\log\left(\frac{\alpha}{x_i}\right) + \alpha\sum_{i\in F}\log\left(\frac{x_i}{\theta}\right) - \frac{1}{2}\sum_{i\in F}\left(\frac{x_i}{\theta}\right)^{2\alpha}$$
$$+ (a-1)\sum_{i\in F}\log\left\{2\Phi\left[\left(\frac{x_i}{\theta}\right)^{\alpha}\right]-1\right\} + (b-1)\sum_{i\in F}\log\left[1-\left\{2\Phi\left[\left(\frac{x_i}{\theta}\right)^{\alpha}\right]-1\right\}^{a}\right]$$
$$+\sum_{i\in C}\log\left\{\pi+(1-\pi)\left[1-\operatorname{erf}\left(\frac{\left(\frac{\mathrm{x_i}}{\theta}\right)^{\alpha}}{\sqrt{2}}\right)^{\mathrm{a}}\right]^{b}\right\}, \qquad (39)$$

where F and C denote the sets of individuals corresponding to lifetime observations and censoring times, respectively, and r is the number of uncensored observations (failures).

13. Applications

Here, for the purpose of illustration, we analyze four data sets. We choose these data because they really show in different fields that it is necessary to have positively skewed distributions with non-negative support.

13.1 Uncensored Data

Description of the data sets.

A1 **Engineering**: The data refer to the failure times of 24 mechanical components as reported in Murthy *et al.* (2004).

A2 **Survival times**: The data analyzed by Kundu *et al.* (2008) and Leiva *et al.* (2009) correspond to 72 survival times of guinea pigs injected with different doses of tubercle bacilli.

A3 **Flood data**: The flood data ($n = 39$) for the Floyd River located in James, Iowa, USA. The Floyd River flood rates (for the years 1935-1973) were reported by Akinsete *et al.* (2008).

Table 5 gives a descriptive summary for these data showing different degrees of skewness and kurtosis.

Table 5: Descriptive statistics for the three data sets

Data	Mean	Median	Mode	SD	Variance	Skewness	Kurtosis	Min.	Max.
A1	22.97	19.24	10.24	10.76	115.67	1.44	1.99	10.24	51.56
A2	99.82	70.0	60.0	81.12	6580.12	1.84	2.89	12.0	376.0
A3	6771.10	3570.0	318.0	11695.68	1E+008	4.74	25.78	318	71500

First, in order to estimate the model parameters, we consider the maximum likelihood estimation method discussed in Section 11. We take the estimates of α and θ from the fitted GHN distribution as starting values for the numerical iterative procedure. The computations were performed using the NLMixed procedure in SAS. Table 6 lists the MLEs of the parameters and the values of the following statistics for some models: Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC) and Bayesian Information Criterion (BIC). The results indicate that the Kw-GHN model has the smallest values of these statistics among all fitted models. So, it could be chosen as the more suitable model.

Table 6: MLEs of the model parameters for the three data sets and the corresponding AIC, CAIC and BIC statistics

Data	Model	a	b	α	θ	AIC	CAIC	BIC
A1	Kw-GHN	133.60	0.2222	0.6035	2.4404	179.0	181.0	183.7
	GHN	1	1	1.6757	28.6011	185.1	185.6	187.4
	HN	1	1	1	25.2703	191.9	192.0	193.0
A2	Kw-GHN	28.4765	0.7361	0.3031	4.3399	788.2	788.8	797.3
	GHN	1	1	1.0163	129.24	807.5	807.6	812.0
	HN	1	1	1	128.27	805.5	805.6	807.8
A3	Kw-GHN	6.0482	14.5335	0.1979	9384.0	768.1	769.3	774.7
	GHN	1	1	0.6436	13384	781.7	782.0	785.0
	HN	1	1	1	13384	799.8	799.9	801.4

A comparison of the proposed distribution with some of its sub-models using LR statistics is shown in Table 7. The p-values indicate that the proposed model yields the best fit to the three data sets.

Table 7: LR statistics for the three data sets

Data	Model	Hypotheses	Statistic w	p-value
A1	Kw-GHN vs GHN	$H_0 : a = b = 1$ vs $H_1 : H_0$ is false	10.10	0.00640
	Kw-GHN vs HN	$H_0 : a = b = \alpha = 1$ vs $H_1 : H_0$ is false	18.90	0.00028
A2	Kw-GHN vs GHN	$H_0 : a = b = 1$ vs $H_1 : H_0$ is false	23.30	<0.0001
	Kw-GHN vs HN	$H_0 : a = b = \alpha = 1$ vs $H_1 : H_0$ is false	25.30	<0.0001
A3	Kw-GHN vs GHN	$H_0 : a = b = 1$ vs $H_1 : H_0$ is false	17.60	0.0002
	Kw-GHN vs HN	$H_0 : a = b = \alpha = 1$ vs $H_1 : H_0$ is false	37.70	<0.0001

In order to assess if the model is appropriate, we show in Figure 7 the histograms of the data sets, the plots of the fitted Kw-GHN and GHN density functions and their estimated survival functions and the plots of the empirical distributions. We can conclude that the new distribution is a very suitable model to fit the three data sets.

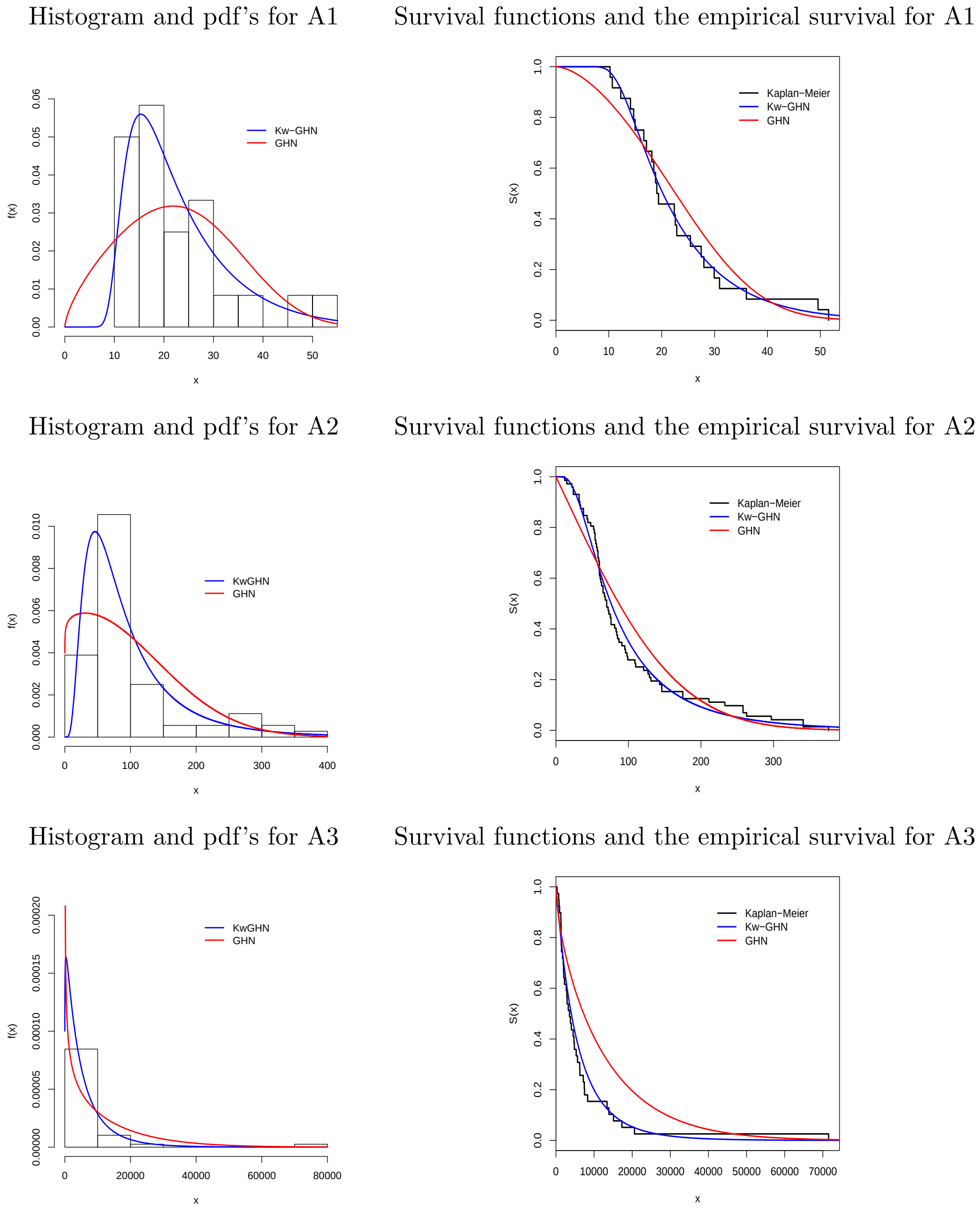

Figure 7: Estimated pdf's, estimated survival functions of the Kw-GHN and GHN models and the empirical survival for the data sets

13.2 Uncensored Data

In this section, the proposed model for survival data with cure fraction is applied to a real data set on cancer recurrence. The data are part of a study on cutaneous melanoma (a type of malignant cancer) for the evaluation of post-operative treatment performance with a high dose of a certain drug (interferon alfa-2b) in order to prevent recurrence. Patients were included in the study from 1991 to 1995 and follow-up was conducted until 1998. The data were collected by Ibrahim *et al.* (2001). The survival time X is defined as the time until the patient's death. The original sample size was $n = 427$ patients, 10 of whom did not present a value for the explanatory variable tumor thickness. When such cases were removed, we obtain a sample of $n = 417$ patients. The percentage of censored observations was 56%. Table 8 lists the MLEs of the model parameters. The values of the AIC, CAIC and BIC statistics are smaller for the Kw-GHN mixture model when compared to those values of the GHN mixture model.

Table 8: MLEs of the model parameters for the melanoma data, the corresponding SEs (given in parentheses) and the AIC, CAIC and BIC statistics

Model	a	b	α	θ	π	AIC	CAIC	BIC
Kw-GHN Mixture	8.8901	1.0304	0.3699	0.3824	0.4872	1059.3	1059.4	1079.4
GHN Mixture	1	1	1.2553	2.5090	0.5150	1074.1	1074.2	1086.2
HN Mixture	1	1	1	2.5975	0.4951	1082.8	1082.9	1090.9

A comparison of the proposed model with some of its sub-models using LR statistics is presented in Table 9. The p-values indicate that the Kw-GHN mixture model yields the best fit to this data set.

Table 9: LR statistics for the melanoma data

Model	Hypotheses	Statistic w	p-value
Kw-GHN vs GHN Mixtures	$H_0 : a = b = 1$ vs $H_1 : H_0$ is false	18.8	<0.0001
Kw-GHN vs HN Mixtures	$H_0 : a = b = \alpha = 1$ vs $H_1 : H_0$ is false	29.6	<0.0001

In Figure 8, we plot the empirical survival function and the estimated survival functions for the Kw-GHN, GHN and HN mixture models. The proportion of cured individuals estimated by the Kw-GHN mixture model ($\pi_{\text{KwGHN}} = 0.4872$) seems more appropriate than that one ($\pi_{\text{GHN}} = 0.5150$) estimated by the GHN mixture model. Further, the Kw-GHN mixture model provides a better fit to these data.

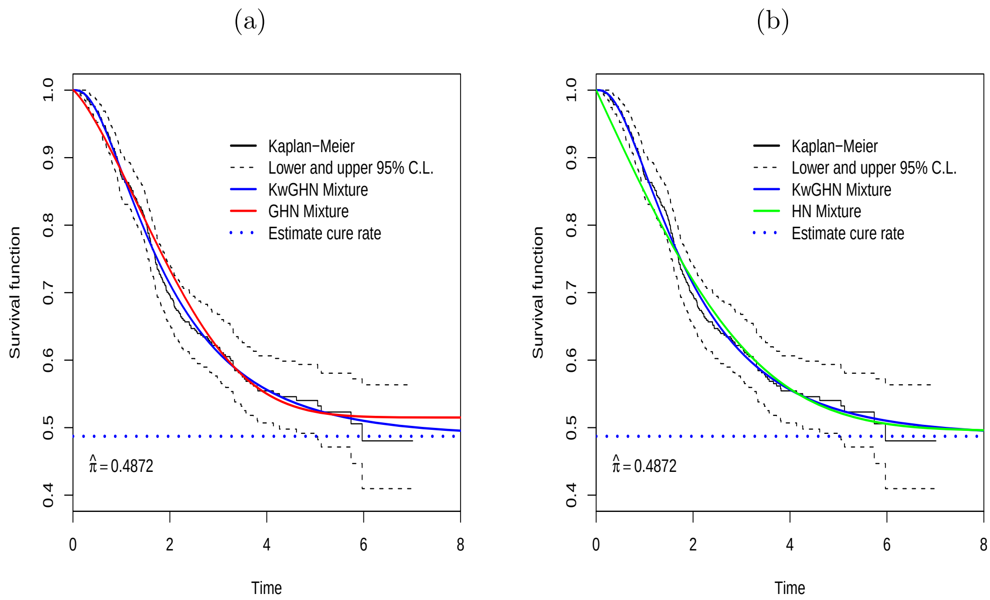

Figure 8: Estimated survival function for the mixtures model and the empirical survival for melanoma data (a) Kw-GHN vs GHN mixtures (b) Kw-GHN vs HN mixtures

14. Conclusions

We propose a new four-parameter distribution called the Kumaraswamy generalized half-normal (Kw-GHN) distribution to extend the half-normal (HN) and generalized half-normal (GHN) (Cooray and Ananda, 2008) distributions. We derive an expansion for the density function and obtain explicit expressions for the moments, quantile and generating functions, mean deviations, density function of the order statistics and their moments. The model parameters are estimated by maximum likelihood and the expected information matrix is determined. We use likelihood ratio statistics to compare the Kw-GHN model with its sub-models. We propose a Kw-GHN model for survival data with cure fraction. Four applications of the new model to real data sets demonstrate that it can be used quite effectively to provide better fits than its main sub-models. We hope this generalization may attract wider applications in survival analysis and biology.

Appendix

The elements of the 4×4 unit expected information matrix are given by

$$\begin{aligned}\kappa_{\alpha,\alpha} &= \frac{1}{\alpha^2} + 2\alpha^2 I_{0,0,0,0,2,0,2} - \frac{2\alpha^2(a-1)}{\sqrt{2\pi}}\left[I_{0,1,0,0,1,1,2} - I_{0,1,0,0,3,0,2}\right] \\ &\quad + \alpha^2 I_{0,2,0,0,2,2,2} - \frac{2\alpha^2 a(1-b)}{\sqrt{2\pi}}\left[(a-1)I_{1,2,0,1,2,2,2}\right. \\ &\quad \left. + I_{1,1,0,1,3,1,2} + I_{1,1,0,1,1,1,2} + aI_{2,2,0,2,2,2,2}\right],\end{aligned}$$

$$\begin{aligned}\kappa_{\alpha,\theta} &= -\frac{1}{\theta} I_{0,0,0,0,2,0,0} - 2\left(\frac{\alpha^2}{\theta}\right) I_{0,0,0,0,2,0,1} + \frac{1}{\theta} \\ &- \frac{2(1-a)}{\sqrt{2\pi}}\left[-\frac{\alpha^2}{\theta} I_{0,1,0,0,3,1,1} + \frac{1}{\theta} I_{0,1,0,0,1,1,0} + \frac{\alpha^2}{\theta}\left(I_{0,1,0,0,1,1,1} - I_{0,2,0,0,2,1,1}\right)\right] \\ &+ \frac{2a(1-b)}{\sqrt{2\pi}}\left[\frac{\alpha^2}{\theta}\left((a-1)I_{1,2,0,1,2,1,1} - I_{1,1,0,1,3,1,1} - I_{2,2,0,2,2,2,1}\right) + \frac{1}{\theta} I_{1,1,0,1,1,1,0}\right],\end{aligned}$$

$$\begin{aligned}\kappa_{\alpha,a} &= -\frac{2\alpha}{\sqrt{2\pi}} I_{0,1,0,0,1,1,1} - \frac{2\alpha(1-b)}{\sqrt{2\pi}} I_{1,1,0,1,1,1,1} - \frac{2a\alpha(1-b)}{\sqrt{2\pi}} \\ &\quad \times \left[I_{1,1,1,1,1,1,1} - I_{2,1,1,2,1,1,1}\right],\end{aligned}$$

$$\begin{aligned}\kappa_{a,a} &= \frac{1}{a^2} - (1-b)I_{1,0,2,1,0,0,0} - I_{2,0,2,2,0,0,0}, \kappa_{\alpha,b} = \frac{2a\alpha}{\sqrt{2\pi}} I_{1,1,0,1,1,1,1}, \\ \kappa_{a,b} &= I_{1,0,1,1,0,0,0}, \\ \kappa_{\theta,\theta} &= \frac{\alpha(1+2\alpha)}{\theta^2} I_{0,0,0,0,2,0,0} - \frac{\alpha}{\theta} \\ &\quad - \frac{2\alpha(1-a)}{\sqrt{2\pi}}\left[\left(\frac{\alpha}{\theta}\right)^2 I_{0,1,0,0,3,1,0} - \frac{\alpha(1+\alpha)}{\theta^2}\ I_{0,1,0,0,1,1,0} - \left(\frac{\alpha}{\theta}\right)^2 I_{0,2,0,0,2,2,0}\right] \\ &\quad - \frac{2a(1-b)}{\sqrt{2\pi}}\left[\frac{(a-2)\alpha^2}{\theta^2} I_{1,2,0,1,2,1,0} + \left(\frac{\alpha}{\theta}\right)^2\left(I_{1,1,0,1,3,1,0} - I_{2,2,0,2,2,2,0}\right)\right. \\ &\quad \left. - \frac{\alpha(1+\alpha)}{\theta} I_{1,1,0,1,1,1,0}\right], \quad \kappa_{\theta,b} = \frac{2a\alpha}{\theta\sqrt{2\pi}} I_{1,1,0,1,1,1,0}, \quad \kappa_{b,b} = \frac{1}{b^2},\end{aligned}$$

$$\kappa_{\theta,a} = \frac{2\alpha}{\theta\sqrt{2\pi}}\left\{I_{0,1,0,0,1,1,0} - (1-b)\left[I_{1,1,0,1,1,1,0} - a\left(I_{1,1,1,1,1,1,0} - I_{2,1,1,2,2,1,0}\right)\right]\right\}.$$

Here, we assume that $T = 2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right] - 1$, and define the expected value

$$I_{i,j,k,l,m,n,p} \quad = \quad E\left\{\frac{T^{ia-j}\Big(\log T\Big)^{k}\left[\Phi^{-1}\left(\frac{T+1}{2}\right)\right]^{m}\exp\Big\{-\frac{n}{2}\left[\Phi^{-1}\left(\frac{T+1}{2}\right)\right]^{2}\Big\}}{\Big(1-T^{a}\Big)^{l}\Big\{\log\left[\Phi^{-1}\left(\frac{T+1}{2}\right)\right]\Big\}^{-p}}\right\}.$$

These expected values can be determined numerically using maple and mathematica for any a and b. For example, for $a = 1.5$ and $b = 3$, we easily calculate all I's in the information matrix:

$I_{0,0,0,0,2,0,2} = 0.02752020$, $I_{0,1,0,0,1,1,2} = 0.07229$, $I_{0,1,0,0,3,0,2} = 0.04376158$, $I_{0,2,0,0,2,2,2} = 0.1537331$, $I_{0,2,0,1,2,2,2} = 0.01490433$, $I_{1,1,0,1,3,1,2} = 0.004680388$, $I_{1,1,0,1,1,1,2} = 0.009813646$, $I_{2,2,0,2,2,2,2} = 0.002113833$, $I_{0,0,0,0,2,0,0} = 0.1807297$, $I_{0,0,0,0,2,0,1} = -0.0669055$, $I_{0,1,0,0,3,1,1} = -0.09313169$, $I_{0,1,0,0,1,1,0} = 0.5274895$, $I_{0,1,0,0,1,1,0} = -0.1952747$, $I_{0,2,0,0,2,1,1} = -0.5271067$, $I_{1,2,0,1,2,1,1} = -0.05110267$, $I_{1,1,0,1,3,1,1} = -0.01264299$, $I_{1,1,0,1,1,1,0} = 0.07160873$, $I_{1,1,0,1,1,1,1} = -0.02650929$, $I_{1,1,1,1,1,1,1} = 0.02303475$, $I_{2,1,1,2,1,1,1} = 0.003331199$, $I_{1,0,2,1,0,0,0} = 0.03683076$, $I_{2,0,2,2,0,0,0} = 0.005423749$, $I_{1,0,1,1,0,0,0} = 0.03964005$, $I_{0,1,0,0,3,1,0} = 0.2515737$, $I_{0,2,0,0,2,2,0} = 1.121768$, $I_{1,2,0,1,2,1,0} = 0.1380420$, $I_{1,1,0,1,3,1,0} = 0.0341521$, $I_{1,1,1,1,1,1,0} = -0.06222305$ and $I_{2,1,1,2,2,1,0} = -0.006214327$.

Acknowledgements

The authors are grateful to the editor and two referees for helpful comments and suggestions. We gratefully acknowledge grants from CNPq, Brazil.

References

Aarts, R. M. (2000). Lauricella functions. From MathWorld - A Wolfram Web Resource, created by Eric W. Weisstein., (http://mathworld.wolfram.com/LauricellaFunctions.html).

Akinsete, A., Famoye, F. and Lee, C. (2008). The beta-Pareto distribution. *Statistics* **42**, 547-563.

Barakat, H. M. and Abdelkader, Y. H. (2004). Computing the moments of order statistic from nonidentical random variables. *Statistical Methods and Applications* **13**, 15-26.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47**, 501-515.

Cooray, K. and Ananda, M. M. A. (2008). A generalization of the half-normal distribution with applications to lifetime data. *Communication in Statistics - Theory and Methods* **37**, 1323-1337.

Cordeiro, G. M., Ortega, E. M. M. and Nadarajah, S. (2010). The Kumaraswamy Weibull distribution. *Journal of the Franklin Institute* **347**, 1399-1429.

Cordeiro, G. M. and de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation* **81**, 883-898.

Díaz-García, J. A. and Leiva, V. (2005). A new family of life distributions based on elliptically contoured distributions. *Journal of Statistical Planning and Inference* **128**, 445-457.

Exton, H. (1978). *Handbook of Hypergeometric Integrals: Theory, Applications, Tables, Computer Programs.* Halsted Press, New York.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041-1046.

Gradshteyn, I. S. and Ryzhik, I. M. (2000). *Table of Integrals, Series, and Products*, sixth edition. Academic Press, San Diego.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C. and Wallis, J. R. (1979). Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research* **15**, 1049-1054.

Gupta, R. D. and Kundu, D. (2001). Exponentiated exponential family: an alternative to gamma and Weibull distributions. *Biometrical Journal* **43**, 117-130.

Gupta, R. C., Gupta, P. L. and Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Communications Statistics - Theory and Methods* **27**, 887-904.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B* **52**, 105-124.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer, New York.

Kenney, J. F. and Keeping, E. S. (1962). *Mathematics of Statistics*, Part 1, 3rd edition. Van Nostrand, New Jersey.

Kundu, D., Kannan, N. and Balakrishnan, N. (2008). On the hazard function of Birnbaum-Saunders distribution and associated inference. *Computational Statistics and Data Analysis* **52**, 2692-2702.

Leiva, V., Barros, M. and Paula, G. A. (2009). *Generalized Birnbaum-Saunders Models Using R.* XI Escola de Modelos de Regressão, Recife, Brazil.

Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors.* Wiley, New York.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society, Series D* (*The Statistician*) **37**, 25-32.

Murthy, D. N. P., Xie, M. and Jiang, R. (2004). *Weibull Models.* John Wiley, New Jersey.

Mudholkar, G. S., Srivastava, D. K. and Freimer, M. (1995). The exponential Weibull family: a reanalysis of the bus-motor-failure data. *Technometrics* **37**, 436-445.

Nadarajah, S. and Gupta, A. K. (2007). A generalized gamma distribution with application to drought data. *Mathematics and Computers in Simulation* **74**, 1-7.

Nelson, W. B. (2004). *Accelerated Testing: Statistical Models, Test Plans and Data Analysis.* John Wiley, New York.

Ortega, E. M. M., Rizzato, F. B. and Demétrio, C. G. B. (2009). The generalized log-gamma mixture model with covariates: local influence and residual analysis. *Statistical Methods and Applications* **18**, 305-331.

Steinbrecher, G. (2002). Taylor expansion for inverse error function around origin. University of Craiova working paper.

Sy, J. P. and Taylor, M. M. G. (2000). Estimation in Cox a proportional hazards cure model. *Biometrics*, **56**, 227-336

Trott, M. (2006). *The Mathematica Guidebook for Symbolics.* With 1 DVD-ROOM (Windows, Macintosh and UNIX). Springer, New York.

Received February 22, 2011; accepted August 25, 2011.

Gauss M. Cordeiro
Departamento de Estatítica
Universidade Federal de Pernambuco - UFPE
Cidade Universitária, 50740-540 -- Recife, PE, Brazil
gausscordeiro@uol.com.br

Rodrigo R. Pescim
Departamento de Ciências Exatas
Universidade de São Paulo - ESALQ - USP
Av. Pádua Dias 11 - Caixa Postal 9, 13418-900, Piracicaba - São Paulo - Brazil
rrpescim@esalq.usp.br

Edwin M. M. Ortega
Departamento de Ciências Exatas
Universidade de São Paulo - ESALQ - USP
Av. Pádua Dias 11 - Caixa Postal 9, 13418-900, Piracicaba - São Paulo - Brazil
edwin@esalq.usp.br