García, Victoriano J.; Gómez-Déniz, Emilio; Vázquez-Polo, Francisco J.

Article

On modelling insurance data by using a generalized lognormal distribution

Revista de Métodos Cuantitativos para la Economía y la Empresa

Provided in Cooperation with:
Universidad Pablo de Olavide, Sevilla

Suggested Citation: García, Victoriano J.; Gómez-Déniz, Emilio; Vázquez-Polo, Francisco J. (2014) : On modelling insurance data by using a generalized lognormal distribution, Revista de Métodos Cuantitativos para la Economía y la Empresa, ISSN 1886-516X, Universidad Pablo de Olavide, Sevilla, Vol. 18, pp. 146-162

This Version is available at:
https://hdl.handle.net/10419/113882

REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (18). Páginas 146–162.
Diciembre de 2014. ISSN: 1886-516X. D.L: SE-2927-06.
URL: http://www.upo.es/RevMetCuant/art.php?id=100

On Modelling Insurance Data by Using a Generalized Lognormal Distribution

García, Victoriano J.
Departamento de Estadística e Investigación Operativa
Universidad de Cádiz (España)
Correo electrónico: victoriano.garcia@uca.es

Gómez-Déniz, Emilio
Departamento de Métodos Cuantitativos e Instituto TiDES
Universidad de Las Palmas de Gran Canaria (España)
Correo electrónico: egomez@dmc.ulpgc.es

Vázquez-Polo, Francisco J.
Departamento de Métodos Cuantitativos e Instituto TiDES
Universidad de Las Palmas de Gran Canaria (España)
Correo electrónico: fjvpolo@dmc.ulpgc.es

ABSTRACT

In this paper, a new heavy-tailed distribution is used to model data with a strong right tail, as often occurs in practical situations. The distribution proposed is derived from the lognormal distribution, by using the Marshall and Olkin procedure. Some basic properties of this new distribution are obtained and we present situations where this new distribution correctly reflects the sample behaviour for the right tail probability. An application of the model to dental insurance data is presented and analysed in depth. We conclude that the generalized lognormal distribution proposed is a distribution that should be taken into account among other possible distributions for insurance data in which the properties of a heavy-tailed distribution are present.

Keywords: heavy-tailed; insurance; lognormal distribution; loss distribution.
JEL classification: C16.
MSC2010: 60E05; 62H05.

Artículo recibido el 26 de septiembre de 2014 y aceptado el 4 de diciembre de 2014.

Sobre la modelización de datos de seguros usando una distribución lognormal generalizada

RESUMEN

Presentamos una nueva distribución lognormal con colas pesadas que se adapta bien a muchas situaciones prácticas en el campo de los seguros. Utilizamos el procedimiento de Marshall y Olkin para generar tal distribución y estudiamos sus propiedades básicas. Se presenta una aplicación de la misma para datos de seguros dentales que es analizada en profundidad, concluyendo que tal distribución debería formar parte del catálogo de distribuciones a tener cuenta para la modernización de datos en seguros cuando hay presencia de colas pesadas.

Palabras clave: seguros; distribución lognormal; función de pérdidas; colas pesadas.
Clasificación JEL: C16.
MSC2010: 60E05; 62H05.

1 Introduction

According to Klugman (1986), three different risk elements are present in most models of insurance risk. Firstly, whether the covered event occurs or not; secondly, the time at which the insurance settlement is paid and, thirdly, the amount to be paid. A simple model to describe the risk is given by

$$P = W \exp(-\delta t) X,$$

where $W \in \{0, 1\}$ represents the occurrence of a covered event, $t \in \mathbb{R}_+$ represents the time and X is the amount. The parameter δ, which is called the force of interest, is usually assumed as known and constant. Depending on the case in question, each of the three elements (W, t, X) can be assumed as a random variable or not. For instance, in life insurance, the value of W is known to be equal to one, X is determined beforehand and the only random variable to be considered is t. In the case examined in this paper, that of dental insurance claims, the time component can be assumed to be a known fixed period and the variables of interest are then the number of claims and their amount. From the point of view of the company, one of the main results to be obtained from the model would be the expected value of the payment after a given period, $E(P)$. Then, it would be useful to assume that $W \in \mathbb{Z}_+$ gives the number of events covered during the period and that X represents the mean value of the corresponding amounts. If the time and amount variables are assumed to be independent (which does not always hold), the expected value to be estimated is given by

$$\begin{aligned} E(P) &= E(W) E[\exp(-\delta t)] E(X) \\ &= K \cdot E(W) \cdot E(X). \end{aligned}$$

With this notation, $E(X)$ is called the severity and K is a known constant. Note that $E(X)$ is also the expected value of an individual claim. In this paper, we shall focus on the estimation of a model for random variable X.

On many occasions, real data sets show a behaviour with extreme values yielding tails which are heavier than those of standard, well-known statistical distributions. Advances in computation speed have made it possible to develop and use new probabilistic models that, not long ago, would have been difficult to apply to describe any type of data.

The literature contains a vast catalogue of probability distributions to obtain a close data fit but new families of distributions may still be welcome, for various reasons. For instance, the heavy-tailed distributions which are available are now competing to become the simplest and most accurate model in each case.

The heavy-tailed distributions are those whose right tail probabilities are heavier than the exponential one, that is, its survival function $\overline{F}$ verifies

$$\lim_{z \to \infty} \frac{\exp(-\lambda x)}{\overline{F}(x)} = 0, \quad \text{for any } \lambda > 0.$$

See Beirlant *et al.* (2001) for further information. Well-known examples of these kind of distributions are the lognormal, Weibull and Pareto ones, when the shape parameter is smaller than one (see Rolski *et al.* 1999, p.49). In many practical situations, such as reliability analysis (Blishke and Murthy, 2000; Chen, 1995) and lifetime data (Prendergast *et al.*, 2005), the lognormal model is suitable for data fitting. Sobkowicz *et al.* (2013) recently presented an analysis of the length of comments posted in Internet discussion fora and found that the size of messages can be fitted quite well using a lognormal distribution.

In the actuarial context, models with heavy-tailed distributions have been used to provide adequate descriptions of claim size distributions, see Hogg and Klugman (1984) and Klugman *et al.* (2008), among many others. The right tail of a distribution is an important issue in various contexts, but especially concerning issues related to insurance, where it represents the total impact of insurance losses, and in risk theory, where it is associated with the extreme–value theory. Dutta and Perry (2006) presented an empirical analysis of loss distributions in which risk was estimated by different approaches, including Exploratory Data Analysis and other empirical approaches. These authors concluded that "*one would need to use a model that is flexible enough in its structure*" and rejected the use of exponential, gamma and Weibull models because of their poor results.

These results encourage us to search for more flexible probability distributions providing greater accuracy in data-fitting.

In recent years, various techniques for extending heavy-tailed distributions have been proposed. One such, introduced by Marshall and Olkin (1997), was first applied by its authors to extend the exponential family to a generalized exponential. Given a distribution with the survival function $\overline{F}(x) = P(X > x)$, a generalization of the family is obtained by considering the survival function

$$\overline{G}_{\alpha}(x) = \frac{\alpha \overline{F}(x)}{1 - \overline{\alpha}\overline{F}(x)}, \quad \alpha > 0,\ \overline{\alpha} = 1 - \alpha. \tag{1}$$

In the present paper, we denote the above cumulative distribution function (cdf) by G_{α}, obtained from the original lognormal cdf, F, where the parameters μ and σ^2 have been omitted for the sake of simplicity. Henceforth, the new generalized lognormal distribution obtained using the Marshall and Olkin procedure is referred to by GLN. We obtain closed expressions for the probability density function (pdf) and the cdf of the new distribution, from which the moments and quantiles can be easily computed. In addition, the method preserves some properties of the original distribution which are needed for risk models, as shown in the second example of applications. This method has been successfully applied by several authors to extend different distributions to generalized ones. Thus, Ghitany *et al.* (2005), applied the exponential and Weibull generalized models to censored data. Furthermore, Ghitany (2005) introduced a generalized Pareto distribution, and the Lomax

distribution was extended by Ghitany *et al.* (2007). García *et al.* (2010) obtained a generalization of the normal distribution, Gómez-Déniz (2010) obtained a generalization of the discrete geometric distributions and finally, Jose *et al.* (2010) presented a Marshall-Olkin q-Weibull distribution applied to time series analysis.

It is well known, that if a lognormal distribution has its shape parameter smaller than one, then it is a heavy-tailed distribution. Equivalently, its cdf verifies that

$$\lim_{x\to\infty} \frac{\overline{F}(x+y)}{\overline{F}(x)} = 1, \quad \text{for all } y \geq 0.$$

Then, it is easy to see that the corresponding transformed cdf G given by (1) also verifies:

$$\begin{aligned}\lim_{x\to\infty} \frac{G(x+y)}{G(x)} &= \lim_{x\to\infty} \frac{F(x+y)\left(1-\overline{\alpha}\overline{F}(x)\right)}{F(x)\left(1-\overline{\alpha}\overline{F}(x+y)\right)} \\ &= \lim_{x\to\infty} \frac{F(x+y)}{F(x)} \cdot \frac{\left(1-\overline{\alpha}\overline{F}(x)\right)}{\left(1-\overline{\alpha}\overline{F}(x+y)\right)} = 1,\end{aligned}$$

where we use that $\lim_{z\to\infty} \overline{F}(z) = 0$. Then, G is also heavy-tailed. It can also be proven that G is stochastically smaller (larger) than F for $\alpha \leq (\geq) 1$. In other words, $\overline{G}(x) \leq (\geq) \overline{F}(x) \Longleftrightarrow \alpha \leq (\geq) 1$.

The distribution and density functions of the GLN obtained from (1) depend on three parameter, $\mu > 0$, $\sigma > 0$ and the additional $\alpha > 0$. Note that, on the one hand, the original lognormal distribution is obtained for $\alpha = 1$ and, on the other hand, this generalization of the lognormal is different from the given in Martín and Pérez (2009).

A different approach to the problem is the max-stable class of distributions given by $G(x|\eta) = (F(x))^{\eta}$, with $\eta > 0$. This method has been under study by several authors as Lehmann (1959), Gupta *et al.* (1998), Gupta and Kundu (1999) and Sarabia and Castillo (2005). In this paper, we shall not apply this approach.

1.1 Motivation for a new heavy-tailed distribution

The following example is based on real insurance data extracted from Klugman (1986). The data set corresponds to 392 claims from a dental insurance group (basic coverage). The model is to be applied on the amount of each claim, X, but the company may establish an upper bound, c, to the coverage, so it then becomes interesting to estimate the expected value of the amount per claim, $Y_c = \min(X, c)$, given by

$$E(Y_c) = \int_0^c f(t)\, dt + c\,(1 - F(c)), \tag{2}$$

where $f(\cdot)$ and $F(\cdot)$, are the density and distribution functions of X, respectively.

Several heavy-tailed distributions have been postulated, including the Lognormal, Weibull and Generalized Exponential distributions. In Figure 1, the histogram for these data is overprinted on the maximum likelihood (ML) estimated densities of the Lognormal (L), Weibull (W), Marshall-Olkin Generalized Exponential (GE) and the new Generalized Lognormal (GLN), derived from (1), distributions. The first row pictures show a good fit for the lower values but underestimate the right tail; on the other hand, the GE distribution (bottom left) shows a better fit for the tail but underestimates the lower values. The *GLN* picture (bottom right) shows the distribution to be flexible enough to describe the whole range of the variable.

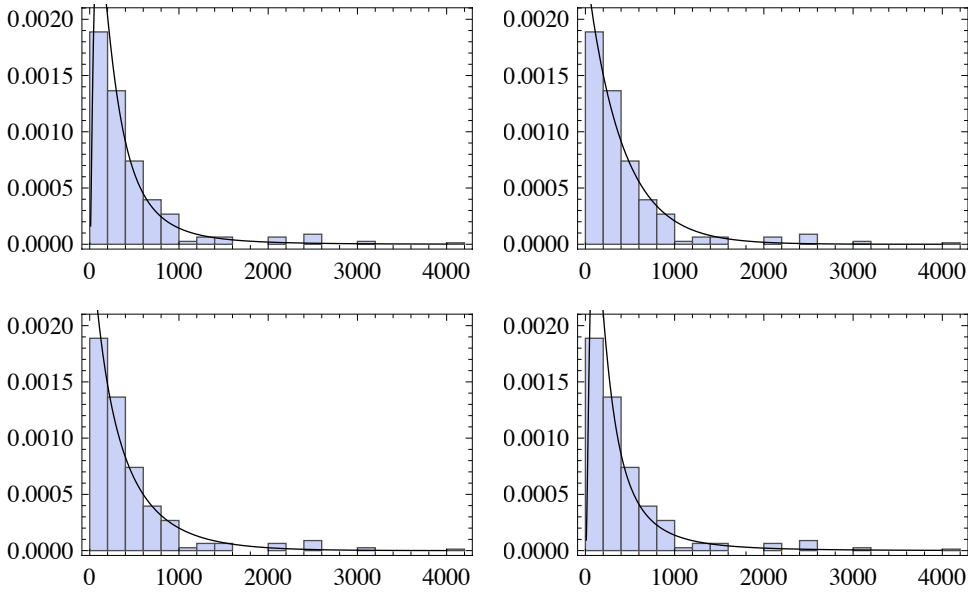

Figure 1: Histogram over densities. From top to bottom, from right to left, Lognormal, Weibull, Generalized Exponential and Generalized Lognormal (GLN).

For comparative purposes, the Akaike Information Criterion (AIC) values were computed, obtaining the following values: $AIC_{GE} = 5487.3$, $AIC_W = 5492.3$, $AIC_L = 5408.2$ and $AIC_{GLN} = 5407.1$. As is well known, a model with a minimum AIC value is to be preferred. In this respect the GLN distribution performs very well in fitting the data distribution, compared to other standard heavy-tailed uniparametric distributions, and also provides a better fit than the Lognormal distribution.

1.2 Outline of the paper

With the aim of predicting the expected value given in (2), we derived the GLN following the procedure suggested by Marshall and Olkin (1997). As shown in the above motivating example, this distribution gives better results

than some well-known models in terms of estimating, and achieves a reasonably good data fit.

This paper presents some novel aspects: (i) the good results of the GLN distribution in fitting certain actuarial data, together with other positive properties; (ii) a simple and easily implementable approach to quantify one of the most important quantities of interest in the insurance scenario, that of the expected value of the amount per claim.

The article is organized as follows. The generalized lognormal distribution, $GLN(\mu, \sigma, \alpha)$, is introduced in Section 2. This section also introduces the moments μ_k of the new distribution and analyses the parameter estimation problem. Furthermore, the motivating example given in Section 1 is continued and developed. Numerical solutions to the ML-estimate problem are obtained using suitable software. Section 3 presents some additional useful properties of the GLN distribution. Finally, in Section 4 some conclusions are drawn and promising areas for further research are proposed.

2 A new heavy-tailed distribution: the generalized lognormal distribution

We propose the GLN, defined as in (1), from an parent lognormal distribution $F(\mu, \sigma)$.

Let be Z_0 a random variable with Gaussian density $\phi_\theta(z)$ and distribution function $\Phi_\theta(z)$, where $\theta = (\mu, \sigma)$, $-\infty < \mu < \infty$ and $\sigma > 0$. Then, the random variable $Z = \exp(Z_0)$ is said to be lognormal distributed. When the lognormal distribution F is used as the parent one, the Marshall and Olkin scheme for generalizing distributions given in (1) leads to the GLN distribution function given by

$$G_{\theta,\alpha}(x) = \frac{\Phi_\theta(\log x)}{1 - \overline{\alpha}\overline{\Phi}_\theta(\log x)},$$

where $\alpha > 0$, $\overline{\alpha} = 1 - \alpha$ and $\overline{\Phi}_\theta = 1 - \Phi_\theta$. By computing the first derivative of $G_{\theta,\alpha}(x)$ we obtain the corresponding density function

$$g_{\theta,\alpha}(x) = \frac{\alpha\phi_\theta(\log x)}{x\left[1 - \overline{\alpha}\overline{\Phi}_\theta(\log x)\right]^2}. \tag{3}$$

The following result is very useful to compute central moments of the GLN distribution.

Proposition 1 *Let be $X \sim GLN(\mu, \sigma, \alpha)$. Then, the k-th moment around the origin of the random variable X is given by*

$$\begin{aligned}\mu_k \doteqdot E\left(X^k\right) &= \int_0^1 \frac{\alpha \exp\left[k\ \Phi_\theta^{-1}(h)\right]}{\left(\overline{\alpha}h + \alpha\right)^2}\, dh \\ &= \alpha\ \exp(k\mu) \int_0^1 \frac{\exp\left[k\sigma\, \mathtt{erf}^{-1}(2h-1)\right]}{\left(\overline{\alpha}h + \alpha\right)^2}\, dh,\end{aligned} \tag{4}$$

where $\texttt{erf}(\cdot)$ *is the error funtion.*

The proof can be found in the Appendix. Table 1 contains values for μ_k $(k = 1, 2, 3)$, the variance and skewness coefficient, Sk, in a $GLN(0, 1, \alpha)$ distribution for several values of α. The skewness coefficient Sk is given by

$$Sk = \frac{E(X - EX)^3}{Var(X)^{3/2}}. \tag{5}$$

Remark: Observe that the computation of $E(Y)$ in the case $c = \infty$ is reduced to find $E(X)$. Such calculus can be obtained with the help of Proposition 1.

Table 1: Moments for $GLN(0, 1, \alpha)$ distributions.

α	μ_1	μ_2	μ_3	Var	Sk
0.1	0.4766	0.9673	9.3901	0.7401	12.9140
0.2	0.6945	1.81272	18.6479	1.3303	10.1278
0.3	0.8662	2.6036	27.8079	1.8532	8.8554
0.4	1.0125	3.3564	36.8851	2.3310	8.0824
0.5	1.1420	4.0795	45.8893	2.7751	7.5472
0.6	1.2592	4.7781	54.8276	3.1924	7.1476
0.7	1.3966	5.4560	63.7055	3.5876	6.8339
0.8	1.4668	6.1158	72.5274	3.9641	6.5789
0.9	1.5604	6.7596	81.2969	4.3245	6.3661
1	1.6487	7.3890	90.0171	4.6707	6.1848
10	4.9431	45.1616	779.502	20.7270	3.7233
20	6.5852	73.7568	1444.70	30.3908	3.3349
30	7.7178	97.1450	2955.61	37.5798	3.1505
40	8.6045	117.5260	2630.04	43.4881	3.0350
50	9.3422	135.8610	3177.03	48.5843	2.9530
60	9.9785	152.6790	3702.12	53.1089	2.8904
70	10.5408	168.3120	4209.03	57.2042	2.8404
80	11.0462	182.9820	4700.42	60.9628	2.7990
90	11.5066	196.8510	5178.32	64.4484	2.7640

2.1 Parameter estimation

From (3), the log-likelihood of the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ reads as follows:

$$l(\boldsymbol{x}; \mu, \sigma, \alpha) = n \log \alpha + \sum_{i=1}^{n} \log \phi_\theta(\log x_i) - 2 \sum_{i=1}^{n} \log \left[1 - \overline{\alpha} \overline{\Phi}_\theta(\log x_i)\right].$$

The normal equations for the maximum likelihood estimation are then obtained by

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{n}{\sigma^2}(\overline{u} - \mu) + 2\overline{\alpha} \sum_{i=1}^{n} \frac{\phi_\theta(u_i)}{D_i} = 0, \\
\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} u_i^2 + \frac{n\mu^2}{\sigma^3} - \frac{2n\mu}{\sigma^3}\overline{u} + \frac{2\overline{\alpha}}{\sigma} \sum_{i=1}^{n} \frac{u_i \phi_\theta(u_i)}{D_i} \\
&\quad - \frac{2\overline{\alpha}\mu}{\sigma} \sum_{i=1}^{n} \frac{\phi_\theta(u_i)}{D_i} = 0, \\
\frac{\partial l}{\partial \alpha} &= \frac{n}{\alpha} - 2 \sum_{i=1}^{n} \frac{\overline{\Phi}_\theta(u_i)}{D_i} = 0,
\end{aligned}$$

where $u_i = \log x_i$, $\overline{u} = n^{-1} \sum_{i=1}^{n} u_i$ and $D_i = 1 - \overline{\alpha}\overline{\Phi}_\theta(u_i)$.

With the help of a proper software, the ML-estimates are easily found. A well established procedure for the numerical resolution of normal equations consists in taking as initial points those obtained by the approximation derived from $\texttt{erf}^{-1}(x) \approx z\sqrt{\pi}/2$ and expanding the kernel of the integrand in (4) by second-order series. Thus

$$\mu_1 \simeq \alpha \exp(\mu) \frac{24\left(1+\alpha^2\right) - 4\sqrt{2\pi}\sigma\left(\alpha^2 - 1\right) + \pi\sigma^2(1+\alpha)^2}{3(1+\alpha)^4},$$

$$\mu_2 \simeq \alpha \exp(2\mu) \frac{4\left(6(1+\alpha)^2 + 2\sqrt{2\pi}\sigma\left(\alpha^2 - 1\right) + \pi\sigma^2(1+\alpha)^2\right)}{3(1+\alpha)^4},$$

$$\mu_3 \simeq \alpha \exp(3\mu) \frac{8\left(1+\alpha^2\right) + 4\sqrt{2\pi}\sigma\left(\alpha^2 - 1\right) + 3\pi\sigma^2(1+\alpha)^2}{(1+\alpha)^4}.$$

The scoring method works now well to solve the above normal equations (see Klugman *et al.*, 2008). These normal equations are solved using the scoring method (see Klugman *et al.*, 2008), and the need to use second order derivatives is avoided by applying the Newton-Raphson method. The Appendix shows the second order derivatives of the log-likelihood which are needed to obtain the Fisher information matrix.

Example [continued] The data set analysed in Klugman (1986) and presented in Section 1 is now revisited. The estimates and the maximized log-likelihood values obtained are shown in Table 2, which shows that the best fit

to the data is provided by the GLN distribution, which produces the lowest AIC value.

Table 2: Estimated parameters, standard errors and AIC for data in example.

Distribution	Parameters	Standard Errors	$L_{\max}$	AIC
GE	$\alpha = 0.67$	0.122		
	$\lambda = 0.002$	0.0002	-2741.63	5487.3
Weibull	$\theta = 0.99$	0.035		
	$\beta = 402.44$	21.71	-2744.13	5492.3
Lognormal	$\mu = 5.50$	0.05		
	$\sigma = 0.974$	0.035	-2702.08	5408.2
GLN	$\mu = 5.874$	0.263		
	$\sigma = 0.972$	0.039		
	$\alpha = 0.51$	0.237	-2700.56	5407.1

The expected value for $E(Y_c)$ is computed for different values of c. The claim amount, X, is therefore described by the GLN, Lognormal, Weibull and GE models, where GE is the generalized exponential distribution from Marshall and Olkin (1997). The results are shown in Table 3. Comparison with the empirical values shows that there is little difference between the fits of the standard lognormal and the GLN distributions, but the latter seems to provide a better fit in the middle part of the data set. This is confirmed by the P-P plot shown in Figure 2.

In summary, the GLN model can adopt a flexible set of density forms, and so it is very suitable for describing a data set like this. It not only improves the data fit with a lognormal distribution, but also the fit with other, alternative models.

3 Some other useful properties of the GLN distribution

Some properties of the GLN distributions are set below. They referred to the most commonly used characteristics of a probability distribution and they could be useful for further applications.

The proof of this result can be found in Appendix.

Proposition 2 *The GLN distribution verifies the following properties:*

1. *It is unimodal, for all feasible values of its parameters $\mu \in \mathbb{R}, \sigma > 0, \alpha > 0$.*

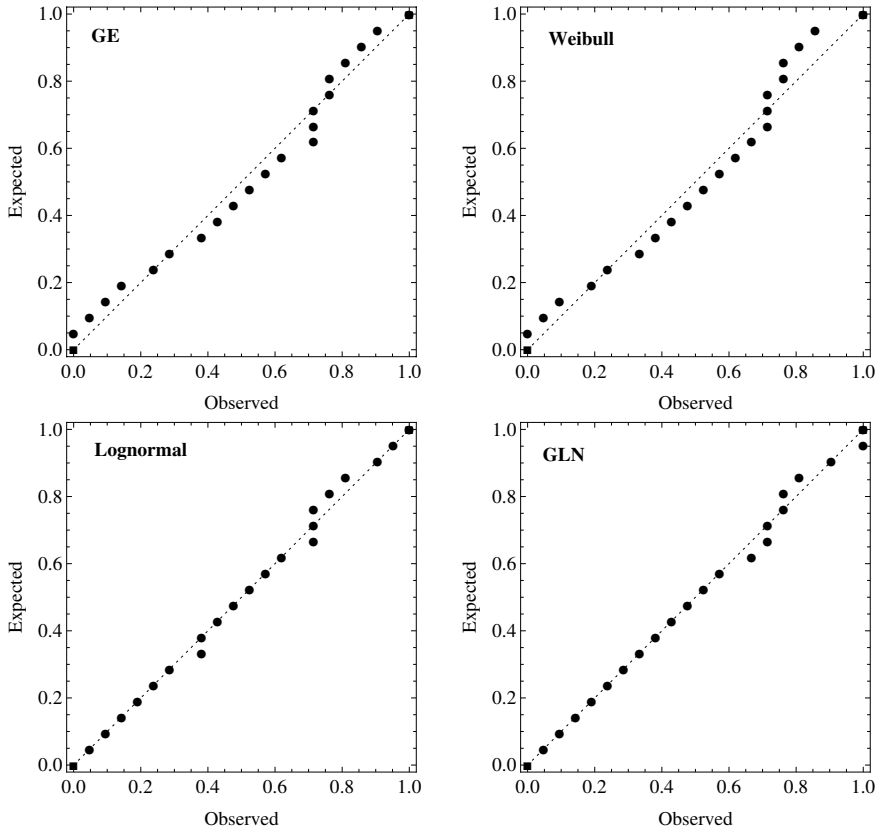

Figure 2: P-P plots for the data set in Klugmann (1986) for the different models considered.

2. It is verified that

$$G_{0,1,\alpha}(x) = 1 - G_{0,1,\alpha^{-1}}\left(x^{-1}\right),$$

for all $x > 0$.

3. Denoting by $\mu_k(\mu,\sigma,\alpha) = E\left(X^k\right)$, *where* X *is GLN distributed, then:*

$$a) \quad \mu_1(\mu,\sigma,\alpha) = \exp(\mu)\,\mu_1(0,\sigma,\alpha),$$

$$b) \quad \mu_k(\mu,\sigma,\alpha) = \exp(k\mu)\,\mu_1(0,k\sigma,\alpha),$$

$$c) \quad Var(\mu,\sigma,\alpha) = \exp(2\mu)\,Var(0,\sigma,\alpha).$$

Finally, the quantiles $\gamma(x_\gamma)$ of a GLN distribution are given by

$$x_\gamma = \exp\left[\Phi_\theta^{-1}\left(\frac{\alpha\gamma}{1-\overline{\alpha}\gamma}\right)\right], \tag{6}$$

Table 3: Limited expected values for the different models considered

i	c_i	Empirical	GE	Weibull	Lognormal	GLN
1	25	25	24.12	24.24	24.94	24.96
2	50	48	46.58	47.05	49.28	49.43
3	75	71	67.53	68.48	72.34	72.67
4	100	91	87.09	88.62	93.81	94.25
5	150	126	153.28	125.34	131.80	132.10
6	200	153	204.97	157.56	163.76	163.50
7	250	177	245.25	186.40	190.66	189.16
8	300	196	277.05	211.66	213.39	211.54
9	400	226	302.31	253.65	249.29	246.09
10	500	250	322.48	286.35	275.93	271.92
11	600	267	338.64	311.82	296.16	291.83
12	700	281	351.63	331.65	311.85	307.58
13	800	292	362.10	347.08	342.22	320.27
14	900	300	362.47	359.09	334.13	330.67
15	1000	306	380.30	368.43	342.17	339.30
16	1250	321	391.04	383.70	356.61	355.43
17	1500	332	401.20	391.84	365.90	366.41
18	2000	349	404.80	398.50	376.55	379.92
19	2500	357	406.09	400.39	382.02	387.50
20	3000	360	406.71	400.93	385.09	392.10
21	4000	361	406.79	401.13	388.11	397.04

where $\Phi_\theta^{-1}(\cdot)$ is the quantile function of the normal distribution function $\Phi_\theta(x)$. In particular, the median, Me, is given by

$$Me = \exp\left[\Phi_\theta^{-1}\left(\frac{\alpha}{1+\alpha}\right)\right]. \tag{7}$$

Numerical values for (6) and (7) can be easily computed by using the instruction `InverseCDF[NormalDistribution[`μ`,`σ`,`γ`]]`, which is available within the `Mathematica`© package. As expression (7) shows, Me increases with α. At the limit values of α, when $\alpha \to \infty$, then $Me \to \infty$; and when $\alpha \to 0$, then $Me \to 0$. Thus, it is shown that there are no limit distributions, and the right or left tail probabilities increase for $\alpha > 1$ and $0 < \alpha < 1$, respectively.

4 Conclusions

In this paper, the generalized lognormal distribution, GLN, has been introduced. The GLN is a generalization of the lognormal distribution, which is contained in the new set for the additional parameter value $\alpha = 1$. The given example shows that this new model can compete with some of the most well-known available models reasonably, and that it should be considered in data-fitting, due to its flexibility.

A generalized distribution by the Marshall and Olkin method preserves some of the properties of the original family. As a consequence, the GLN distribution maintains some desirable properties for risk-theory, as heavy-tailed profile and sub-exponential belonging. This way, its application on fitting data of this kind, the study of properties of its hazard rate function or its extension to higher dimensions could be interesting for future researches.

Acknowledgments

VG is partially funded by Junta de Anadalucía (project SEJ–02814). VG, EGD and FJVP are partially funded by Ministerio de Economía y Competitividad, Spain, under project ECO2013–47092.

References

Beirlant, J., Matthys, G., and Dierckx, G. (2001). Heavy-tailed distributions and rating. *Astin Bulletin*, 31, 1, 37–58.

Blishke, W. and Murthy, D. (2000). *Reliability: Modeling, Prediction, and Optimization.* Wiley.

Chen, G. (1995). Generalized log-normal distributions with reliability application. *Computational Statistics and Data Analysis*, 19, 309–319.

Dutta, K. and Perry, J. (2006). A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital. *Federal Reserve Bank of Boston, Working Paper* , 06–13, 2006 *Series.*

García, V., Gómez-Déniz, E., and Vázquez-Polo, F.J. (2010). A new skew generalization of the Normal distribution: properties and applications. *Computational Statistics and Data Analysis*, 54, 2021–2034.

Ghitany, M.E. (2005). Marshall-Olkin extended Pareto distribution and its applications. *International Journal of Applied Mathematics*, 18, 17–32.

Ghitany, M.E., Al-Awadhi, F.A., and Alkhalfan, L.A. (2007). Marshall-Olkin extended Weibull distribution and its applications to censored data. *Communications to Statistics: Theory and Methods*, 36, 1855–1866.

Ghitany, M.E., Al-Hussaini, E.K., and Al-Jarallah, R.A. (2005). Marshall-Olkin extended Lomax distribution and its applications to censored data. *Journal of Applied Statistics*, 32, 1025–1034.

Gómez-Déniz, E. (2010). Another generalization of the geometric distribution. *Test*, 19, 399–415.

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

Gupta, R.D. and Kundu, D. (1999). Generalized Exponential Distributions. *Australian and New Zealand Journal of Statistics*, 41, 2, 173–188.

Hogg, R.V. and Klugman, S.A. (1984). *Loss Distributions.* Wiley Series in Probability and Mathematical Statistics.

Jose, K.K., Naik, S.R., and Ristić, M.M. (2010). Marshall-Olkin q-Weibul distribution and max-min processes. *Statistical Papers*, 51, 837–851.

Klugman, S.A. (1986). Loss distributions. In *Actuarial Mathematics. Proceedings of Symposia in Applied Mathematics. American Mathematical Society*, pp. 31–55.

Klugman, S.A., Panjer, H.H., and Willmot, G.E. (2008). *Loss models: from data to decisions*, Wiley.

Lehmann, E.L. (1959). The power of rank test. *Annals of Mathematical Statistics*, 24, 23–43.

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

Martín, J. and Pérez, C.J. (2009). Bayesian analysis of a generalized lognormal distribution. *Computational Statistics and Data Analysis*, 53, 1377–1387.

Prendergast, J., O'Driscoll, E., and Mullen, E. (2005). Investigation into the correct statistical distribution for oxide breakdown over oxide thickness range. *Microelectronics Reliability*, 45, 5–6, 973–977.

Rolski, T., Schmidli, H. Schmidt, V., and Teugel, J. (1999). Stochastic processes for insurance and finance. John Wiley & Sons.

Sarabia, J.M. and Castillo, E. (2005). About a class of max-stable families with applications to income distributions. *Metron*, LXIII, 3, 505–527.

Sobkowicz, P; Thelwall, M.; Buckley, K.; Paltoglou, G., and Sobkowicz, A. (2013). Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law? *EPJ Data Science* 2013, 2:2. Available at `http://www.epjdatascience.com/content/2/1/2`.

Appendix

A1: Proof of Proposition 1

The k-th moment around the origin is defined by

$$\mu_k = \int_0^\infty x^k g_{\theta,\alpha}(x)\, dx.$$

We term $h = \Phi_\theta(\log x)$, and thus $dh = \phi_\theta(t)\, dt$ and $t = \Phi_\theta^{-1}(h)$. Then, we can write

$$\mu_k = \int_0^1 \frac{\alpha \exp\left[k\ \Phi_\theta^{-1}(h)\right]}{\left(\overline{\alpha}h + \alpha\right)^2} dh.$$

In order to prove the second expression (4), it is only necessary to consider the relationship

$$\Phi_\theta^{-1}(h) = \mu + \sigma \mathtt{erf}^{-1}(2h - 1),$$

and the proof is completed.

A2: Proof of Proposition 2

Here, we provide a brief sketch of the proof of Proposition 2.

By taking the derivative of (3) and equating to zero, we obtain the equation,

$$\left(\mu - \log Mo - \sigma^2\right)\left[1 - \overline{\alpha}\overline{\Phi}_\theta(\log Mo)\right] - 2\overline{\alpha}\sigma^2\phi_\theta(\log Mo) = 0. \tag{8}$$

Now, assume the function

$$\Psi(x) = \left(\mu - \log x - \sigma^2\right)\left[1 - \overline{\alpha}\overline{\Phi}_\theta(\log x)\right] - 2\overline{\alpha}\sigma^2\phi_\theta(\log x).$$

It is simple to verify that $\Psi(0^+) = \infty$, $\Psi(\infty) = -\infty$ and $\Psi(x)$ is a continuous function. Then, we obtain that

$$\Psi'(x) = \frac{1}{x}\left[\overline{\alpha}\left(\overline{\Phi}_\theta(\log x) + \left(\log x - \left(\mu + \sigma^2\right)\right)\phi_\theta(\log x)\right) - 1\right].$$

As $\overline{\alpha} < 1$, the desired result follows if it can be shown that

$$\overline{\Phi}_\theta(z) + \phi_\theta(z)\left(z - \left(\mu + \sigma^2\right)\right) < 1,$$

or equivalently

$$\phi_\theta(z)\left(z - \left(\mu + \sigma^2\right)\right) < \Phi_\theta(z), \tag{9}$$

with $z = \log x \in \mathbb{R}$. To this end, we firstly consider $z > \mu + \sigma^2$ and then observe that the left-hand side in (9) coincides with the area of a rectangle

with base $\left[\mu+\sigma^2, z\right]$ and height $\phi_\theta(z)$, constructed under pdf curve $\phi_\theta(z)$. Trivially, this area is smaller that the total area under $\phi_\theta(z)$ in the interval $(-\infty, z]$. Secondly, if $z \leq \mu+\sigma^2$, then $\phi_\theta(z)\left(z-\left(\mu+\sigma^2\right)\right)<0$. Therefore, we conclude that $\Psi'(x)<0$, so then $\Psi(x)$ is a decreasing function on x and the solution to equation (8) is unique and the unimodality is proven.

The proof of (6) is direct, from (3). From (4), expression (6) is direct. Hence, expression (6) is derived. Finally, note that

$$\begin{aligned} Var(\mu,\sigma,\alpha) &= \mu_2(\mu,\sigma,\alpha) - \mu_1^2(\mu,\sigma,\alpha) \\ &= \exp(2\mu)\,\mu_2(0,\sigma,\alpha) - \exp(2\mu)\,\mu_1^2(0,\sigma,\alpha), \end{aligned}$$

and the proof is completed.

A3: Second-order derivatives of the log-likelihood function

After some algebraic simplication, they can be written as

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} + \frac{2\overline{\alpha}}{\sigma^2}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)}{D_i} - \frac{2\overline{\alpha}\mu}{\sigma^2}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} + 2\overline{\alpha}^2\sum_{i=1}^{n}\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2.$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \mu \partial \sigma} = &-\frac{2n}{\sigma^3}(\overline{u}-\mu) - \frac{2\overline{\alpha}}{\sigma}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} + \frac{2\overline{\alpha}}{\sigma^3}\sum_{i=1}^{n}\frac{u_i^2\phi_\theta(u_i)}{D_i} + \frac{2\overline{\alpha}\mu}{\sigma^3}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} \\ &-\frac{4\overline{\alpha}\mu}{\sigma^3}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)}{D_i} + \frac{2\overline{\alpha}^2}{\sigma}\sum_{i=1}^{n}u_i\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2 - \frac{2\overline{\alpha}\mu}{\sigma}\sum_{i=1}^{n}\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2. \end{aligned}$$

$$\frac{\partial^2 l}{\partial \mu \partial \alpha} = -2\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} - 2\overline{\alpha}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)\,\overline{\Phi}_\theta(u_i)}{D_i^2}.$$

$$
\begin{aligned}
\frac{\partial^2 l}{\partial \sigma^2} = & \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n} u_i^2 - \frac{3n\mu^2}{\sigma^4} + \frac{6n\mu}{\sigma^4}\overline{u} - \frac{4\overline{\alpha}}{\sigma^2}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)}{D_i} + \frac{2\overline{\alpha}}{\sigma^4}\sum_{i=1}^{n}\frac{u_i^3\phi_\theta(u_i)}{D_i} \\
& + \frac{2\overline{\alpha}\mu^2}{\sigma^4}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)}{D_i} - \frac{4\overline{\alpha}\mu}{\sigma^4}\sum_{i=1}^{n}\frac{u_i^2\phi_\theta(u_i)}{D_i} + \frac{2\overline{\alpha}}{\sigma^4}\sum_{i=1}^{n}\left(\frac{u_i\phi_\theta(u_i)}{D_i}\right)^2 \\
& - \frac{4\overline{\alpha}^2\mu}{\sigma^2}\sum_{i=1}^{n} u_i\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2 + \frac{4\overline{\alpha}\mu}{\sigma^2}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} - \frac{2\overline{\alpha}\mu}{\sigma^4}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} \\
& + \frac{2\overline{\alpha}\mu^2}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2 \\
\frac{\partial^2 l}{\partial \sigma \partial \alpha} = & -\frac{2}{\sigma}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)}{D_i} + \frac{2\mu}{\sigma}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)}{D_i} - \frac{2\overline{\alpha}}{\sigma}\sum_{i=1}^{n}\frac{u_i\phi_\theta(u_i)\,\overline{\Phi}_\theta(u_i)}{D_i^2} \\
& + \frac{2\overline{\alpha}\mu}{\sigma}\sum_{i=1}^{n}\frac{\phi_\theta(u_i)\,\overline{\Phi}_\theta(u_i)}{D_i^2}. \\
\frac{\partial^2 l}{\partial \alpha^2} = & -\frac{n}{\alpha^2} + 2\sum_{i=1}^{n}\left(\frac{\phi_\theta(u_i)}{D_i}\right)^2.
\end{aligned}
$$

The elements of the observed information matrix are minus the second-order partial derivatives fo the log-likelihood with respect to the parameters and the elements of the expected information matrix are the expected values of their corresponding above elements.