The Power Function Lognormal Distribution: A New Probability Model for Insurance Claims Data

Chao Wang

Abstract—**This paper introduces the power function lognormal (PFL) distribution, which assumes a power function distribution up to a specified threshold, and a lognormal distribution beyond it. It can be positively or negatively skewed, and/or leptokurtic/platykurtic. Maximum likelihood, moment, nonlinear least squares, and Bayes estimators are obtained. A simulation study is performed, numerical computations are carried out to display the performance of the proposed method, and insurance claims data are analyzed for illustrative purposes.**

Index Terms—**composite distribution, power function lognormal distribution, positively skewed, negatively skewed, parameter estimation, claims data.**

I. Introduction

INSURANCE claims and loss data are typically very highly right-skewed, and many long-tailed, skewed probability distributions have been considered to modeling them (see, for example, [1], [2]). [3] proposed a composite lognormal-Pareto model to analyze Danish fire claim data, which has been widely discussed in actuarial literature. Their model is based on a lognormal density up to an unknown threshold, and two-parameter Pareto density beyond it. They imposed continuity and differentiability conditions at the threshold to obtain a smooth density function and reduce the number of parameters. [4] and [5] improved the composite models by allowing flexible mixing weights, replacing a constant weight applied by [3], resulting in a better fit to the loss data. Various composite distributions have been developed over the past 15 years, including the composite lognormal-Pareto ([3], [4]), composite Weibull-Pareto ([6], [7], [8]), composite inverse Weibull-Weibull ([9]), composite exponential-Pareto ([10], [11]), truncated composite lognormal-Pareto model([12]), truncated composite Weibull-Pareto ([13]), composite lognormal-Pareto model with random threshold([14]), composite Weibull-Burr ([15]), composite Stoppa ([16]), composite Pareto- Arctan ([17]), and composite log-Gauss-Pareto ([18]). These distributions have shown superior performance to classical models such as lognormal, Pareto, Inverse Gaussian, Gamma, and Weibull when modeling insurance loss/claims data. [19] and [20] derived robust and Bayesian estimators, respectively, of the composite lognormal-Pareto distribution, and [21] used four composite distributions to model claim severity in the presence of extreme values in the non-life insurance industry.

Composite distributions have also been found useful in modeling city sizes. [22] examined the distribution of the sizes of all French settlements (communes) by means of a three-parameter composite lognormal-Pareto distribution. [23] proposed a two-Pareto tail-lognormal distribution, consisting of a lower-tail Pareto, lognormal body, and upper-tail Pareto, to estimate the size distribution of U.S. cities. Through a literature review, we found that published papers have not considered the shapes, including skewness and kurtosis, of composite distributions, possibly because, as they are used to model insurance claims/loss data, they are all right-skewed.

Manuscript received Feb 09, 2021; revised Aug 18, 2021. This research was funded by the National Social Science Fund of China, Grant Number 19BTJ051.

Chao Wang is an Associate Professor in School of Mathematics and Statistics, Anyang Normal University, Anyang, 455000, China (e-mail: aynuchw@126.com).

Similar to [3], we propose a composite power function lognomal (PFL) distribution that can be used on both highly positively or negatively skewed data, and we extensively address its shapes.

Different from the literature, we propose to fit the logs of insurance claims, and not the original data, to a PFL distribution, which we believe is the first such use of a composite distribution.

The rest of this paper is organized as follows. Section 2 introduces the PFL distribution and describes its properties and shapes. Parameter estimation methods are derived, and simulation studies are presented in section 3. Probability plot method is proposed in section 4, and section 5 provieds an illustrative example. Section 6 presents some conclusions.

II. The power function Lognormal distribution and its properties

A. Power function lognormal distribution

Based on [3], let X be a random variable with probability density function (pdf)

$$f(x)=\begin{cases} wf_1(x), & 0<x\le\theta \\ wf_2(x), & \theta\le x<\infty, \end{cases} \tag{1}$$

where $0<w<1$ is the normalizing constant, and $f_1(x)$ and $f_2(x)$ have the forms of a power function distribution and lognormal distribution, respectively, i.e.

$$f_1(x)=\frac{\alpha x^{\alpha-1}}{\theta^\alpha}, 0<x\le\theta, \tag{2}$$

$$f_2(x)=\frac{1}{\sqrt{2\pi}x\sigma}\exp(-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})^2), x>0. \tag{3}$$

We note that the power function is the inverse of the Pareto random variable ([24], [25]), and is also known as the inverse Pareto ([26]) and the inverse power distribution ([27]). For the power function, $\alpha>0$ is a shape parameter and $\theta>0$ is a scale parameter; for the lognormal distribution, $\mu\in\Re$ is a location parameter and $\sigma>0$ is a scale parameter. To impose continuity and differentiability conditions on θ, we have

$$f_1(\theta)=f_2(\theta), f_1^{'}(\theta)=f_2^{'}(\theta),$$

where $f_1^{'}(\theta)$ and $f_2^{'}(\theta)$ are the first derivatives of $f_1(x)$ and $f_2(x)$, respectively, evaluated at θ. We get $\mu = \ln(\theta) + \alpha\sigma^2$, and $\alpha\sigma = (2\pi)^{-1/2}\exp(-0.5(\alpha\sigma)^2)$. Since $\int_0^{+\infty} f(x)\,\mathrm{d}x = 1$, we get $w\int_0^{\theta} f_1(x)\,\mathrm{d}x + w\int_{\theta}^{+\infty} f_2(x)\,\mathrm{d}x = 1$. Thus this composite density can be reparameterized and rewritten as

$$f(x) = \begin{cases} \dfrac{1}{1+\Phi(k)}\dfrac{\alpha x^{\alpha-1}}{\theta^{\alpha}}, & 0 < x \le \theta, \\ \dfrac{1}{1+\Phi(k)}\dfrac{\alpha x^{\alpha-1}}{\theta^{\alpha}} & \\ .\exp\left\{-\dfrac{1}{2}\dfrac{\alpha^2}{k^2}\ln^2(\dfrac{x}{\theta})\right\}, & \theta \le x < \infty, \end{cases} \tag{4}$$

where $\Phi(.)$ is the cumulative distribution function(cdf) of the standard normal distribution, and k is a known constant given by the positive solution of the equation $k = (2\pi)^{-1/2}\exp(-0.5k^2)$, which is $k = 0.372238898$. Here, $k = \alpha\sigma$ and $w = 1/[1+\Phi(k)] = 0.60785008$. Therefore, this composite probability density has only two unknown parameters, $\alpha > 0$ and $\theta > 0$. We refer to (4) as the power function lognormal model (PFL), PFL(α,θ), whose cdf, $F(x)$, and the quantile function, $X(p)$, are respectively given by

$$F(x) = \begin{cases} \dfrac{1}{1+\Phi(k)}\left(\dfrac{x}{\theta}\right)^{\alpha}, & 0 < x \le \theta, \\ \dfrac{1}{1+\Phi(k)} & \\ \left\{\Phi(k) + \Phi\left[\dfrac{\alpha}{k}\ln\left(\dfrac{x}{\theta}\right) - k\right]\right\}, & \theta \le x < \infty \end{cases} \tag{5}$$

and

$$X(p) = \begin{cases} \theta\left[(1+\Phi(k))p\right]^{1/\alpha} \\ ,0 < p \le \dfrac{1}{1+\Phi(k)} \\ \theta\exp\left\{\dfrac{k}{\alpha}\left[k + \Phi^{-1}\left((1+\Phi(k))p - \Phi(k)\right)\right]\right\}, \\ \dfrac{1}{1+\Phi(k)} \le p < 1 \end{cases} \tag{6}$$

It is clearly shown that the median of PFL(α,θ) is $X(0.5) = \theta[(1+\Phi(k))/2)]^{1/\alpha} < \theta$. Its moments can be obtained as

$$E(X^r) = \frac{\theta^r}{1+\Phi(k)}.$$
$$\left\{\frac{\alpha}{(\alpha+r)} + \exp\left[r(\alpha+0.5)\left(\frac{k}{\alpha}\right)^2\right]\Phi\left[\frac{k(\alpha+r)}{\alpha}\right]\right\} \tag{7}$$

B. *Properties of* PFL$(\alpha,\theta$)

For $\alpha < 1$, PFL(α,θ) has no modes. For $\alpha > 1$, PFL(α,θ) has a unique mode at $x = x_0 = \exp(\mu - \sigma^2) = \theta\exp[(\alpha-1)\sigma^2]$, which is larger than the threshold value θ. For $\alpha = 1$, since the power function distribution is simplified to a uniform distribution, PFL(α,θ) has many modes in the interval (0, θ]. Fig 1 demonstrates the shape of the pdf of the PFL distribution. It is easy to see how many behaviors it can have, as positive or negative skewness, and leptokurtic or platykurtic qualities change the values of the parameters. In other words, the flexibility of PFL model can give it broader applicability.

C. *Skewness and kurtosis*

It can be proved that the skewness coefficient (Sk) and kurtosis coefficient (Ku) of PFL distributions are only related to the parameter α. Table I gives some key values of Sk and Ku of PFL distributions under different α. For the skewness coefficients, it is symmetric for $\alpha = 1.722$, positively skewed for $\alpha < 1.722$, and negatively skewed for $\alpha > 1.722$. It can also be proved that when α tends to zero and positive infinity, the skewness coefficient tends to positive infinity and -1.8006, respectively. For the kurtosis coefficients, when α is between 1.2568 and 3.0214, the kurtosis coefficient is less than 3, and when α is less than 1.2568 or greater than 3.0214, it is greater than 3. When α tends to zero and positive infinity, the kurtosis coefficient tends to positive infinity and 8.2508, respectively.

Fig 2 and Fig 3 show plots of Sk(α) and Ku(α), respectively, of the PFL distribution. Fig 2(a) and 2(b) show plots of Sk(α) for $1 < \alpha < 5$ and $100 < \alpha < 1000$, respectively. Calculations show that the first derivative of Sk(α) is less than 0, so Sk(α) decreases on (0,+∞), and the minimum skewness is -1.8006. Fig 3(a) and 3(b) show plots of Ku(α) for $1 < \alpha < 5$ and $5 < \alpha < 200$, respectively. Ku(α) decreases for $\alpha < 1.8374$, and increases for $\alpha > 1.8374$. Ku(α) ranges from 2.6025 to 8.2508. Hence the skewness and kurtosis coefficients of PFL distribution can vary greatly in magnitude.

III. Parameter Estimation

A. *Maximum Likelihood Estimation*

Let $X_1, X_2, \cdots, X_n$, be a random sample from PFL(α,θ) . Without loss of generality, we can assume that this is an ordered sample, i.e., $x_1 \le x_2 \le x_3 \le \cdots \le x_n$. Suppose the unknown parameter θ is between the m^{th} and $m+1^{th}$ observation, i.e., $x_m \le \theta \le x_{m+1}$. Then the likelihood function can be written as

$$L(\boldsymbol{x},\boldsymbol{\omega}) = \prod_{i=1}^{m}\frac{1}{(1+\Phi(k))}\frac{\alpha x_i^{\alpha-1}}{\theta^{\alpha}} \cdot \prod_{i=m+1}^{n}\frac{1}{(1+\Phi(k))}\frac{\alpha x_i^{\alpha-1}}{\theta^{\alpha}}\exp\left[-\frac{1}{2}\frac{\alpha^2}{k^2}\ln^2\left(\frac{x_i}{\theta}\right)\right] \tag{8}$$

The log-likelihood function is

$$\ln L(\boldsymbol{x},\boldsymbol{\omega}) = -n\ln(1+\Phi(k)) + n\ln\alpha - n\alpha\ln\theta + (\alpha-1)\sum_{i=1}^{n}\ln x_i - \frac{\alpha^2}{2k^2}\sum_{i=m+1}^{n}(\ln x_i - \ln\theta)^2, \tag{9}$$

where $\boldsymbol{\omega} = (\alpha,\theta)^T$, since θ can only occur between x_1 and x_n. Therefore, the maximum likelihood estimators (MLEs) of ω can be obtained numerically as follows.

Step 1: Numerically find the value of α and θ that maximizes lnL(x,ω) for a given θ in the interval $(x_m, x_{m+1}), m = 1, 2, \cdots, n-1$. Thus, we obtain $(\hat{\alpha}_1, \hat{\theta}_1)$, $(\hat{\alpha}_2, \hat{\theta}_2)$,$\cdots$, $(\hat{\alpha}_{n-1}, \hat{\theta}_{n-1})$.

Step 2: Compute the corresponding log-likelihood function values from $n-1$ groups of MLEs.

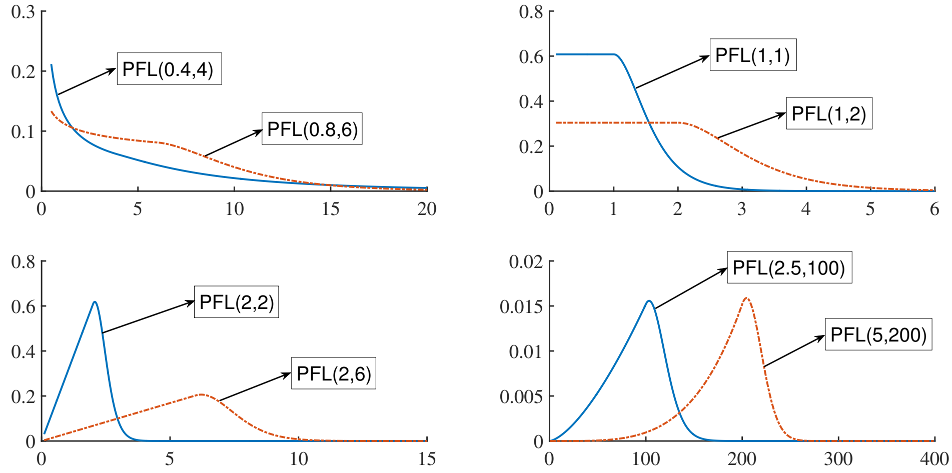

Fig 1. Plots of PFL distribution for selected parameter values

Table I: Skewness and Kurtosis of PFL distribution

Skewness	Kurtosis
$\alpha = 1.722$, Sk $= 0$	$\alpha \in (0, 1.2568) \cup (3.0214, +\infty)$, Ku > 3
$\alpha < 1.722$, Sk > 0	$\alpha \in (1.2568, 3.0214)$, Ku < 3
$\alpha > 1.722$, Sk < 0	$\alpha = 1.8347$, Ku$= 2.6025$(minimum)
$\alpha \to 0$, Sk $\to +\infty$	$\alpha \to 0$, Ku $\to +\infty$
$\alpha \to +\infty$, Sk $\to -1.8006$	$\alpha \to +\infty$, Ku $\to 8.2508$

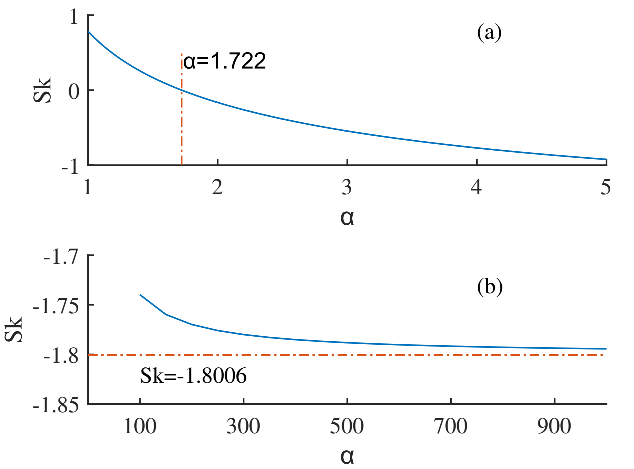

Fig 2. Plots of Sk(α) of PFL distribution

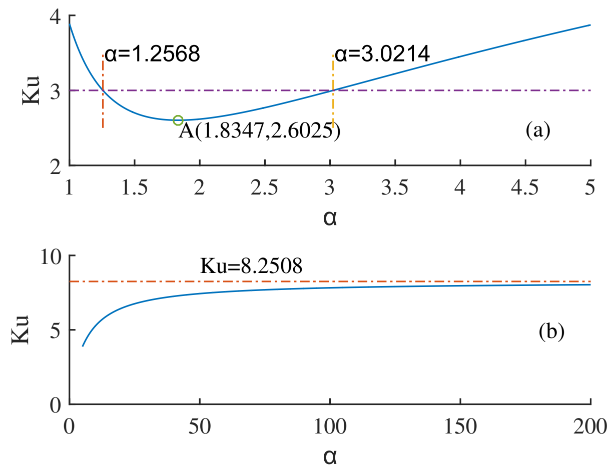

Fig 3. Plots of Ku(α) of PFL distribution

Step 3: Obtain the MLEs of PFL parameters (α, θ), denoted by $(\hat{\alpha}_{mle}, \hat{\theta}_{mle})$, by choosing the estimates with the largest log-likelihood function values among all $n-1$ groups of MLEs.

The asymptotic variance and covariance of the MLE's of the parameters are given by the inverse of the Fisher information matrix I(ω),

$$
\begin{aligned}
\mathrm{I}(\omega) &= -E\frac{\partial^2 \ln L(x;\omega)}{\partial\omega\partial\omega^T} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \\
&= E\begin{bmatrix} -\frac{\partial^2 \ln L(x;\omega)}{\partial\alpha^2} & -\frac{\partial^2 \ln L(x;\omega)}{\partial\alpha\partial\theta} \\ -\frac{\partial^2 \ln L(x;\omega)}{\partial\theta\partial\alpha} & -\frac{\partial^2 \ln L(x;\omega)}{\partial\theta^2} \end{bmatrix}
\end{aligned}
$$

The elements of I($\boldsymbol{\omega}$) are

$$\begin{aligned} I_{11} &= \frac{1}{\alpha^2(1+\Phi(k))} \\ &\quad \left\{m + (n-m)\Phi(k) + (n-m)\left[(1+k^2)\Phi(k) + k^2\right]\right\} \\ I_{12} &= I_{21} = \frac{(n-m)}{\theta} \\ I_{22} &= \frac{\alpha^2}{k^2\theta^2(1+\Phi(k))}\left[(n-m)\Phi(k) - (n-2m)k^2\right] \end{aligned}.$$

B. Method of Moments

The method of moments (MM) is a technique to construct estimators of the parameters based on matching sample moments with corresponding distribution moments. From (7), one can easily obtain the first and second raw moment of PFL. Under the method of moments, we equate E(X), E(X^2) to sample moments $\frac{1}{n}\sum_{i=1}^{n} x_i$ and $\frac{1}{n}\sum_{i=1}^{n} x_i^2$, respectively, and obtain a nonlinear system of equations,

$$\begin{cases} \frac{\theta}{1+\Phi(k)}\frac{\alpha}{(\alpha+1)} + \frac{\theta}{1+\Phi(k)} \\ \left\{\exp\left[(\alpha+0.5)\left(\frac{k}{\alpha}\right)^2\right]\Phi\left[\frac{k(\alpha+1)}{\alpha}\right]\right\} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \frac{\theta^2}{1+\Phi(k)}\frac{\alpha}{(\alpha+2)} + \frac{\theta^2}{1+\Phi(k)} \\ \left\{\exp\left[2(\alpha+0.5)\left(\frac{k}{\alpha}\right)^2\right]\Phi\left[\frac{k(\alpha+2)}{\alpha}\right]\right\} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases}, \tag{10}$$

which we can iteratively solve with respect to (α, θ) to obtain the MM estimators (MMEs) of α, θ, i.e., $\hat{\alpha}_{mm}, \hat{\theta}_{mm}$, respectively.

C. Nonlinear Least Squares Estimation

Nonlinear least squares (NLS) based on the relationship between the empirical cumulative distribution function and order statistics is frequently used to estimate parameters of distributions. Taking the logarithm of the cdf of PFL (5), we have

$$\ln F(x) = \begin{cases} -\ln\left(1+\Phi(k)\right) + \alpha\ln x - \alpha\ln\theta, \\ 0 < x \le \theta \\ -\ln(1+\Phi(k)) + \\ \ln\left\{\Phi(k) + \Phi\left[\frac{\alpha}{k}\ln\left(\frac{x}{\theta}\right) - k\right]\right\}, \\ \theta \le x < +\infty \end{cases} \tag{11}$$

In this method, it is necessary to use a plotting position to estimate the distribution function corresponding to the i^{th} order statistic. The most widely used expression is

$$p_i = \frac{i}{n+1}, \tag{12}$$

where n is the sample size, and p_i is the empirical (observed) distribution function corresponding to the i^{th} ordered dataset. The nonlinear least squares estimators (NLSEs) of α and θ, say $\hat{\alpha}_{nls}$, $\hat{\theta}_{nls}$, respectively, can be obtained by minimizing

$$G(\alpha, \theta) = \sum_{i=1}^{n}\left\{\ln p_i - \ln F(x)\right\}^2. \tag{13}$$

This problem is solved similarly to maximum likelihood estimation (9).

D. Bayes estimation

Following the Bayesian paradigm, we assume α and θ are independent, with prior $\pi(\alpha, \theta) \sim c/\theta$, where c is a constant. Thus, the joint density of the data, α and θ can be obtained as

$$\pi^*(\alpha, \theta|data) \propto \alpha^n\theta^{-n\alpha-1} \exp\left[\alpha\sum_{i=1}^{n}\ln x_i - \frac{1}{2}\frac{\alpha^2}{k^2}\sum_{i=m+1}^{n}(\ln x_i - \ln\theta)^2\right] \tag{14}$$

The posterior pdfs, of α conditional on θ and θ conditional on α, are

$$\pi_1^*(\alpha|\theta, data) \propto \alpha^n\theta^{-n\alpha} \exp\left[\alpha\sum_{i=1}^{n}\ln x_i - \frac{1}{2}\frac{\alpha^2}{k^2}\sum_{i=m+1}^{n}(\ln x_i - \ln\theta)^2\right], \tag{15}$$

$$\pi_2^*(\theta|\alpha, data) \propto LN\left(\frac{B_1}{A_1}, \frac{1}{A_1}\right) \tag{16}$$

where $A_1 = \frac{(n-m)\alpha^2}{k^2}, B_1 = \frac{\alpha^2\sum_{i=m+1}^{n}\ln x_i}{k^2} - n\alpha.$

It is clear that the posterior distribution of θ is lognormal, while that of α is unknown. Here, we use Gibbs sampling to generate random numbers from the posterior pdfs of α and θ, as follows:

(1). Start with initial guess $(\alpha_{(0)}, \theta_{(0)})$; we set $\alpha_{(0)} =$ mode(data), $\theta_{(0)} =$ min(data)+2;

(2). Set $j = 1$;

(3). Generate $\theta_{(j)}$ from conditional distribution LN($B_1^*/A_1^*, 1/A_1^*$), described in (15), where $A_1^* = \frac{(n-m)\left[\alpha^{(j-1)}\right]^2}{k^2}, B_1^* = \frac{\left[\alpha^{(j-1)}\right]^2\sum_{i=m+1}^{n}\ln x_i}{k^2} - n\alpha^{(j-1)}$;

(4). Use the Metropolis–Hastings algorithm to generate $\alpha^{(j)}$ from $\pi_1^*(\alpha^{(j-1)}|\theta^{(j-1)}, data)$, with proposed distribution N($\alpha^{(j-1)}, V_\alpha$);

(5). Set $j = j + 1$;

(6). Repeat steps (1)-(5) a large number of times, N, to get chains $\left[\alpha^{(1)}, \alpha^{(2)}, \cdots, \alpha^{(N)}\right]$ and $\left[\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(N)}\right]$;

(7). Obtain Bayes estimators of α and θ, say $\hat{\alpha}_{bayes}$, $\hat{\theta}_{bayes}$, respectively, as

$$\hat{\alpha}_{bayes} = \frac{1}{N-M}\sum_{i=M+1}^{N}\alpha^{(j)}, \hat{\theta}_{bayes} = \frac{1}{N-M}\sum_{i=M+1}^{N}\theta^{(j)},$$

where N is the length of chains and M is the burn-in period; we set $N = 10000$, $M = 5000$.

E. Simulation study

We conducted a Monte Carlo simulation study to compare the performance of the MLEs, MMEs, NLSEs, and Bayes estimators in terms of biases and mean square errors (MSEs) using 10,000 replications, examining sample sizes ranging from very small ($n = 10$) to very large ($n = 500$). Results were obtained for PFL (2.5, 8) and PFL (0.8, 10). We note that PFL (2.5, 8) is platykurtic with kurtosis coefficient 2.7704, and left-skewed with skewness coefficient -0.3843, and PFL (0.8, 10) is leptokurtic with kurtosis coefficient 5.8302, and right-skewed with skewness coefficient 1.2667.

To compute Bayes estimators, we assume the hyperparameter $V_\alpha = 1$. All computations were performed using MATLAB R2015b, programs are available from the authors upon request. The simulation results of parameter estimation for the two distributions are shown in Tables II to IX. We make the following observations from the simulation results.

(1) As n increases, the biases for MMEs, NLSEs, and Bayes estimators decrease in most cases. In all cases, MLEs have larger biases and MSEs than MMEs, NLSEs, and the Bayes estimators, i.e., MLEs perform worst among the four parameter estimation methods;

(2) It can also be shown that MMEs perform better than NLSEs in terms of both biases and MSEs, and Bayes estimators perform better than NLSEs in terms of both biases and MSEs when the sample size n is larger than 50;

(3) In most cases, MMEs perform better than Bayes estimators in terms of biases , and Bayes estimators better than MMEs in terms of MSEs. An interesting observation is that the MMEs are mainly positively biased, and the Bayes estimators mainly negatively biased;

The Bayes estimators show smaller absolute values of biases; hence, they work better than MMEs. From the presented results, Bayes estimators are recommended for parameter estimation of the PFL distribution in practical application.

Table II: Biases and MSEs of MLEs for PFL(2.5,8)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	-1.8677	0.0349	1.3086	0.0242
20	-1.5523	0.0482	1.9741	0.0874
50	-0.7974	0.0335	2.6895	0.3789
100	-0.8053	0.0708	2.8009	0.8022
200	-0.7998	0.1394	2.9093	1.7112
500	-0.8322	0.3591	2.9764	4.4502

Table III: Biases and MSEs of MLEs for PFL(0.8,10)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	-0.3234	0.0011	0.5274	0.2365
20	-0.1734	0.0010	3.2140	1.3982
50	-0.2035	0.0029	4.0183	4.1918
100	-0.2264	0.0060	4.4265	9.0424
200	-0.2347	0.0120	4.7939	19.4182
500	-0.2427	0.0307	5.1077	51.4917

Table IV: Biases and MSEs of MMEs for PFL(2.5,8)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	0.5246	0.0193	-0.0590	0.0053
20	0.2128	0.0104	-0.0266	0.0051
50	0.0804	0.0073	-0.0286	0.0051
100	0.0367	0.0069	-0.0069	0.0050
200	0.0156	0.0061	-0.0021	0.0052
500	0.0105	0.0065	-0.0041	0.0052

Table V: Biases and MSEs of MMEs for PFL(0.8,10)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	0.1753	0.0013	0.1653	0.0848
20	0.0882	0.0010	0.0208	0.0792
50	0.0290	0.0007	-0.0449	0.0775
100	0.0190	0.0007	0.0176	0.0764
200	0.0080	0.0007	-0.0035	0.0780
500	0.0040	0.0007	0.0167	0.0785

Table VI: Biases and MSEs of NLSEs for PFL(2.5,8)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	-0.1294	0.0204	0.0738	0.0076
20	-0.2544	0.0132	0.1066	0.0082
50	-0.1810	0.0139	0.0994	0.0092
100	-0.1381	0.0149	0.0881	0.0097
200	-0.0948	0.0148	0.0582	0.0099
500	-0.0519	0.0152	0.0405	0.0108

Table VII: Biases and MSEs of NLSEs for PFL(0.8,10)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	-0.0196	0.0021	0.6497	0.1194
20	-0.0646	0.0014	0.5887	0.1329
50	-0.0587	0.0014	0.5133	0.1649
100	-0.0507	0.0016	0.3890	0.1706
200	-0.0360	0.0017	0.3027	0.1589
500	-0.0191	0.0016	0.1630	0.1591

Table VIII: Biases and MSEs of Bayes estimators for PFL(2.5,8)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	-0.5189	0.0044	-0.1235	0.0065
20	-0.2869	0.0049	-0.1048	0.0030
50	-0.1100	0.0050	-0.0983	0.0049
100	-0.0498	0.0053	-0.0557	0.0048
200	-0.0252	0.0051	-0.0218	0.0048
500	-0.0127	0.0053	-0.0158	0.0050

Table IX: Biases and MSEs of Bayes estimators for PFL(0.8,10)

n	bias(α)	mse(α)	bias(θ)	mse(θ)
10	0.0210	0.0006	-0.1366	0.0202
20	0.0098	0.0005	-0.1187	0.0315
50	0.0067	0.0006	-0.1084	0.0753
100	0.0051	0.0006	-0.0851	0.0739
200	-0.0015	0.0005	-0.0676	0.0699
500	0.0009	0.0006	0.0548	0.0720

IV. Probability plot

The use of probability plots to evaluate distributional assumptions for a given sample has been discussed by a number of studies (e.g., [28], [29]). The basic idea is that

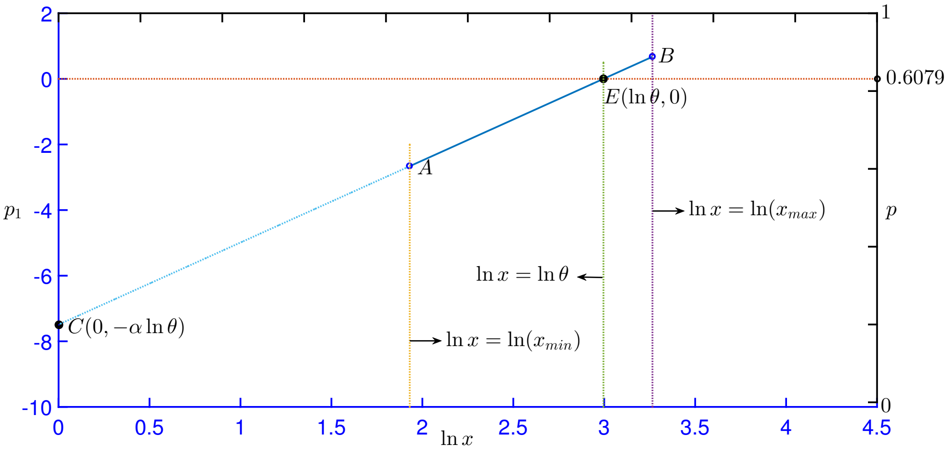

Fig 4. Probability plot for 100 sets of simulated PFL(2.5,20) data

the quantiles of the theoretical distribution when ploted against the sample order statistics will tend to yield a set of approximately linear points.

The Quantile function of the PFL is given by (6). Taking its logarithm of (6), we have

$$\ln x(p) = \begin{cases} \ln\theta + 1/\alpha\left[\ln(1+\Phi(k)) + \ln p\right] \\ , 0 < p \le \dfrac{1}{1+\Phi(k)} \\ \ln\theta + 1/\alpha\left\{k^2 + k\Phi^{-1}\left[(1+\Phi(k))p - \Phi(k)\right]\right\} \\ , \dfrac{1}{1+\Phi(k)} \le p < 1 \end{cases} \tag{17}$$

Equation (17) can be written as

$$\begin{cases} \ln(1+\Phi(k)) + \ln p = -\alpha\ln\theta + \alpha\ln x \\ , 0 < x \le \theta \\ k^2 + k\Phi^{-1}\left[(1+\Phi(k))p - \Phi(k)\right] = -\alpha\ln\theta + \alpha\ln x \\ , x \ge \theta \end{cases} \tag{18}$$

where $k = 0.372239$. By letting p_1 the following formula

$$p_1 = \begin{cases} \ln(1+\Phi(k)) + \ln p, \\ 0 < x \le \theta \\ k^2 + k\Phi^{-1}\left[(1+\Phi(k))p - \Phi(k)\right] \\ , x \ge \theta \end{cases} \tag{19}$$

Thus, (18) represents a linear relationship between p_1 and $\ln x$, with intercept $-\alpha\ln\theta$, and slope α. It can also be seen that when $p = 1/(1+\Phi(k)) = 0.6079$, $p_1 = 0$ and $x = \ln\theta$. The probability plot of 100 sets of simulated data from PFL(2.5, 20) with the random seed 1234 is shown in Fig 4. The solid line AB in Fig 4 is a probability plot drawn according to (17). The starting point A is the natural logarithm of the minimum value of the simulated data, denoted by $\ln(x_{min})$, and the ending point B is the natural logarithm $\ln(x_{max})$ of the maximum value of the simulated data. In this example, $\ln(x_{min}) = 1.9306, \ln(x_{max}) = 3.2651$. Point $C(0, -\alpha\ln\theta)$ is the intersection of AB and the vertical axis, and point $E(\ln\theta, 0)$ is the intersection of $\ln x = \ln\theta$ and $p_1 = 0$, which exactly corresponds to $p = 1/(1+\Phi(k)) = 0.6079$.

The line $p_1 = 0$ divides AB into two parts, for the upper part AE, p_1 adopts the formula $\ln(1+\Phi(k)) + \ln p$, for the lower part BE, p_1 adopts the formula $k^2 + k\Phi^{-1}\left[(1+\Phi(k))p - \Phi(k)\right]$. The straight line $p = 0.6079$ also divides AB into two parts, $p < 0.6079$ and $p > 0.6079$.

We determin whether the PFL distribution can be used to fit one dataset by probability plot as follows.

(1). Sort the data in ascending order, $x = (x_1, x_2, \cdots, x_n)$;

(2). Compute $p_i = \dfrac{i}{n+1}, i = 1, 2, \cdots, n$;

(3). Compute p_1 from (19): $p_{1i} = \ln(1+\Phi(k)) + \ln p_i$ when $p_i < 0.6079$, and $p_{1i} = k^2 + k\Phi^{-1}\left[(1+\Phi(k))p_i - \Phi(k)\right]$ when $p_i > 0.6079$;

(4). Plot p_1 verse $\ln x$; one can use PFL to model the dataset if the plot is approximately a straight line.

V. An Illustrative Example

We use a real insurance dataset to demonstrate the above methods. The data present basic dental claims on a policy with a deductible of 50 USD. This dataset may be found in the “actuar” add-on package for R (https://CRAN.R-project.org/package=actuar). There are 10 observations listed below

141 16 46 40 351 259 317 1511 107 567

Table X shows the basic descriptive statistics of the data. CLAIM clearly shows positive skewness, and is leptokurtic, which is typical of insurance claims data, while the logged CLAIM appears to be negative-skewed and leptokurtic. Hence, a negatively skewed distribution will be needed to model model the logged CLAIM, which is not common in insurance claims data. The same feature can be found in Fig 5, which shows the histogram of the CLAIM data that is heavily right-skewed, while the logs are obviously left-skewed.

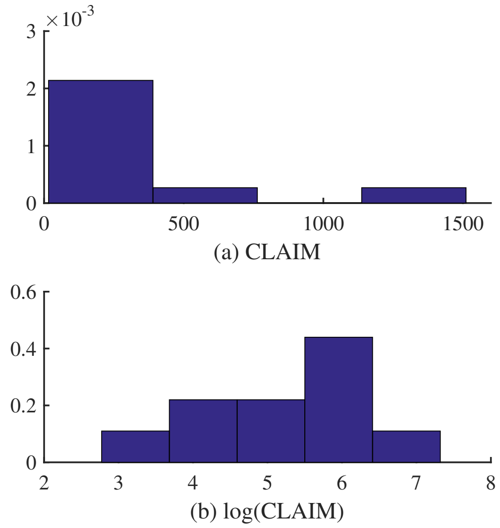

Fig 5. Histograms of CLAIM data

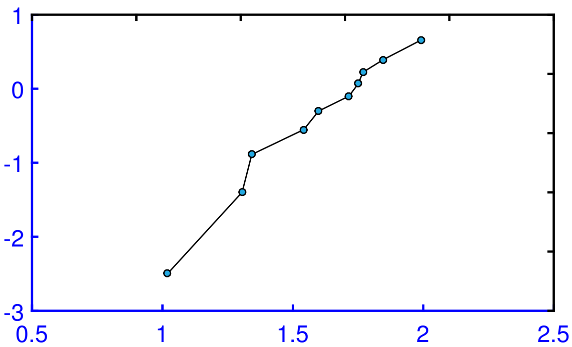

Fig 6. Probability plot for CLAIM data

Table X: Descriptive statistics of CLAIM data

	CLAIM	Log(CLAIM)
mean	335.5	2.204
standard error	447.8443	0.5953
minimum	16	1.2041
median	200	2.2813
maximum	1511	3.1793
skewness	2.3724	-0.1402
kurtosis	6.1947	-0.4541

Fig 6 shows the probability plot of the CLAIM data, which suggests that a PFL distribution will provide an adequate approximation.

For the CLAIM data, the Bayes estimates of α, θ, are 2.0428, 5.1947 respectively. Table XI reports test statistics and P-values (in brackets) for the three GoF tests commonly used in statistical literature. We consider Kolmogorov-Smirnov (K-S), Cramer-von-Mises (C-vM), and Anderson-Darling (A-D). P-values are all larger than 0.05, meaning that (at a 5% significance level) the logs of CLAIM data can be statistically described by a PFL distribution.

Table XI: EDF goodness-of-fit measures for fitted PFL to CLAIM data

K-S	C-vM	A-D
0.2021(0.8089)	0.1110(0.4181)	0.6205(0.5571)

VI. Conclusion

In this study, we proposed the power function lognormal distribution composite model, using the idea of [3] to model insurance claims data. In theory, the PFL distribution consists of a power function distribution before a specified threshold, and a lognormal distribution beyond it. This distribution can be used with highly positively or negatively skewed data. Parameter estimation methods were investigated, and Monte Carlo simulations carried out to display their performance. Probability plot method was also considered. Different from most literature, we find that the logged claims data, not the original claims data, can be properly described by a PFL distribution on the basis of probability plot and goodness-of-fit (GoF) tests. This provides a new idea for modeling insurance claims data.

References

[1] S. A. Klugmann, H. H. Panjer and G. E. Willmot, *Loss Models: From Data to Decisions*, 2nd ed. New York, NY, USA: John Wiley & Sons, 2004.

[2] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*. New York, NY, USA: John Wiley & Sons, 2003.

[3] K. Cooray and M.M. A. Ananda, “Modeling actuarial data with a composite Lognormal– Paretomodel,” *Scandinavian Actuarial Journal*, vol. 2005, no. 5, pp. 321-334, Sep. 2005.

[4] D. P. M. Scollnik, “On composite lognormal-Pareto models,” *Scandinavian Actuarial Journal*, vol. 2007, no. 1, pp. 20-33, 2007.

[5] S. Nadarajah and S. A. A. Bakar,“ New composite models for the Danish fire insurance data,” *Scandinavian Actuarial Journal*, vol. 2014, no. 2, pp. 180-187, 2014.

[6] V. Preda and R. Ciumara, “On Composite Models: Weibull-Pareto and Lognormal-Pareto. - A comparative study,” *Romanian Journal of Economic Forecasting*, vol. 3, no. 2, pp. 32-46, 2006.

[7] K. Cooray, “The Weibull–Pareto Composite Family with Applications to the Analysis of Unimodal Failure Rate Data,” *Communication in Statistics- Theory and Methods*, vol. 38, no. 11, pp. 1901 -1915, May. 2009.

[8] D. P. M. Scollnik and C. Sun, “Modeling with Weibull-Pareto Models,” *North American Actuarial Journal*, vol. 16, no. 2, pp. 260-272, 2012.

[9] K. Cooray, S. Gunasekera and M.M. A. Ananda, “Weibull and inverse Weibull composite distribution for modeling reliability data,” *Model Assisted Statistics and Applications*, vol. 5, no. 2, pp. 109-115, 2010.

[10] S. Teodorescu and R. Vernic, “A composite exponential-Pareto distribution,” *Analele Stiintifice ale Universitatii Ovidius Constanta*, vol. 14, no. 1, pp. 99-108, 2006.

[11] S. Teodorescu and R. Vernic, “Some composite exponential-Pareto models for actuarial prediction,” *Romanian Journal of Economic Forecasting*, vol. 12, no. 4, pp. 82-100, 2009.

[12] S. Teodorescu, “On the truncated composite lognormal-Pareto model,” *Mathematical Reports*, vol. 12, no. 1, pp. 71-84, 2010.

[13] S. Teodorescu and E. Panaitescu,“On the truncated composite Weibull-Pareto model,” *Mathematical Reports*, vol. 11, no. 61, pp. 259-273, 2009.

[14] M. Pigeon and M. Denuit,“Composite Lognormal–Pareto model with random threshold,” *Scandinavian Actuarial Journal* , vol. 2011, no. 3, pp. 177-192, 2011.

[15] E. Calderín-Ojeda,“On the Composite Weibull–Burr Model to describe claim data,” *Communications in Statistics - Case Studies, Data Analysis and Applications*, vol. 1, no. 1, pp. 59-69, 2015.

[16] E. Calderín-Ojeda and C. F. Kwok,“Modeling claims data with composite Stoppa models,” *Scandinavian Actuarial Journal*, vol. 2016, no. 9, pp. 817-836, 2016.

[17] E. Gómez-Déniz and E. Calderín-Ojeda,“Modeling insurance data with the Pareto Arctan distribution,” *ASTIN Bulletin*, vol. 45, no. 3, pp. 639-660, 2015.

[18] I. I. Eliazar and M. H. Cohen,“ A Langevin approach to the log-Gauss-Pareto composite statistical structure,” *Physica A Statistical Mechanics & Its Applications*, vol. 391, no. 22, pp. 5598-5610, 2012.

[19] R. J.Pak,“ A Robust Estimation for the Composite Lognormal-Pareto Model,” *Communications for Statistical Applications & Methods* , vol. 20, no. 4, pp. 311-320, 2013.

[20] K. Cooray and C. I.Cheng,“Bayesian estimators of the lognormal-Pareto composite distribution,” *Scandinavian Actuarial Journal*, vol. 2015, no. 6, pp. 500-515, 2015.

[21] Y. Wang,I. Hobk Haff and A. Huseby, “ Modelling extreme claims via composite models and threshold selection methods,” *Insurance: Mathematics and Economics*, vol.91, pp.257-268, 2020.

[22] E. Calderín-Ojeda,“The distribution of all French communes: A composite parametric approach,” *Physica A Statistical Mechanics & Its Applications*, vol.450, pp.385-394, 2016.

[23] J. Luckstead and S. Devadoss,“Pareto tails and lognormal body of US cities size distribution,” *Physica A Statistical Mechanics & Its Applications*, vol.465, pp.573-578, 2017.

[24] P. Rider,“Distribution of product and of quotient of maximum values in samples from a Power-function population,” *The Journal of American Statistical Association*, vol.59, pp.877-880, 1964.

[25] M. Meniconi and D. M.Barry,“The power function distribution: A useful and simple distribution to assess electrical component reliability,” *Microelectronics Reliability*, vol.36, pp.1207-1212, 1996.

[26] T. Moothathu,“A Characterization of the a-Mixture Pareto Distribution through a Property of Lorenz Curve,” *Sankhya the Indian Journal of Statistics*, vol.55, pp. 130-134, 1993.

[27] I. Malinowska and D. Szynal,“On characterization of certain distributions of kth lower (upper) record values,” *Applied Mathematics and Computation*, vol.202, pp.338-347, 2008.

[28] B. F. Kimbal,“On the Choice of Plotting Positions on Probability Paper,” *Journal of the American Statistical Association*, vol.55, no.291, pp. 546-560, 1960.

[29] W. L. Stephen and R. G. Thomas,“Probability Plotting Positions and Goodness of Fit for the Normal Distribution,” *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol.34, no.3, pp. 297-303, 1985.