Modelling Insurance Losses Using Contaminated Generalised Log Moyal Distribution

Project Report

submitted in partial fullfillment of the requirements for the award of degree of

Integrated

Master of Sciences

in

Statistics

Supervisor:

Dr. Deepesh Bhati

(Assistant Professor)

Submitted By:

Monika Gupta

Department of Statistics

School of Mathematics, Statistics and Computational Sciences

Central University of Rajasthan, Ajmer - 305817

2016-2021

Supervisor Name
Designation
Department of Statistics

Email :
Website : http://www.curaj.ac.in

राजस्थान कैन्द्रीय विश्वविद्यालय
Central University of Rajasthan
(संसद के अधिनियम के तहत स्थापित कैन्द्रीय विश्वविद्यालय)
(A Central University by an Act of Parliament)
NH-8, Bandarsindri, 305801
Kishangarh (Ajmer), Rajasthan, INDIA
Phone (Office): +91-1463-238755/260200
Telefax: +91-1463-238722

Certificate

This is to certify that the work embodied in the accompanying project report entitled ***"Modelling Insurance Losses Using Contaminated Generalised Log Moyal Distribution"*** has been successfully carried by ***Monika Gupta***, a 10^{th} Semester student, of the Department of Statistics, Central University of Rajasthan, under my guidance and supervision.

She worked for about 8 weeks and her work carried out is satisfactory.

Place: Bandarsindri

Date:

Dr. Deepesh Bhati
(Supervisor)

TO WHOMSOEVER IT MAY CONCERN

This is to certify that ***Monika Gupta***, Enroll. No 2016IMSST008, Integrated M.Sc. Statistics student of Department of Statistics has done project work ***"Modelling Insurance Losses Using Contaminated Generalised Log Moyal Distribution"*** under the guidance of ***Dr. Deepesh Bhati***, Department of Statistics, Central University of Rajasthan towards the partial fulfilment of the award of "Integrated Master of Science in Statistics" during the period January 2021 to July 2021.

Head
Department of Statistics
Central University of Rajasthan

Declaration

I Monika Gupta, hereby declare that the project work entitled **"Modelling Insurance Losses Using Contaminated Generalised Log Moyal Distribution"** submitted to the Department of Statistics, Central University of Rajasthan as a partial fulfilment of requirements of X-Semester examination is a bonafide record of work undertaken by me, under the supervision of **Dr. Deepesh Bhati**, Designation at Department of Statistics, Central University of Rajasthan and it is not formed the basis for the award of any other Degree/Associateship/Fellowship by any University.

Signature of Candidate

Student Name: Monika Gupta

Enroll. No.:2016IMSST008

Place: ....

Date: ....

Acknowledgement

To get us acquainted with the application in Statistics, the Dept. of Statistics, Central University Of Rajasthan has introduced project work to be carried out by each student for the Integrated M.Sc 10th-semester exam. I have been carrying out my project work on "**Modelling Insurance losses using Contaminated Generalised log Moyal Distribution**" under the supervision of **Dr Deepesh Bhati** (Assistant Professor) of our department. I would like to express my special thanks of gratitude to my Supervisor, Dr Deepesh Bhati, for the continuous support, guidance, inspiration, and immense knowledge, suggestions that helped me throughout project work. To all my batch mates of the Department of Computer Science at Central University of Rajasthan, thank you for your support and for helping me always! Finally, I would like to thank again to all my friends and faculty.

Monika Gupta
Department of Statistics
Enroll. No.:2016IMSST008

Contents

1 Abstract

The generalised log-Moyal (GlogM) has been proposed for modelling heavy-tailed data. In this project, we extend the GlogM family to the contaminated GlogM family. Properties of the contaminated distribution are derived. We further derive the Value at Risk (VaR) and Tail Conditional Expectation (TCE) risk measures (Jorion, 1997; Artzner et al., 1999; Landsman and Valdez, 2005). These measures are helpful in assessing model performance. The Expected Maximization algorithm method has been utilized for the estimation of parameters. Finally, a well-known fire insurance data set is used to analyze this newly proposed model.

2 Introduction

Often in the field of fire, motor, third party liability, catastrophe and other general insurance, the claim size distribution is modelled by heavy-tailed distribution or by a distribution having a tail heavier than exponential. Hence it is always of interest to propose a new class of distributions having a tail heavier than exponential. The two-parameter model generalised log-Moyal, introduced by Bhati and Ravi (2018), has been obtained by transformation and has uni-modality, right skewness and has tail heavier than exponential, which are desirable properties. The generalized log-Moyal distribution is denoted by GlogM(μ, σ). GlogM(μ, σ) is the generalization of the continuous Moyal distribution (Moyal, 1955) using the transformation method, exhibiting unimodality and right skewness with the right tail being heavier than the exponential model. The Generalized Log-Moyal distribution is suitable for modelling heavy-tailed data. This distribution can be related to some well-known distributions like Moyal, folded-normal,skew-normal and chi-square.

In this project, our interest focuses on the underlying heterogeneous groups of small and large claims. In this regard, a mixture distribution is defined as:

$$f(y) = (1 - w)f_1(y) + wf_2(y) \tag{1}$$

Where w is the unknown mixture weight and $f_i(y)$ are the mixture component densities. The derivatives of the log-likelihood function based on (1) is complicated, so the expectation-maximisation(EM) provides efficient ways to tackle the problem.

In the data analysis, we consider the Danish fire insurance loss data set. The Danish fire insurance data have been used for many newly developed statistical distributions involving heavy-tailed assumptions. In particular, it has been used to illustrate the fit of composite distributions. Some of the composite distributions fitted to the Danish fire insurance data are the composite log-normal Pareto distribution due to Cooray and Ananda (2005), the composite log-normal Pareto distribution due to Scollnik (2007), the composite log-normal–Pareto distribution due to Pigeon and Denuit (2011), the composite Weibull–Pareto distribution due to Scollnik and Sun (2012), the composite log-normal distributions due to Nadarajah and Bakar (2014), the composite Weibull distributions due to Bakar et al. (2015), and others.

2.1 Generalized Log-Moyal Distribution:

Definition:

The density function of GlogM(μ, σ) is given by:

$$f_Y(y) = \frac{\sqrt{\tau}}{\sqrt{2\pi}\sigma y}\left(\frac{1}{y}\right)^{1/2\sigma} e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}}, y > 0, \mu > 0, \sigma > 0 \qquad (2)$$

Where $\tau = \mu^{1/\sigma}$. and μ and σ are the scale and the shape parameters respectively. If the random variable $X = \sqrt{\tau}\left(\frac{1}{Y}\right)^{1/2\sigma}$, with random variable (rv) Y having pdf f , then X has a standard half-normal distribution.
Also , if rv Z has the pdf $g(.)$ proposed by Moyal (1995) : g(z)= $\frac{1}{\sqrt{2\pi}}exp(-(z+exp(-z))/2)$, z ϵ (-∞ , ∞), then Y = $\tau^{\sigma}e^{\sigma Z}$. We call the pdf f in (1) as the **generalized log-Moyal** pdf and denote it by GlogM(μ, σ) .

Distribution function:

The distribution function (df) of the generalized log-Moyal distribution is given by

$$F_Y(y) = \frac{1}{\Gamma\left(\frac{1}{2}\right)}\Gamma\left(\frac{1}{2}, \frac{\tau}{2}\left(\frac{1}{y}\right)^{1/2\sigma}\right) = erfc\left(\sqrt{\frac{\tau}{2}}\left(\frac{1}{y}\right)^{1/2\sigma)}\right), y > 0 \qquad (3)$$

Where erfc(.) is the complementary error function given as

$$erfc(z) = \frac{2}{\sqrt{\pi}}\int_z^{\infty} \exp(-t^2)dt, z > 0 \qquad (4)$$

2.2 Relationship of the generalised log-Moyal distribution with other distributions

The following proposition gives the relationships between GlogM(μ, σ).Some well known families of distributions are the following :

(a) If the rv Z follows the standard Moyal distribution then the rv Y = $\mu e^{\sigma Z} \sim GlogM(\mu, \sigma)$

b (b) If the rv X follows the half-normal (0 , σ^2) distribution ,then the rv Y = $\mu\left(\frac{\sigma}{X}\right)^{2\sigma} \sim GlogM(\mu, \sigma)$.

(c) If the rv X follows the generalized half-normal (θ, α) distribution ,(Cooray and Ananda ,2008),then the rv Y = 1/X $\sim GlogM(\mu = 1/\theta, \sigma = 1/2\alpha)$.

(d) If the rv $X \sim \chi^2_1$,then the rv $Y = \mu X^{-\sigma} \sim GlogM(\mu, \sigma)$

2.3 Distributional Properties of GLogM Distribution:

Here, statistical properties of the $GLogM(\mu, \sigma)$ distributions including moments, quantile function and mode are discussed.

2.3.1 Moments:

The r-th moment about origin of $Y \sim GLogM(\mu, \sigma)$ is given by :

$$E(Y^r) = \int_0^\infty y^r f(y)dy = \frac{\tau^{r\sigma}}{2^{r\sigma}\sqrt{\pi}}\Upsilon_\sigma(r), \qquad r < \frac{1}{2\sigma} \tag{5}$$

where $\Upsilon_\sigma(r) = \Gamma(\frac{1}{2} - r\sigma)$.

Hence mean of Y is

$$E(Y) = \frac{\tau^{\sigma}}{2^{\sigma}\sqrt{\pi}}\Upsilon_\sigma(1), \qquad \sigma < 1/2 \tag{6}$$

And variance of Y is

$$V(Y) = \frac{\tau^{2\sigma}}{2^{2\sigma}\pi}\left(\Upsilon_\sigma(2) - \frac{1}{\sqrt{\pi}}\Upsilon^2_\sigma(1)\right), \qquad \sigma < 1/4 \tag{7}$$

So the mean and variance of the GlogM(μ, σ) distribution exist if and only $\sigma < 1/2$ and $\sigma < 1/4$, respectively

Skewness of Y is

$$\gamma_0 = \frac{\mu_3}{{\mu_2}^{\frac{3}{2}}} = \frac{\dfrac{c_3}{\sqrt{\pi}} - \dfrac{3c_1c_2}{\pi} + \dfrac{2{c_1}^3}{\pi\sqrt{\pi}}}{\left(\dfrac{c_3}{\sqrt{\pi}} - \dfrac{{c_1}^2}{\pi}\right)^{3/2}} \tag{8}$$

Kurtosis of Y is

$$\kappa_0 = \frac{\mu_4}{{\mu_2}^2} = \frac{\dfrac{c_4}{\sqrt{\pi}} - \dfrac{4c_1c_3}{\pi} + \dfrac{6c_1^2c_2}{\pi\sqrt{\pi}} - \dfrac{3{c_1}^4}{\pi^2}}{\left(\dfrac{c_3}{\sqrt{\pi}} - \dfrac{{c_1}^2}{\pi}\right)^{2}} \tag{9}$$

where $c_r = \dfrac{\Upsilon_\sigma(r)}{2^{r\sigma}}$

2.3.2 Quantile function of GlogM distribution

The ϱ-th quantile $y_{\varrho} = min\{y : F(y) \geq \varrho\},\ 0 < \varrho < 1$, of GlogM(μ, σ) is obtained by inverting the df(2) and is given by

$$y_{\varrho} = \tau^{\sigma} \left(\sqrt{2} erfc^{\leftarrow}(\varrho) \right)^{-2\sigma} \tag{10}$$

Where $erfc^{\leftarrow}$ is the inverse of $erfc$. Then the median of Y is obtained by setting ϱ =1/2 , and is given by

$$y_{1/2} = \tau^{\sigma} \left(\sqrt{2} erfc^{\leftarrow}(1/2) \right)^{-2\sigma} = \tau^{\sigma} (0.6745)^{-2\sigma}$$

2.3.3 Uni modality of the GlogM distribution

The mode of GlogM(μ, σ) can be find by taking the logarthim of the pdf of GlogM distribution and equating it to zero.
Let y_0 be the mode of GlogM distribution, then the y_0 will be:

$$y_0 = \frac{\tau^{\sigma}}{(1 + 2\sigma)^{\sigma}} \tag{11}$$

This shows that GlogM(μ, σ) is uni-modal.

2.4 Actuarial Measures of the GlogM distribution

Two important actuarial measures of GlogM distribution are as follow:

(i)**Value at Risk(VaR)**:

It is a measure of the risk of loss for investments. It estimates how much a set of investments might lose (with a given probability), given normal market conditions, in a set time period such as a day.
Mathematically, VaR of a random variable Y is the q-th quantile of its distribution function, i.e.,
$VaR_q(Y) = min\{y : F(y) \geq q\},\ 0 < q < 1$ If the random variable $Y \sim GlogM(\mu, \sigma)$, then

$$VaR_q(Y) = \tau^{\sigma} \left(\sqrt{2} erfc^{\leftarrow}(q) \right)^{-2\sigma} \tag{12}$$

(ii)**Tail value at risk(TVaR)**:

It is also known as tail conditional expectation (TCE) or conditional tail expectation (CTE). It is a risk measure associated with the more general value at risk. It quantifies the expected value of the loss given that an event outside a given probability level has occurred.
Mathematically, TVaR is defined as:

$$TVaR = E(Y|Y > VaR_q(Y)) \tag{13}$$

If the random variable $Y \sim GlogM(\mu, \sigma)$, then

$$\begin{aligned} TVaR_q(Y) &= \frac{1}{1-q}\int_{VaR_q(Y))}^{\infty} y f_Y(y) dy \\ &= \frac{(\tau/2)^{\sigma}}{1-q}\left(\Upsilon_{\sigma}(1) - \Gamma\left(\frac{1}{2}, (erfc^{\leftarrow}(1-\varrho))^2)\right)\right), \;\; \sigma < 1/2 \end{aligned} \tag{14}$$

3 Contaminated Generalised log-Moyal distribution

3.1 Density function of Contaminated Generalized Log-Moyal Distribution:

.The contaminated generalised log-Moyal distribution is a mixture of two generalised log-Moyal distributions with mixing probabilities (1 - w) and w, where $w \in (0.0.5)$The density function of CGlogM(τ, σ, w, k) is given by:

$$\begin{aligned}
f_{CGlogM}(y|\tau,\sigma,w,k) &= (1-w)f_{GlogM}(y|\tau,\sigma) + wf_{GlogM}(y|k\tau,\sigma) \\
&= (1-w)\frac{\sqrt{\tau}}{\sqrt{2\pi\sigma y}}\left(\frac{1}{y}\right)^{1/2\sigma} e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} + w\frac{\sqrt{k\tau}}{\sqrt{2\pi\sigma y}}\left(\frac{1}{y}\right)^{1/2\sigma} e^{-\frac{k\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} \\
&= \frac{(1-w)\sqrt{\tau}}{\sqrt{2\pi\sigma}}\left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1} e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} + \frac{w\sqrt{k\tau}}{\sqrt{2\pi\sigma y}}\left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1} e^{-\frac{k\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}}
\end{aligned}$$

$$f_{CGlogM}(y|\tau,\sigma,w,k) = \frac{\sqrt{\tau}}{\sqrt{2\pi\sigma y}}\left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1}\left[(1-w)e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} + we^{-\frac{k\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}}\right] \tag{15}$$

where $f_{GlogM}(.)$ is given by (11). w is the parameter for the degree of contamination, $w \in (0.0.5)$ and k is the contaminating dispersion parameter k is set such that k>1 to avoid identifiability problem. The parameter k indicates the amount of inflated dispersion for the contaminated mixture component. GlogM is a special case of CGlogM when k=1. We call the pdf f in (11) as the **Contaminated generalized log-Moyal** pdf and denote it by CGlogM(τ, σ, w, k) .

Distribution function:

The distribution function (df) of the contaminated generalized log-Moyal distribution is given by

$$F_{CGlogM}(y|\tau,\sigma,w,k) = (1-w)\int_0^y f_{GlogM}(t|\tau,\sigma)dt + w\int_0^y f_{GlogM}(t|k,\tau,\sigma)dt$$

$$F_Y(y) = (1-w)\ erfc\left(\sqrt{\frac{\tau}{2}}\left(\frac{1}{y}\right)^{1/2\sigma)}\right) + w\ erfc\left(\sqrt{\frac{k\tau}{2}}\left(\frac{1}{y}\right)^{1/2\sigma)}\right), y > 0 \tag{16}$$

3.2 Distributional Properties of CGlogM Distribution:

Here, statistical properties of the $CGlogM(\tau, \sigma, w, k)$ distributions including moments, quantile function and mode are discussed.

3.2.1 Moments:

The r-th moment about origin of $Y \sim CGlogM(\tau, \sigma, w, k)$ is given by :

$$\begin{aligned}
E(Y^r) &= \int_0^\infty y^r f_{CGlogM}(y|\tau, \sigma, w, k)dy \\
&= \int_0^\infty y^r[(1-w)f_{GlogM}(y|\tau, \sigma) + wf_{GlogM}(y|k, \tau, \sigma)]dy \\
&= \int_0^\infty y^r(1-w)f_{GlogM}(y|\tau, \sigma)dy + \int_0^\infty y^r w f_{GlogM}(y|k, \tau, \sigma)]dy \\
&= \int_0^\infty y^r \frac{(1-w)\sqrt{\tau}}{\sqrt{2\pi}\sigma}\left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1} e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} dy + \int_0^\infty y^r \frac{w\sqrt{k\tau}}{\sqrt{2\pi}\sigma y}\left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1} e^{-\frac{k\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} \\
&= (1-w+wk^r)\frac{\tau^{r\sigma}}{2^{r\sigma}\sqrt{\pi}}\Upsilon_\sigma(r), \qquad r < \frac{1}{2\sigma} \qquad (17)
\end{aligned}$$

where $\Upsilon_\sigma(r) = \Gamma(\frac{1}{2} - r\sigma)$.

Hence mean of Y is

$$E(Y) = (1-w+wk)\frac{\tau^{\sigma}}{2^{\sigma}\sqrt{\pi}}\Upsilon_\sigma(1), \qquad \sigma < 1/2 \qquad (18)$$

And variance of Y is

$$V(Y) = (1-w+wk^2)\frac{\tau^{2\sigma}}{2^{2\sigma}\pi}\Upsilon_\sigma(2) - (1-w+wk)^2\frac{\tau^{2\sigma}}{2^{2\sigma}\pi\sqrt{\pi}}\Upsilon_\sigma^2(1), \qquad \sigma < 1/4 \qquad (19)$$

So the mean and variance of the CGlogM(μ, σ) distribution exist if and only $\sigma < 1/2$ and $\sigma < 1/4$, respectively.

Skewness of Y is

$$\begin{aligned}
\gamma_0 &= \frac{\mu_3}{{\mu_2}^{\frac{3}{2}}} \\
&= \frac{(1-w+wk^3)\frac{c_3}{\sqrt{\pi}} - 3(1-w+wk)(1-w+wk^2)\frac{c_1c_2}{\pi} + 2(1-w+wk)^3\frac{{c_1}^3}{\pi\sqrt{\pi}}}{\left((1-w+wk^2)\frac{c_2}{\sqrt{\pi}} - (1-w+wk)^2\frac{{c_1}^2}{\pi}\right)^{3/2}} \qquad (20)
\end{aligned}$$

Kurtosis of Y is

$$\begin{aligned}\kappa_0 &= \frac{\mu_4}{{\mu_2}^2} \\ &= \frac{(1-w+wk^4)\dfrac{c_4}{\sqrt{\pi}} - 4(1-w+wk)(1-w+wk^3)\dfrac{c_1c_3}{\pi}}{\left((1-w+wk^2)\,\dfrac{c_3}{\sqrt{\pi}} - (1-w+wk)^2\,\dfrac{{c_1}^2}{\pi}\right)^2} \\ &+ \frac{6(1-w+wk)^2(1-w+wk^2)\dfrac{c_1^2c_2}{\pi\sqrt{\pi}} - 3(1-w+wk)^4\dfrac{{c_1}^4}{\pi^2}}{\left((1-w+wk^2)\,\dfrac{c_3}{\sqrt{\pi}} - (1-w+wk)^2\,\dfrac{{c_1}^2}{\pi}\right)^2} \qquad (21)\end{aligned}$$

where $c_r = \frac{\Upsilon_\sigma(r)}{2^{r\sigma}}$

The limits for γ_0 and κ_0 when k tends to ∞ are:

$$\gamma_1 = \lim_{k\to\infty}\gamma_0 = \frac{\dfrac{c_3}{\sqrt{\pi}} - \dfrac{3w\;c_1c_2}{\pi} + \dfrac{2w^2{c_1}^3}{\pi\sqrt{\pi}}}{w^{1/2}\left(\dfrac{c_2}{\sqrt{\pi}} - \dfrac{w{c_1}^2}{\pi}\right)^{3/2}} \qquad (22)$$

$$\kappa_1 = \lim_{k\to\infty}\kappa_0 = \frac{\dfrac{c_4}{\sqrt{\pi}} - \dfrac{4w\;c_1c_3}{\pi} + \dfrac{6w^2c_1^2c_2}{\pi\sqrt{\pi}} - \dfrac{3w^3{c_1}^4}{\pi^2}}{w\left(\dfrac{c_2}{\sqrt{\pi}} - \dfrac{w{c_1}^2}{\pi}\right)^2} \qquad (23)$$

łAs large claims are rare, w often tends to zero when k tends to infinity. Under these conditions, (21) and (22) show that γ_1 and κ_1 tends to infinity but at different rates

3.2.2 Mode of the contaminated CGlogM distribution

The mode of CGlogM(τ, σ, w, k) can be find by taking the logarithm of the pdf of CGlogM distribution and equating it to zero. Mode for CGlogM can be obtained by solving:

$$(1-w)\left[\frac{\tau}{2\sigma}\left(\frac{1}{y}\right)^{\frac{1}{\sigma}} - \frac{1}{2\sigma} - 1\right] + w\;\left[\frac{k\tau}{2\sigma}\left(\frac{1}{y}\right)^{\frac{1}{\sigma}} - \frac{1}{2\sigma} - 1\right] = 0$$

$$\Rightarrow (1-w)\left[\frac{1}{2\sigma}\left(\frac{\tau}{2\sigma}\left(\frac{1}{y}\right)^{\frac{1}{\sigma}} - 1\right) - 1\right] + w\;\left[\frac{1}{2\sigma}\left(\frac{k\tau}{2\sigma}\left(\frac{1}{y}\right)^{\frac{1}{\sigma}} - 1\right) - 1\right] = 0 \qquad (24)$$

3.3 Actuarial Measures of the CGlogM distribution

One important actuarial measure of CGlogM distribution is as follow:

(i)**Tail value at risk(TVaR)**:

If the random variable $Y \sim CGlogM(\tau, \sigma, k, w)$, then

$$\begin{aligned}
TVaR_{CGogm.q} &= \frac{1-w}{1-q}\int_{VaR_q(Y))}^{\infty} y f_Y(y)dy + \frac{w}{1-q}\int_{VaR_q(Y))}^{\infty} y f_Y(y)dy \\
&= \frac{(1-w)2^{-\sigma}\tau^{\sigma}}{1-q}\left(\Upsilon_{\sigma}(1) - \Gamma\left(\frac{1}{2}, (erfc^{\leftarrow}(1-\varrho))^2)\right)\right) \\
&+ \frac{w\ 2^{-\sigma}\tau^{\sigma}}{1-q}\left(\Upsilon_{\sigma}(1) - \Gamma\left(\frac{1}{2}, (erfc^{\leftarrow}(1-\varrho))^2)\right)\right) \\
&= \frac{(1-w+wk)2^{-\sigma}\tau^{\sigma}}{1-q}\left(\Upsilon_{\sigma}(1) - \Gamma\left(\frac{1}{2}, (erfc^{\leftarrow}(1-\varrho))^2)\right)\right), \quad \sigma < 1/2
\end{aligned} \tag{25}$$

Now, plotting the pdf when one parameter remains fixed to show how the parameters affect the shape of CGlogM distribution. Figure 1 plots the density graphs for different CGlogM distribution when one parameter varies, keeping other parameters fixed.

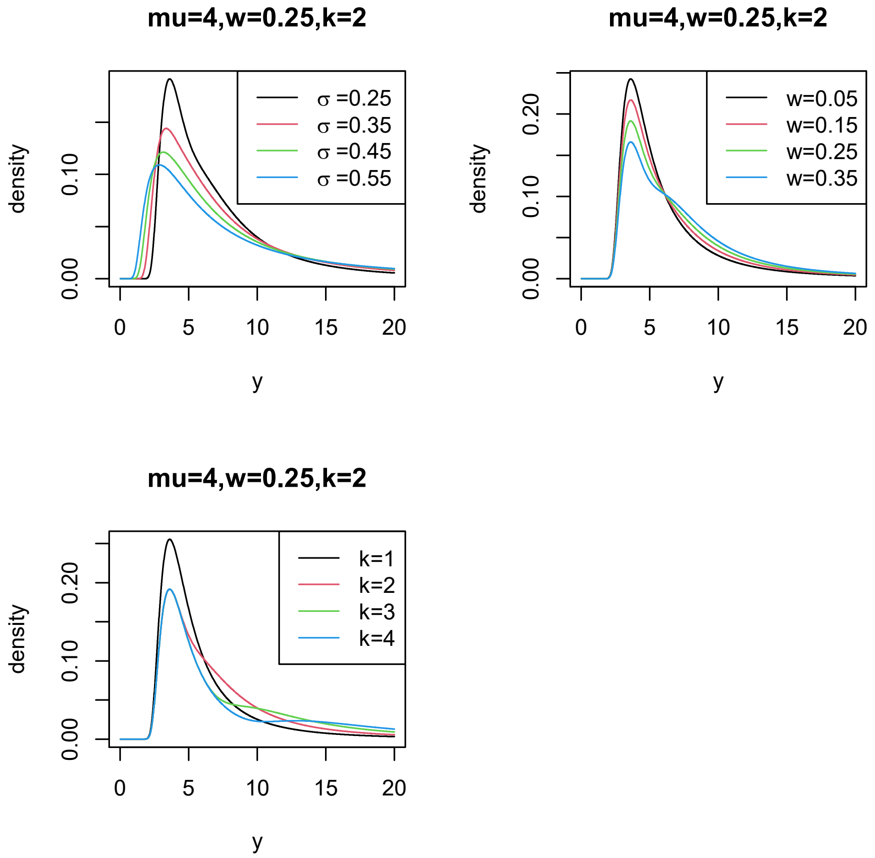

Figure 1: Density graphs for different CGlogM distribution

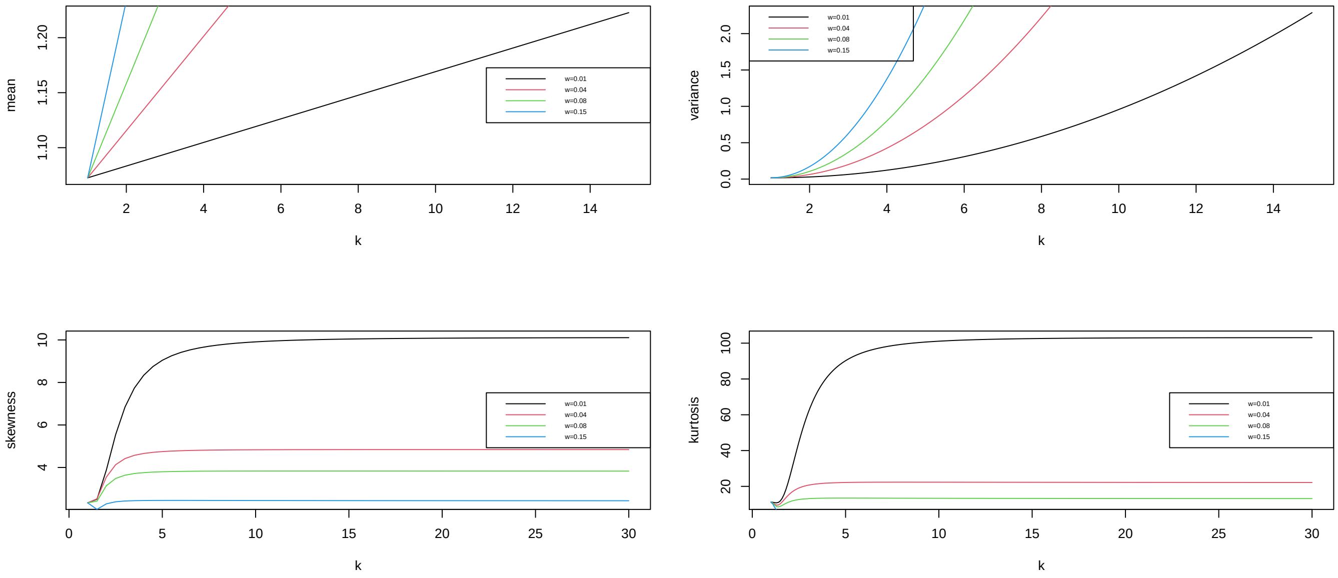

Figure 2: change of mean,var,skew,kurt against k for various levels of w

4 Estimation

4.1 EM algorithm:

The EM algorithm was proposed in 1997 by Arthur Dempster 9 Laird and Donald Rubin. The EM algorithm is an approach for obtaining maximum likelihood estimates of parameters of a statistical model in case there are latent variables, i.e. unobserved variables are present, or some of the data is missing, or data is incomplete. The latent variables are variables that are not directly observable. In such a case, we apply the EM algorithm. So the EM algorithm follows the steps to find the relevant model parameters in the presence of latent variables. The EM algorithm has many applications throughout bayesian statistics. This algorithm is often used in data clustering in machine learning and data mining applications and in Bayesian statistics, where it is used to obtain the model of the posterior marginal distributions of parameters.

Firstly, we assume that the complete data-set consists of $Z = (X, Y)$ but that only X is observed. The complete-data log-likelihood is then denoted by $l(\theta; X, Y)$ where θ is the unknown parameter vector for which we wish to find the MLE.

E-step: The E-step is used to estimate the values of the missing values in the data. It computes the expected value of $l(\theta; X, Y)$ given the observed data, X, and the current parameter estimate, θ_{old} say. In particular, we define

$$Q(\theta; \theta_{old}) := E[l(\theta; X, Y)|X, \theta_{old}] = \int l(\theta; X, y)p(y|X, \theta_{old})dy \tag{26}$$

where $p(|X, \theta_{old})$ is the conditional density of Y given the observed data, X,and assuming $\theta = \theta_{old}$.

M-Step: The M-step consists of maximizing over θ the expectation computed in (1). This step generates complete data after the expectation step and updates the missing values in the data. That is, we set

$$\theta_{new} = maxQ(\theta; \theta_{old}) \tag{27}$$

We then set $\theta_{old} = \theta_{new}$. These two E and M steps are repeated as necessary until the sequence of θ_{new}'s converges. If the log-likelihood function has multiple local maximums; then the this algorithm should be run many times, using a different starting value of θ_{old} on each occasion. Then the ML estimate of θ is taken as the best of the set of local maximums obtained from the various runs of the EM algorithm.

4.2 EM algorithm for contaminated GLogM distribution:

As,the density function of CGlogM(τ, σ, w, k) is given by:

$$f_{CGlogM}(y|\tau, \sigma, w, k) = \frac{\sqrt{\tau}}{\sqrt{2\pi\sigma y}} \left(\frac{1}{y}\right)^{\frac{1}{2\sigma}+1} \left[(1-w)e^{-\frac{\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}} + we^{-\frac{k\tau}{2}\left(\frac{1}{y}\right)^{1/\sigma}}\right] \tag{28}$$

In order to estimate the parameter of (26), usual MLE will not be obtained in closed form, and hence numerical procedures such as the newton-raphson method or others need to be followed. But taking advantage of its stochastic representation. The expectation-Maximization method can be used as follows:

Let $Y_i|Z_i = 0 \sim GlogM(\tau, \sigma)$ and $Y_i|Z_i = 1 \sim GlogM(k\tau, \sigma)$, so the marginal of Y_i is

$$P(Y_i = y) = (1-w)P(Y_i = y|Z_i = 0) + w\ P(Y_i = y|Z_i = 1) \tag{29}$$

$$= (1-w)f_{GlogM}(y|\tau, \sigma) + wf_{GlogM}(y|k\tau, \sigma) \tag{30}$$

Similarly the joint probability of observations $Y_1, Y_2, ..., Y_n$ is therefore

$$P[Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n] = \prod_{i=1}^{n}[(1-w)f_{GlogM}(y_i|\tau, \sigma) + w\ f_{GlogM}(y_i|k\tau, \sigma)] \tag{31}$$

Our unknown parameters $\underline{\theta} = (w, \tau, k, \sigma)$, which need to be estimated.

Intuitively, the latent variable Z_i should help us finding the MLE. We first try to compute the posterior distribution of Z_i given the observations.

$$P(Z_i|X_i) = \frac{P(Y_i|Z_i = z)P(Z_i = z)}{P(Y_i)} \tag{32}$$

$$= \begin{cases} \frac{(1-w)f_Y(y_i|\tau,\sigma)}{f_Y(y)} & ;\ z = 0 \\ \frac{wf_Y(y_i|k\tau,\sigma)}{f_Y(y)} & ;\ z = 1 \end{cases}$$

Here $f_Y(y_i)$ is density defined in (27).

The complete likelihood takes the form

$$\mathcal{L}_{Y,Z}(y_i, z_i) = \prod_{i=1}^{n}(1-w)^{1-z_i}w^{z_i}(f_Y(y_i|\tau,\sigma))^{1-z_i}(f_Y(y_i|k\tau,\sigma))^{z_i} \tag{33}$$

Taking log both sides,

$$log\mathcal{L}(w,\tau,\sigma,k) = \sum_{i=1}^{n}(1-z_i)log(1-w)+\sum_{i=1}^{n}z_ilogw+\sum_{i=1}^{n}(1-z_i)logf_Y(y_i|\tau,\sigma)+\sum_{i=1}^{n}z_ilogf_Y(y_i|k\tau,\sigma) \tag{34}$$

$$\begin{aligned} l(w,\tau,\sigma,k) = &\sum_{i=1}^{n}(1-z_i)log(1-w) + logw\sum_{i=1}^{n}z_i + \sum_{i=1}^{n}(1-z_i)\frac{log\tau}{2} - \sum_{i=1}^{n}(1-z_i)\frac{log2\pi}{2} \\ &-log\sigma\sum_{i=1}^{n}(1-z_i) - \left(\frac{1}{2\sigma}+1\right)\sum_{i=1}^{n}(1-z_i)logy_i - \frac{\tau}{2}\sum_{i=1}^{n}(1-z_i)\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} \\ &+\sum_{i=1}^{n}z_i\frac{logk}{2} + \sum_{i=1}^{n}z_i\frac{log\tau}{2} - \sum_{i=1}^{n}z_i\frac{log2\pi}{2} - log\sigma\sum_{i=1}^{n}z_i \\ &-\left(\frac{1}{2\sigma}+1\right)\sum_{i=1}^{n}z_ilogy_i - \frac{k\tau}{2}\sum_{i=1}^{n}z_i\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} \end{aligned} \tag{35}$$

Now, differentiating with respect to $\underline{\theta}$,

$$\frac{\partial l}{\partial w} = -\frac{1}{1-w}\sum_{i=1}^{n}(1-z_i) + \frac{1}{w}\sum_{i=1}^{n}z_i = 0 \ gives \ \widehat{w} = \frac{1}{n}\sum_{i=1}^{n}z_i \tag{36}$$

$$\frac{\partial l}{\partial \tau} = \frac{1}{2\tau}\sum_{i=1}^{n}(1-z_i) - \frac{1}{2}\sum_{i=1}^{n}(1-z_i)\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} + \frac{1}{2\tau}\sum_{i=1}^{n}z_i - \frac{k}{2}\sum_{i=1}^{n}z_i\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} = 0$$

$$= \frac{1}{2\tau}\left[\sum_{i=1}^{n}(1-z_i) + \sum_{i=1}^{n} z_i\right] - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} - \frac{(k-1)}{2}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} = 0$$

$$\Rightarrow \widehat{\tau} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} + \frac{(k-1)}{n}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}}} \tag{37}$$

$$\begin{aligned}
\frac{\partial l}{\partial \sigma} &= \frac{1}{\sigma}\sum_{i=1}^{n}(1-z_i) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(1-z_i) log y_i - \frac{1}{\sigma}\sum_{i=1}^{n} z_i + + \frac{1}{2\sigma^2}\sum_{i=1}^{n} z_i log y_i \\
&- \frac{\tau}{2}\sum_{i=1}^{n}(1-z_i)\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} \frac{log\ y_i}{\sigma^2} - \frac{k\tau}{2}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} \frac{log\ y_i}{\sigma^2} \\
&= \frac{n}{\sigma} + \frac{1}{2\sigma^2}\sum_{i=1}^{n} log y_i - \frac{\tau}{2\sigma^2}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} log\ y_i - \frac{(k-1)\tau}{2\sigma^2}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} log\ y_i = 0 \\
\Rightarrow \widehat{\sigma} &= \frac{1}{2}\sum_{i=1}^{n} log\ y_i - \frac{\tau}{2}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} log\ y_i - \frac{(k-1)\tau}{2}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} log\ y_i && (38) \\
\frac{\partial l}{\partial k} &= \frac{1}{2k}\sum_{i=1}^{n} z_i - \frac{\tau}{2}\sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}} = 0 \\
\Rightarrow \widehat{k} &= \frac{\sum_{i=1}^{n} z_i}{\tau \sum_{i=1}^{n} z_i \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma}}} && (39)
\end{aligned}$$

E-step:

$$E(Z|Y_i = y_i, \underline{\theta}) = w_i^{(r)} = \frac{w_i^{(r)} f(y_i|k^{(r)}, \tau^{(r)}, \sigma^{(r)})}{f(y_i|k^{(r)}, \tau^{(r)}, \sigma^{(r)})} \tag{40}$$

M-step:

$$w^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(r)} \tag{41}$$

$$\tau^{(r+1)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma^r}} + \frac{(k^{(r)}-1)}{n}} \tag{42}$$

$$k^{(r+1)} = \frac{\sum_{i=1}^{n} w_i^{(r)}}{\tau^{(r)} \sum_{i=1}^{n} w_i^{(r)} \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma^r}}} \tag{43}$$

$$\sigma^{(r+1)} = \frac{1}{2}\sum_{i=1}^{n} log\ y_i - \frac{\tau^{(r)}}{2}\sum_{i=1}^{n}\left(\frac{1}{y_i}\right)^{\frac{1}{\sigma^{(r+1)}}} log\ y_i - \frac{(k^{(r)}-1)\tau^{(r)}}{2}\sum_{i=1}^{n} w_i^{(r)} \left(\frac{1}{y_i}\right)^{\frac{1}{\sigma^{(r+1)}}} log\ y_i \tag{44}$$

5 Data analysis

In this section we fit the $CGlogM(\tau, \sigma, w, k)$ to the Danish fire insurance data set. The Danish fire insurance data is available in the "SMPracticals" add-on packages of R. This data set consists of 2492 losses in millions of Danish Kroner during the period from 1980 to 1990. The Danish data set has been used as the main example data for many newly developed statistical distributions involving heavy-tailed assumptions.

In this regard, we consider two discrimination measures such as the Akaike information criterion (AIC), Bayesian information criterion (BIC). These measures are given by

$$AIC = -2logL + 2p$$
$$BIC = -2logL + plog(n)$$

Where L denotes the likelihood evaluated by the EM algorithm, p is the number of model parameters, and n is the sample size of the data set. For our model p is 4 and sample size of data set is 2492.

Table 1: Estimated parameter values GlogM model for the Danish fire insurance losses data set

parameter	Initial values	Estimated values
$\hat{\mu}$	1.312	0.013
$\hat{\sigma}$	0.321	0..005

Table 2: Estimated parameter values of the proposed model for the Danish fire insurance losses data set

parameter	Initial values	Estimated values
$\hat{\tau}$	2.33	2.057
$\hat{\sigma}$	0.321	0.318
$\hat{k}$	1.1	1.315
$\hat{w}$	0.5	0.508

Table 3: Analytical measures of the proposed and GlogM model for the Danish fire insurance losses data set

Model → measures ↓	$GlogM$	$CGlogM$
$Log - likelihood$	-3932.99	-3437.343
AIC	7869.99	6882.685
BIC	7872.78	6905.969

We compare the $CGlogM(\tau, \sigma, w, k)$ with $GlogM(\mu, \sigma)$. The parameter estimates, log-likelihood values, values of the Akaike information criterion(AIC) and the Bayesian information criterion(BIC) are computed and estimated parameter values are tabulated in the table 1 and 2. The log-likelihood values as well as the AIC and BIC values are given in Table 3.

Maximum value for log-likelihood and minimum values for AIC and BIC give evidence that the $CGlogM(\tau, \sigma, k, w)$ distribution gives a better fit as compared to the $GlogM(\mu, \sigma)$.

6 Conclusion:

In this report, a family for modelling insurance losses is proposed. Some of its properties are derived. Using EM algorithm the model parameters are obtained. A practical application to the insurance loss data set is analysed. The data applicability of this new family of distributions has been illustrated using Danish fire insurance data set and the model performs well as compared to $GlogM(\mu, \sigma)$ model.

References

[1] Asgharzadeh, A., Nadarajah, S., and Sharafi, F., (2017). Generalized inverse Lindley distribution with application to Danish fire insurance data. *Communications in Statistics-Theory and Methods*, 46 (10), 5000-5021.

[2] BAKAR, S.A., HAMZAH, N., MAGHSOUDI, M. and NADARAJAH, S. (2015) Modeling loss data using composite models. *Insurance: Mathematics and Economics*, 61, 146–154.

[3] BHATI, D. and RAVI, S. (2018), "On generalized log-moyal distribution: A new heavy tailed size distribution".*Insurance: Mathematics and Economics*, 79, 247–259.

[4] CALDERÍN-OJEDA, E., FERGUSSON, K. and WU, X. (2017) An EM algorithm for DoublePareto-Lognormal generalized linear model applied to heavy-tailed insurance claims. *Risks*, 5, 60.

[5] CHAN, J., CHOY, S., MAKOV, U. and LANDSMAN, Z. (2018) Modelling insurance losses using contaminated generalised Beta Type-II distribution. *ASTIN Bulletin: The Journal of the IAA*,48, 871–904.

[6] Coorey, K., and Ananda, M.M.,(2005). Modelling acturial data with a composite log-normal-Pareto model.*Scandinavian Actuarial Journal*, 5, 321-334.

[7] CUMMINS, J.D., DIONNE, G., MCDONALD, J.B. and PRITCHETT, B.M. (1990) Applications of the GB2 family of distributions in modeling insurance loss processes. *Insurance: Mathematics and Economics*, 9, 257–272.

[8] MILJKOVIC, T. and GRÜN, B. (2016) Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics*, 70, 387–396.

[9] Nadarajah, S. and Bakar, S. A. A., (2014). New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal*, 2, 180-187.

[10] Scollnik, D. P., (2007). On composite Log-normal-Pareto models. *Scandinavian Actuarial Journal*, 1, 20-33.