Mathematica Aeterna, Vol. 8, 2018, no. 4, 239 - 247

Modeling fire loss data using mixture of Burr and log-normal distributions

Xiaolin Song

School of Mathematics,
Liaoning Normal University, Dalian 116029, China
Email: 1263259717@qq.com

Abstract

Fire loss data usually has the characters of leptokurtic and fat tail, and it is fitted well by the mixture models. In his paper, we construct mixture of Burr and log-normal distributions. Some related statistical properties are investigated including the analytical expressions for the hazard function, each order moment and the mean deviation. The parameters are estimated via EM algorithm. We illustrate an application to the Chinese fire loss data. The parameter estimation and model checking of simulation data are given using the R language. Finally, we investigate risk measures VaR and $TVaR$, the theoretical values and empirical values are compared. The results show that the mixture of Burr and log-normal model gives better fit.

Mathematics Subject Classification: 62P05, 91B30

Keywords: Burr and log-normal mixture model; EM algorithm; Chinese fire loss data; Risk measures.

1 Introduction

Insurance payments data in actuarial industries are typically highly positively skewed and distributed with a larger upper tail. It is an interesting topic to model insurance loss data with a heavy tailed distribution. Several heavy tailed models have been discussed in the literature including the Pareto, log-normal, Weibull, gamma and Inverse Gaussian distribution. However, these classical simple models based on single parametric distributions do not provide a reasonable fit for many applications.

Recently, many researchers investigate the flexible finite mixture approach for modeling insurance losses using suitable parametric distributions. The concept of finite mixture distribution was pioneered by Newcomb [1] as a model for outliers. Lee and Lin [2] proposed modeling and evaluating insurance losses via mixtures of Erlang distributions using the EM algorithm for estimation. Verbelen et al. [3] generalized the mixture method in literature [2] and used the mixture distribution to simulate the censoring data and truncating data. Miljkovic and Grn [4] developed six models with components from parametric, non-Gaussian families of distributions previously used in actuarial modeling: Burr, Gamma, Inverse Burr, Inverse Gaussian, Log-normal, and Weibull. It is showed that the mixture models provide the best fitting for modeling Danish Fire insurance losses. Ghosh et al. [5] considered a finite mixture of two absolutely continuous distribution, a two parameter Weibull distribution and a three parameter Pareto (IV) distribution. Hyppolite [6] presented nine alternative mixture models that may be of interest for making inference from available economic panel data sets.

In this paper we consider a finite mixture of Burr and log-normal distributions and the relevant statistical properties of the model are investigated. The rest of this paper is structured as follows. In Section 2, we give the mixture model and study the distributional properties including the hazard function, moments and deviation. In Section 3, we estimate the model parameters and provide the application of the mixture of Burr and log-normal distribution to a real data sets in Section 4.

2 The mixture of Burr and log-normal distributions

Assume that $f_1(x)$ is the probability density function (pdf) of Burr distribution with

$$f_1(x) = \alpha\gamma\left(\frac{x}{\theta}\right)^{\gamma}\left\{x\left[1+\left(\frac{x}{\theta}\right)^{\gamma}\right]^{\alpha+1}\right\}^{-1}, x>0, \alpha>0, \gamma>0, \theta>0, \qquad (1)$$

where α and γ are the shape parameters and θ is the scale parameter. On the other hand, let $f_2(x)$ be the pdf of the log-normal distribution with

$$f_2(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right), x>0. \qquad (2)$$

Similar to [5], we consider the following model

$$f(x) = pf_1(x) + (1-p)f_2(x), \qquad (3)$$

where $f_1(x)$ and $f_2(x)$ are densities defined by (1) and (2), respectively. Clearly, p and $1-p$ are mixture weights.

The pdf in (3) is called the mixture of Burr and log-normal (hereafter MBLN in short) distribution. There are six parameters in the MBLN distribution. Figure (1) depicts the pdf of MBLN distribution varying with the parameter p, other parameter values in the figure are $\alpha = 2, \gamma = 2, \theta = 2, \sigma = 0.1, \mu = -0.1$, respectively. Figure (2) depicts the pdf of MBLN distribution about the parameter α, other parameter values in the figure are $\gamma = 2, \theta = 2, \sigma = 0.1, \mu = -0.1, p = 0.5$. Similarly, we can give the graphs for the pdf changing on other parameters.

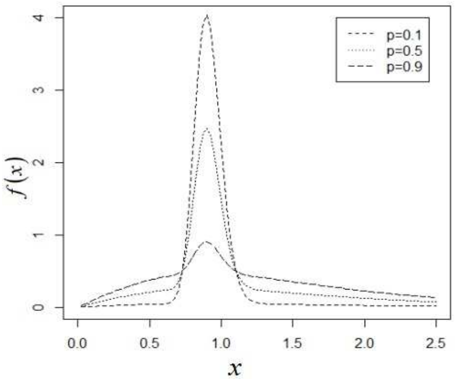

Figure 1: Density curves with different p.

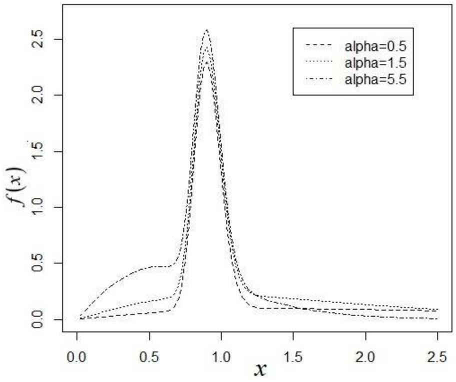

Figure 2: Density curves with different α.

In what follows, we give some structure properties for MBLN distribution. Let $F(x)$ be the corresponding distribution function of MBLN distribution. According to equations (1)-(3), the associated hazard function is

$$\begin{aligned} h_f(x) =& \frac{f(x)}{1-F(x)} \\ =& \left\{ p\alpha\gamma \left(\frac{x}{\theta}\right)^{\gamma} \left\{ x\left[1+\left(\frac{x}{\theta}\right)^{\gamma}\right]^{\alpha+1} \right\}^{-1} + (1-p)\frac{1}{\sqrt{2\pi}\sigma x} e^{\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right)} \right\} \\ & * \left\{ p\left[1+\left(\frac{x}{\theta}\right)^{\gamma}\right]^{-\alpha} + (1-p)\phi\left(-\frac{\log x-\mu}{\sigma}\right) \right\}^{-1}. \end{aligned}$$

On the other hand, suppose that a random variable X follows the mixture distribution defined in equation (3). For any $k \geq 1$, we have

$$\begin{aligned} E(X^k) =& p\int_0^{+\infty} x^k \alpha\gamma \left(\frac{x}{\theta}\right)^{\gamma} \left\{ x\left[1+\left(\frac{x}{\theta}\right)^{\gamma}\right]^{\alpha+1}\right\}^{-1} dx \\ &+ (1-p)\int_0^{+\infty} x^k \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right) dx \\ =& pJ_1 + (1-p)J_2, \end{aligned} \tag{4}$$

where

$$J_1 = \alpha\theta^k B\left(\frac{k}{\gamma}+1, \alpha-\frac{k}{\gamma}\right), J_2 = \exp\left(k\mu + \frac{k^2\sigma^2}{2}\right).$$

and the Beta function $B(a,b)$ is defined as

$$B(a,b) = \int_0^1 (1-t)^{a-1} t^{b-1} dt.$$

Specially, let $k=1$ in (4), we obtain the mean of X as

$$E(X) = p\alpha\theta B\left(\frac{1}{\gamma}+1, \alpha-\frac{1}{\gamma}\right) + (1-p)\exp\left(\mu+\frac{\sigma^2}{2}\right), \alpha > \frac{1}{\gamma}.$$

Similarly, we can get explicit expressions for other order moments. In terms of these expressions, the formulae for skewness and kurtosis can be obtained directly.

The deviation from the mean can be used to measure the dispersion for the random variable. After some algebras, it leads to

$$D(u) = E|X-EX| = u(K_1-K_2) - K_3 + K_4,$$

where

$$\begin{aligned} K_1 &= p - p\left[1+\left(\frac{u}{\theta}\right)^{\gamma}\right]^{-\alpha} + (1-p)\phi\left(\frac{\log u - \mu}{\sigma}\right), \\ K_2 &= p\left[1+\left(\frac{u}{\theta}\right)^{\gamma}\right]^{-\alpha} + (1-p)\phi\left(\frac{-\log u + \mu}{\sigma}\right), \\ K_3 &= p\alpha\theta B_{\left[\left(\frac{u}{\theta}\right)^{\gamma}+1\right]^{-1}}\left(\frac{1}{\gamma}+1, \alpha-\frac{1}{\gamma}\right) + (1-p)e^{\left(\mu+\frac{1}{2}\sigma^2\right)}\phi\left(\frac{\log u - \mu - \sigma^2}{\sigma}\right), \\ K_4 &= E(X) - K_3, B_x(a,b) = \int_x^1 (1-t)^{a-1} t^{b-1} dt. \end{aligned}$$

3 Parameter estimation

In this section, we discuss the parameter estimation for the MBLN distribution. Let $X_1, X_2, \cdots, X_n$ be a random sample of size n drawn from the density in (3). Let $\theta_1 = (\beta, \gamma)$ and $\theta_2 = (\delta, \alpha, \sigma)$ represent the parameters of $f_1(x)$ and $f_2(x)$ respectively. Given a list of observations $(x_1, x_2 \cdots x_n)$, let $\theta = (p, \theta_1, \theta_2)$ represent the complete parameter to be estimated. For clarity, we write $f_1(x)$ as $f_1(x|\theta_1)$, and $f_2(x)$ as $f_2(x|\theta_2)$.

We use EM algorithm to estimate parameters given in [5]. Considering $u_i(=0,1), i = 1, 2, \cdots, n$ as missing value with i^{th} observation x_i drawn from $f_1(x|\theta_1)$ if $u_i = 1$, and $f_2(x)$ as $f_2(x|\theta_2)$ if $u_i = 0$. Then the likelihood function is given by

$$L_c(\theta) = \prod_{i=1}^{n} (pf_1(x_i|\theta_1))^{u_i} [(1-p)f_2(x_i|\theta_2)]^{1-u_i}.$$

Denoting $\theta^{(k)}$ as the k^{th} iterative solution, the E-step is given by

$$\begin{aligned} E[\log L_c(\theta)|x, \theta^k] &= \sum_{i=1}^{n} E\left(u_i|x, \theta^{(k)}\right) \log f_1(x_i|\theta_1) + \sum_{i=1}^{n} \left[1 - E\left(u_i|x, \theta^{(k)}\right)\right] \\ * \log f_2(x_i|\theta_2) &+ \sum_{i=1}^{n} E\left(u_i|x, \theta^{(k)}\right) \log p + \sum_{i=1}^{n} \left[1 - E\left(u_i|x, \theta^{(k)}\right)\right] \log(1-p), \end{aligned}$$

where

$$E\left(u_i|x, \theta^{(k)}\right) = \frac{p^{(k)} f_1\left(x_i|\theta^{(k)}\right)}{p^{(k)} f_1\left(x_i|\theta^{(k)}\right) + \left(1 - p^{(k)}\right) f_2\left(x_i|\theta^{(k)}\right)} \equiv p_i^{(k)}.$$

Now following the M-step, take the derivatives of the solution $E[\log L_c(\theta)|x, \theta^k]$ with respect to the parameter $p, \beta, \gamma, \delta, \alpha, \sigma$ and set them equal to 0, then we get the parameter estimate of the $k+1^{th}$

$$p^{(k+1)} = \frac{1}{n} \sum_{i=1}^{n} p_i^{(k)}, \tag{5}$$

$$\alpha = \frac{\sum_{i=1}^{n} p_i^{(k)}}{\sum_{i=1}^{n} p_i^{(k)} \log\left(1 + \left(\frac{x_i}{\theta}\right)^{\gamma}\right)}, \tag{6}$$

$$\alpha = \frac{\sum_{i=1}^{n} p_i^{(k)} - \sum_{i=1}^{n} p_i^{(k)} \left[1 + \left(\frac{x_i}{\theta}\right)^{\gamma}\right]^{-1} \left(\frac{x_i}{\theta}\right)^{\gamma}}{\sum_{i=1}^{n} p_i^{(k)} \left[1 + \left(\frac{x_i}{\theta}\right)^{\gamma}\right]^{-1} \left(\frac{x_i}{\theta}\right)^{\gamma}}, \tag{7}$$

$$\gamma = \frac{\sum_{i=1}^{n} p_i^{(k)}}{\sum_{i=1}^{n} p_i^{(k)}(\alpha+1)\left[1+\left(\frac{x_i}{\theta}\right)^{\gamma}\right]^{-1}\left(\frac{x_i}{\theta}\right)^{\gamma}\log\left(\frac{x_i}{\theta}\right) - \sum_{i=1}^{n} p_i^{(k)}\log\left(\frac{x_i}{\theta}\right)}, \tag{8}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(\log x_i - \mu)^2}{\sum_{i=1}^{n}\left(1-p_i^{(k)}\right)}, \tag{9}$$

$$\mu = \frac{\sum_{i=1}^{n}\left(1-p_i^{(k)}\right)(\log x_i - 1)\log x_i}{\sum_{i=1}^{n}\left(1-p_i^{(k)}\right)(\log x_i - 1)}. \tag{10}$$

To solve (5)-(10), we consider the following steps:

Step 1: Generate a random sample of size n drawn from the MBLN distribution.

Step 2: Given the initial value $\theta^{(0)} = \left(p^{(0)}, \beta^{(0)}, \gamma^{(0)}, \delta^{(0)}, \alpha^{(0)}, \sigma^{(0)}\right)$.

Step 3: The initial values in the second step are respectively substituted into the iterative formula (5)-(10) to obtain $\theta^{(1)}$.

Step 4: Repeat the above steps until the iterative equation converges.

4 The empirical analysis

In this section, we use Chinese fire loss data for empirical analysis. The data recorded the fire insurance claims in China from 1999 to 2012 (in tens of millions CNY). The related summary statistics of the loss data are shown in Table (1).

Table 1: Descriptive statistic for Chinese fire data.

Statistic	Statistic value
minimum	0.2377
maximum	32.3631
mean	5.0829
standard deviation	4.1711
skewness	2.0311
kurtosis	6.8343
1/4 quantile	2.3648
3/4 quantile	6.7646

The maximum likelihood estimates, the loglikelihood value, the Akaike information criterion (AIC) for the selected distributions are reported in Tables 2. The selected distributions are Exponentiated Exponential distribution, Weibull distribution, log-normal distribution, Pareto (IV) distribution, MEEP(IV) distribution, MWP (IV) distribution and MBLN distribution. The fitting effect for MBLN distribution is depicted in figure (3).

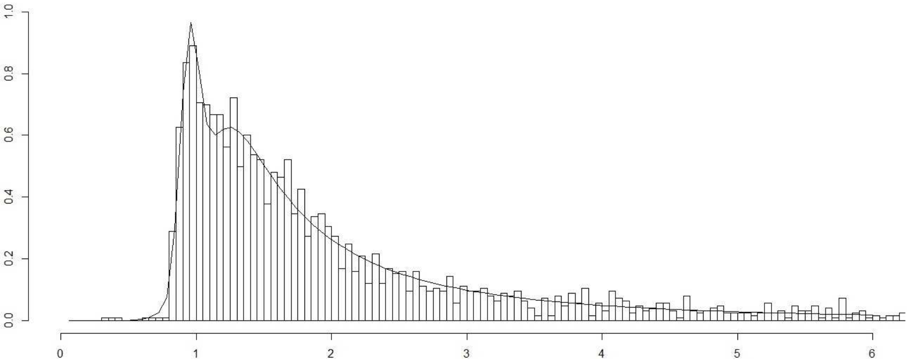

Figure 3: Comparison of mixture model and histogram for China fire losses.

Based on the NLL and AIC values, we see that the mixture models give the better fit than single models. It is clearly evident that the MBLN distribution as a more desirable distribution for the given data.

In recent years, VaR (value-at-risk) and $TVaR$ (tail-value-at-risk) have been investigated deeply in financial and insurance fields. According to Klugman et al. [7], if π_p represents the $100p$ quantile of random variable X, then $VaR_p(X)$ is equivalent to π_p and satisfies

$$P(X > \pi_p) = 1 - p.$$

In the case of mixture models, $VaR_p(X)$ does not have a closed form solution and requires a numerical solution of the following equation

$$F_X(\pi_p) = p,$$

which can be done in R using the function uniroot() from the base package stats.

The risk measure $TVaR_p(X)$ can be solved by the following equation

$$TVaR_p(X) = E(X|X > \pi_p) = \frac{\int_{\pi_p}^{+\infty} xf(x)dx}{1 - F_X(\pi_p)} = \frac{\int_{\pi_p}^{+\infty} xf(x)dx}{1 - p}.$$

Since the conditional expectation is linear, the $TVaR_p(X)$ of the mixture distribution can be obtained by calculating the corresponding value of each component.

Table (3) presents the theoretical and empirical estimations for three mixture models at 99% significance level. It can be seen that for MBLN mixture distribution, the relative difference between the empirical value of $VaR_p(X)$ and $TVaR_p(X)$ is the smallest, which indicates that the risk of using MBLN mixture model to simulate the given fire loss data is the lowest.

Table 2: Parameter estimates for different models.

Distribution	Parameter estimates	NLL	AIC	BIC
Exponentiated Exponential	$\hat{\beta} = 1.64760$ $\hat{\gamma} = 0.26563$	1115.194	2234.389	2242.534
Weibull	$\hat{\beta} = 1.29056$ $\hat{\gamma} = 5.51113$	1118.138	2240.276	2248.422
log-norma	$\hat{\mu} = 1.27773$ $\hat{\sigma} = 0.91883$	1133.615	2271.23	2279.376
Pareto(IV)	$\hat{\alpha} = 4.95550$ $\hat{\delta} = 0.67937$ $\hat{\sigma} = 14.69751$	1114.509	2235.018	2241.164
MEEP(IV)	$\hat{p} = 0.9508$ $\hat{\alpha} = 2267$ $\hat{\delta} = 0.1845$ $\hat{\sigma} = 2.148$ $\hat{\beta} = 2.039$ $\hat{\gamma} = 0.2852$	1106.356	2224.711	2224.858
MWP(IV)	$\hat{p} = 0.91462$ $\hat{\alpha} = 1.96885$ $\hat{\delta} = 0.49763$ $\hat{\sigma} = 6.94338$ $\hat{\beta} = 3.16543$ $\hat{\gamma} = 0.66199$	1104.08	2220.161	2220.306
MBLN	$\hat{p} = 0.15974$ $\hat{\alpha} = 0.31125$ $\hat{\gamma} = 4.84594$ $\hat{\theta} = 0.45712$ $\hat{\sigma} = 0.63771$ $\hat{\mu} = 1.55965$	1102.338	2216.675	2216.822

Table 3: Chinese Fire losses: Summary of risk measures.

Empirical estimates	$VaR(0.99)$	$TVaR(0.99)$
	20.36	24.92
mixture models	$VaR(0.99)$	$TVaR(0.99)$
MWP(IV)	36.59	190.95
MEEP(IV)	33.62	195.04
MBLN	25.83	101.14

ACKNOWLEDGEMENTS. This research was supported by the Ministry of Education of Humanities and Social Science Project (15YJC910001); Science Research Foundation of Educational Department of Liaoning Province of China (NO. W201783664).

References

[1] NEWCOMB S. A Generalized Theory of the Combination of Observations so as to Obtain the Best Result, American Journal of Mathematics, 1886,8(4):343-366.

[2] LEE S, LIN X. Modeling and Evaluating Insurance Losses Via Mixtures of Erlang Distributions, North American Actuarial Journal, 2010, 14(1):107-130.

[3] VERBELEN R, GONG L, ANTONIO K, BADESCU A, LIN S. Fitting mixtures of erlangs to censored and truncated data using the em algorithm, Astin Bulletin, 2015, 45(3):729-758.

[4] MILJKOVIC T, GRUN B. Modeling loss data using mixtures of distributions, Insurance Mathematics & Economics, 2016, 70:387-396.

[5] GHOSH D, HAMEDANI G, BANSAL N, MADOLIAT M. On the Mixtures of Weibull and Pareto(IV) Distribution: An Alternative to Pareto Distribution, Working paper, 2016.

[6] HYPPOLITE J, Alternative approaches for econometric modeling of panel data using mixture distributions, Journal of Statistical Distributions and Applications, 2017, 4(1).

[7] KLUGMAN S, PANJER H, WILLMOTt G. Loss Models: From Data to Decisions, fourth ed, John Wiley & Sons, Hobuken, NJ, 2012.

Received: July 02, 2018