Journal of Mathematics and Statistics 8 (1): 49-56, 2012
ISSN 1549-3644

Fitting of Finite Mixture Distributions to Motor Insurance Claims

Sattayatham, P. and T. Talangtam
Institute of Science, School of Mathematics,
Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

Abstract: Problem statement: The modeling of claims is an important task of actuaries. Our problem is in modelling the actual motor insurance claim data set. In this study, we show that the actual motor insurance claim can be fitted by a finite mixture model. **Approach:** Firstly, we analyse the actual data set and then we choose the finite mixture Lognormal distributions as our model. The estimated parameters of the model are obtained from the EM algorithm. Then, we use the K-S and A-D test for showing how well the finite mixture Lognormal distributions fit the actual data set. We also mention the bootstrap technique in estimating the parameters. **Results:** From the tests, we found that the finite mixture lognormal distributions fit the actual data set with significant level 0.10. **Conclusion:** The finite mixture Lognormal distributions can be fitted to motor insurance claims and this fitting is better when the number of components (k) are increase.

Key words: Bootstrap, claim size distribution, EM algorithm, finite mixture models, lognormal distribution, loss distribution

INTRODUCTION

Introduction and motivation: Many problems in actuarial science involve the building of a mathematical model that can be used to forecast or predict insurance costs. So modeling is an important procedure for actuaries so that they can estimate the degree of uncertainty as to when a claim will be made and how much will be paid. In particular, the modeling of claims and outstanding claims lead to the pricing of insurance premiums and an estimation of claim reserving, respectively. The most useful approach to uncertainty representation is through probability, so we will concentrate on probability models.

Losses depend on two random variables, i.e., the number of losses and the amount of loss which will occur in a specified period. The number of losses (claim number) is referred to as the frequency of loss (claim frequency) and the probability distribution is called the frequency distribution. The amount of loss (claim size) is referred to as the severity of loss (claim severity) and its probability distribution is called the severity distribution. Loss distribution and its modeling are described in detail in the book of Klugman *et al.* (2008) and paper of Janczuraa and Weron (2010). The severity distribution is solely considered for this study.

The mixture of distributions is sometime called compounding, which is extremely important as it can provide a superior fit. A successful use of this technique is illustrated in Hewitt and Lefkowitz (1979). In the 1960s and 1970s, finite mixture models appeared in the statistical literature and they proved to be useful for modeling discrete unobserved heterogeneity in the population. Since there are many different modes for claim possibilities, a finite mixture model should work well.

The Expectations-Maximization (EM) algorithm is provided to fit the model that introduces unobserved indicators with the goal of maximizing the complete likelihood function. The EM algorithm is also applicable for parameter estimation of mixture models. For more detail, McLachlan and Peel (2000); Aitkin and Rubin (1985); Hogg *et al.* (2004) and Hogg and Klugman (1984).

The bootstrap process is a tool for fitting and it is not complicated to implement. Usually, the bootstrap process involves resampling with replacements from the residual more than the data themselves. We apply the bootstrap technique to recalculate the estimated parameters for model fitting. For more detail, Efron and Tibshirani (1993).

The purpose of this study is find a statistical model for the claim severity. Many authors investigate some special distributions of the severity claims and apply them to calculate the insurance premium. Recently, Mohamed *et al.* (2010) investigated a model of severity claims which has Pareto distribution and they used it to calculate insurance premiums under the retention limit. Moreover, Brazauskas *et al.* (2009) suggest the Method of Trimmed Moments (MTM) in the case of loss

Corresponding Author: Sattayatham, P., Institute of Science, School of Mathematics, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand

distribution of Lognormal and Paerto and analyze a real data set concerning hurricane damage in the United States. But in our work, we work in the opposite direction, i.e., we shall find a model that is fitted to the empirical data.

We considers the data from a set of motor insurance claims from the top three non-life insurance public companies in Thailand. A mixture model is fitted to the data and the estimated parameters for the model are calculated by the EM algorithm. We also use the bootstrap technique to fit the data and show that the bootstrap sample for observation can be applicable to the estimated parameters.

MATERIALS AND METHODS

We present the statistical modeling for a finite mixture of Lognormal distributions, the EM algorithm is explained and the bootstrap technique is demonstrated.

Statistical modeling: The skewed right distribution such as Gamma, Lognormal, Weibull and Pareto distribution have often been used by actuaries to fit claim sizes; see Klugman *et al*. (2008). In insurance companies, there are 2 types of claim data recording, i.e., individual and group data. We model the individual claim data. Some assumptions and symbols are specified as below.

Assumption 1 (Policy independence): Consider n different policies (contracts). Let X_i denote the response for policy i. Then $X_1,\dots, X_n$ are independent.

Assumption 2 (Time independence): Consider n disjointed time intervals. Let X_i denote the response in time interval i. Then $X_1,\dots, X_n$ are independent.

Assumption 3 (Homogeneity): Consider any two policies in the same tariff cell, having the same exposure. Let X_i denote the response for policy i. Then X_1 and X_2 have the same probability distribution.

Assumption 4: Severity losses are non-catastrophe losses.

Assumption 5: A recorded claim is equal to an actual claim (observation).

Single parametric distribution: On the basis of the analyst's knowledge, experience and statistical test, the Lognormal distribution is our selection for modeling and fitting to the data set. The Maximum Likelihood Estimation (MLE) is provided for estimation of the parameters.

The model: Assume that X ~ Lognormal (μ,σ), abbreviates $X \sim LN(\mu,\sigma)$, with density:

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

$$\mu \in R,\ \ \sigma > 0,\ \ x > 0$$

Estimation for the model: Let $x = (x_1,\dots, x_n)$ be an independent observation. Consider the amount x_i paid for the i^{th} contract. We fit the Lognormal distribution in Eq. 1 to the data set by MLE.

The likelihood function $L = \prod_{i=1}^{n} f_X(x_i)$ then:

$$\begin{aligned}
\ln L &= \ln\prod_{i=1}^{n} f_X(x_i)\\
&= \sum_{i=1}^{n} \ln f_X(x_i)\\
&= \sum_{i=1}^{n} \ln\left[\frac{1}{x_i\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln x_i-\mu)^{\,2}}{2\sigma^2}\right)\right]\\
&= \sum_{i=1}^{n}\left[-(\ln\sigma+\ln x_i)-\frac{1}{2}\ln 2\pi-\frac{1}{2\sigma^2}(\ln x_i-\mu)^{\,2}\right]
\end{aligned}$$

We estimate $\hat{\mu}$ and $\hat{\sigma}$ for μ and σ respectively by $\frac{\partial \ln L}{\partial\mu} = 0$ and $\frac{\partial \ln L}{\partial\sigma} = 0$.

We obtain maximum likelihood estimates for the parameter μ and the parameter σ as follows:

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\ln x_i}{n}$$

and:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(\ln x_i-\hat{\mu})^{\,2}}{n}}\text{, respectively}$$

Finite mixture models: We consider the second-order and more than second-order finite mixture model. We aim to find the mixing weights according to the number of Lognormal distributions and estimated parameters by the MLE via EM algorithm.

The model:

$$\text{Let } X \sim \tau_1 LN(X|\ \mu_1,\sigma_1)+\dots+\tau_k LN(X|\ \mu_k,\sigma_k) \qquad (2)$$

Then its probability density function is:

$$f_X(x) = \tau_1 f_1(x)+\dots+\tau_k f_k(x)$$

$$= \frac{1}{x\sqrt{2\pi}} \begin{pmatrix} \tau_1 \frac{1}{\sigma_1} \exp\left(-\frac{(\ln x - \mu_1)^2}{2\sigma_1^2} \right) + \cdots \\ + \tau_k \frac{1}{\sigma_k} \exp\left(-\frac{(\ln x - \mu_k)^2}{2\sigma_k^2} \right) \end{pmatrix}$$

where, $0<\tau_j <1$ for $j=1, ..., k$ and $\tau_1+\ldots\tau_k = 1$. The likelihood function can be written as follows:

$$L = \prod_{i=1}^{n} \frac{1}{x_i\sqrt{2\pi}} \begin{pmatrix} \tau_1 \frac{1}{\sigma_1} \exp\left(-\frac{(\ln x - \mu_1)^2}{2\sigma_1^2} \right) + \cdots \\ + \tau_k \frac{1}{\sigma_k} \exp\left(-\frac{(\ln x - \mu_k)^2}{2\sigma_k^2} \right) \end{pmatrix}$$

and the log-likelihood function is in form:

$$\ln L = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{k} \tau_j \frac{1}{x_i\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(\ln x_i - \mu_j)^2}{2\sigma_j^2} \right) \right]$$

Estimation for the model: EM algorithm is a powerful algorithm for data arising from mixtures. Assume that the data set of motor insurance claim is produced according to model Eq. 2.

Let $z = (z_{ij},\ i=1,...,n;\ j=1,...,k)$ be the latent (unobservable) variables that determine the components from which the observation originates. The values z_{ij} are indicators defined as:

$$z_{ij} = \begin{cases} 1 & \text{, observation } x_i \text{ comes from the distribution } f_j \\ 0 & \text{, elsewhere} \end{cases}$$

The complete likelihood takes quite a simple form:

$$L_c(\psi \mid x) = \prod_{i=1}^{n} \prod_{j=1}^{k} \begin{pmatrix} \tau_j \frac{1}{x_i\sqrt{2\pi}\sigma_j} \\ \exp\left(-\frac{(\ln x_i - \mu_j)^2}{2\sigma_j^2} \right) \end{pmatrix}^{z_{ij}}$$

The complete log-likelihood function is Eq. 3:

$$\ln L_c(\psi \mid x) = \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij} \begin{bmatrix} \ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln(2\pi) \\ -\frac{1}{2\sigma_j^2}(\ln x_i - \mu_j)^2 \end{bmatrix} \qquad (3)$$

Set $\psi = (\theta, \tau)$, $\tau = (\tau_1, ..., \tau_{k-1})$ and $\theta = (\mu_1, ..., \mu_k, \sigma_1, ..., \sigma_k)$.

For each k components, there are 3k-1 unknown parameters that will be estimated by EM algorithm. We use a computer for the calculation of the parameters and visualization as a way to see its modeling. The proper number of components to be included in the mixture model will be considered.

E-step: replacing z_{ij} in Eq. 3 by its expected value $E[z_{ij}] := T_{ij}$, yields the expected complete log-likelihood:

$$E[\ln L_c(\theta \mid x, \tau)] = \sum_{i=1}^{n} \sum_{j=1}^{k} T_{ij} \begin{bmatrix} \ln\tau_j - \ln x_i - \ln\sigma_j \\ -\frac{1}{2}\ln(2\pi) - \frac{1}{2\sigma_j^2}(\ln x_i - \mu_j)^2 \end{bmatrix} \qquad (4)$$

Note that T_{ij} is the marginal probability that an observation x_i comes from the j^{th} component. By Baye's theorem, the marginal probability T_{ij} is given by:

$$\begin{aligned} T_{ij} &= P(z_{ij} = 1 \mid X_i = x_i; \psi) \\ &= \frac{\tau_j f_j(x_i;\ \theta_j)}{\sum_{j=1}^{k} \tau_j\ f_j(x_i;\ \theta_j)} = \frac{\tau_j f_j(x_i;\ \theta_j)}{f_X(x_i)} \end{aligned}$$

M-step: we maximize Eq. 4 to estimate ψ. Firstly, we solve the first order conditions:

$$\frac{\partial\ E[\ln L_c(\theta \mid x, \tau)]}{\partial\tau_j} = 0$$

$$\frac{\partial}{\partial\tau_j} \sum_{i=1}^{n} \sum_{j=1}^{k} T_{ij} \begin{bmatrix} \ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln(2\pi) \\ -\frac{1}{2\sigma_j^2}(\ln x_i - \mu_j)^2 \end{bmatrix} = 0$$

$$\frac{\partial}{\partial\tau_j} \sum_{i=1}^{n} [\ T_{i1}\ (\ln\tau_1) + \cdots + T_{ik}\ (\ln\tau_k)\ ] = 0$$

$$\frac{\partial}{\partial\tau_j} \left\{ \left[\sum_{j=1}^{k} \left(\sum_{i=1}^{n} T_{ij} \right) \ln\tau_j \right] \right\} = 0$$

This has the same form as the MLE for the multinomial distribution, so:

$$\hat{\tau}_j = \frac{\sum_{i=1}^{n} T_{ij}}{\sum_{j=1}^{k} \left(\sum_{i=1}^{n} T_{ij} \right)} = \frac{\sum_{i=1}^{n} T_{ij}}{\sum_{i=1}^{n} \left(\sum_{j=1}^{k} T_{ij} \right)} = \frac{1}{n} \sum_{i=1}^{n} T_{ij}$$

Secondly, We solve the equation $\frac{\partial E\left[\ln L_c\left(\theta \mid x, \tau\right)\right]}{\partial \theta_j} = 0$ for estimated parameters of $\theta_j = \left(\mu_j, \sigma_j\right)$, j = 1,2,…,k. Consider $\theta_1 = \left(\mu_1, \sigma_1\right)$.

We will estimate θ_1 by solving:

$$\frac{\partial E\left[\ln L_c\left(\theta \mid x, \tau\right)\right]}{\partial \mu_1} = 0$$

and:

$$\frac{\partial E\left[\ln L_c\left(\theta \mid x, \tau\right)\right]}{\partial \sigma_1} = 0$$

Note that the relation $\frac{\partial E\left[\ln L_c\left(\theta \mid x, \tau\right)\right]}{\partial \mu_1} = 0$ and Eq. 4 imply:

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}\frac{\partial}{\partial \mu_1}\begin{bmatrix}\ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln\left(2\pi\right)\\ -\frac{1}{2\sigma_j^2}\left(\ln x_i - \mu_j\right)^2\end{bmatrix} = 0$$

$$\sum_{i=1}^{n} T_{i1}\frac{\partial}{\partial \mu_1}\begin{bmatrix}\ln\tau_1 - \ln x_i - \ln\sigma_1 - \frac{1}{2}\ln\left(2\pi\right)\\ -\frac{1}{2\sigma_1^2}\left(\ln x_i - \mu_1\right)^2\end{bmatrix} = 0$$

$$\sum_{i=1}^{n} T_{i1}\left(\ln x_i - \mu_1\right) = 0$$

$$\sum_{i=1}^{n} T_{i1}\ln x_i - \sum_{i=1}^{n} T_{i1}\mu_1 = 0$$

$$\mu_1 = \frac{\sum_{i=1}^{n} T_{i1}\ln x_i}{\sum_{i=1}^{n} T_{i1}}$$

$$\frac{\partial E\left[\ln L_c\left(\theta \mid x, \tau\right)\right]}{\partial \sigma_1} = 0$$

$$\sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}\frac{\partial}{\partial \sigma_1}\begin{bmatrix}\ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln\left(2\pi\right)\\ -\frac{1}{2\sigma_j^2}\left(\ln x_i - \mu_j\right)^2\end{bmatrix} = 0$$

$$\sum_{i=1}^{n} T_{i1}\frac{\partial}{\partial \sigma_1}\begin{bmatrix}\ln\tau_1 - \ln x_i - \ln\sigma_1 - \frac{1}{2}\ln\left(2\pi\right)\\ -\frac{1}{2\sigma_1^2}\left(\ln x_i - \mu_1\right)^2\end{bmatrix} = 0$$

$$\sum_{i=1}^{n} T_{i1}\left[-\frac{1}{\sigma_1} + \frac{1}{\sigma_1^3}\left(\ln x_i - \mu_1\right)^2\right] = 0$$

$$\sum_{i=1}^{n} T_{i1}\left[-1 + \frac{1}{\sigma_1^2}\left(\ln x_i - \mu_1\right)^2\right] = 0$$

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{n} T_{i1}\left(\ln x_i - \mu_1\right)^2 = \sum_{i=1}^{n} T_{i1}$$

$$\sigma_1 = \sqrt{\frac{\sum_{i=1}^{n} T_{i1}\left(\ln x_i - \mu_1\right)^2}{\sum_{i=1}^{n} T_{i1}}}$$

Similarly, one can show that:

$$\mu_j = \frac{\sum_{i=1}^{n} T_{ij}\ln x_i}{\sum_{i=1}^{n} T_{ij}}$$

and:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} T_{ij}\left(\ln x_i - \mu_j\right)^2}{\sum_{i=1}^{n} T_{ij}}},$$
$$j = 1,\ 2, ...,\ k.$$

In summary, we obtain that:

$$\hat{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} T_{ij}\ , \quad \mu_j = \frac{\sum_{i=1}^{n} T_{ij}\ln x_i}{\sum_{i=1}^{n} T_{ij}}$$

and:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} T_{ij}\left(\ln x_i - \mu_j\right)^2}{\sum_{i=1}^{n} T_{ij}}},\ j = 1,\ 2, ...,\ k$$

Note that the expected complete log-likelihood function is given by:

$$E\left[\ln L_c\left(\theta \mid x, \tau\right)\right] = \sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}\begin{bmatrix}\ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln\left(2\pi\right)\\ -\frac{1}{2\sigma_j^2}\left(\ln x_i - \mu_j\right)^2\end{bmatrix}$$

For a given set of parameters ψ, i.e., $\theta_j = (\mu_j, \sigma_j)$, j = 1,2,…, k and $\tau = (\tau_1,\ldots, \tau_{k-1})$, the E-step consists of calculating T_{ij} and τ_j for M-step. Given τ_j, the M-step consists of maximizing the expected complete log-likelihood function. The E-step and M-step are repeated in an alternating fashion until the expected complete log-likelihood fails to increase. At this point, we conduct a final M-step in which the set of parameters ψ is estimated. Otherwise, we return to the E-step for the next iteration. In the final step after the m^{th} iteration, the EM algorithm is produced as below:

E-step: Given our current estimation of the parameters $\psi^{(m)}$ after the m^{th} iteration. Thus the E-step results in the function:

$$Q\left(\psi \mid \psi^{(m)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{k} T_{ij}^{(m)} \begin{bmatrix} \ln\tau_j - \ln x_i - \ln\sigma_j - \frac{1}{2}\ln(2\pi) \\ -\frac{1}{2\sigma_j^2}\left(\ln x_i - \mu_j\right)^2 \end{bmatrix} \quad (5)$$

M-step: Maximizing ψ. That is:

$$\tau^{(m+1)} = \underset{\tau}{\arg\max}\ Q\left(\psi \mid \psi^{(m)}\right)$$

And

$$\left(\mu_j^{(m+1)}, \sigma_j^{(m+1)}\right) = \underset{\mu_j,\sigma_j}{\arg\max}\ Q\left(\psi \mid \psi^{(m)}\right).$$

By taking partial derivative Eq. 5 with respect to ψ and by equating to zero, one gets:

$$\tau_j^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} T_{ij}^{(m)} \quad \mu_j^{(m+1)} = \frac{\sum_{i=1}^{n} T_{ij}^{(m)} \ln x_i}{\sum_{i=1}^{n} T_{ij}^{(m)}}$$

and:

$$\sigma_j^{(m+1)} = \sqrt{\frac{\sum_{i=1}^{n} T_{ij}^{(m)}\left(\ln x_i - \mu_j\right)^2}{\sum_{i=1}^{n} T_{ij}^{(m)}}}$$

Bootstrap technique: We are interested in the bootstrap sample for observation and residual. We shall recalculate the estimated parameters of the Lognormal distribution by using the bootstrap technique and MLE. One of advantage of the bootstrap technique is that we can calculate as many replications of the sample as we want.

Observation bootstrap:

Define $\quad x^* = \left(x_1^*, x_2^*, \ldots, x_n^*\right) \quad (6)$

The bootstrap data points $x_1^*, x_2^*, \ldots, x_n^*$ are a random sample of size n with replacement from the observation of n objects $\left(x_1, x_2, \ldots, x_n\right)$. Then we recalculate the estimated parameters, $\hat{\mu}^*$ and $\hat{\sigma}^*$, by MLE based on x^*.

Residual bootstrap: There are many forms of the residual definition and it is important to use an appropriate residual definition for the determination of each problem. We have already considered some forms of residual definitions, such as the unscaled Pearson residual and the unscaled Anscombe residual. But these forms of residual are not suitable for our data. Hence, we consider the residual form $\hat{\mu}$, that is, we define the form of the residual as follows:

$$\varepsilon_i = \ln x_i - \hat{\mu}$$

where, ε_i is the residual (i = 1,2,…, n) and $\hat{\mu}$ comes from Eq. 6.

Let $\varepsilon = \left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)$ and let $\varepsilon^* = \left(\varepsilon_1^*, \varepsilon_2^*, \ldots, \varepsilon_n^*\right)$ be the resample residual.

By using the bootstrap technique, we obtain a resample ε^* and the bootstrap data samples Eq. 7:

$$\ln x_i^* = \varepsilon_i^* + \hat{\mu}, i = 1, 2, \ldots, n \quad (7)$$

We recalculate the estimated parameters, $\hat{\mu}^*$ and $\hat{\sigma}^*$ by MLE based on $\ln x_i^*$, i = 1, 2,…, n.

Goodness of fit test: The Goodness of Fit (GOF) test measures the compatibility of a random sample with a theoretical probability distribution function. We use the Kolmogorov-Smirnov test (K-S test) and the Anderson-Darling test (A-D test) for showing how well the distribution fits our data set.

The K-S test is used to decide if a sample comes from a hypothesized continuous distribution. It is based on the Empirical Cumulative Distribution Function (ECDF) and denoted by:

$$F_X^n(x) = \frac{1}{n}\left[\text{Number of observations } \le x\right]$$

The K-S test statistic is defined by:

$$D = \sup_x \left| F_X^n(x) - F_X^*(x) \right|$$

The A-D test is a general test to compare the fit of an observed cumulative distribution function to an expected cumulative distribution function. This test gives more weight to the tails than the K-S test.

The A-D test statistic is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F_X^*(x_i) - \ln\left\{1 - F_X^*\left(x_{n-i+1}\right)\right\}\right]$$

where, F_X^* is the theoretical cumulative distribution of the distribution being tested.

RESULTS

The finite mixture of Lognormal distributions is applied to the actual set of claims data and the bootstrap procedure is analyzed. An analysis and some comparisons are shown with respective to statistical tests**.**

An application: We fitted the finite mixture of Lognormal distributions to the data set which was provided by a non-life insurance company in Thailand. We considered it for both a whole portfolio and various types of product coverages. The Kolmogorov-Siminov test and the Anderson-Darling test are statistical tests for model fitting.

Motor insurance data set: We consider the data set of motor insurance claims for the year 2009; all types of vehicle, i.e. automobiles, lorries and motorcycles are included. The total of each claim amount is paid by the insurer. The data set is classified by product coverage type-i for i = 0, 1,…, 5. There are 1,296 observations of type-5 that meet the mixture Lognormal distributions. The historical data of sevirity claim and histogram of severity claim (log scale) are illustrated in Fig. 1 and 2, repectively.

Table 1-2 show the statistical test value for fitting the finite mixture Lognormal distributions to the data set. For both the K-S and A-D test consideration, the summaries are as the following cases.

Case 1: at a significant level of $\alpha = 0.05$. We obtain the estimated parameters, $\hat{\mu} = 8.9672$ and $\hat{\sigma} = 1.1804$, that the Lognormal distribution does not fit to type-5. While the mixture Lognormal distributions are fitted to type-5 as k components greater than or equal to 20.

Case 2: at a significant level of $\alpha = 0.10$. The mixture Lognormal distributions are fitted to type-5 as k components equal to 25 and over. Mostly, k components of the mixture Lognormal distributions are better fit to the type-5 while k are increased. The maximum numbers of components is 130, since over 130 components are not applicable to k mean clustering.

From Table 2, by the A-D test. We can see that A^2 value are reduced when k are increased as the D value is not.

Figure 3-4 show probability density function (p.d.f.) of Lognormal distribution (k=1, with $\hat{\mu} = 8.9672$ and $\hat{\sigma} = 1.1804$) and mixture Lognormal distributions when k=100, respectively.

Figure 5-6, solid line, show the distribution functions (d.f.) of finite mixture Lognormal when k=1 and k=100, respectively. The dashed line is ECDF.

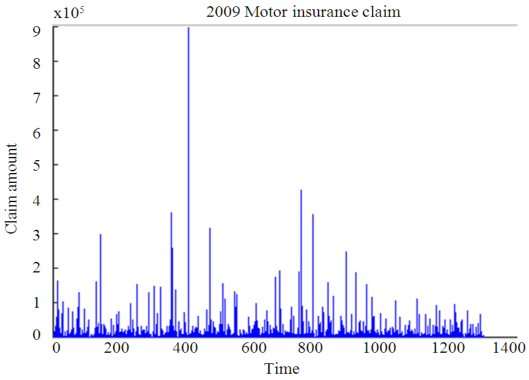

Fig. 1: Historical data 1,296 observations

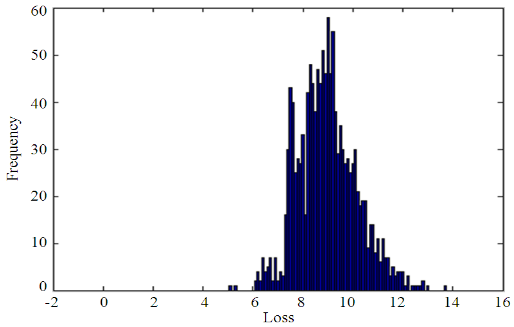

Fig. 2: Histogram (log scale)

Table 1: The Lognormal distribution

Single parametric distribution	K-S test		A-D test	
	D Value	P Value	A^2 Value	P Value
Lognormal	0.0466	p<0.01	3.3770	0.0241

Table 2: The finite mixture Lognormal distributions

	K-S test		A-D test	
k-components	D value	P value	A^2 value	P value
15	0.0430	0.0215	3.1900	0.0296
20	0.0355	0.0793	2.0373	0.0907
25	0.0330	p>0.1	1.6118	p>0.1
30	0.0261	p>0.1	1.1829	p>0.1
35	0.0264	p>0.1	1.0348	p>0.1
40	0.0217	p>0.1	0.7989	p>0.1
50	0.0247	p>0.1	0.6193	p>0.1
62	0.0234	p>0.1	0.5447	P>0.1
65	0.0247	p>0.1	0.4594	P>0.1
76	0.0239	p>0.1	0.4094	p>0.1
78	0.0224	p>0.1	0.3454	p>0.1
88	0.0216	p>0.1	0.3401	p>0.1
100	0.0216	p>0.1	0.3029	p>0.1

Figure 7-8 show the P-P plot of finite mixture Lognormal distributions when k=1 and k=100, respectively.

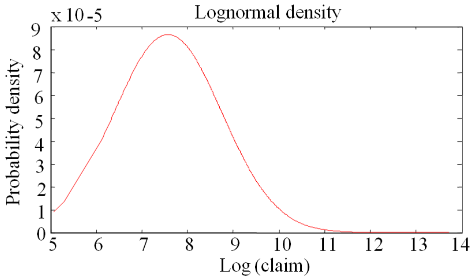

Fig. 3: P.d.f. of Lognormal distribution

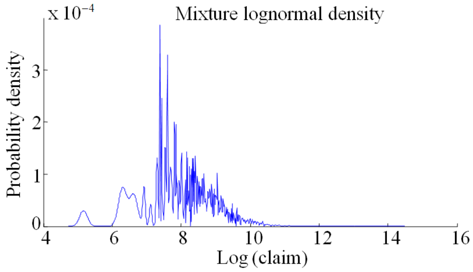

Fig. 4: k = 100

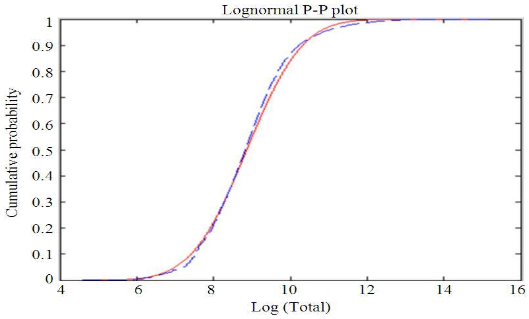

Fig. 5: k = 1

A bootstrap data sample can be calculated by using Eq. 6 and 7 for observation and residual respectively. The Lognormal distribution was fitted to the data set, when we recalculated the new estimated parameters respective to the bootstrap process. We have found that the Lognormal distribution is fitted to type-5 at a significant level of α=0.01 for both the K-S and A-D test. By K-S test, the Lognormal distribution can be fitted to type-5 at a significant level of α=0.10. We can see some examples of this from Table 3.

From Table 3, we can see that the bootstrap technique can be applicable to refitting the model of the data set. Note that the residual bootstrap provides better A^2 values in a shorter time of a computer run than the observation bootstrap.

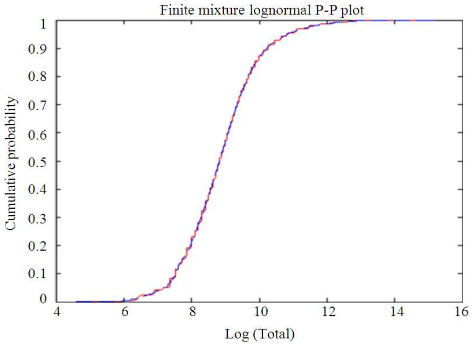

Fig. 6: k = 100

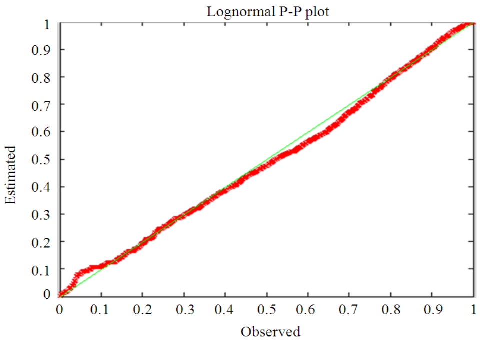

Fig.7: k = 1

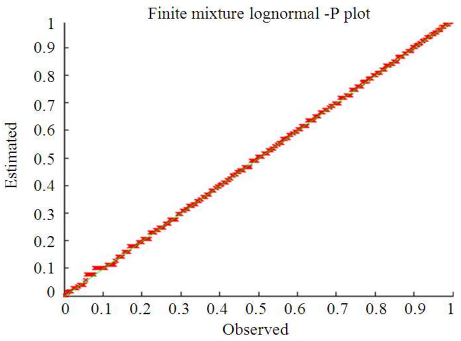

Fig. 8: k = 100

Table 3: Recalculation of the estimated parameters based on data and residual bootstrap

Bootstrap and MLE	$\hat{\mu}^*$	$\hat{\sigma}*$	K-S test D value	 P value	A-D test A^2 value	 P value
Data	8.9024	1.1654	0.0427	0.0238	3.2188	0.0287
	8.9339	1.1607	0.0377	0.0510	2.7781	0.0416
	8.9433	1.1185	0.0331	p>0.1	3.3329	0.0255
	8.9154	1.1102	0.0309	p>0.1	3.6200	0.0170
	8.9336	1.1094	0.0289	p>0.1	3.5141	0.0201
Residual	8.9182	1.1656	0.0406	0.0350	2.8866	0.0384
	8.9384	1.1541	0.0359	0.0714	2.8051	0.0408
	8.9334	1.1313	0.0324	p>0.1	3.0150	0.0347
	8.9355	1.1215	0.0307	p>0.1	3.2072	0.0290
	8.9249	1.1095	0.0295	p>0.1	3.5309	0.0196

DISCUSSION

We should consider the infinite mixture Lognormal distributions (uncountable family) for reducing the problem of the number of components and it should be considered for the fitting of truncated and/or censored data sets in further research.

The model presented fitted the claim amount. It can be used for actuaries to determine which estimated parameters are acceptable or distribution functions are suitable for their work. The bootstrap technique can estimate the parameters easily and quickly. The finite mixture model makes the approach moderately useful for heavy tail (fat tail) distribution.

CONCLUSION

The finite mixture of Lognormal distributions can be fitted to the set of actual claim data while the Lognormal distribution cannot be fitted. The mixture of Lognormal distributions fit very well to product type-5. The limitation of the finite mixture model is the number of components that depend on a mean clustering. So we should be careful to consider the number of components used for computing the estimated parameters. The estimated parameter of Lognormal distribution by using the bootstrap method is fitted to the data according to K-S test. Although the bootstrap process is not as good for fitting in a tail as the finite mixture of Lognormal distributions is.

ACKNOWLEDGMENT

This research is (partially) supported by the center of excellence in mathematics, the Commission of Higher Education (CHE), Sriayudthaya Road, Bangkok 10140, Thailand.

We are grateful for the advice and help we obtained from discussions and from sample data sets provided by a Thai actuary.

REFERENCES

Aitkin, M. and D.B. Rubin, 1985. Estimation and hypothesis testing in finite mixture models. J.R. Statist. Soc. B., 47: 67-75.

Brazauskas, V., B.L. Jones and R. Zitikis, 2009. Robust fitting of claim severity distributions and the method of trimmed moments. J. Stat. Plann. Infer. 139: 2028-2043.

Efron, B. and R. Tibshirani, 1993. An Introduction to the Bootstrap. 1st Edn., Chapman and Hall, London, ISBN: 0412042312, pp: 436.

Hewitt, C.C. Jr., and B. Lefkowitz, 1979. Methods for fitting distributions to insurance loss data. Proc. Casualty Actuarial Soc., 125: 139-160.

Hogg, R.V. and S.A. Klugman, 1984. Loss Distributions. 1st Edn., By John Wiley Sons, New York, ISBN: 0471879290, pp: 235.

Hogg, R.V., A. Craig and J.W. McKean, 2004. Introduction to Mathematical Statistics. 6th Edn, Prentice Hall, Upper Saddle River, NJ., ISBN-10: 0130085073, pp: 692.

Janczuraa, J. and R. Weron, 2010. An empirical comparison of alternate regime-switching models for electricity spot prices. MPRA, 32: 1059-1073.

Klugman, S.A., H.H. Panjer and G.E. Willmot, 2008. Loss Models: From Data to Decisions, 3rd Edn., John Wiley Sons, Hoboken, ISBN: 9780470187814, pp: 726.

McLachlan, G.J. and D. Peel, 2000. Finite Mixture Models. 1st Edn., John Wiley Sons, Inc., New York, ISBN: 0471006262, pp: 419.

Mohamed, M.A., A.M. Razali and N. Ismail, 2010. Approximation of aggregate losses using simulation. J. Math. Stat., 6: 233-239. DOI: 10.3844/jmssp.2010.233.239