Modelling Insurance Losses and Calculating Risk Measures via a Mixture of Erlangs

Article

2 authors, including:

X. Sheldon Lin
University of Toronto

Modelling Insurance Losses and Calculating Risk Measures via a Mixture of Erlangs

Simon Lee [*]and X. Sheldon Lin[†]

Abstract

In this paper, we propose a mixture of Erlangs with common scale parameter to fit insurance data. Advantages of using a mixture of Erlangs include (i) that it can fit heavy tailed irregular data very well; (ii) easy calculation of VaR and CTE of aggregate losses, among others. We present an EM algorithm for the estimation of its parameters and demonstrate the efficiency of the algorithm using data generated from commonly used distributions as well as the extremely irregular PCS loss data. Closed-form expressions of VaR and CTE are given for individual and aggregate losses.

Keywords: misture of Erlangs, EM algorithm, aggregate losses, PCS catastrophe losses, value at risk, conditional tail expectation

1 Introduction

Data modeling has been a challenging and yet inevitable problem to actuaries and statisticians. On one hand, models help us to visualize the system we are interested in and link the real world phenomenon to decision making. On the other hand, it's not possible to find a model that can mimic the whole system. It can actually be reflected in George Box's infamous motto:

All models are wrong, but some models are useful.

With useful models in place, it facilitates users to make decisions. Especially in the context of actuarial science, a good model can help, for instance, determine the optimum retention level in reinsurance, calculate

[*]Simon Lee, A.S.A, is a Ph.D candidate in Department of Statistics at University of Toronto, Toronto, Ontario M5S 3G3, Canada, e-mail: simonlee@utstat.utoronto.ca

[†]X. Sheldon Lin, Ph.D., A.S.A., corresponding author, is a Professor of Actuarial Science in Department of Statistics at University of Toronto, Toronto, Ontario M5S 3G3, Canada, e-mail: sheldon@utstat.utoronto.ca

risk measures (value at risk and conditional tail expectation for examples) and design efficient simulation algorithms for solvency tests and required capitals. In many practical situations, users are faced with heavy tail data. It is particularly the case in reinsurance or catastrophe insurance. Thus, the goodness of fit on large claims is of critical importance. Therefore, a useful model should be able to reproduce the behaviour of the body as well as the tail of insurance claims data.

Heavy-tailed data are commonly modeled by theoretical heavy-tailed distributions such as lognormal, Weibull, or Pareto distributions. However, these theoretical heavy-tailed representations are intractable and are difficult to work with when data are needed to be aggregated. Hyper-exponentials are another commonly used distribution class because of its well understood properties and ease of manipulation (Rachid et al. (2003)):

$$F(t) = 1 - \sum_{i=1}^{\infty} \alpha_i e^{-t/\theta_i} \tag{1.1}$$

where α_j are non-negative and sum up to 1. However, hyper-exponential distributions have two major drawbacks: its requirement on the distribution to have complete monotonicity (Wang et al., 2006), and a coefficient of variation that is less than one (Thummler et al., 2006). With these restrictions, the fitted distributions will not have any interior modes. It is particularly insufficient when we deal with property and casualty claims. Property and casualty losses usually come from multiple sources. As a result, this type of loss data often exhibits a multimodal characteristic and hence multimodal models are necessary.

It is well accepted in physical science to apply combination of exponentials to fit distributions (Smith et al. , 1967). Its application in ruin theory is also popular (see Gerber et al. (1987), Dufresne and Gerber (1991), Lin and Willmot (1999), Lin and Willmot (2000), Lin and Willmot (2003) and Willmot and Lin (2001) for examples.) The distribution function of this type is the same as distribution (1.1) except that α_j are not restricted to be non-negative. This is an extension of hyper-exponentials distribution mentioned above (Willmot and Lin, 2001). It releases the restriction for hyper-exponentials and is dense in the space of positive continuous distributions. However, since some α_j's are allowed to be negative, negative probability density could likely be admitted in numerical implementation. Moreover, in order to solve the parameters (α_i and θ_i), techniques involving inverse of Laplace transform are applied (Smith et al. , 1967). This technique will involve solving polynomials. This will become very difficult if either the order of polynomial is high, roots are not robust or too close to each other. Furthermore, the roots might even be complex, leading to inappropriate density. Dufresne (2007) proposed a solution to this by introducing Jacobian polynomials. Instead of solving roots of Laplace transform, Dufresne obtains the parameters by integrating different polynomials. Thus, the above mentioned problems can be avoided. However, in Dufresne's model, several parameters are needed to be specified upfront. There are no specific selection rules on the optimal parameters. Thus, fitting results are not necessarily unique due to the freedom of parameter selection. Moreover, the convergence to some important distributions, uniform distribution for instance, is not satisfactory.

Klugman and Rioux (2006) suggested the augmented mixture of exponentials for approximation, having the following cumulative distribution function:

$$F_A(x) = mF_m(x;\alpha,\theta,k) + gF_G(x) + lF_L(x) + pF_P(x) \tag{1.2}$$

where $F_m(x;\alpha,\theta,k)$ is a distribution of hyper-exponentials in (1.1), and m, g, l, p are non-negative weights that sum to 1, and with either $g = 0$ or $l = 0$. $F_G(x)$, $F_L(x)$ and $F_P(x)$ are the cumulative distribution function of the Gamma distribution, lognormal distribution, and Pareto distribution respectively. Distribution (1.2) is thus a mixture of exponentials with the addition of the lognormal or Gamma distribution and Pareto distribution. The additions were proposed to allow for an interior mode (Klugman and Rioux, 2006). As we will show in Appendix, this mixture can have at most two modes only. Again, due to the incorporation of Log-normal and Pareto distribution, the problem of intractability arises in this method.

In this paper, we propose the use of a class of mixtures of Erlang distributions. In this class, all the underlying Erlang distributions share a common scale parameter. It is shown in Tijms (1994) that this class of distributions can approximate any positive absolutely continuous distribution to any accuracy. In other words, it is dense in the space of positive continuous distributions. Although Tijms (1994) provides a parameter estimation that is easy to work with, the parameters are not optimum and encounter the risk of overfitting. Our proposed method improves the accuracy of the fitting by employing EM algorithm. This algorithm is fast and efficient due to the fact that the approximation in Tijms (1994) is served as a good initial estimate of our model. We will also demonstrate that the proposed distribution will overcome some of the shortcomings in Dufresne (2007) and Klugman and Rioux (2006).

Modeling will only be useful if it sheds light on decision making. In order to maintain companies' solvency, interests are expressed in quantities like risk-based capital (RBC), value at risk(VaR) and conditional tail expectation (CTE) in risk management. Therefore, a tractable model is necessary to be in place to calculate the measures. Besides having a density function that can be written in the form of (2.1), a mixture of Erlangs can also be interpreted as a compound distribution with secondary distribution being an exponential distribution. Due to the simplicity of the secondary distribution, any quantiles and conditional distributions can be written in a linear combination of gamma kernels. Thus calculating VaR and CTE is trivial in our modeling. To the extent of authors' knowledge, there is no other common class of positive continuous distribution, having similar characteristics. As a consequence, mixtures of Erlangs with common scale parameter are also closed in mixing, convolution and compounding. These features will become particularly appealing when a clear picture of the distribution of claim counts can be obtained.

Similar to many data fitting procedures, the study of the performance of our proposed distributions in this paper follows the procedures below:

- *Parameter estimation.* Maximum log likelihood estimation will be used by employing the Expectation Maximization (EM) algorithm tailored to the proposed distribution.
- *Graphical comparison of fitted and empirical distribution.* Several plots will be used to compare the goodness of fit of the proposed to the empirical, and to the augmented mixture of exponentials respectively.
- *Statistical test of goodness of fit.* Hypothesis testing will be performed to quantitatively compare the goodness of fit of the proposed to the empirical, and of the augmented mixture of exponentials to the empirical.

The outline of the paper is as follows. We first introduce the proposed mixture of Erlangs with common scale parameter in Section 2. In Section 3, the details of the proposed EM algorithm for parameter estimation as tailored to the mixture of Erlangs with common scale parameter are presented and its computation efficiency is discussed. Section 4 describes the graphical comparison methods as well as the statistical tests of goodness of fit to be used. Uniform distribution has long been one of the most popular benchmark yet difficult distribution to approximate. In Section 5, we attempt to apply our proposed model to approximate uniform distribution and evaluated the fitness. In Section 6, a brief description of the fitness of our model to some other common distributions is presented to further support our proposed methods. With the generous contribution by ISO, we are able to model a set of catastrophe data in US. In Section 7, analyses about the fitness to this data are evaluated. With useful models in place, actuaries can work on premium setting and risk management. In Sections 8 to 9, analytical formulae for the stop-loss premium, the value-at-risk and the conditional tail expectation are derived for individual and aggregate losses. Conclusions are drawn in the last section of the paper.

2 The proposed model: mixture of Erlangs with common scale parameter

We propose to use mixtures of Erlang distributions to fit positive data. The density of this class of distribution is:

$$f(x|\theta, \alpha) = \sum_{i=1}^{k} \alpha_i \frac{x^{i-1} e^{-x/\theta}}{\theta^i (i-1)!} \tag{2.1}$$

where α_i is the non-negative weight of the ith Erlang distribution in the mixture, and θ is the common scale parameter.

Tijms (1994) showed that this distribution class is dense in the space of positive continuous densities. For any given positive distribution function $F(x)$, let :

$$\hat{f}(x|\theta) = \sum_{i=1}^{\infty} (F(i\theta) - F((i-1)\theta)) p_i(x) \tag{2.2}$$

where

$$p_i(x) = \frac{x^{i-1} e^{-x/\theta}}{\theta^i (i-1)!} \tag{2.3}$$

Then, as shown in Tijms (1994), $\lim_{\theta \to 0} \hat{F}(x|\theta) = F(x)$ for all continuous points x, where $\hat{F}(x)$ is the distribution function of $\hat{f}(x)$. Thus, the density function in (2.1) can approximate any positive continuous distribution. The accuracy can be improved by increasing k and/or decreasing θ in theory.

Although the above formula can approximate any positive continuous distribution in theory, the parameters are not optimized. Improving the accuracy by increasing the number of mixtures is a dangerous action as it may risk the problem of overfitting in many situations. Our proposed method is to optimize the parameters by applying EM algorithm and attempt to reduce the number of parameters needed for modeling.

As introduced in Section 1, our model has another distinctive characteristic to solve for risk measures or other statistics analytically. Any mixture of Erlang distributions with a common scale parameter can also be written as a compound distribution with secondary distribuiton being an exponential distribution. More precisely, if X has a density function in (2.1), then

$$X = \sum_{i=1}^{N} Z_i \tag{2.4}$$

where N is a counting random variable with $P(N = i) = \alpha_i$ and $Z_i, \quad$ where $i = 1, 2, \ldots$ is independent and identically distributed exponential distribution with scale parameter θ. Thus, convolution of random variables in this class that have the same scale θ will be again in the class. This is especially important when we want to aggregate the losses in different pools. Therefore, if company-wide risk measures are needed, our proposed models can aggregate all the heterogeneous pools easily and the measures of interest can be calculated without difficulties. Moreover, conditional distribution for our proposed model is again a linear combination of Erlangs (Willmot and Lin, 2001). This characteristic allows the censored or truncated statistics be calculated analytically and will be elaborated in more details in Sections 8 to 9.

3 EM Algorithm for a Mixture of Erlangs with common scale parameter

In this section, we present an EM algorithm that is tailored to our proposed distribution class. The proposed algorithms can be decomposed into three parts: the EM algorithm, parameters initialization, adjustment of shape parameters.

3.1 The EM Algorithm

The EM algorithm was proposed in Dempster et al. (1997) and it is an iterative algorithm for finding the maximum likelihood estimate of the parameters of an underlying distribution from a set of incomplete data. Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ be the incomplete data generated from a pair of random variable (X, Y) with joint density $p(x, y|\Phi)$, where Y is the hidden random variable and Φ is the set of parameters to be estimated. The corresponding complete data is then $\{(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)\}$, where $Y_i, i = 1, 2, \cdots, n$ are unobservable and hence are random. The complete-data log-likelihood is given by

$$l(\Phi|\boldsymbol{x}, \boldsymbol{Y}) = \sum_{i=1}^{n} \ln p(x_i, Y_i|\Phi) \tag{3.1}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$. Given the observed data $\boldsymbol{x}$ and the current estimates of the parameters $\Phi^{(k-1)}$, the distribution of $Y_i = y_i$ is given by

$$q(y_i|x_i, \Phi^{(k-1)}) = \frac{p(x_i, y_i|\Phi^{(k-1)})}{\sum_{j=1}^{n} p(x_j, y_i|\Phi^{(k-1)})} \tag{3.2}$$

Thus, the expectation of the complete-data log-likelihood is

$$Q(\Phi|\Phi^{(k-1)}) = \sum_{i=1}^{n} E\{\ln p(x_i, Y_i|\Phi)\} = \sum_{i=1}^{n} \int [\ln p(x_i, y_i|\Phi)] q(y_i|x_i, \Phi^{(k-1)}) dy_i \tag{3.3}$$

The procedure of the calculation of the expectation is called the E-Step.

The next step (M-Step) is to maximize the expectation:

$$\Phi^{(k)} = \max_{\Phi} Q(\Phi|\Phi^{(k-1)}) \tag{3.4}$$

The EM algorithm is particularly useful in estimating the parameters of a finite mixture, including its mixing weights. In the context of our model, $\Phi = (\theta, \alpha_1, \alpha_2, \cdots, \alpha_M)$, and Y is a counting random variable with values $1, 2, \cdots, M$. If $Y_i = l$, then it indicates that x_i follows Erlang(l) distribution.

$$p(x,y|\Phi) = \alpha_y \frac{x^{y-1}e^{-x/\theta}}{\theta^y (y-1)!} \tag{3.5}$$

E-step specifies conditional distribution $Y_i|x_i, \Phi$,

$$q(y_i|x_i, \Phi) = \alpha_{y_i} \frac{x_i^{y_i-1}e^{-x_i/\theta}}{\sum_{j=1}^n x_j^{y_i-1}e^{-x_j/\theta}} \tag{3.6}$$

and the corresponding loglikelihood,

$$Q(\Phi|\Phi^{(k-1)}) = \sum_{i=1}^n \sum_{l=1}^M \left(-\frac{x_i}{\theta} - l\ln\theta\right) q(l|x_i, \Phi^{(k-1)}) + C \tag{3.7}$$

where C is a constant independent of Φ.

The parameters that maximize the loglikelihood in (3.7) are:

$$\alpha_l = \frac{1}{n}\sum_{i=1}^n q(l|x_i, \Phi^{(k-1)}) \tag{3.8}$$

$$\theta = \frac{\sum\limits_{l=1}^M \sum\limits_{i=1}^n x_i q(l|x_i, \Phi^{(k-1)})}{\sum\limits_{l=1}^M l \sum\limits_{i=1}^n q(l|x_i, \Phi^{(k-1)})} \tag{3.9}$$

The iterations will continue until a convergence of loglikelihood or a predefined tolerance is reached. This process will locally maximize the estimates of the parameter θ and α_i on the given set of shape parameters r_i.

3.2 Initialization of Parameters

To begin this iterative algorithm, initial values for the parameters that needed to be estimated are required, namely those for θ and α_m. As stated in Section 2, (2.1) can approximate any non-negative distribution to any desired accuracy by increasing m and/or decreasing θ . However, with no restriction on the number of Erlangs in the mixture, it may lead to an issue of overfitting. In order to avoid the situation, one can start by specifying the number of Erlangs, k, needed for approximation. Then (2.1) is used to initialize the parameters. In particular, the initial values of α_i will be equal to $F(i\theta) - F((i-1)\theta)$. If it turns out that $F(i\theta) - F((i-1)\theta) = 0$ for some i, according to equation (3.6) and (3.8), α_i will always be zero in all the iterations. Hence, they can be deleted upfront for a more efficient use of computation time. Thus, at the end, the parameters needed to be estimates will be reduced.

It is noted that although decreasing θ can improve the accuracy of the fit on data with numerous modes, a smaller θ also causes a faster decay at the tail which is not desirable in fitting heavy tail data. Users can attempt to balance this trade-off by starting with appropriate initial guesses.

A program in R has been developed to implement the above procedure and will be used for parameter estimation in the subsequent studies.

3.3 Adjustment of shape parameters

In statistical modelling, overfitting is always a concern. Overfitting arises when too many parameters are used to fit a distribution. In this section, a parameter reduction procedure is introduced.

In Section 3.2, it is mentioned that some of the parameters could be deleted upfront. Thus, the resulting density is not necessarily a mixture of k Erlang distributions (may be a number less than k). Also, as pointed in previous subsections, the parameters only locally maximizing the likelihood. With this rationale, the overall fitness can be improved by adjusting shape parameters. The algorithm is as follows: initially, the set of shape parameters is $\{r_1, r_2, ..., r_n\}$. EM algorithm in Section 3.1 will be run again for $\{r_1+1, r_2, ..., r_n\}$. If the resulting likelihood is higher, then it will replace the old set. The process will go on until the likelihood for $\{r_1 + 1, r_2, ..., r_n\}$ is lower than that of $\{r_1, r_2, ..., r_n\}$. At that instant, we will apply similar procedures to the 2nd parameter until all the parameters are treated. We then run the whole process again in a similar fashion but instead of increasing the parameters, we will decrease them. For instance, we will compare the likelihood for $\{r_1, r_2, ..., r_n\}$ and $\{r_1 - 1, r_2, ..., r_n\}$. After a set of above adjustment, some of the parameters may be found insignificant to improve the log-likelihood. In this case, a trigger, l, can be set such that when $\alpha_i < l$, r_i, α_i will be deleted. l can be set inversely related to the number of shape parameters. $l = \frac{0.01}{\text{number of shape parameters}}$ is an instance. This can reduce the problem of overfitting. Overfitting will become a more important issue when irregular data is fitted. It is because a lot of Erlang distributions will be needed for fitting. Thus, it is more desirable to delete insignificant parameters to avoid the issue of overfitting. The whole process will be repeated until the likelihood cannot be improved. The resulting set of parameters will be maximizing the likelihood in a larger subspace.

3.4 Efficiency

Although both our proposed method and Dufresne (2007) work can approximate any non-negative random variable to any desired accuracy, Dufresne's method involves users providing educated guess of a few parameters. Optimizing such values is a non-trivial task. It also involves solving many integrals, depending on the number of exponentials in the combination, using numerical procedures. In contrast, our method only requires finding the probabilities of various intervals. As such, the computations are more direct and involve little complicated numerical algorithms. Moreover, as stated in Dufresne (2007), the method may not work well on approximating some distributions. It is due to the fact that the coefficient of variation for an exponential is 1. For modeling distribution with coefficient of variation that is varying or far from unity, a lot of exponentials will be needed. On the contrary, our proposed method has a bigger freedom on this problem as each Erlang distribution has a unique coefficient of variation. Thus, in order to obtain the same accuracy with our proposed method, Dufresne's model may require a large number of exponentials, inducing the problem of overfitting. The difference will become significant when efficiency or computing time is a concern. A comparison between the two models will be made in Section 5.

Another notable characteristic of the proposed algorithm is its method of parameter initialization. Good initial estimates are generally one of the most critical elements in deciding computational speed of an algorithm as they dictate the number of iterations the algorithm needs to run through.

4 Evaluation of Goodness-of-Fit

Evaluation of goodness of fit can be conveyed by two means, qualitatively by graphical plots, and quantitatively by statistical hypothesis testing and other measures. They will be used in the studies in Sections 5 and 7.

4.1 Graphical comparison of Goodness-of-Fit

Although subjective, it is always good to start with observing the plot of the observed distribution and fitted distribution. In the two studies that follow, three plots will be used for this purpose: fitted distribution curve with empirical histogram overlay, the probability-probability (P-P) plot, and the quantile-quantile (Q-Q) plot. The procedure for constructing each is based on Law and Kelton (1995).

4.1.1 Fitted Distribution Curve with Empirical Histogram Overlay

Let X_i 's be the set of observed data with a fitted density function $\widehat{f}(x)$, $i = 1, 2, 3 \cdots, N$. We define a set of n histogram interval with endpoints d_j's such that each interval has width $\Delta : \{[d_0, d_1), [d_1, d_2), ..., [d_{n-1}, d_n)\}$; a set of m_j's is defined as the midpoint of jth interval. Let O_j be the proportion of observed data that fall in interval $[d_{j-1}, d_j)$, and E_j be the expected proportion of data that fall in that interval if data were generated from the fitted density function. O_j is determined by counting the number of observed data that fall in interval j and dividing by N; E_j is approximated by $\Delta \hat{f}(m_j)$. E_j's are then plotted as a curve with O_j's as an histogram overlay. If the fitted distribution is indeed a good fit of the true underlying distribution of the observed data, the curve will closely envelop the histogram.

4.1.2 P-P Plot

While the plot in 3.1.1 makes use of the probability density function, the P-P plot and the Q-Q plot make use of the cumulative distribution function. Let X_i's be the set of observed data with a fitted cdf $\widehat{F}(x), i = 1, 2, 3, ..., N$, where:

$$\widehat{F}(x) = 1 - \sum_{i=1}^{M} \alpha_i e^{-x/\theta} \sum_{n=0}^{\alpha_i} \frac{x^n}{\theta^n n!} \tag{4.1}$$

We sort the X_i's in ascending order, and let $X_{(j)}$ be the jth order statistic of the X_i's. According to Chakravarti et al. (1967), define the empirical cumulative density function as:

$$\widetilde{F}(X_{(j)}) = \frac{(j - 0.5)}{N} \tag{4.2}$$

The P-P plot consists of plotting the fitted $\widehat{F}(x_{(j)})$ versus the empirical $\widetilde{F}(x_{(j)})$; more explicitly, the P-P plot compares the cdf values of the empirical versus the fitted at the same$X_{(j)}$. If the fitted distribution is indeed a good fit of the true underlying distribution of the observed data, the plotted curve will be approximately linear with slope 1 and intercept at 0. One advantage of the P-P plot over the fitted curve with histogram overlay is its ease of perception, namely, deviations from a straight line is easier to percept than deviations from an irregular curve.

4.1.3 Q-Q Plot

Similar to the P-P plot, the Q-Q plot is based on the cumulative distribution function. In contrast, the Q-Q plot compares the quantile values of the empirical versus the fitted at the same cdf value. Define $X_{O(j)}$ as the jth order statistic of the observed data, and $X_{O(j)}$ as the empirical counterpart where:

$$X_{e(j)} = \widetilde{F}^{-1}(\widehat{F}(X_{O(j)})) \tag{4.3}$$

The Q-Q plot consists of plotting the fitted $X_{e(j)}$ versus the observed$X_{O(j)}$. If the fitted distribution is indeed a good fit, the plotted curve will be approximately linear with slope 1 and intercept at 0. While the P-P plot is good at detecting deviations in the body of the data, the Q-Q plot amplifies the deviations that exist at the tails between the fitted and the empirical distributions (Law and Kelton (1995)).

4.2 Statistical Tests of Goodness-of-Fit

Statistical tests or more technically hypothesis tests always start with two hypotheses, namely null hypothesis H_0 and alternative hypothesis H_a.

H_0: The data follow the fitted distribution

H_a: The data do not follow the fitted distribution

If the fitted distribution closely resemble the underlying distribution of the observed data, then H_0 will be accepted. In this paper, we use 5% significant level for the test.

4.2.1 Chi-Square Goodness-of-Fit Test

The Chi-square test is widely used to test for goodness-of-fit. Its notable advantage is its ease of use in two respects: firstly, the test statistic for the observed data can be computed without excessive effort, and secondly, the critical values of the test statistic is independent of the fitted distribution and are readily tabulated. To compute the test statistic, divide the distribution range into n intervals with endpoints d_j's: $\{[d_0, d_1), [d_1, d_2), ..., [d_{n-1}, d_n)\}$. Let O_j be the number of observed data that fall in interval $[d_{j-1}, d_j)$, and E_j be the expected number of data that fall in that interval if data were generated from the fitted density function. According to Smith et al. (1967), E_j is calculated as follows:

$$E_j = n \int_{d_{j-1}}^{d_j} \widehat{f}(x)dx$$

The test statistic is then:

$$\chi^2 = \sum_{j=1}^{n} \frac{(O_j - E_j)^2}{E_j} \tag{4.4}$$

The null hypothesis is rejected if $\chi^2 > \chi^2_{n-m-1}$.

4.2.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test is based on the empirical distribution function (ECDF).
The Kolmogorov-Smirnov test statistic is defined as:

$$D = \max_{1 \le i \le N} \left(\hat{F}(Y_i) - \frac{i-1}{N}, \frac{i}{N} - \hat{F}(Y_i) \right)$$

where $\hat{F}$ is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution

An attractive feature of this test is that the distribution of the K-S test statistic itself does not depend on the underlying cumulative distribution function being tested. Another advantage is that it is an exact test (the chi-square goodness-of-fit test depends on an adequate sample size for the approximations to be valid).

4.2.3 Anderson-Darling Test

It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. This characteristic is of particular importance as one of the main goals of our proposed model is to fit the tail of heavy-tail distribution well. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution.

The Anderson-Darling test statistic is defined as

$$A = -n - S \tag{4.5}$$

where n = number of observation and

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \Big(ln\hat{F}(Y_i) + ln(1 - \hat{F}(Y_{n+1-i}) \Big)$$

4.3 Moments for Goodness-of-Fit

Another quantitative measure of goodness-of-fit is by comparing the raw moments of the empirical data and that of the fitted distribution. If the fitted distribution is a good fit, the moments will coincide with the moments of the empirical distribution. In particular, at high order moments, the tails contributes more significantly. Thus, deviations at the tail are amplified at high order moments (Wang et al. (2006)). For a mixture of Erlangs with the common scale parameter, the nth moment can be simply calculated as follows:

$$E(X^n) = \theta^n \sum_{i=1}^{\infty} \alpha_i \frac{(r_i - 1 + n)!}{(r_i - 1)!} \tag{4.6}$$

5 Fitting Uniform Data

The uniform distribution is a benchmark to test whether a fitting algorithm is of high quality. The uniform distribution has a flat density and sharp decay at the end points. It is usually difficult to fit the uniform density using the smooth distribution with a mode or modes. A set of 5000 data between 1 and 2 were generated uniformly for study. In this study, we will use Dufresne's result as a comparison.

The EM algorithm is applied to the data. A mixture of 19 Erlangs with a fixed θ of 0.000498 fits the data well with each Erlang having significant weight. The parameters that maximize the loglikelihood function can be found in Appendix 2. The resulting loglikelihood value is -83.18.

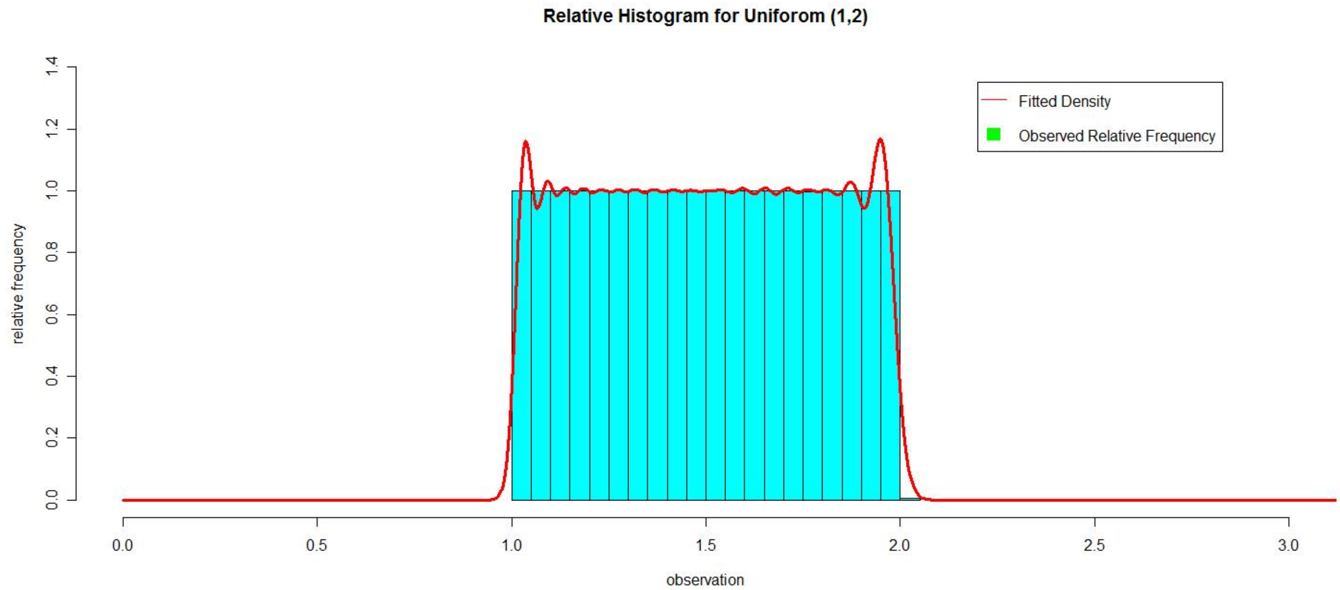

Figure 1: Histogram of uniform distribution and line for the fitted distribution

5.1 Graphical Comparison of the fitted distribution and underlying distribution

The fitted distribution is graphed with the empirical histogram overlay in the Figure 1. It can be observed that the fitted curve tightly envelops the relative histogram with slight overshoot near $x = 1^-$ and $x = 2^+$. The QQ plot in Figure 2 provide another evidence of overshooting at the extremes. The PP plot suggests that the proposed model has a good fitness.

5.2 Statistical Tests

The follwing table summarizes the results of the statistical tests suggested in the previous section.

Test	Statistic	p-value	Accepted at 5% significant level?
Chi Square Test	61.1416	1	Yes
K-S Test	0.0062	1	Yes
AD Test	-1.05754	0.64383	Yes

Table 1: Results for statistical tests for the fitness of the proposed model

The fitted distribution passed all three tests with significant margin. It implies that the fitted distribution is a good representation of the uniform distribution. Dufresne (2007) uses a mixture of exponentials to fit

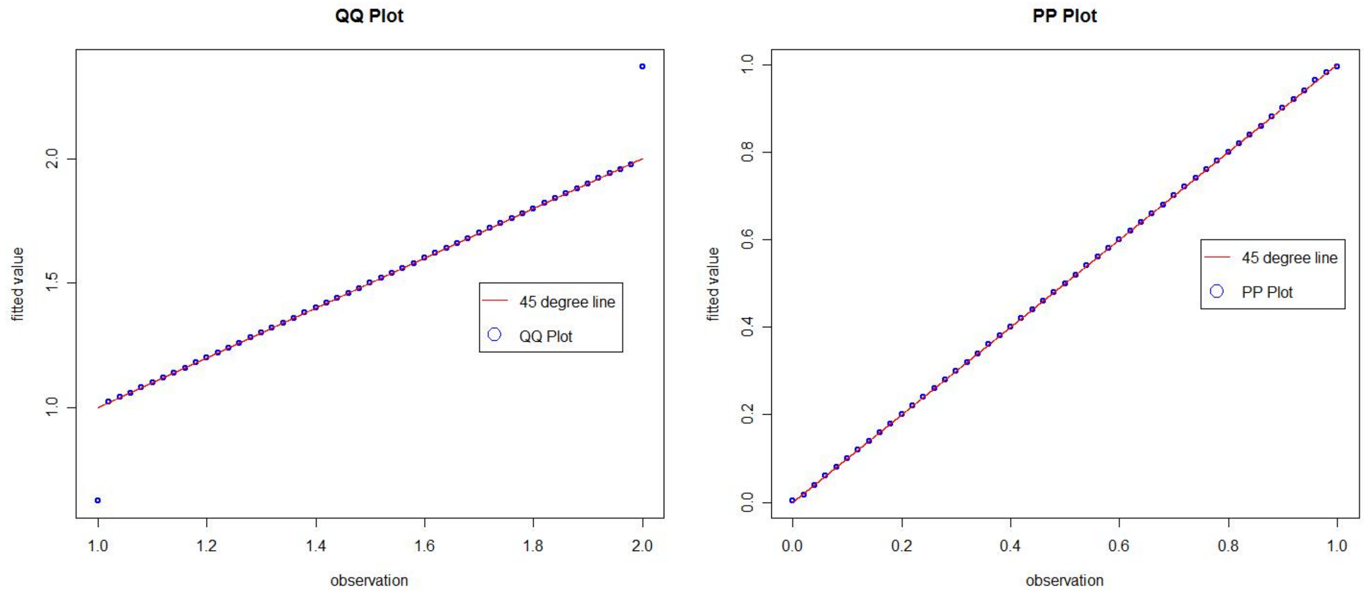

Figure 2: PP and QQ plot for uniform (1,2) and the fitted distribution

the same the distribution. Dufrense used Kolmogorov-Smirnov statistics as the measure of goodness-of-test.

The statistic for a mixture of 20 Exponentials in Dufresne (2007) is 0.038 while the statistic for a mixture of 19 Erlang distributions is 0.006. Note that the number of parameters estimated for Dufrense's model and our model are 39 and 38 respectively. Thus, by forgoing similar degree of freedom, our proposed method performs better in fitting the uniform distribution.

5.3 Relative Moments for Goodness-of-Fit

n	Empirical Distribution	Fitted Distribution	Fitted/ Empirical	Percentage Difference (%)
1	1.5	1.5	1.00000	0.0000%
2	2.333	2.333	1.00001	0.0006%
3	3.750	3.750	1.00002	0.0018%
4	6.200	6.2001	1.00004	0.0040%
5	10.501	10.502	1.00008	0.0075%

Table 2: Uniform Data- Relative Moments

Table 2 compares the 1-5th order moment of the empirical distribution and the fitted distribution for the uniform data. Recall that higher order moments amplify the deviations at the upper tail. The percentage difference of the 5th order moment is 0.0083% which is negligible. Again, it shows that the fitted distribution is a good fit to the data.

The results of qualitative graphical comparisons, quantitative statistical test, and relative moments all suggest that the fitted distribution is suitable approximation to the uniform data.

6 Fitting of other common distribution

This section provides examples of some other common distributions with light tail(Weibull) and heavy tail(Weibull, Pareto and lognormal) distributions without detailed quantitative tests.

A Weibull distribution can be either light tail or heavy tail depending on the shape parameter. If the shape parameter is bigger than 1, then it is a light tail. The opposite comes true when the parameter is smaller than 1. In the first two examples of this section, we will introduce the graphs for Weibull with scale parameter 2 and shape parameter 2(Example 1) and 0.8(Example 2) respectively.

Example 1: Light tail Weibull distribution

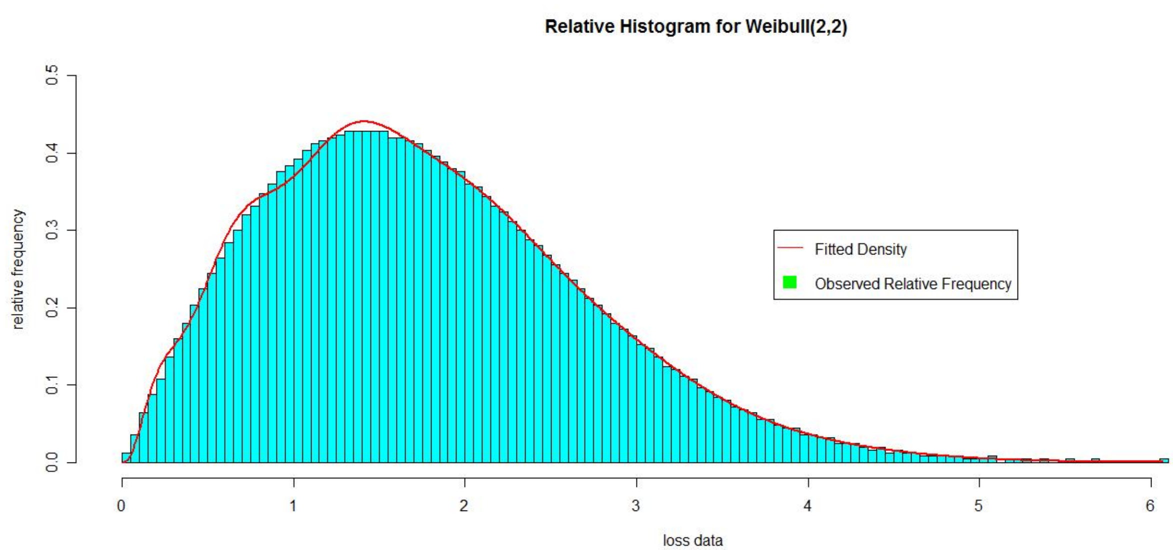

Figure 3: histogram for Weibull (2,2) and the fitted density using a mixture of 6 Erlangs

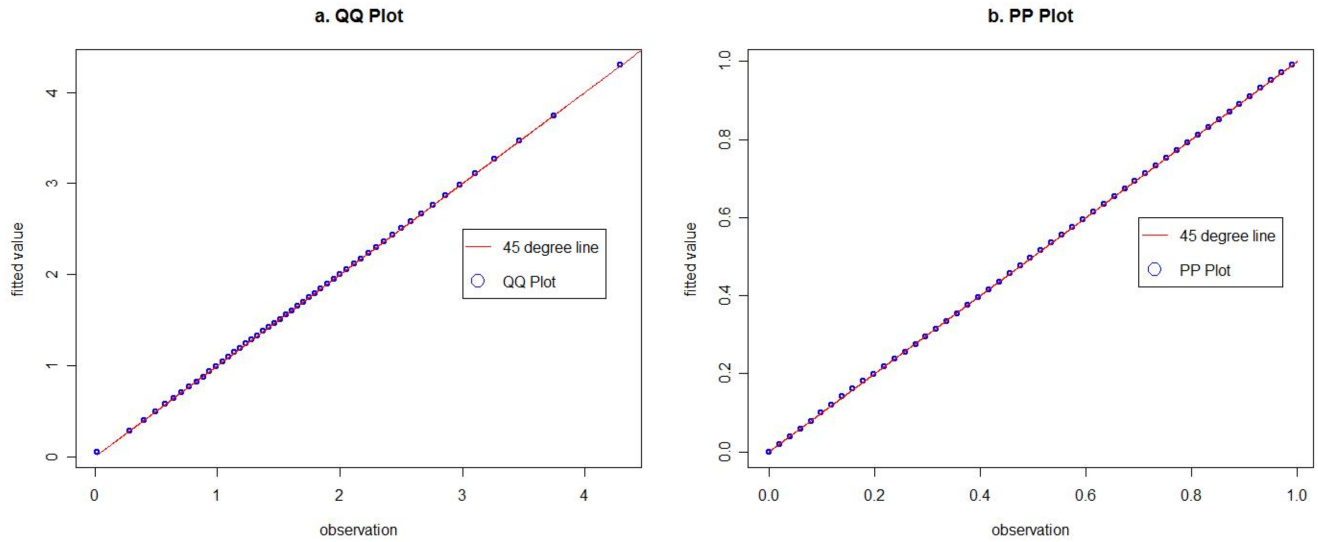

Figure 4: PP and QQ plots for Weibull (2,2) and the fitted distribution

The plots in figures 3 and 4 show our proposed model has an almost complete fit for the light tail Weibull distribution.

Example 2: Heavy tail Weibull distribution

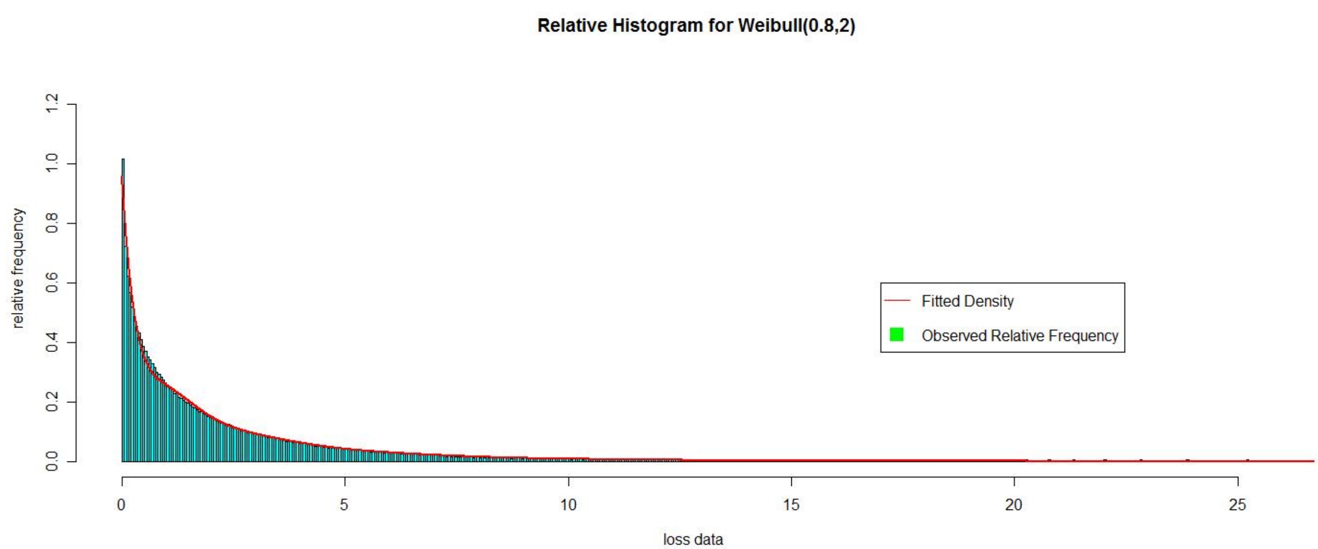

Figure 5: Histogram for Weibull (0.8,2) and the fitted density using a mixutre of 10 Erlangs

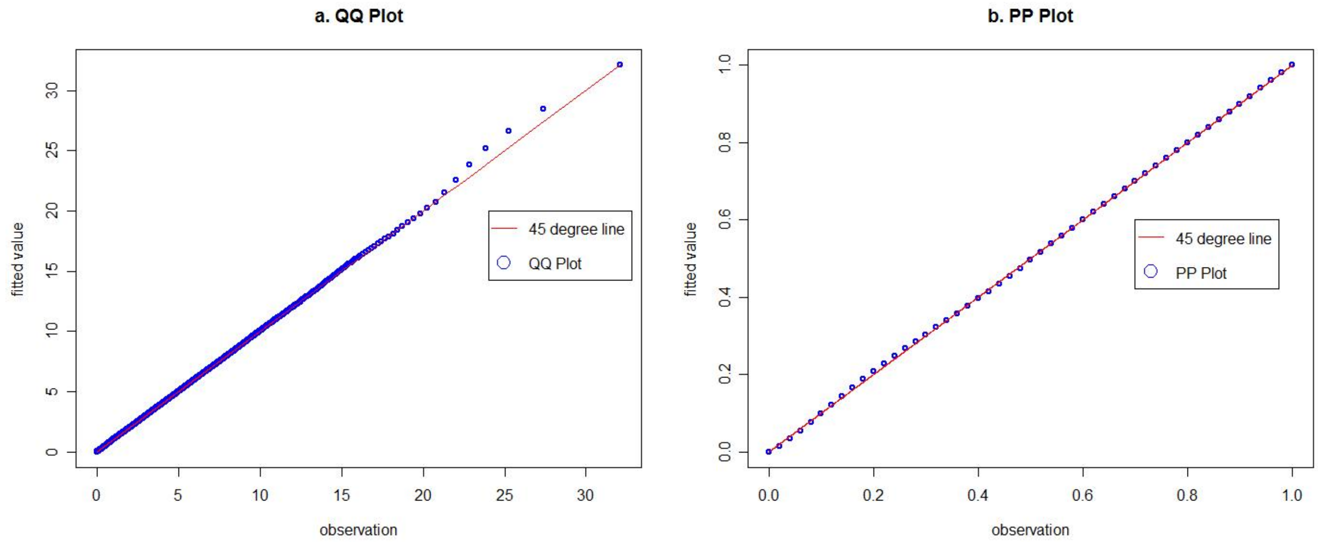

Figure 6: PP and QQ plots for Weibull (0.8,2)

The plots in figures 5 and 6 show our proposed model again performs well in fitting for the heavy tail Weibull distribution. The QQ plot suggests that the fitness of the tail is not perfect. However, the problem can be solved by using more Erlangs. The effect of increasing the number of Erlangs will be shown in the example for lognormal distribution.

Example 3: Pareto distribution

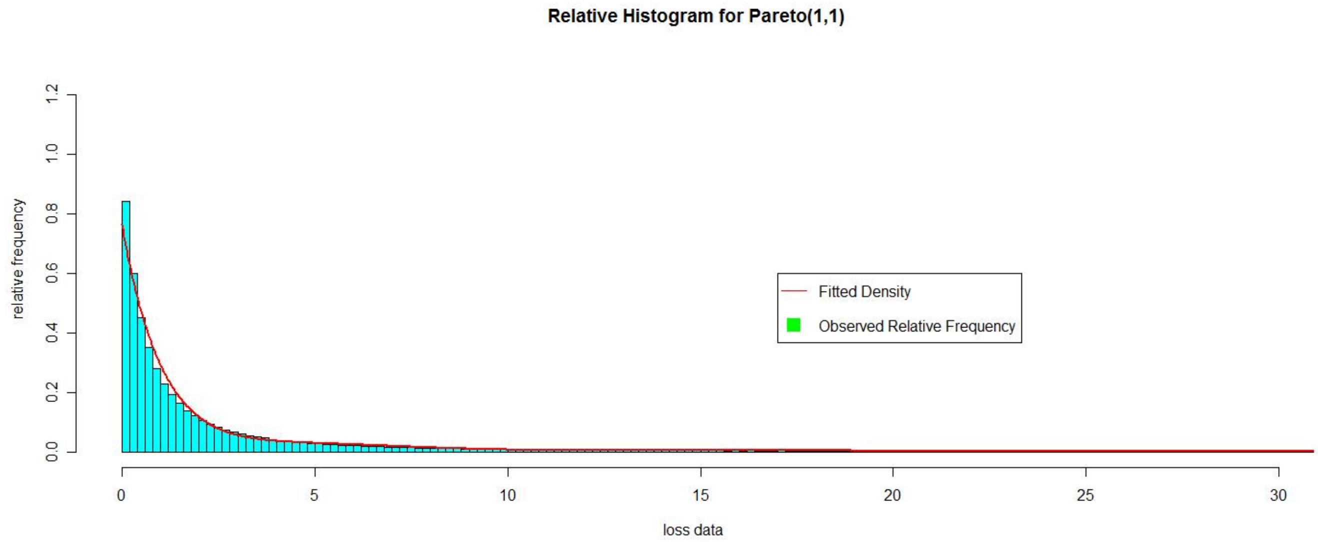

Figure 7: Histogram for Pareto(1,1) and the fitted desnity using a mixutre of 5 Erlangs

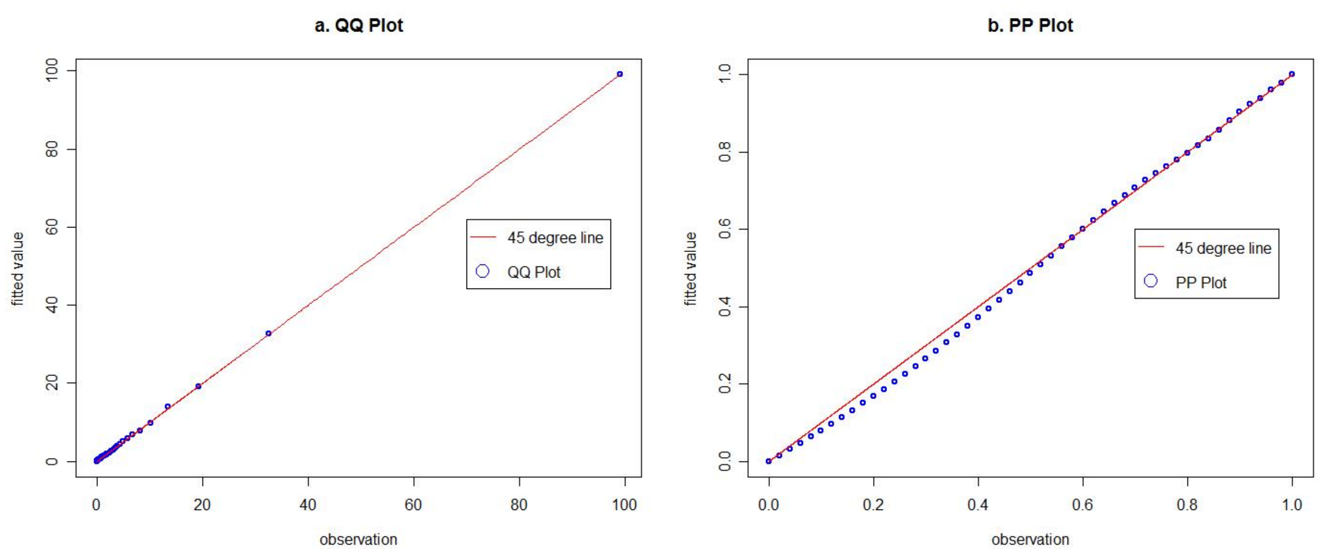

Figure 8: PP and QQ plots for Pareto (1,1)

As shown in figures 7 and 8, the plots again show our model has a good fitness for Pareto distribution.

Example 4: Lognormal distribution

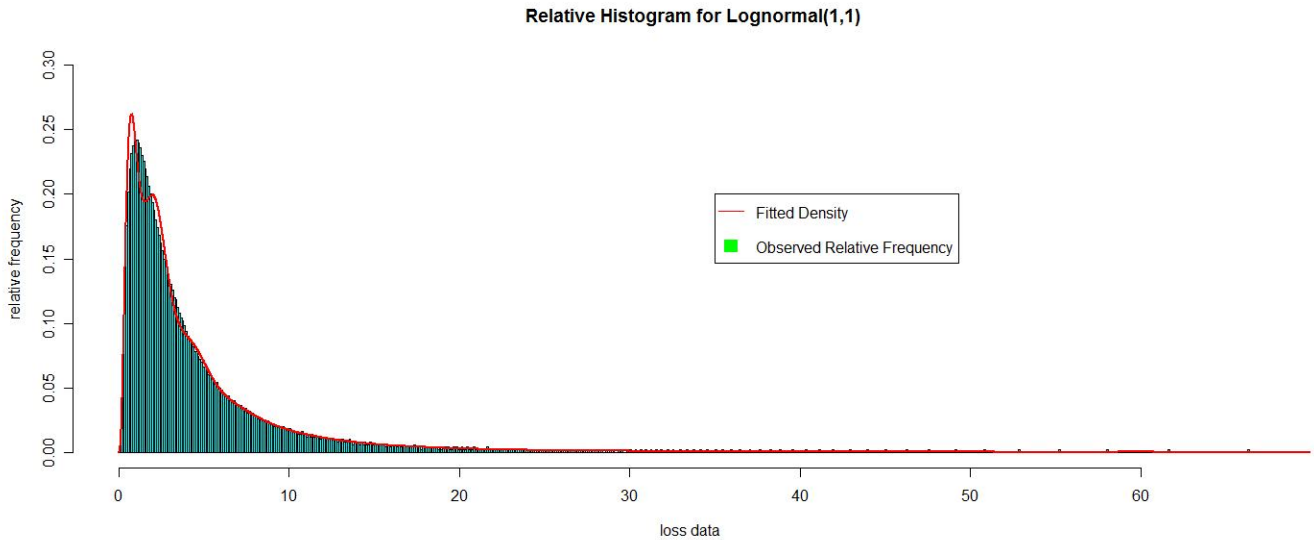

Figure 9: Histogram for lognormal(1,1) and the fitted desnity using a mixutre of 15 Erlangs

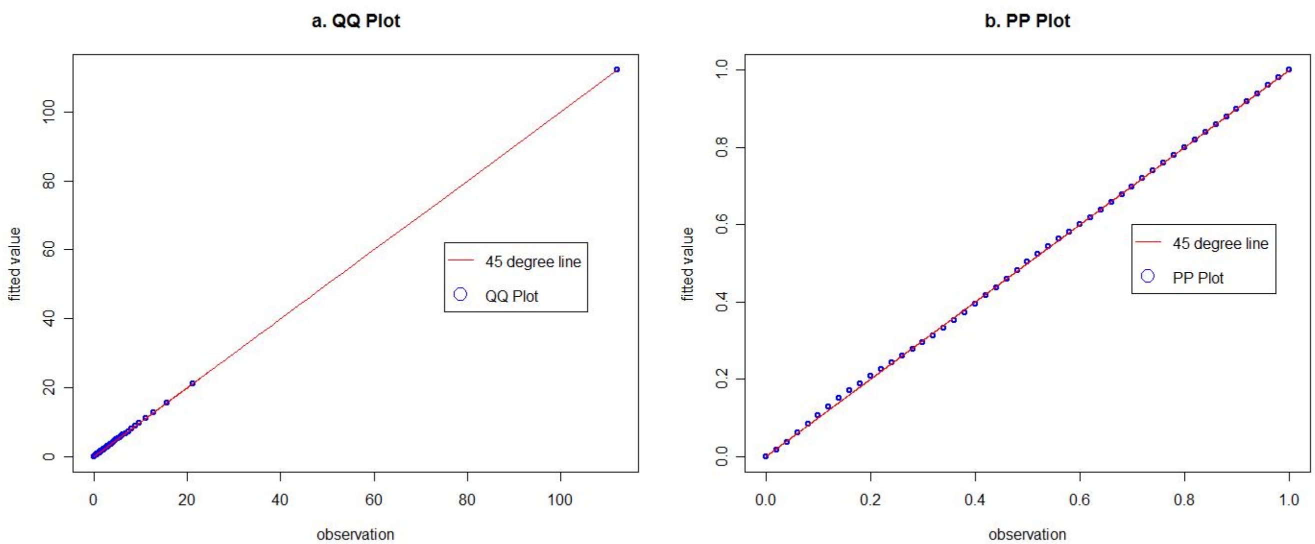

Figure 10: PP and QQ plots for lognormal (1,1)

By increasing the number of Erlangs used for fitting to 15, the fitness for heavy tail distribution improves significantly. As shown in figures 9 and 10, the plots again show our model has an almost perfect fitness for lognormal distribution.

7 PCS Catastrophe Data

7.1 Data and parameter estimation

In this study, we apply the proposed mixture of Erlangs with common scale parameter to a set of 1271 catastrophe loss data in US from 1997 to 2005. The data is generously supplied by ISO, a leading source of information about risk.

Due to confidentiality, explicit data cannot be provided. However, some features can be outlined as follows:

1. The maximum value of the data is 247 times of the mean
2. There is 9.13% of the observations categorized as outliers if 1.5 IQR rule is used.
3. The skewness and kurtosis for the data are 23.04 and 619.63
4. 56% of the data is smaller than 1/1000 of the maximum value while 96.6% of the data is smaller than 1/100 of the maximum value

All points above suggest that the data is heavy-tailed.

A mixture of 52 Erlangs with a common scale parameter of 257683 fits the data well with each Erlang having significant weight. The fitting results and computing times[1] are presented in Table 3.

Methods	Tijm (Step 1)	EM (Step 2)	Adjustment (Step 3)
Parameters	188	188	104
Log-likelihood	-23321	-23297	-23258
Computing time	1 sec	25 sec	15 min

Table 3: Summary of efficiency of various methods.

Although the number of parameters in the model seems to be large, when taken into account about the data characteristics, it is not a surprise that such a large number of parameters are needed. The data contains largely separate values. If only few parameters are fed, the fitted model may not be able to convey correct messages about the underlying phenomenon. Moreover, the procedure described in Section 3.3 can effectively reduce the number of parameters. The number of parameters used is reduced by more than 40%. Thus, the issue of overfitting can be lessened.

[1] The authors used a PC with 2.10GHz processor and 8.00GB RAM

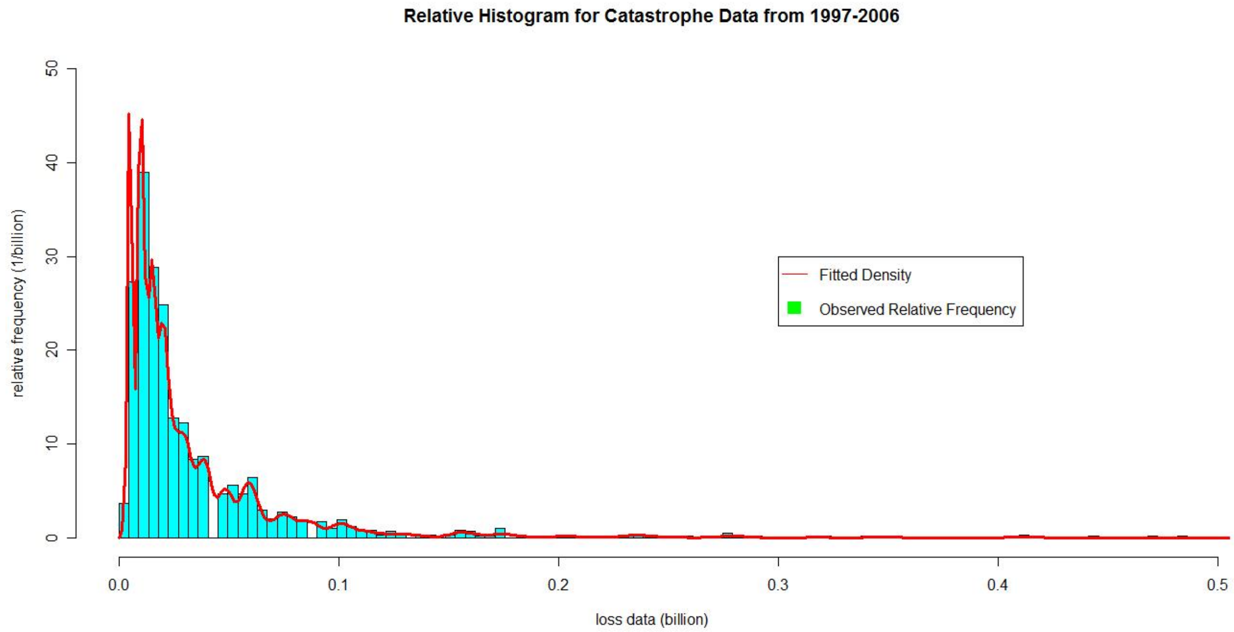

Figure 11: Histogram of observed loss and line for the fitted distribution

7.2 Graphical comparison of Goodness-of-Fit

From figures 11 and 12, it is observed that the fit is satisfactory. The application of a lot of Erlangs is justified as the data is highly irregular. If only few parameters are used, modes in the observed loss will disappear and may risk misleading information.

In the PP plot, points in the body do not lie on the 45 degree line. It is due to the fact that the observations are concentrated in a few ranges. In each range, the observations are closed but not located at the same point. Thus, a small percentage increase in the observation will cause a significant change in percentile. This problem is common in modeling. Again, this problem can be reduced by increasing the number of Erlangs.

7.3 Statistical Test

Unlike the previous example of uniform distribution, the catastrophe data is highly irregular. Because of this characteristic, a high number of parameters are needed. However, when many parameters are used, it is generally not appropriate to use Chi-Square test to test the goodness-of-fit. Thus, the test is omitted for the distribution.

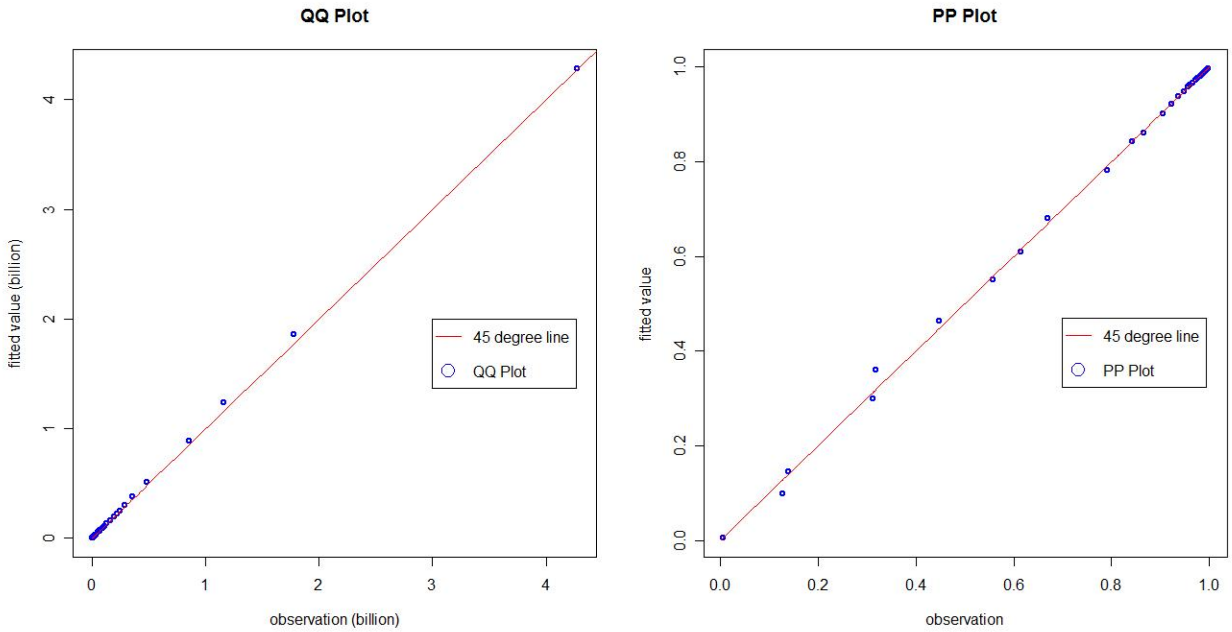

Figure 12: Histogram of observed loss and line for the fitted distribution

Test	Statistic	p-value	Accepted at 5% significant level?
Chi Square Test	Not available	Not available	Not available
K-S Test	0.0373	1	Yes
AD Test	-1.26239	0.69898	Yes

Table 4: Results for statistical tests for catastrophe data

Again, as shown in table 4, the fitted distribution performs well on statistical tests.

7.4 Relative Moments for Goodness-of-Fit

nth moment	Empirical	Fitted	Fitted/Empirical	Percentage Difference (%)
1	$9.833 * 10^{7}$	$9.833 * 10^{7}$	1.0000	0.00%
2	$6.917 * 10^{17}$	$6.908 * 10^{17}$	0.9987	-0.13%
3	$1.317 * 10^{28}$	$1.315 * 10^{28}$	0.9983	-0.17%
4	$2.932 * 10^{38}$	$2.926 * 10^{38}$	0.9979	-0.21%
5	$6.857 * 10^{48}$	$6.840 * 10^{48}$	0.9975	-0.25%

Table 5: Catastrophe Data- Relative Moments

Quantities	Empirical	Fitted	Fitted/Empirical	Percentage Difference (%)
Mean	98.33 million	98.33 million	1.0000	0.00%
Standard Deviation	825.85 million	825.31 million	0.9993	-0.07%
Skewness	23.03	23.04	1.0003	0.03%
Kurtosis	619.28	619.63	1.0006	0.06%

Table 6: Catastrophe Data- Relative Moments

Table 5 compares the 1st-5th order moment of the empirical distribution and the fitted distribution for the catastrophe data. Recall that higher order moments amplify the deviations at the upper tail. Even in a higher moment, there is only negligible difference. Considering that there are some extremely value in the data set, the difference indicates a good fit to the data. Table 6 provides popular descriptitve quantities about the general shape of the distributions. Again, the quantities for the fitted distribution closely ensemble the values for the empirical data.

The results of qualitative graphical comparisons, quantitative statistical test, and relative moments all suggest that the fitted distribution is a suitable approximation to the Catastrophe data.

8 Calculation of risk measures for individual loss model

In previous sections, the performance of the proposed fitting distribution on two sample datasets was evaluated. In this section, a few unique properties of our proposed models are employed to derive closed-form expressions for insurance risk measures.

Calculation of risk measures is of central importance to the risk management of insurance companies. In this and next sections, we derive analytical expressions for the value at risk, the stop-loss premium and the conditional tail expectation using the proposed distribution for individual loss models and aggregate loss models.

8.1 Value at Risk (VaR)

Let X be a loss random variable and $F_X(x)$ be its distribution function. The value at risk at confidence level p, denoted by V_p, is the $100p$-th percentile of the distribution and hence is define as

$$F_X(V_p) = p. \tag{8.1}$$

When the loss is modelled by a mixture of Erlangs with density (2.1), the calculation of V_p is straightforward. It is easy to see that its distribution function is given by

$$F_X(x) = 1 - e^{-x/\theta} \sum_{i=1}^{M} \alpha_i \sum_{j=0}^{i-1} \frac{x^j}{\theta^j j!} \tag{8.2}$$

Thus, V_p is the solution of

$$e^{-V_p/\theta} \sum_{i=0}^{M} Q_i \frac{V_p^i}{\theta^i i!} = 1 - p \tag{8.3}$$

where $Q_i = \sum_{j=i+1}^{M} \alpha_j$. Any numerical method will work to find its value.

8.2 Stop loss premium

The stop-loss premium $E(X-d)^+$ with deductible d may be calculated by

$$\begin{aligned} E(X-d)^+ &= \int_0^\infty [1 - F_X(y)]dy = \int_d^\infty e^{-y/\theta} \sum_{i=0}^{M} Q_i \frac{y^i}{\theta^i i!} dy \\ &= \theta e^{-d/\theta} \sum_{i=0}^{M} Q_i^* \frac{d^i}{\theta^i i!} \end{aligned} \tag{8.4}$$

where $Q_i^* = \sum_{j=i}^{M} Q_j = \sum_{j=i+1}^{M} (j-i)\alpha_j$

As a result, the expected loss under various policy modifications and the loss elimination ratio are obtainable explicitly.

8.3 Conditional Tail Expectation

Conditional Tail Expectation is also called conditional VaR (CVaR), Tail VaR (TVaR) or Expected Shorfall. It plays a similar role as VaR and is a risk measure that may be used for the determination of economic capital.

The conditional tail expectation at confidence level p is usually denoted as $CTE(100p)$ and defined as

$$\text{CTE}(100p) = E(X|X > V_p). \tag{8.5}$$

Using the mixture of Erlangs model, it can be calculated easily:

$$\begin{aligned} \text{CTE}(100p) &= \frac{E(X - V_p)^+}{1-p} + V_p \\ &= \frac{\theta e^{-V_p/\theta}}{1-\beta} \sum_{i=0}^{\infty} Q_i^* \frac{V_p^i}{\theta^i i!} + V_p. \end{aligned} \tag{8.6}$$

9 Calculation of risk measures for aggregate loss models

As mentioned earlier, a great advantage of a mixture of Erlangs model is that it enables us to calculate many risk measures for aggregated risks in a fairly simple way. In this section, we will demonstrate how to calculate VaR and CTE for aggregate loss models.

Let the counting random variable N with $P(N = n) = p_n,\ n = 0, 1, \cdots$, represents the claim count from an insurance portfolio, and a sequence of positive, independent and identically distributed random variables $X_n,\ n = 1, 2, \cdots$, represents successive claim amounts arising from the portfolio. It is now assumed that the distribution of X_n has the density given by (2.1). The amount of the aggregate losses is thus

$$S = \sum_{n=1}^{N} X_n \tag{9.1}$$

where $S = 0$, if $N = 0$. The positive portion of the distribution of S is again a mixture of Erlangs with the same scale parameter. More precisely, the defective density of $S,\ S > 0$, is

$$f_S(x) = \sum_{k=1}^{\infty} \eta_k \frac{x^{i-1} e^{-x/\theta}}{\theta^i (i-1)!}. \tag{9.2}$$

The coefficients $\eta_k,\ k = 0, 1, \cdots$, are the coefficients of the power series $P_N(P_\alpha(z))$, where $P_N(z)$ is the probability generating function (PGF) of N and $P_\alpha(z) = \sum_{i=1}^{M} \alpha_i z^i$. Furthermore, if N belongs to the (a,b,1) class, that is,

$$p_n = \left(a + \frac{b}{n}\right) \cdot p_{n-1}, \quad n = 2, 3, 4, \ldots,$$

η_k can be calculated recursively using the Panjer recursion: for $k = 1, 2, \cdots$,

$$\eta_k = [p_1 - (a+b)\, p_0]\, \alpha_k + \sum_{j=1}^{k} \left(a + \frac{b \cdot j}{k}\right) \cdot \alpha_j \cdot \eta_{k-j} \tag{9.3}$$

with $\eta_0 = p_0$. See Klugman et al (2004).

9.1 Value at Risk

Similar to Equation (8.1), the $100p\%$ VaR, V_p^* can be derived by solving the following equation:

$$e^{-V_p^*/\theta} \sum_{i=0}^{\infty} P_i \frac{V_p^{*i}}{\theta^i i!} = 1 - p, \tag{9.4}$$

where $P_i = \sum_{k=i+1}^{\infty} \eta_k$.

9.2 Stop- Loss Premium

The stop-loss premium with deductible d can be evaluated in a similar fashion:

$$E\left(\sum_{j=1}^{N} X_j - d\right)^+ = \theta e^{-d/\theta} \sum_{i=0}^{\infty} P_i^* \frac{d^i}{\theta^i i!}. \tag{9.5}$$

where $P_i^* = \sum_{j=i}^{\infty} P_i = \sum_{k=i+1}^{\infty} (k-i)\eta_k$

9.3 Conditional Tail Expectation

The formula for CTE below is another analog for $100p\%$ conditional tail distribution for individual loss:

$$\mathrm{CTE}(100p) = \frac{\theta e^{-V_p^*/\theta}}{1-p} \sum_{i=0}^{\infty} P_i^* \frac{V_p^{*i}}{\theta^i i!} + V_p^*. \tag{9.6}$$

10 Concluding Remarks

We have demonstrated in this paper that the proposed mixture of Erlangs model can fit insurance data well, especially irregular heavy tailed data and data with multiple modes. We also showed that many useful risk measures can be easily calculated under this model for aggregate risks. The calculation of the risk measures can be extended to a model of the form:

$$S = S_1 + S_2 + \cdots + S_n$$

where n is a fixed integer and $S_j, j = 1, \cdots, n$, are not identical. Here, each risk S_j can be either individual or aggregate. Thus, there is a potential application of the mixture of Erlangs model in enterprise risk management (ERM) when company-wide risk measures are required. Another potential application is to price mortality-linked bonds and securities. Note that the mixture of Erlangs model can be used to model a projected mortality curve. A time-change technique will allow us to incorporate stochastic mortality and to produce a feasible stochastic mortality model for valuation purposes. See Lin and Liu (2007) and Liu (2007) for details.

Appendix Parameters for Fitted Distributions

i	r_i	α_i	i	r_i	α_i
1	2075	0.06310472	11	3019	0.04511672
2	2187	0.05125336	12	3111	0.04734103
3	2285	0.04710777	13	3211	0.05219257
4	2378	0.04573512	14	3319	0.05492296
5	2469	0.04496167	15	3430	0.05522278
6	2559	0.04487514	16	3540	0.05456390
7	2650	0.04589960	17	3651	0.05683452
8	2743	0.04649021	18	3773	0.06592862
9	2836	0.04608899	19	3922	0.08679154
10	2928	0.04556878			

Table 7: Estimated parameters for Uniform(1,2)

i	r_i	α_i	i	r_i	α_i
1	1	0.06310472	6	26	0.04511672
2	4	0.05125336	7	29	0.04734103
3	9	0.04710777	8	30	0.05219257
4	15	0.04573512	9	41	0.05492296
5	19	0.04496167	10	64	0.05522278

Table 8: Estimated parameters for Weibull(0.8,2)

i	r_i	α_i	i	r_i	α_i
1	4	0.05418686	6	25	0.04123584
2	9	0.17517153	7	30	0.03164086
3	14	0.08680983	8	33	0.11732483
4	16	0.22324938	9	44	0.02349588
5	23	0.24688501			

Table 9: Estimated parameters for Weibull(2,2)

i	r_i	α_i
1	1	0.795859740
2	6	0.136637782
3	17	0.042771535
4	37	0.016830619
5	70	0.007900324

Table 10: Estimated parameters for Pareto(1,1)

i	r_i	α_i	i	r_i	α_i
1	4	0.3143974132	10	91	0.0040564368
2	10	0.3217014101	11	99	0.0020862136
3	19	0.1746491672	12	103	0.0025648730
4	29	0.0734593526	13	124	0.0038877842
5	37	0.0363721817	14	156	0.0022681105
6	47	0.0311374107	15	198	0.0011988609
7	60	0.0192800280	16	255	0.0005988198
8	75	0.0095227295	17	370	0.0003883261
9	78	0.0024308822			

Table 11: Estimated parameters for Weibull(1,1)

References

Asmussen, S., O. Nerman and M. Olsson (1996). "Fitting phase-type distributions via the EM algorithm," *Scandinavian Journal of Statistics* 23(4), 419-441.

Chakravarti, I. M., R. G. Laha and J. Roy (1967). *Handbook of Methods of Applied Statistics,* Volume I, John Wiley and Sons, 392-394.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B (Methodological),* 39(1), 1-38.

Dufresne, D. (2007). "Fitting combinations of exponentials to probability distributions," *Applied Stochastic Models in Business and Industry,* 23(1), 23-48.

Dufresne, F. and H.U. Gerber (1991). “Risk theory for the compuund Poisson process that is perturbed by diffusion,” *Insurance: Mathematics and Economics,* 10, 51-59.

Gerber, H., M. Goovaerts and R. Kass (1987). “On the probability and severity of ruin,” *ASTIN Bulletin,* 17, 151-163.

Klugman, S., H.H. Panjer and G.E. Willmot (2004). *Loss Models: from Data to Decision,* second edition, Wiley and Sons, Hoboken, NJ.

Klugman, S. and R. Jacques (2006). “Toward a unified approach to fitting loss models,” *North American Actuarial Journal,* 10(1), 63-83.

Law, A.M. and D.W. Kelton (1995). “Simulation modeling and analysis,” *Interfaces,* 285(5), 214-215.

Lin, X.S. and X. Liu (2007). “Markov aging process and phase-type law of mortality,” *North American Actuarial Journal*, 11(4), 92-109.

Lin, X.S. and G.E. Willmot (1999). “Analysis of a defective renewal equation arising in ruin theory,” *Insurance: Mathematics and Economocis,* 25, 63-84.

Lin, X.S. and G.E. Willmot (2000). “The moments of the time of ruin, the surplus before ruin, and the deficit at ruin,” *Insurance: Mathematics and Economocis,* 27, 19-44.

Lin, X.S., G.E. Willmot and S. Drekic (2003). “The classical risk model with a constant dividend barrier: analysis of the Gerber-Shiu discounted penalty function,” *Insurance: Mathematics and Economocis,* 33, 551-566.

Liu, X. (2007). *Stochastic Mortality Modelling,* PhD Thesis, Department of Statistics, University of Toronto.

Rachid, El A-K., S. Ramin and B.R. Haverkort (2003). “Fitting World-Wide Web request traces with the EM-Algorithm,” *Performance Evaluation,* 52, 175-91.

Smith, M.R., S. Cohn-Sfetcu and H.A. Buckmaster (1967). *Decomposition of Multicomponent Exponential Decays by Spectral Analytic TechniquesDecomposition of Multicomponent Exponential Decays by Spectral Analytic Techniques.* Technometrics, volume 18, number 4, pp. 467-482

Thummler, A., P. Buchholz and M. Telek (2006). “A novel approach for phase-type fitting with the EM algorithm,” *IEEE Transactions on Dependable and Secure Computing,* 3(3), 245-58.

Tijms, H. (1994) *Stochastic Models: An Algorithm Approach.* John Wiley, Chiester.

Wang, JF, H.X. Zhou, M.T. Zhou and L. Li (2006). “A General model for long-tailed network traffic approximation,” *Journal of Supercomputing,* 38(2), 155-172.

Willmot, G.E. and X.S. Lin (2001). *Lundberg Approximations for Compound Distributions with Insurance Applications.* Springer, New York.