Insurance: Mathematics and Economics 51 (2012) 538–550

Contents lists available at SciVerse ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Modeling insurance claims via a mixture exponential model combined with peaks-over-threshold approach

David Lee*, Wai Keung Li, Tony Siu Tung Wong

The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received April 2012
Received in revised form
July 2012
Accepted 20 July 2012

JEL classification:
C12
C13
C51

Keywords:
Mixture exponential distribution
Extreme value theory
Threshold model
Mixture component testing
Insurance claims modeling

ABSTRACT

We consider a model which allows data-driven threshold selection in extreme value analysis. A mixture exponential distribution is employed as the thin-tailed distribution in view of the special structure of insurance claims, where individuals are often grouped into categories. An EM algorithm-based procedure is described in model fitting. We then demonstrate how a multi-level fitting procedure will substantially reduce computation time when the data set is large. The fitted model is applied to derive statistics such as return level and expected tail loss of the claim distribution, and ruin probability bounds are obtained. Finally we propose a statistical test to justify the choice of mixture exponential distribution over the homogeneous exponential distribution in terms of improved fit.

1. Introduction

The peaks-over-threshold (POT) method is frequently used to model extreme data. For a sufficiently high threshold u and under some regularity conditions, Pickands (1975) showed that, if the maxima of a continuous random variable X, suitably normalized, converges to a non-degenerate distribution, then the conditional distribution of X given $X > u$ converges asymptotically to the generalized Pareto distribution (GPD) with distribution function

$$G_u(x;\gamma,\sigma) = \begin{cases} 1-\left[1+\dfrac{\gamma(x-u)}{\sigma}\right]^{-1/\gamma}, & \gamma \neq 0, \\ 1-\exp\left\{-\dfrac{x-u}{\sigma}\right\}, & \gamma = 0 \end{cases}$$

with support (u,∞) if $\gamma \geq 0$ and $(u, u-\frac{\sigma}{\gamma})$ otherwise. Here γ is the shape parameter or extreme value index while $\sigma > 0$ is the scale parameter. Traditional POT model fitting involves selecting a suitable u, perhaps using the mean residual life plot which charts the mean excesses over different candidate values $\tilde{u}$ against $\tilde{u}$ themselves as illustrated by, for example, Coles (2001). Then G_u is fitted to the k exceedances over u. However, one drawback of this method is that there is no universal way of determining u, and in practice different choices of u will often result in very different parameter estimates and inferences. Wong and Li (2010) proposed a model in which the full data set is assumed to follow the distribution

$$F(x;\theta,\gamma,\sigma) = \begin{cases} P(x;\theta), & x \leq u, \\ P(u;\theta) + (1-P(u;\theta))G_u(x;\gamma,\sigma), & x > u \end{cases} \quad (1)$$

where $P(\cdot)$ is the distribution function specified by the user with parameters θ. A simple application of the model is to set P to be exponential, so that $P(x;\lambda) = 1-e^{-\lambda x}$ and the density function of (1) becomes

$$f(x;\lambda,\gamma,\sigma) = \begin{cases} \lambda e^{-\lambda x}, & x \leq u, \\ e^{-\lambda u} g_u(x;\gamma,\sigma), & x > u \end{cases} \quad (2)$$

where g_u is the density function of the GPD. In essence, the portion of observations below u is fitted with the exponential distribution and for the part above u we fit it with the GPD. For a random sample $\{x_1,\ldots,x_n\}$ with ordered values $x_{(1)} \leq \cdots \leq x_{(n)}$, finding the estimate of u, denoted by $\hat{u}$, can be done by the maximum likelihood (ML) method, where the likelihood function

$$L(\lambda,\gamma,\sigma;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\lambda,\gamma,\sigma)$$

* Correspondence to: Room 502, Meng Wah Complex, The University of Hong Kong, Pokfulam Road, Hong Kong. Tel.: +852 2859 2466; fax: +852 2858 9041.

E-mail addresses: dav001@gmail.com, davlee@hku.hk (D. Lee), hrntlwk@hku.hk (W.K. Li), wongtonyst@hku.hk (T.S.T. Wong).

0167-6687/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2012.07.008

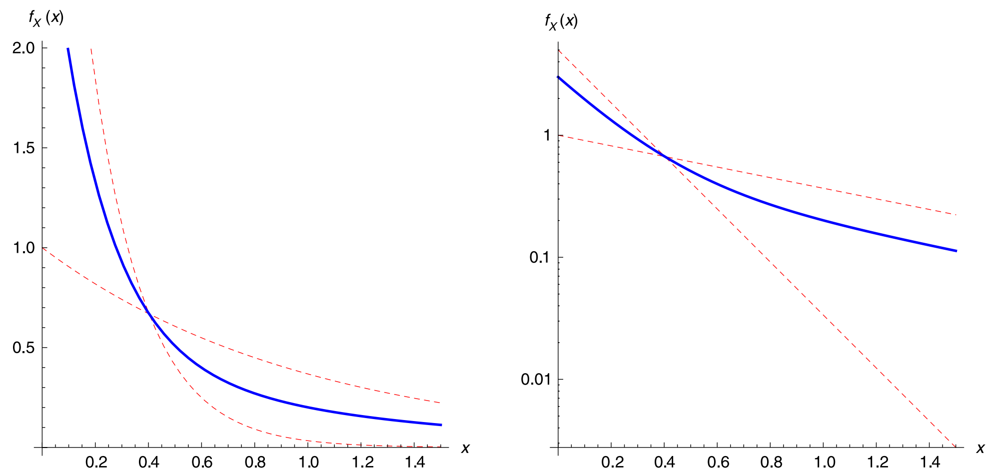

Fig. 1. Plots of exponential densities with rates 1 and 5 (Dashed) and mixture exponential density with component rates 1 and 5 (Solid). A log y-axis is used in the plot on the right.

or log-likelihood function $\log L$ is maximized for each value of $u = x_{(n-k)}$. In practice we can restrict the number of threshold exceedances k to be at most $\lfloor n/4 \rfloor$. The value $\hat{u}$ together with its corresponding parameters that give the largest maximized likelihood value will be chosen, and the ML estimate is given by $(\hat{u}, \hat{\lambda}, \hat{\gamma}, \hat{\sigma})$.

The main advantage of such modeling is that the estimation of u is data-driven. Wong and Li (2010) chose the Secura Belgian Re automobile claims data for demonstration, in which a satisfactory result was obtained when P takes the exponential form.

However, despite its elegance and simplicity, the exponential distribution does not offer much flexibility and its shape is rather limited. Consider the mixture exponential distribution as a generalization of the exponential distribution. A random variable Y has a mixture exponential distribution or hyperexponential distribution with m stages when $Y = X_i$ with probability p_i for $i = 1, \ldots, m$ such that the p_i's are positive and sum to one, and that each X_i follows an exponential distribution with rate parameter λ_i. The density function of Y is given by $f_Y(y) = \sum_{i=1}^{m} p_i \lambda_i e^{-\lambda_i y}$.

In the context of insurance, individuals are often grouped according to certain characteristics. For instance, in life insurance applicants are often classified according to their age and disease history, while in automobile insurance the classification criteria may be accident history. Such categorization is crucial as the claim distributions of individuals falling into various subgroups can be vastly different, which in turn affects pricing. If we assume that individual claim amounts in a particular group follow the exponential distribution, for example claim amounts from diseased and healthy individuals are exponentially distributed with respective means $1/\lambda_1$ and $1/\lambda_2$ with $\lambda_1 < \lambda_2$, and that each claim can be from group i with probability p_i, then we arrive at the mixture exponential distribution when we consider the overall claim amount distribution. Recently, Lee and Lin (2010) also employed mixture distributions in modeling which caters for the nature of insurance losses. It is often the case that we are interested in the overall claim distribution, and in particular we consider extremes that may affect the financial stability of the insurer regardless of which category such a massive claim comes from. Hence a GPD fitted to the tail will summarize such information in terms of tail properties expressed by the extreme value index γ.

Mixture exponential distribution is also widely used in basic ruin theory, in which we are interested in the probability of ruin $\psi(v)$ given initial surplus v in the surplus process

$$U(t) = v + ct - S(t), \quad t > 0$$

with aggregate claims up to time t, $S(t)$, being a compound Poisson process. When the claim amount is distributed as mixture exponential, there exists an explicit expression for $\psi(v)$ (Dufresne and Gerber, 1988; Gerber et al., 1987). Due to the introduction of the GPD component, there may not exist a closed form for $\psi(v)$, but we can still investigate numerically how this will change compared to the mixture exponential case.

Even when the claim distribution itself does not have a direct relation for subgroup partition, the mixture exponential distribution provides better control on the degree of curvature of the density function that will allow more flexible fitting. Fig. 1 shows the plots of two exponential densities and a mixture exponential density. The left panel shows both axes in the original scale while the right panel uses a logarithmic scale on the y-axis. It can be seen that each exponential density approximates a particular part of the mixture exponential distribution to some degree, but fails to model the rest of it. Meanwhile, the mixture exponential is a generalization of the exponential distribution, as we can simply set the mixture components to be the same in order to retrieve the latter. We will also see in Section 2 that, if we employ the EM algorithm in likelihood maximization, there exist closed-form solutions for the parameter estimates in each iteration for any proposed threshold, so that computational efficiency can be improved.

In this paper we attempt to use a mixture exponential distribution as the thin-tailed distribution P, and explore the flexibility it provides. Section 2 presents our model and fitting procedures using the EM algorithm. A simulation study is provided in Section 3, while in Section 4 two real data sets are fitted and further comments are given. Section 5 is devoted to a likelihood ratio test statistic based on which we test the justification of P being a mixture exponential against a homogeneous exponential distribution, where the test is also applied on the two real data sets. Concluding remarks are given in Section 6.

2. Model formulation and parameter estimation

In applying the mixture exponential distribution to $P(\cdot)$ in the threshold model (1), we consider the case where $m = 2$. The generalization is straightforward when $m > 2$, but as we will address in Section 4, a further increase in the number of components will not offer much improvement to justify the additional time spent in modeling. With this specification, the density function of the model is

$$f(x; p, \lambda, \gamma, \sigma) = \begin{cases} p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}, & x \le u, \\ \left(p e^{-\lambda_1 u} + (1-p) e^{-\lambda_2 u}\right) g_u(x; \gamma, \sigma), & x > u \end{cases} \quad (3)$$