An. Şt. Univ. Ovidius Constanţa Vol. 14(1), 2006, 99–108

A composite Exponential-Pareto distribution

Sandra Teodorescu and Raluca Vernic

Abstract

In this paper we introduce a composite Exponential-Pareto model, which equals an exponential density up to a certain threshold value, and a two parameter Pareto density for the rest of the model. Compared with the exponential, the resulting density has a similar shape and a larger tail. This is why we expect that such a model will be a better fit than the exponential one for some heavy tailed insurance claims data (e.g. with extreme values).

Subject Classification: 60E05, 62F10.

1 Introduction

It is known that usually, insurance claims have skewed and heavy tailed distributions. Therefore, researchers tend to use heavy tailed distributions to model these claims, like e.g. Gamma, LogNormal or Pareto. Unfortunately, such distributions often lead to complicate models, that need to be simplified in various ways. In theory, such models are often studied using the alternative Exponential distribution instead of a heavy tailed one, since this distribution has nice properties that make it very tractable.

Though not very realistic from a practical point of view, a model based on the exponential distribution can be of great importance to provide an insight into the phenomena, and also for pedagogical reasons. For example, it is known that analytical methods to compute ruin probabilities exist only for claims distributions that are mixtures and combinations of exponential distributions.

Key Words: Exponential and Pareto distributions; Composite Exponential-Pareto model; Parameter estimation.
Received: 24 February, 2006

Another nice property is that all gamma distributions with an integer scale parameter are limits of densities that are combinations of exponential distributions (see e.g. Kaas et al., 2001).

Also, when classifying the distributions with right-infinite domain into "light" and "heavy" tailed, the exponential distribution plays a central role: the distributions that are less spread out in the right tail than the exponential model are "light-tailed", while the others are "heavy-tailed".

This is why, based on the simplicity and nice properties of the exponential distribution, in this paper we suggest a composite Exponential-Pareto density, which equals an exponential density up to a certain threshold value, and a two parameter Pareto density for the rest of the model. The resulting density has a larger tail than the exponential one, as well as a smaller tail than the corresponding Pareto density; its density shape is similar to the exponential. The idea of such a composite model comes from Cooray and Ananda (2005).

The paper is structured as follows: in section 2 we present the derivation of the composite Exponential-Pareto model and illustrate its behavior, in section 3 we discuss the parameter estimation, and in section 4 we give numerical examples.

2 The composite Exponential-Pareto model

Let X be the random variable (r.v.) with density

$$f(x) = \begin{cases} cf_1(x) & if\ 0 < x \leq \theta \\ cf_2(x) & if\ \theta \leq x < \infty \end{cases}, \tag{1}$$

where f_1 is an exponential density, f_2 a two-parameter Pareto density, and c the normalizing constant. Hence,

$$\begin{aligned} f_1(x) &= \lambda e^{-\lambda x},\ x > 0 \\ f_2(x) &= \frac{\alpha\theta^\alpha}{x^{\alpha+1}},\ x > \theta, \end{aligned}$$

where $\lambda > 0, \alpha > 0, \theta > 0$ are unknown parameters.

In order to obtain a composite smooth density function, we impose continuity and differentiability conditions at the threshold point θ, i.e.

$$f_1(\theta) = f_2(\theta) \text{ and } f_1'(\theta) = f_2'(\theta),$$

where f' is the first derivative of f. These two restrictions give

$$\begin{cases} \lambda e^{-\lambda\theta} = \dfrac{\alpha}{\theta} \\ \lambda^2 e^{-\lambda\theta} = \dfrac{\alpha(\alpha+1)}{\theta^2} \end{cases},$$

and, after some calculation, we obtain

$$\begin{cases} \alpha = \lambda\theta - 1 \\ \lambda\theta\left(e^{-\lambda\theta} - 1\right) + 1 = 0 \end{cases}.$$

Solving the second equation by numerical methods, it results the solution

$$\begin{cases} \lambda\theta = 1.35 \\ \alpha = 0.35 \end{cases}.$$

Hence, the number of unknown parameters is reduced from 3 to 1.

In order to find the normalizing constant, we impose the condition $\int_0^\infty f(x)\,dx = 1$, which gives

$$c = \frac{1}{2 - e^{-\lambda\theta}} = 0.574.$$

So the composite density (1) becomes

$$f(x) = \begin{cases} \dfrac{0.775}{\theta} e^{-\dfrac{1.35\,x}{\theta}} & if\ 0 < x \leq \theta \\ \\ 0.2\,\dfrac{\theta^{0.35}}{x^{1.35}} & if\ \theta \leq x < \infty \end{cases}. \tag{2}$$

The cumulative distribution function (c.d.f.) of this composite model is

$$F(x) = \begin{cases} 0.574\left(1 - e^{-\dfrac{1.35\,x}{\theta}}\right) & if\ 0 < x \leq \theta \\ \\ 1 - 0.574\left(\dfrac{\theta}{x}\right)^{0.35} & if\ \theta \leq x < \infty \end{cases}. \tag{3}$$

A composite Exponential-Pareto density is illustrated in Figure 1 with solid line. This model is obtained by joining the exponential density in dashed blue line with the Pareto one in dotted red line, at point $\theta = 10$. It is easy to see that the composite density does not fade away to zero as quickly as the exponential one.

Figure 2 shows the variation of the composite Exponential-Pareto density with parameter θ. We can see that the tail becomes heavier as θ increases.

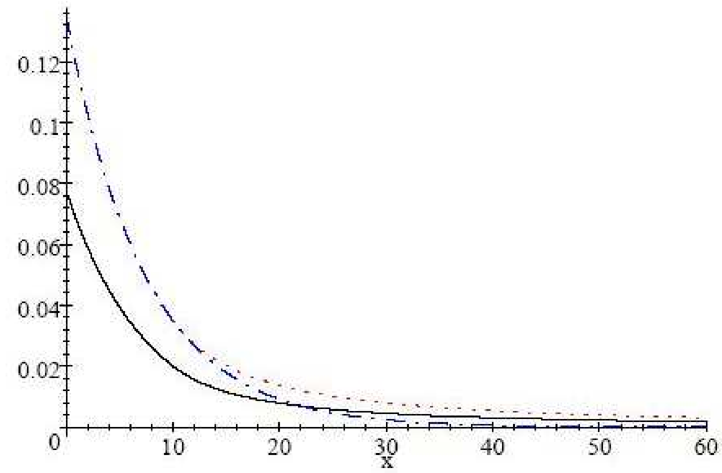

Figure 1: Exponential (dashed blue line), Pareto (dotted red line) and composite Exponential-Pareto (solid line) density curves for $\theta = 10$.

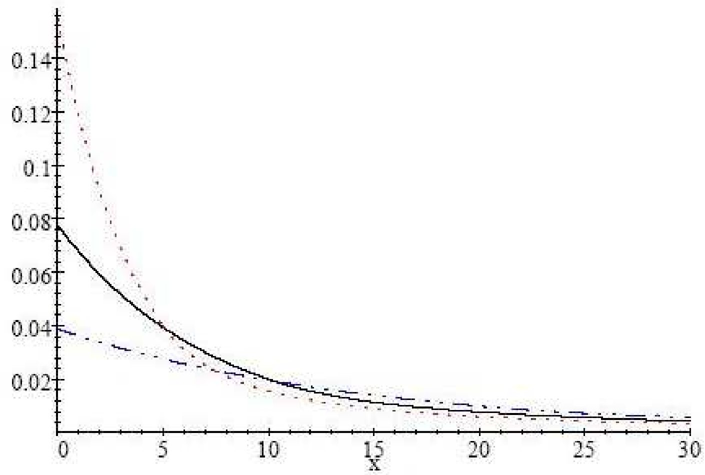

Figure 2: The composite Exponential-Pareto density curves for $\theta = 5$ (dotted red line), $\theta = 10$ (solid black line) and $\theta = 20$(dashed blue line).

3 Parameter estimation

In this section, we will present two methods for the estimation of the unknown parameter θ.

3.1 An ad-hoc procedure based on percentiles

The following *ad-hoc* procedure provides a closed form for the parameter θ, estimated using percentiles. To describe the procedure, let $x_1 \leq x_2 \leq ... \leq x_n$ be an ordered sample from the composite Exponential-Pareto model (2). We assume that the unknown parameter θ is in between the m^{th} observation and $m+1^{th}$ observation, i.e. $x_m \leq \theta \leq x_{m+1}$.

Based on percentiles, the parameter θ can be estimated as the p^{th} percentile, where $p = F(\theta)$. Here the distribution function F is given by (3), so that we have

$$p = 0.574 \left(1 - e^{-\dfrac{1.35\,\theta}{\theta}}\right) = 0.574 \left(1 - e^{-1.35}\right) \simeq 0.425.$$

From Klugman et al. (1998), we have a smooth empirical estimate of the p^{th} percentile given by

$$\tilde{\theta} = (1-h)\,x_m + hx_{m+1},$$

with

$$\begin{cases} m = [(n+1)\,p] \\ h = (n+1)\,p - m \end{cases} . \tag{4}$$

Here $[a]$ indicates the greatest integer smaller or equal with a.

Note that if $\tilde{\theta}$ is closer to x_1 or x_n, then Pareto or exponential will respectively be a superior model than the composite one.

3.2 Maximum likelihood estimation (MLE)

As before, let $x_1 \leq x_2 \leq ... \leq x_n$ be an ordered sample from the composite Exponential-Pareto model (2). In order to evaluate the likelihood function, we must have an idea of where is the unknown parameter θ situated correspondingly to this sample, so assume again that θ is in between the m^{th} observation and $m+1^{th}$ observation, i.e. $x_m \leq \theta \leq x_{m+1}$. Then the likelihood function is

$$\begin{aligned} L(x_1, ..., x_n; \theta) &= \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{m} f(x_i) \prod_{i=m+1}^{n} f(x_i) \overset{(2)}{=} \\ &= k\,\theta^{0.35n-1.35m}\, e^{-1.35\theta^{-1}\sum\limits_{i=1}^{m} x_i}, \end{aligned}$$

with $k = \dfrac{0.775^m 0.2^{n-m}}{\prod_{i=m+1}^{n} x_i^{1.35}}$. Denoting $\bar{x}_{(m)} = \frac{1}{m}\sum_{i=1}^{m} x_i$ the partial sample mean, and differentiating $\ln L$ with respect to θ gives

$$\frac{\partial \ln L}{\partial \theta} = \frac{0.35n - 1.35m}{\theta} + \frac{1.35m\bar{x}_{(m)}}{\theta^2}.$$

Hence the solution of the likelihood equation $\dfrac{\partial \ln L}{\partial \theta} = 0$ is

$$\hat{\theta} = \frac{1.35\, m\, \bar{x}_{(m)}}{1.35\, m - 0.35\, n}. \tag{5}$$

Since this estimator needs the value of m, we suggest the following algorithm:

Algorithm 1.

Step 1. Evaluate m as in previous section, from (4).

Step 2. Evaluate $\hat{\theta}$ from (5).

Step 3. Check if $\hat{\theta}$ is in between $x_m \leq \hat{\theta} \leq x_{m+1}$. If yes, then $\hat{\theta}$ is the maximum likelihood estimator. If no, then try algorithm 2.

An alternative algorithm would be to replace Step 1 with considering all possible values for m and performing for each one the checking in Step 3:

Algorithm 2.

Step 1. For each m $(m = 1, 2, ..., n-1)$, evaluate $\hat{\theta}_m$ from (5).
Check if $\hat{\theta}_m$ is in between $x_m \leq \hat{\theta} \leq x_{m+1}$. If yes, then $\hat{\theta}_m$ is the maximum likelihood estimator. If no, then go to next m.

Step 2. If there is no solution for θ, then try another model.

4 Numerical examples

In order to illustrate the procedures described in section 3, we will consider two data samples generated from the Exponential-Pareto model. The generating algorithm used is based on the inversion of the c.d.f. (3). When writing this algorithm, we took into account the fact that (3) has two different formulas, so that for a random number u one uses the first inverted formula if $u \leq F(\theta)$, and the second inverted formula if $u > F(\theta)$. The algorithm was written in Pascal and the data analysis was realized using Excel.

4.1 First example

The first data set consisting of 100 values was sampled from an Exponential-Pareto population with parameter $\theta = 5$ (see Table 1).

Table 1. 100 Exponential-Pareto values for $\theta = 5$

0.0151	0.0211	0.0721	0.0955	0.1730	0.3707	0.4240	0.4736	0.6076	0.6265
0.7335	0.7902	0.8178	0.9568	0.9993	1.3108	1.4076	1.4499	1.5756	1.6316
1.7033	1.7877	1.9217	1.9284	2.0464	2.1284	2.1509	2.3048	2.3785	2.5746
2.5750	3.0746	3.5561	4.0450	4.3008	4.3293	4.3664	4.5329	4.7059	5.4451
5.6778	5.8756	6.7573	6.9894	7.2925	7.8400	8.4130	8.5263	9.1961	9.5696
10.041	10.287	10.930	11.504	12.532	13.860	15.052	15.160	16.457	18.041
18.072	18.243	19.414	21.366	22.400	24.773	25.913	26.424	27.214	35.016
46.339	51.071	53.470	64.477	68.493	75.489	86.625	98.765	104.76	106.45
150.25	181.60	182.85	186.51	208.39	213.64	221.23	312.16	346.28	376.76
430.84	451.42	452.96	545.39	625.33	993.23	1170.0	3457.0	6842.0	7929.2

The estimated values of the parameter are:
- by the ad-hoc procedure based on percentiles: $\hat{\theta}_1 = 6.691$
- by MLE algorithm 1: $\hat{\theta}_2 = 5.472$
- by MLE algorithm 2: $\hat{\theta}_3 = 5.427$.

We notice that, as expected, algorithm 2 gives a more accurate value.

We also applied the χ^2 test to check the distribution fitting, and the results for $\hat{\theta}_3$ are given in Table 2. The χ^2 distances calculated for the three estimated values of the parameters are

$$\begin{aligned} d^2(\hat{\theta}_1) &= 10.25 \\ d^2(\hat{\theta}_2) &= 10.98 \\ d^2(\hat{\theta}_3) &= 11.05, \end{aligned}$$

which means that the χ^2 test accepts the Exponential-Pareto model for all three values of the parameter as expected. The interesting thing is that, unlike expected, $d^2(\hat{\theta}_1)$ is minimum, and a motive could be the errors due to the generating process.

Table 2. Grouped data and χ^2 test (columns 2 and 3 result form the data sample, $f_i = n_i/n$; column 4 is calculated using the Exponential-Pareto c.d.f.)

Classes	Frequencies, n_i	Relative freq., f_i	Theoretical freq., p_i	$\frac{n\,(f_i\text{-}p_i)^2}{p_i}$
[0, 1)	15	0.15	0.1263	0.4409
[1, 4)	18	0.18	0.2353	1.3019
[4, 8)	13	0.13	0.1371	0.0369
[8, 15)	10	0.10	0.0989	0.0010
[15, 30)	13	0.13	0.0866	2.1709
[30, 100)	9	0.09	0.1085	0.3154
[100, 300)	9	0.09	0.0660	0.8651
[300, 500)	6	0.06	0.0230	5.9089
[500, 7930)	7	0.07	0.0730	0.0128
$\sum$	n =100	1	χ^2 distance:	11.054

4.2 Second example

We also considered a set of $n = 500$ data sampled from the Exponential-Pareto model (2) with $\theta = 10$. For these data, the estimated values of the parameter are:
- by the ad-hoc procedure based on percentiles: $\hat{\theta}_1 = 8.22$
- by MLE algorithm 1: $\hat{\theta}_2 = 9.15$
- by MLE algorithm 2: $\hat{\theta}_3 = 9.10$.

We notice that this time, algorithm 1 gives a more accurate value.

Table 3. Grouped data and χ^2 test (the columns significance is the same as in Table 2)

Classes	Frequencies, n_i	Relative freq., f_i	Theoretical freq., p_i	$\frac{n\,(f_i\text{-}p_i)^2}{p_i}$
[0, 3)	93	0.186	0.2060	0.9766
[3, 6)	78	0.156	0.1320	2.1645
[6, 9)	48	0.096	0.0846	0.7582
[9, 14)	47	0.094	0.0833	0.6755
[14, 25)	33	0.066	0.0906	3.3621
[25, 40)	34	0.068	0.0611	0.3844
[40, 100)	54	0.108	0.0938	1.0716
[100, 300)	24	0.048	0.0792	6.1486
[300, 600)	16	0.032	0.0363	0.2647
[600, 1500)	14	0.028	0.0363	0.9617
[1500, 2500)	11	0.022	0.0157	1.2422
[2500, 10000)	15	0.030	0.0309	0.0136
[10000, 30000)	11	0.022	0.0158	1.2143
[30000, 10^5)	11	0.022	0.0115	4.6748
[10^5,21806000)	11	0.022	0.0187	0.2804
$\sum$	$n = 500$	1	χ^2 distance:	24.193

The χ^2 distances calculated for the three estimated values of the parameters are

$$\begin{aligned} d^2(\hat{\theta}_1) &= 26.43 \\ d^2(\hat{\theta}_2) &= 24.14 \\ d^2(\hat{\theta}_3) &= 24.19, \end{aligned}$$

and the χ^2 test accepts the Exponential-Pareto model for all three values of the parameter as expected. This time, as expected, since $\hat{\theta}_2$ is the best estimation, $d^2(\hat{\theta}_2)$ is minimum. It seems that the errors due to the generating process are smaller when the data volume is bigger.

In Table 3 one can see the data grouping and χ^2 test values for $\hat{\theta}_3$.

Acknowledgment. Raluca Vernic acknowledges the support of the research grant GAR 12/2005.

References

[1] Cooray, K. and Ananda, M. (2005) - Modeling actuarial data with a composite lognormal-Pareto model. Scandinavian Actuarial Journal 5, 321-334.

[2] Kaas, R.; Goovaerts, M.; Dhaene, J.; Denuit, M. (2001) - Modern Actuarial Risk Theory. Kluwer, Boston.

[3] Klugman, S.A., Panjer, H.H. and Willmot, G.E. (1998) - Loss Models: From Data to Decisions. Wiley, New York.

Sandra Teodorescu
Ecological University of Bucharest
Faculty of Mathematics and Computer Science
Romania
e-mail: tcalin@xnet.ro

Raluca Vernic "Ovidius" University of Constanta
Department of Mathematics and Informatics,
900527 Constanta, Bd. Mamaia 124
Romania
e-mail: rvernic@univ-ovidius.ro