ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 16, Number 1, 2012
Available online at www.math.ut.ee/acta/

On estimation of loss distributions and risk measures

Meelis Käärik and Anastassia Žegulova

Abstract. The estimation of certain loss distribution and analyzing its properties is a key issue in several finance mathematical and actuarial applications. It is common to apply the tools of extreme value theory and generalized Pareto distribution in problems related to heavy-tailed data.

Our main goal is to study third party liability claims data obtained from Estonian Traffic Insurance Fund (ETIF). The data is quite typical for insurance claims containing very many observations and being heavy-tailed. In our approach the fitting consists of two parts: for main part of the distribution we use lognormal fit (which was the most suitable based on our previous studies) and a generalized Pareto distribution is used for the tail. Main emphasis of the fitting techniques is on the proper threshold selection. We seek for stability of parameter estimates and study the behaviour of risk measures at a wide range of thresholds. Two related lemmas will be proved.

1. Introduction

The estimation of loss distributions has several practical and theoretical aspects, first of them being the choice of theoretical candidate distributions. There are few intuitive choices like lognormal, gamma, log-gamma, Weibull and Pareto distributions, but it is not rare that mentioned distributions do not fit very well. This work is a follow-up to our preliminary research (Käärik and Umbleja, 2010, 2011) where we established that lognormal distribution had best fit among the candidates. But the lognormal assumption is too strong and the tail behaviour needs to be revisited. Therefore, we focus on a model where the main part of the data follows a (truncated) lognormal distribution and the tail is fitted by generalized Pareto distribion (for brevity,

Received October 28, 2011.
2010 *Mathematics Subject Classification.* 91B30, 97M30, 62E20.
Key words and phrases. Insurance mathematics, extreme value theory, generalized Pareto distribution, composite distributions, risk measures.

the term composite lognormal/generalized Pareto distribution is also used). The choice of generalized Pareto distribution for tail fit is based on a well-known result from extreme values theory, the Pickands–Balkema–de Haan's theorem, which states that for a reasonably large class of distributions the conditional distribution of values exceeding certain threshold is close to a generalized Pareto distribution. The idea of using certain composite model is not new, there are several studies conducted in this field (see, e.g., Cooray and Ananda, 2005; Cooray, 2009; Pigeon and Denuit, 2010, Rooks, et al., 2010; Scollnik, 2007; Teodorescu and Vernic, 2009).

Our first task is to recall the relevant results from the theory of extreme values, certain properties of generalized Pareto distribution and the most common threshold selection methods. We will also provide an alternative threshold selection method, which is based on the risk measures, and therefore should be especially suitable for insurance data.

2. Preliminaries

In this section, we give a short overview about the required tools from the theory of extreme values. We refer to Beirlant et al. (2004), Coles (2001), Embrechts et al. (1997), McNeil et al. (2005) and McNeil (1999) for the results in the following subsections unless specifically stated otherwise.

2.1. Extreme value theory.

2.1.1. *Generalized Pareto distribution.*

Definition 2.1 (Generalized Pareto distribution)**.** Let us have a nonnegative random variable X with distribution function G. X is said to follow generalized Pareto distribution, $X \sim GPD(\sigma, \xi)$ if

$$G(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}, \quad y > 0,$$

where σ is a scale parameter and ξ is a shape parameter.

The shape parameter ξ determines the upper bound of the distribution: if $\xi < 0$ then upper bound exists and is equal to $u - \frac{\sigma}{\xi}$, if $\xi > 0$ then there is no upper bound. In case $\xi = 0$ there is also no upper bound, but it can be seen easily that the limit of the distribution function is $G(y) = 1 - \exp(-\frac{y}{\sigma})$, $y > 0$, i.e., the distribution function of an exponential distribution.

The expectation of the generalized Pareto distribution is given by

$$E(X) = \begin{cases} \frac{\sigma}{1-\xi} & \text{if } \xi < 1, \\ \infty & \text{if } \xi \geq 1. \end{cases} \tag{2.1}$$

The next definition is also required to build up the framework.

Definition 2.2 (Conditional tail distribution)**.** Let us have a nonnegative random variable X with distribution function F. Then for each threshold u the corresponding conditional tail distribution is defined by

$$F_u(y) = \mathbf{P}\{X - u \leq y | X > u\} = \frac{F(y+u) - F(u)}{1 - F(u)}, \tag{2.2}$$

with $0 \leq y < x_0$, where $x_0 \leq \infty$.

A useful property of generalized Pareto distribution is that for any thresholds u and u_0, $u > u_0$, the conditional tail distribution $F(u)$ for a generalized Pareto distribution can be calculated as

$$F_u(y) = 1 - \left(1 + \frac{\xi_{u_0} y}{\sigma_{u_0} + \xi_{u_0} u}\right)^{-\frac{1}{\xi_{u_0}}}. \tag{2.3}$$

The outcome is again a generalized Pareto distribution with parameters ξ_{u_0} and $\sigma_u = \sigma_{u_0} + \xi_{u_0} u$, which means that the shape parameter does not depend on the threshold u and the scale parameter depends linearly from threshold u. This result is useful for finding a suitable threshold point later on.

2.1.2. *Pickands–Balkema–de Haan's theorem.* Our main motivation to use a generalized Pareto distribution for fitting the tail is explained by the following theorem.

Theorem 2.1 (Pickands–Balkema–de Haan)**.** *For a sufficiently large class of distributions there exists a function $\sigma(u)$ such that the following equation*

$$\lim_{u \to x_0} \sup_{0 \leq y < x_0 - u} |F_u(y) - G_{\xi, \sigma(u)}(y)| = 0$$

holds.

The exact class of distributions on which this theorem can be applied is not of interest, but it is important to note that most distributions used in actuarial models belong to this class (see, e.g., McNeil, 1999, Embrechts et al., 1997).

An important step from practical perspective is the estimation of the parameters of the distribution, which can be done, e.g., using the method of maximum likelihood. As it is not possible to maximize the likelihood analytically, various numerical methods can be applied.

2.2. Composite models.

2.2.1. *Composite lognormal/Pareto distribution.* In our article, the main emphasis is on combining lognormal and generalized Pareto distribution. A particularly interesting research in this area is done by Cooray and Ananda (2005), who combined lognormal and Pareto distributions with certain differentiability and continuity restrictions at the threshold point. We now briefly recall this setup and reveal its main strengths and weaknesses.

Let X be a random variable with the probability density function

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

where c is the normalizing constant, $f_1(x)$ has the form of the two-parameter lognormal density, and the $f_2(x)$ has the form of the two-parameter Pareto density, i.e.

$$f_1(x)=\frac{(2\pi)^{-\frac{1}{2}}}{x\sigma}\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right], \quad x>0,$$

$$f_2(x)=\frac{\alpha\theta^2}{x^{\alpha+1}}, \quad x\geq\theta,$$

where $\theta>0, \mu\in\mathbf{R}, \sigma>0$ and $\alpha>0$ are unknown parameters. Thus we can say X follows a four-parameter composite lognormal/Pareto distribution, $X\sim LNP(\theta,\mu,\sigma,\alpha)$. Also, to get a smooth probability density function, the following continuity and differentiability conditions need to be fulfilled:

$$f_1(\theta)=f_2(\theta), \qquad f_1^{'}(\theta)=f_2^{'}(\theta). \tag{2.4}$$

Conditions (2.4) imply that $\ln\theta-\mu=\alpha\sigma^2$ and $\exp(-\alpha^2\sigma^2)=2\pi\alpha^2\sigma^2$. This leads to

$$\int_0^\theta f_1(x)dx=\Phi(\alpha\sigma) \quad \text{and} \quad c=\frac{1}{1+\Phi(\alpha\sigma)},$$

finally resulting that $\alpha\sigma$ and c are constants. Thus $\alpha\sigma$ and c do not depend on the values of the parameters, $\alpha\sigma\approx 0.372$ and $c\approx 0.608$. See Cooray and Ananda (2005) for more details.

The importance of the result is that it allows to reparametrize the distribution and reduce the number of parameters from four to two. But it also fixes the proportions of lognormal and Pareto parts (approximately 0.392 and 0.608, respectively) regardless of the values of the mixture parameters. This simplification makes the construction very appealing in case the fixed proportions are realistic for given problem. The downside is that this cannot be always assured (in our example in Section 4 all thresholds of interest were greater than the 0.9-quantile of the lognormal distribution), and either regular lognormal or Pareto distribution or a mixture with different proportions can yield better results. The shortcomings and possible extensions of this

approach are addressed in (Scollnik, 2007) and (Pigeon and Denuit, 2010). As the threshold of this model does not suit our data, we focus on a different model, specified in the following subsection.

2.2.2. *Composite lognormal/generalized Pareto distribution.* Our research is motivated by the Pickands–Balkema–de Haan's theorem and thus we choose the model where the main part of the distribution is lognormal and after certain threshold u it is truncated and tail is substituted with generalized Pareto distribution, resulting in certain composite lognormal/generalized Pareto model.

By Theorem 2.1, the conditional tail distribution F_u has the following form:

$$F_u(y) = G_{\xi,\sigma(u)}(y),$$

where $\sigma(u) = \sigma + \xi u$. Defining $x := u + y$ and using the last result together with (2.2) implies

$$F_u(x-u) = \frac{F(x) - F(u)}{1 - F(u)} = G_{\xi,\sigma(u)}(x-u),$$

which leads us to

$$F(x) = (1 - F(u))G_{\xi,\sigma(u)}(x-u) + F(u) = 1 - (1 - F(u))\left(1 + \xi\frac{x-u}{\sigma(u)}\right)^{-\frac{1}{\xi}}, \tag{2.5}$$

where $x > u$.

Note that this is a general result, there are no additional assumptions made for the main part distribution (i.e., for the value of $F(u)$). If we assume now that up to threshold u the distribution is lognormal, say, with parameters μ_l and σ_l, then formula (2.5) modifies to

$$F(x) = 1 - \left(1 - \Phi\left(\frac{\ln u - \mu_l}{\sigma_l}\right)\right)\left(1 + \xi\frac{x-u}{\sigma + \xi u}\right)^{-\frac{1}{\xi}}. \tag{2.6}$$

We denote the corresponding distribution by $LNGP(u, \mu_l, \sigma_l, \xi, \sigma)$, i.e., if a random variable X has distribution function (2.6), we write $X \sim LNGP(u, \mu_l, \sigma_l, \xi, \sigma)$.

2.3. Threshold selection techniques. Since we only want to fit the (conditional) tail by generalized Pareto distribution, the most important thing is to choose the right threshold, for our particular case data below threshold is fitted by a lognormal distribution.

Also, although all the techniques rely on mathematical tools, there is always a subjectivity factor involved. Therefore, we test different methods and choose threshold which seems acceptable by all methods. More details can be found, e.g., in Coles (2001), Čižek et al. (2005) and Ribatet (2006).

2.3.1. *Mean excess function.* In this section, we recall the definition of mean excess function and provide some of its relevant basic properties.

Definition 2.3. For any random variable X, the mean excess function $e(x)$ is defined by

$$e(x) = E(X - x|X > x).$$

If we apply this result to the generalized Pareto distribution, then by equation (2.1) we can write the conditional expectation of values exceeding threshold u_0 as

$$e(u_0) = E(X - u_0|X > u_0) = \frac{\sigma_{u_0}}{1-\xi},$$

where $\xi < 1$ and σ_{u_0} is a scale parameter corresponding to values exceeding threshold u_0. The last result together with equation (2.3) yields that for all $u > u_0$ we have

$$e(u) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi}. \tag{2.7}$$

This means the mean excess function of a GPD-distributed random variable is linear. Similarly, the empirical mean excess function is calculated as

$$e_n(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u),$$

where $x_{(1)}, \ldots, x_{(n_u)}$ are the n_u observations exceeding u. The threshold selection based on empirical mean excess function consists in analyzing the empirical function $e_n(u)$ and determining the value u_0 starting from which this function stays linear.

2.3.2. *Threshold choice plot.* From formula (2.3), we know that if the values exceeding threshold u_0 follow a generalized Pareto distribution, then the same holds for any higher thresholds u, $u > u_0$. Moreover, the shape parameters are equal for all u and the scale parameter is a linear function of u. By a simple reparametrization of the scale parameter we obtain the so-called modified scale parameter:

$$\sigma^* = \sigma_u - \xi u,$$

which does not depend on threshold u anymore. In summary, we have reparametrized the distribution in such way that for all thresholds $u > u_0$ the parameters of the distributions remain constant, providing us another tool for selecting a proper threshold. In threshold choice plot the maximum likelihood estimates for the shape parameter ξ and the modified scale parameter σ^* are plotted against the thresholds.

In practical situations, we cannot expect that the fitted parameters remain constant, because they are estimated from a sample. Nevertheless, we could also estimate the corresponding confidence intervals and choose the threshold from where the confidence intervals remain constants (or close to constants).

3. Risk measures: value at risk and expected shortfall

3.1. Definitions. Since we are only dealing with continuous and strictly increasing distributions there exists an inverse F^{-1} of distribution function F. So, the value at risk can be simply defined as q-quantile of corresponding distribution as follows (see, e.g., Artzner et al., 1999; Kaas et al., 2008).

Definition 3.1. Value at risk (VaR) for random variable X (with continuous and strictly increasing distribution function F) at given confidence level $q \in (0,1)$ is defined as

$$VaR(q) = F^{-1}(q). \tag{3.8}$$

Definition 3.2 (Expected shortfall)**.** Let us have a random variable X with distribution function F and with $E(|X|) < \infty$. The expected shorfall of X at confidence level $q \in (0,1)$ is defined as

$$ES(q) = \frac{1}{1-q}\int_q^1 F^{-1}(l)dl.$$

We can also derive a simple but useful formula describing the connection between expected shortfall and value at risk:

$$ES(q) = E(X|X > VaR(q)) \tag{3.9}$$

or, equivalently,

$$ES(q) = VaR(q) + E(X - VaR(q)|X > VaR(q)) = VaR(q) + e(VaR(q)). \tag{3.10}$$

If a theoretical distribution fits the data, then the values of risk measures based on empirical and theoretical distributions should be close as well. This argumentation motivates us to formulate another method for threshold selection: if the values of risk measures for the theoretical distribution (in our example lognormal) at some point are too different from the corresponding values from data, then this point should be chosen as threshold, and the tail part will be substituted with generalized Pareto distribution.

More formally, from our data (or corresponding empirical distribution) we can always find estimates for value at risk and expected shortfall (denote them by $\widehat{VaR}_{emp}$ and $\widehat{ES}_{emp}$) for any confidence level q and compare these values with corresponding values of proposed theoretical distributions $\widehat{VaR}_{th}$ and $\widehat{ES}_{th}$. From the insurance perspective, the theoretical values should not be too optimistic compared to the empirical ones.

In the following subsections we study more closely the calculation of VaR and ES with distributions of our special interest: lognormal distribution and composite lognormal/generalized Pareto distribution.

3.2. Estimation for lognormal distribution. Suppose now that X is a lognormally distributed random variable and study the behaviour of VaR and ES in this case.

Lemma 3.1. *Value at risk and expected shortfall for a lognormally distributed random variable $X \sim LN(\mu_l, \sigma_l)$ have the following forms:*

$$VaR(q) = \exp(\mu_l + \sigma_l \Phi^{-1}(q)) \tag{3.11}$$

and

$$ES(q) = \frac{e^{\mu_l + \frac{\sigma_l^2}{2}}}{1-q}(1 - \Phi(\Phi^{-1}(q) - \sigma_l)), \tag{3.12}$$

where $\Phi^{-1}(q)$ is the q-quantile of the standard normal distribution

Proof. The results follow from the definition of lognormal distribution and from the fact that lognormal distribution is strictly increasing, which allows us to use formula (3.8) to calculate the value at risk. The calculation of expected shortfall is straightforward using formula (3.9). Details are omitted.

3.3. Estimation for composite lognormal/generalized Pareto distribution. Now, we turn our attention to the composite model defined in (2.5) and, more precisely, to the composite lognormal/generalized Pareto distribution. Let us note that the calculation of $F(x)$ in (2.5) requires besides estimating the parameters ξ and σ of generalized Pareto distribution also the estimation of $F(u)$. There are two main approaches for the estimation of $F(u)$. Empirical method uses empirical estimate $(n - N_u)/n$, where n is the number of observations and N_u is the number of observations exceeding threshold u. The other idea is to use the value of proposed theoretical distribution (in our case lognormal, which gives us the $LNGP$-model) at u as estimate.

Substituting ξ and σ with estimates $\hat{\xi}$ and $\hat{\sigma}$ and also $x = \widehat{VaR}(q)$ and $\hat{F}(x) = q$ into equation (2.5) we get the following formulas for estimating the value at risk:

a) the semi-parametric estimate using the empirical method (also called historical simulation method) (see, e.g., McNeil, 1999)

$$\widehat{VaR}(q) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{n}{N_u}(1-q)\right)^{-\hat{\xi}} - 1\right); \tag{3.13}$$

b) the parametric estimate with the value of $F(u)$ from proposed theoretical distribution

$$\widehat{VaR}(q) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{1-q}{1-F(u)}\right)^{-\hat{\xi}} - 1\right). \tag{3.14}$$

In case of our special interest, i.e., if the observable variable follows $LNGP$-distribution, the proposed theoretical distribution in (3.14) is lognormal. Thus the value of $F(u)$ in Equation (3.14) is calculated as $F(u) = \Phi(\frac{\ln u - \hat{\mu}_l}{\hat{\sigma}_l})$, where $\hat{\mu}_l$ and $\hat{\sigma}_l$ are the (maximum likelihood) estimates for parameters of the fitted lognormal distribution.

By construction of the composite distribution, equations (3.13) and (3.14) hold only for $q > F(u)$. When $q \leq F(u)$, the estimate of $VaR(q)$ equals to the q-quantile of non-truncated distribution, calculated by (3.11).

Let us now find the formula for expected shortfall, assuming that after certain threshold the tail follows generalized Pareto distribution. Then, from Theorem 2.1 and formula (2.3), it follows that

$$(X - VaR(q)|X > VaR(q)) \sim GPD(\xi, \sigma + \xi(VaR(q) - u)).$$

Now assuming $\xi < 1$, we apply the result about the expectation of generalized Pareto distribution (2.1) to formula (3.10). We get

$$\begin{aligned} ES(q) &= VaR(q) + E(X - VaR(q)|X > VaR(q)) \\ &= VaR(q) + \frac{\sigma + \xi(VaR(q) - u)}{1 - \xi} = \frac{VaR(q)}{1 - \xi} + \frac{\sigma - \xi u}{1 - \xi} \end{aligned} \tag{3.15}$$

and similarly for the estimates

$$\widehat{ES}(q) = \frac{\widehat{VaR}(q)}{1 - \hat{\xi}} + \frac{\hat{\sigma} - \hat{\xi} u}{1 - \hat{\xi}}, \tag{3.16}$$

where $\hat{\xi}$, $\hat{\sigma}$ and $\widehat{VaR}(q)$ are estimates for GPD parameters and for the value at risk, respectively (see also McNeil, 1999). Depending on the calculation of $\widehat{VaR}(q)$ (see formulas (3.13) and (3.14)), formula (3.16) may give us parametric or semi-parametric estimate for $ES(q)$.

Similarly to the formulas for value at risk, the formulas (3.15) and (3.16) only hold for $q > F(u)$. When $q \leq F(u)$, the estimate of $VaR(q)$ equals to the q-quantile of lognormal distribution, the calculation of expected shortfall is addressed in the next section.

3.4. Expected shortfall for composite lognormal/generalized Pareto distribution when $q \leq F(u)$. As already mentioned, by the construction of the composite distribution, the general formulas for expected shortfall (3.15) and (3.16) are only applicable for $q > F(u)$. On the other hand, in many situations it is important to know the value of expected shortfall for lower confidence levels as well. Let us study this situation more thoroughly.

Lemma 3.2. *Consider the composite lognormal/generalized Pareto distribution (up to threshold u it is lognormal, and conditionally generalized Pareto after). Let μ_l and σ_l be the parameters of the lognormal distribution and let*

ξ and σ_u be the parameters for generalized Pareto distribution. Then the corresponding expected shortfall $ES(q)$ for $q \le F(u)$ can be calculated as

$$ES(q) = \frac{1}{1-q}\left(e^{\mu_l+\frac{\sigma_l^2}{2}}\left(\Phi\left(\frac{\ln u - \mu_l - \sigma_l^2}{\sigma_l}\right)\right.\right. \tag{3.17}$$
$$\left.\left. -\Phi\left(\frac{\ln(VaR(q)) - \mu_l - \sigma_l^2}{\sigma_l}\right)\right)\right)$$
$$+\frac{1}{1-q}\left(\left(1-\Phi\left(\frac{\ln u - \mu_l}{\sigma_l}\right)\right)\left(u+\frac{\sigma_u}{1-\xi}\right)\right). \tag{3.18}$$

Proof. By construction, the expected shortfall can be written as

$$ES(q) = \frac{1}{\mathbf{P}\{X > VaR(q)\}}\left(\int_{VaR(q)}^{u} x f_l(x)dx + \mathbf{P}\{X \ge u\}E(X|X \ge u)\right),$$

where $f_l(\cdot)$ is the probability density function of lognormal distribution, i.e. $f_l(x) = \frac{1}{x}\phi(\frac{\ln x - \mu_l}{\sigma_l})$ with $\phi(\cdot)$ being the standard normal probability density function.

Now the first (integral) term can be simplified using the properties of lognormal distribution together with equation (3.12) and to the second term we can apply formula (2.7). This implies (3.17), the lemma is proved. □

A result similar to (3.17) can be obtained for the estimate $\widehat{ES}(q)$ as well, one can simply substitute the estimates for parameter values and $VaR(q)$ into (3.17). As for value at risk, it is also possible to use the empirical estimator of the value at risk when estimating the expected shortfall. In that case

$$\widehat{ES}(q) = \frac{1}{\mathbf{P}\{X > \widehat{VaR}(q)\}}\Big(\mathbf{P}\{\widehat{VaR}(q) < X < u\}\bar{x}_*$$
$$+\mathbf{P}\{X \ge u\}E(X|X \ge u)\Big),$$

where $\bar{x}_*$ is the arithmetic mean over the sample values which are greater than $\widehat{VaR}(q)$ but less than u. The weight probabilities can be found from

$$\mathbf{P}\{\widehat{VaR}(q) < X < u\} = \frac{\#\{x|\widehat{VaR}(q) < x < u\}}{n} =: w_1$$

and

$$\mathbf{P}\{X \ge u\} = \frac{\#\{x|x \ge u\}}{n} =: w_2,$$

resulting in the following final formula for estimation of expected shortfall (with $F(u)$ estimated by empirical method):

$$\widehat{ES}(q) = \frac{1}{1-q}\left(w_1 \cdot \bar{x}_* + w_2\left(u + \frac{\hat{\sigma}_u}{1-\hat{\xi}}\right)\right). \tag{3.19}$$

4. Case study: Estonian traffic insurance

4.1. Description of data. The data is provided by Estonian Traffic Insurance Fund (ETIF) and contains Estonian third party liability insurance claims from 01.07.06 - 30.06.07. There are 39 306 claims in total, with average claim being 22 450 EEK (Estonian kroons) and most frequent claims are between 5000 and 15 000 EEK. Short summary of characteristics for claim severity is provided in Table 1.

Min.	1st Quartile	Median	Mean	3rd Quartile	Max.
80	6 725	11 800	22 450	22 090	5 258 000

TABLE 1. Descriptive statistics for claim severity (in EEK)

There are also 8 different types of vehicles, with cars being prevalent (77.1%), followed by small trucks, trucks, buses, etc.

4.2. Results. According to our preliminary research (Käärik and Umbleja, 2011), where several classical distributions (Weibull, gamma, beta, lognormal, Pareto) were used to fit the given data, lognormal distribution with parameters $\mu_l = 9.4$ and $\sigma_l = 1.1$ had the best fit. It is important to remember that nevertheless both Kolmogorov-Smirnov and χ^2-test rejected all distributions and the tail behaviour was clearly too optimistic.

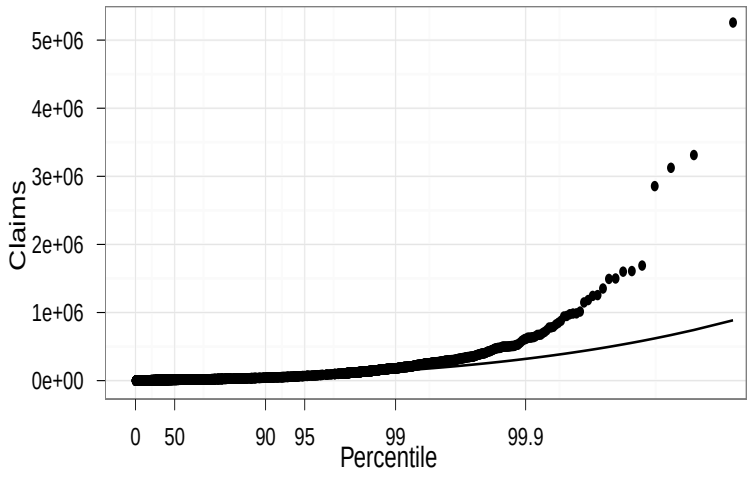

FIGURE 1. Probability plot for lognormal distribution

On Figure 1, we observe that the q-quantile of the fitted lognormal distribution underestimates the observed one, starting from $q \geq 0.99$.

The problem of data having a heavy tail is actually quite widespread problem in insurance field and thus one of the main motivators of this study was

to modify the tail estimate and to obtain a more conservative results. We will use the composite lognormal/generalized Pareto model described before, and search for a proper threshold that divides the main and tail part of the data. This will be done by applying the results and methods from Sections 2 and 3 to given data.

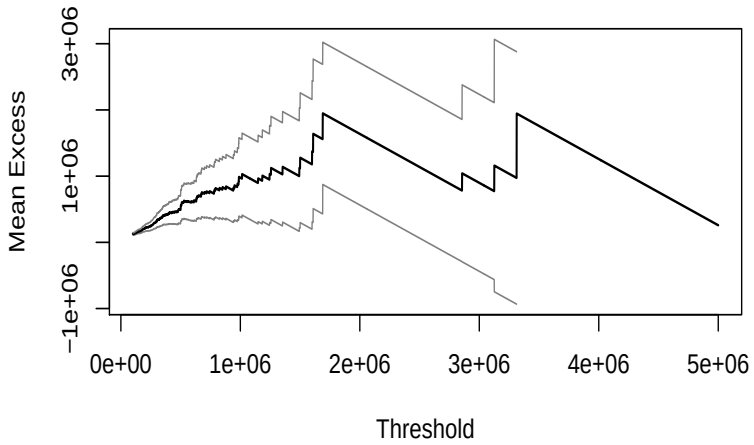

FIGURE 2. Empirical mean excess function

Based on the behaviour of the empirical mean excess function (Figure 2) and the parameter estimates (Figure 3), a suitable threshold would be $u = 500\ 000$, while the quantile fitting of risk measures (by formulas (3.13)–(3.19)) also proposes 0.98-quantile $u_* = 121\ 729$ as a possible candidate.

q	0.8	0.9	0.95	0.98	0.99	0.999
$\widehat{VaR}_{emp}(q)$	26 847	42 975	68 607	121 730	180 959	606 138
$\widehat{VaR}_{LN}(q)$	29 694	47 452	69 810	107 870	144 175	325 047
$\widehat{VaR}_{LN}^{121\ 729}(q)$	29 694	47 452	69 810	107 870	182 165	643 763

TABLE 2. VaRs for candidate distributions on different confidence levels q

The relevant values for value at risk and expected shortfall are given in the Tables 2 and 3, where $\widehat{VaR}_{emp}$, $\widehat{VaR}_{LN}$ and $\widehat{VaR}_{LN}^{121\ 729}$ denote VaR of empirical distribution, VaR of lognormal distribution and VaR of $LNGP$-distribution with threshold 121 729, respectively. Wa also note that the value at risk for composite $LNGP$-distribution with threshold 500 000 is not

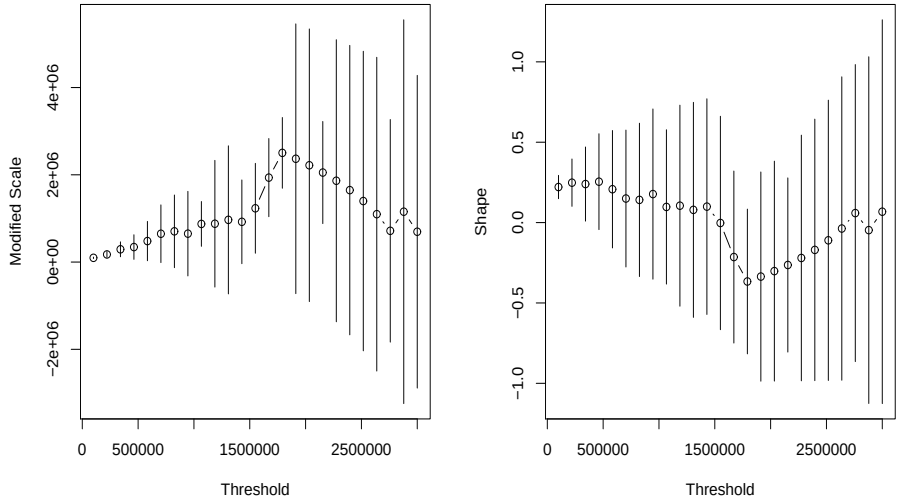

FIGURE 3. Maximum likelihood estimates for parameterers of generalized Pareto distribution

present in the Table 2 as it equals to $\widehat{VaR}_{LN}(q)$ for all values of q (since the threshold 500 000 exceeds all quantiles of lognormal distribution given in Table 2).

The notation for expected shortfall is similar, with $\widehat{ES}_{emp}^{500\ 000}$ being expected shortfall for distribution with empirical estimate for main part and conditional generalized Pareto tail from threshold 500 000.

q	0.8	0.9	0.95	0.98	0.99	0.999
$\widehat{ES}_{emp}(q)$	69 600	105 564	158 195	261 627	375 106	1 215 392
$\widehat{ES}_{LN}(q)$	62 810	88 492	119 939	172 172	220 982	456 734
$\widehat{ES}_{emp}^{500\ 000}(q)$	67 335	101 124	149 136	238 997	329 903	1 417 023
$\widehat{ES}_{LN}^{121\ 729}(q)$	76 409	115 627	174 335	308 161	379 574	973 689

TABLE 3. Expected shortfalls for candidate distributions on different confidence levels q

It can be seen that the best performing distribution overall is the *LNGP*-distribution with threshold 121 729, but for especially conservative results on high quantiles, the estimates with threshold 500 000 and empirical method may be useful.

Most of the calculations are done using R statistical software (R Development Core Team, 2012) package *actuar* (Dutang et al., 2008), figures are created with R statistical software package *POT* (Ribatet, 2006).

4.3. Conclusions. The following findings can be pointed out. The proposed idea of threshold selection using the values of risk measures provides valuable information from a different viewpoint than the classical methods. The difference between proposed candidate thresholds is large, confirming the fact that the threshold selection is still a very subjective task. The best candidate distribution for a given data set is the composite lognormal/generalized Pareto distribution $LNGP(121\ 729, 9.4, 1.1, 0.22, 1.4 \cdot 10^5)$, i.e., a lognormal distribution with parameters $\mu_l = 9.4$ and $\sigma_l = 1.1$ for the main part and generalized Pareto with parameters $\sigma = 1.4 \cdot 10^5$ and $\xi = 0.22$ for the (conditional) tail part, with threshold $u_* = 121\ 729$ dividing the main and tail parts. For especially conservative estimates for high quantiles ($q \geq 0.999$) one can use the estimates obtained by empirical method with threshold $u = 500\ 000$.

Acknowledgements

The work is supported by Estonian Science Foundation Grant 7313 and by Targeted Financing Project SF0180015s12. The authors also thank the two anonymous referees for their helpful comments and suggestions that certainly improved the quality of the paper.

References

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). *Coherent measures of risk*, Math. Finance **9**(3), 203–228.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*, Wiley & Sons, Chichester.

Čižek, P., Härdle, W., and Weron, R. (2005), *Statistical Tools for Finance and Insurance*, Springer, Berlin – Heidelberg.

Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, Springer, London.

Cooray, K. (2009), *The Weibull-Pareto composite family with applications to the analysis of unimodal failure rate data*, Comm. Statist. Theory Methods **38**, 1901–1915.

Cooray, K., and Ananda, M. (2005), *Modeling actuarial data with a composite lognormal-Pareto model*, Scand. Actuar. J. **5**, 321–334.

Dutang, C., Goulet, V., and Pigeon, M. (2008), *actuar: An R package for actuarial science*, J. Statist. Software **25**(7), 1–37.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer, New York – Berlin – Heidelberg – Tokyo.

Käärik, M., and Umbleja, M. (2010), *Estimation of claim size distributions in Estonian traffic insurance*; In: Selected Topics in Applied Computing. Proceedings of Applied Computing Conference (ACC '10), Timisoara, Romania, pp. 28–32.

Käärik, M., and Umbleja, M. (2011), *On claim size fitting and rough estimation of risk premiums based on Estonian traffic insurance example*, Internat. J. Math. Models Methods Appl. Sci. **5**(1), 17–24.

Kaas, R., Goovaerts, M., Dhaene, J., and Denuit, M. (2008), *Modern Actuarial Risk Theory Using R*. Springer, Heidelberg.

McNeil, A. (1999), *Extreme value theory for risk managers*; In: Internal Modelling and CAD II, London, pp. 93–113.
Available: `http://riskbooks.com/internal-modelling-and-cad-ii`

McNeil, A., Frey, R., and Embrechts, P. (2005), *Quantitative Risk Management: Concepts, Techniques, and Tools*, Princeton University Press, Princeton.

Pigeon, M., and Denuit, M. (2010), *Composite lognormal-Pareto model with random threshold*, Scand. Actuar. J. **10**, 49–64.

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available: `http://www.R-project.org`

Ribatet, M. (2006), *A User's Guide to the POT Package (Version 1.4)*, University of Montpellier II, France.

Rooks, B., Schumacher, A., and Cooray, K. (2010), *The power Cauchy distribution: derivation, description, and composite models*, NSF-REU Program Reports. Available: `http://www.cst.cmich.edu/mathematics/research/REU_and_LURE.shtml`

Scollnik, D. P. M. (2007), *On composite lognormal-Pareto models*, Scand. Actuar. J. **7**, 20–33.

Teodorescu, S., and Vernic, R. (2009), *Some composite exponential-Pareto models for actuarial prediction*, Romanian J. Econom. Forecasting **12**, 82–100.

Institute of Mathematical Statistics, University of Tartu, Tartu, Estonia
E-mail address: `meelis.kaarik@ut.ee`
E-mail address: `anastassia.zhegulova@ut.ee`