Analysis of truncated data with application to the operational risk estimation

Petr Volf[1]

Abstract. Researchers interested in the estimation of operational risk often face problems arising from the structure of available data. The present contribution deals with the problem of left truncation, which means that the values (e.g. the losses) under certain threshold are not reported. Simultaneously, we have to take into account possible occurrence of heavy-tailed distribution of loss values. We recall briefly the methods of incomplete data analysis, then we concentrate to the case of fixed left truncation and parametric models of distribution. The Cramér-von Mises, Anderson-Darling, and the Kolmogorov-Smirnov minimum distance estimators, the maximum likelihood, and the moment estimators are used, their performance is compared, with the aid of randomly generated examples covering also the case of heavy-tailed distribution. Higher robustness of some distance-based estimators is demonstrated. The main objective is to propose a method of statistical analysis and modeling for the distribution of sum of losses over a given period, particularly of its right quantiles.

Keywords: operational risk, severity distribution, truncated data, statistical analysis.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 Introduction, the problem of incomplete data

The most traditional field of statistical analysis where the methodology dealing with incomplete data (caused by censoring or truncation) has been developed systematically is the area of statistical survival analysis. While the censoring means that the data values are hidden in known intervals, the truncation arises when some results, though relevant for the analysis, are not reported at all (i.e. we even do not know the number of such lost data). As a rule, there are thresholds (which could be individual and taken as random, or fixed equal for the whole set of observations) such that the values under them (in the case of left truncation) or above them (the right truncation) are not included in available data. It has been shown, for instance already in [7], that when the design of truncation threshold is such that the values from the whole data region are allowed (can be obtained), consistent non-parametric estimation of data distribution is possible. The result has later been extended to regression setting adapting the approach based on counting processes and hazard rate models, the overview is e.g. in [1].

The fixed truncation means that there are no data observed under (or above) a given threshold, therefore only the information on a conditional distribution is available and, in order to fit the complete distribution to such data, its parametric form has to be assumed. The present contribution deals with the case of left truncation. It is inspired by the problem how to estimate the operational risk regulatory capital on the basis of available data-base when the loss data of our interest are truncated from below at a fixed threshold. It is caused by an attempt to avoid recording and storing too many small loss events. However, omitting a part of data makes the problem of modeling operational risk accurately rather difficult [3]. In [6] the authors give an overview of different challenges connected with such an analysis, besides the problem of missing data also the problem of possible heavy-tailed nature of the losses distribution.

The structure of the paper is the following: In the next section the problem will be further specified, the structure of data described, and four methods of losses distribution estimators presented, namely the maximum likelihood estimator (MLE), the moment method (MM), then the Cramér-von Mises (CvM), Anderson-Darling (AD), and Kolmogorov-Smirnov (KS) minimum distance estimators. The methods will be examined on randomly generated data and their performance compared, in particular their reaction to the presence of a part of data coming from a heavy-tailed distribution. It is necessary to emphasize here that the main goal is a reliable estimation (and then the prediction) of the sum of values (losses) over certain period, not only the estimation of parameters of distribution of losses itself. And that the difficulties of analysis are caused principally by two aspects: The truncation (a set of values, though small ones, not recorder at all) and the accidental presence of very high values, outliers from the statistical point of view, which, however, must not be omitted.

[1] Department of Stochastic Informatics, UTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic, volf@utia.cas.cz

2 The problem of heavy tails

In statistics, the robustness of a method (for instance of an estimator) means that its performance is not influenced much by a presence of (a small) portion of outlied values contaminating the regular data. There exists a set of characteristics quantifying the reliability (stability) of a robust method, e.g. the breakdown point or the empirical influence function (cf. [4]). Thus, even in the setting considered here, from the robust statistics point of view, the aim is to estimate well the underlying basic distribution, when it is contaminated by (mixed with) a certain portion of a distribution with heavy tails. To this end, both in [3] and [6] the empirical influence functions for several estimators are derived, showing highly non-robust behavior of the MLE and moment estimators and at least partial robustness of the Cramér-von Mises method (see also [2]). Let us recall here that, in general, a heavy tail of a distribution means that it is not exponentially bounded. In fact, we shall consider here a sub-class of so called ”fat-tailed” distributions having its right tail $P(X > x)$ comparable with x^{-a}, for some $a > 0$, as $x \to \infty$.

More specifically, the situation is as follows: We assume that certain parametric distribution type is the baseline model. Further, it is assumed that a realistic model of the data arises from the mixture with another distribution having heavier right tail. In fact, its type can also be specified, still we face a difficult task to estimate parameters of both distributions and the rate of mixture. As the case is further complicated by missing part of data, in general such a problem has no unique solution. Fortunately, in the left truncation case considered here, just certain portion of small values is missing, high values remain available in observed data sets. Then, the main condition of successful model identification is a sufficiently robust method of estimation of the baseline distribution parameters. Hence, the estimators will be compared also from this point of view.

The robustness can be further improved with the aid of convenient robust estimator. In [6] the authors use so called ”optimally bias-robust estimator” (OBRE) set of estimators. On the other hand, the structure of left truncated data suggests the use of so called trimmed estimator of the location parameter, i.e. a very simple robust estimation method. That is why we considered such a kind of estimator as a tool for improving the estimation results. However, the improvement was rather negligible, therefore the method is not considered in the follow up.

Proposed estimation procedure has in fact two stages. In the first, the parameters of the baseline distribution are estimated. To do it reliably, sufficiently robust estimator should be employed. Then, on the basis of well estimated parameters of the baseline distribution, the second component of the mixture and the mixture rate can be estimated, which is crucial for the main goal of the analysis, namely for prediction of aggregated losses. This stage, on the contrary, has to use an estimator sensitive to all values, in order to distinguish both mixture components.

In the sequel we shall consider, similarly as [3] and [6], the log-normal baseline distribution of losses, as it is a model convenient both from practical and theoretical point of view. Further, its right part will be contaminated by the Pareto distribution as a model of possible occurrence of large values, as it is commonly considered to be a reasonable choice [5]. Again, let us recall here briefly that the Pareto (or also ”power law”) distribution has distribution function $F_p(x) = 1 - (A/x)^{\lambda}$ for $x > A > 0$, $F_p(x) = 0$ for $x \leq A$, $\lambda > 0$ is its shape parameter.

3 The model and estimators

It is assumed that a positive random variable X is observed just when its value is above a given threshold T. Hence, the data consist of a random sample $X_i, i = 1, .., N_1$, all $X_i > T$. The part under T is not observed, nor its frequency N_2 is known. Denote the density function of X $f(x)$, distribution function $F(x)$. It is further assumed that this distribution is a mixture, namely $f(x) = (1 - \alpha) \cdot f_0(x) + \alpha \cdot f_1(x)$, where the basic part $f_0(x)$ is given by a log-normal distribution with unknown parameters μ_0, σ_0, and is contaminated by a Pareto distribution, with density function $f_1(x)$ and with appropriate parameters. As it has been said above, both its parameters and the rate of contamination are also the object of estimation. We assume that the contamination rate α is not large, we have examined its influence for $\alpha \in [0, 0.2]$. Thus, the first goal is to estimate parameters of $f_0(x)$. As it has been said above, the aim of this first stage is to use a sufficiently robust procedure. Just for comparison, we shall deal with cases both with and without contamination, examining the behavior of several estimators. Namely the MLE, moment estimator and three distance-based methods.

Remark: The assumption of log-normal distribution allows to work with normal distribution model for logarithmized data. Hence the methods described above can be used for transformed data, it can simplify numerical procedures. As regards the contamination, let us recall that logarithmized Pareto distribution yields the exponential one. This connection will be used later for the random generation of data.

3.1 Estimation methods

In the case of full data, we can construct the full empirical distribution function, as a reliable non-parametric distribution estimate. Under the assumption of parametrized distribution, let us denote its density $f(x; \theta)$, distribution function $F(x;\theta)$, set of parameters hidden in θ should be estimated. From the fixed truncation it follows that the part of distribution above threshold T is given by the density and distribution functions, resp., both for $x > T$:

$$f_T(x;\theta) = \frac{f(x;\theta)}{1 - F(T;\theta)}, \quad F_T(x;\theta) = \frac{F(x;\theta) - F(T;\theta)}{1 - F(T;\theta)}.$$

1. Maximum likelihood estimator. The likelihood based on observed data has the form

$$L(\theta, \boldsymbol{x}) = \prod_{i=1}^{N_1} f_T(X_i;\theta)$$

and we search for θ maximizing its logarithm.

2. Moment estimator. Let us compute conditional first 2 moments of $(X|X > T)$ and compare them with empirical moments obtained from observed data. Namely, we shall compute

$$E_T^{(k)}(\theta) = \int_T^\infty x^k\, f_T(x;\theta)\, dx, \quad \bar{X^k} = \frac{1}{N_1}\sum_{i=1}^{N_1} X_i^k.$$

The best θ should minimize a distance of them, in the simplest case $\sum_{k=1}^2 (E_T^{(k)}(\theta) - \bar{X^k})^2$.

3. Cramér-von Mises estimator. It minimizes the distance between the empirical and assumed distribution function on (T, ∞), namely we search for θ minimizing

$$\sum_{i=1}^{N_1} (F_{emp,T}(X_i) \; - \; F_T(X_i;\theta))^2,$$

where $F_{emp,T}(x)$ is the empirical distribution function computed from data observed above T. Namely, the simplest form is $F_{emp,T}(X_{(i)}) = i/N_1$, $i = 1, ..., N_1$, where $X_{(1)} \leq X_{(2)} \leq, ..., \leq X_{(N_1)}$ denote ordered observations. We use the following variant: $F_{emp,T}(X_{(i)}) = (2i-1)/2N_1$.

4. Anderson-Darling estimator is a weighted variant of the CvM estimator giving the data-points weights corresponding to the variance of empirical distribution function. Hence, it minimizes

$$\sum_{i=1}^{N_1} (F_{emp,T}(X_i) \; - \; F_T(X_i;\theta))^2 \cdot \frac{1}{w_i},$$

where $w_i = F_T(X_i;\theta) \cdot (1 - F_T(X_i;\theta))$. The weighting results in a higher sensitivity to small and large data, hence also in smaller robustness compared to the CvM method. However, still its influence function is bounded. This difference actually will lead us to the estimator choice, on the basis of following Monte Carlo study. In the first stage, where rare outlying data should have small influence, the CvM estimator will be preferred. Further, however, when the model should describe well also the source of contamination, the AD estimator will be utilized.

5. Kolmogorov-Smirnov estimator is based on minimizing the maximal distance between empirical and model distribution functions, i.e. it minimizes

$$\max_{X_i} |F_{emp,T}(X_i) \; - \; F_T(X_i;\theta)|.$$

It is evident that in all cases the estimation has to be solved with the aid of a convenient numerical optimization procedure, the moment method evaluation includes also numerical integration.

4 Monte Carlo study

The study is based on K-times repeated generation of data sets of extent N. Each such set is taken as representing the loss data over certain period. The data have been generated from normal distribution with parameters μ_0, σ_0 and mixed with values from exponential distribution with parameter λ shifted by a constant a, i.e. having distribution function $F_e(x) = 1 - \exp(-\lambda \cdot (x - a))$ for $x \geq a$. The mixture (contamination) rate α was selected from $[0, 0.3]$. Such data represented logarithms of losses, they then were truncated from the left side by a threshold T_0. Hence, the losses were given by values coming from the mixture of log-normal distribution (with μ_0 and σ_0) with the Pareto distribution having distribution function $F_p(x) = 1 - (A/x)^\lambda$ for $x \geq A = \exp(a)$. The values of losses were truncated by threshold $T = \exp(T_0)$.

The set of truncated data then contained just $N_1 \leq N$ values greater than the threshold, it was assumed that the number of omitted data as well as their values were not known. In fact, as the data were prepared artificially, we knew them and could use them as a benchmark for comparison of performance of estimation methods and examination of information loss caused by the truncation. As in each Monte Carlo study, the repetition of analysis enabled us to construct empirical distribution of estimates, to study their bias and variability, and, later on, to analyze and compare distributions of sums reconstructed on the basis of different estimation methods.

The example provided here uses the following values: $\mu_0 = 2$, $\sigma_0 = 0.5$, $\lambda = 1$, $a = 2$, hence $A \doteq 7.39$. Further $T_0 = 1.3$, $\alpha = 0$ or 0.1, $N = 1000$, $K = 1000$ were selected. From such a choice it follows that the basic log-normal distribution had expectation $\doteq 8.4$ and standard deviation $\doteq 4.5$, while the Pareto distribution with parameter $\lambda = 1$ had infinite all moments. Threshold $T = \exp(1.3) \doteq 3.67$, the proportion of data truncated off was about 8%. Just for comparison, the 95% quantiles were 16.8 and 147.8 for these log-normal and Pareto distributions, respectively, 99% quantiles were 23.6 and 738.9.

4.1 Results of parameters estimation

The first case examined was the case without contamination, the data were generated just to correspond the log-normal distribution with given parameters μ_0, σ_0. Data were then truncated and parameters estimated from truncated samples by three methods. As the data generation was repeated K times, K estimates were obtained for each parameter and each method. Figure 1 displays these sets of estimates in a form of boxplots. The first correspond to the MLE from complete data, the other three then to the CvM estimator, the MLE and to moment estimator. It is seen that their performance is comparable, bias negligible and variability increased (compared to estimates from full data) due a loss of information caused by the truncation. Other estimators (KS and AD) performed very similarly.

In the second case presented here the log-normal (μ_0, σ_0) data were mixed with values generated from the Pareto distribution, their proportion was $\alpha = 0.1$. As it was said, during this stage of analysis the data were still treated as coming from log-normal distribution with unknown parameters μ, σ. Figure 2 again shows the results of estimation, in K repetitions, first the MLE from full data, then the results of 3 selected estimation methods used to truncated data. Now the pattern is different. First, as the contamination has caused a number of large, outlying values in data, the consequence is that the estimates are shifted, namely estimated standard deviation is increased and estimate of μ biased even in the case of the MLE from full data. Further, reactions of examined estimation methods to contaminated and truncated data differ. As expected, both the MLE and moment estimators react by even more increased both bias and variability of values, relative to estimates obtained from full data. On the other hand, in order to cope with heavier right tail of the data, the CvM method yielded a slightly increased estimates of both μ andσ. Simultaneously, variability of estimates did not increase significantly, which indicates a consistency of method. Such an phenomenon can be related to findings in [6] concluding that the CvM method is much more robust (having bounded empirical distribution function) than the other two. Further, as regards the other distance-based estimators, the result is collected in Table 1. It is seen that the KS method yielded results quite comparable with those of the CvM, while the AD estimators showed a stronger reaction to right tail data, it was biased and had larger variability similarly like the MLE and the moment method.

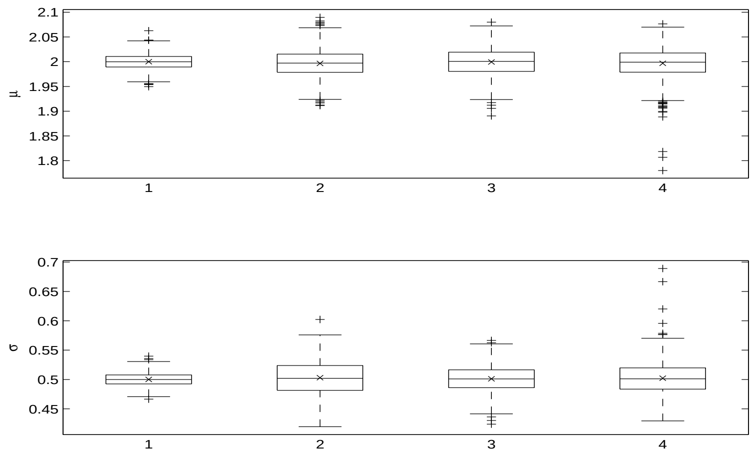

Figure 1 Estimated μ (above) and σ (below) in the case of no contamination: 1–MLE estimates from complete data, 2–CvM estimator, 3–MLE, 4–moment method.

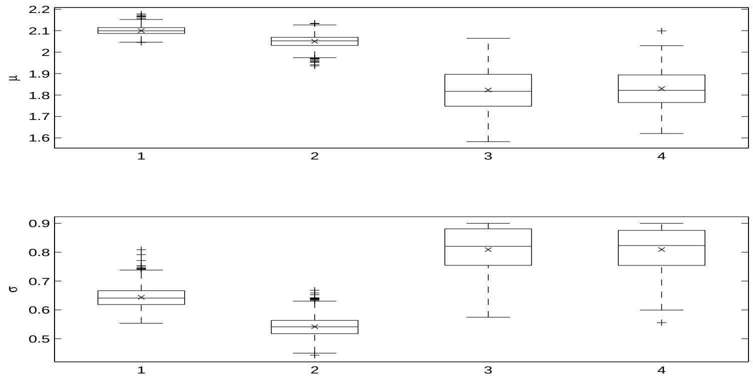

Figure 2 Estimated μ (above) and σ (below) when contamination rate was $\alpha = 0.1$: 1–MLE estimates from complete data, 2–CvM estimator, 3–MLE, 4–moment method.

	estimated:	μ			estimated:	σ		
Method	mean	median	Q(0.05)	Q(0.95)	mean	median	Q(0.05)	Q(0.95)
CvM	2.0493	2.0512	1.9952	2.0954	0.5431	0.5416	0.4900	0.6007
KS	2.0441	2.0460	1.9841	2.0966	0.5533	0.5529	0.4930	0.6195
AD	2.2426	2.1097	1.7212	2.9800	0.8682	0.8918	0.7494	0.8995
MLE	1.8230	1.8132	1.6812	1.9806	0.8073	0.8215	0.6708	0.8963
Moment	1.8298	1.8219	1.6984	1.9778	0.8092	0.8230	0.6749	0.8952

Table 1 Empirical characteristics of estimates obtained from different methods.

4.2 Analysis of contamination

In the second estimation stage the aim is to identify the heavy-tailed component of the mixture and estimate its parameters, when the Pareto model is assumed.Hence, the method should be sensitive to all observed values, giving an appropriate weights also to right tails of data. After a set of experiments we decided to prefer the AD estimator meeting best such requirements. The numerical example presented here, again based on K sets of N data (partly left-truncated), and using μ andσ estimated in the first stage, yielded the estimates which empirical characteristics (from K repetitions) are summarized in Table 2.

Parameter	mean	median	Q(0.05)	Q(0.95)
a	1.695	1.521	1.321	2.982
λ	0.839	0.903	0.121	1.602
α	0.094	0.091	0.014	0.217

Table 2 Empirical characteristics of estimates.

It is seen that empirical distribution of estimates is not symmetric, still rather wide, but at least the mean or median values providing acceptable results. Simultaneously, certain trade-off among parameters can be traced. For instance, smaller λ leads to longer right tail, while smaller a shifts the whole distribution left.

4.3 Estimated distribution of sums

As it has been said, this task is the main and final objective of the study. In particular, we are interested in how well the methods are able to model (and then to predict) upper right end quantiles of distribution of sums. This distribution is very sensitive even to just small changes of parameters, hence also to their unperfect estimates. And we have seen how rather complicated the estimation procedure is. Simultaneously, the results depends also on the number of losses during given period. This point is not considered here, we just try to estimate the distribution of convolution of a fixed number, D, of i.i.d. random variables representing the losses. The recommended approach to the operational risk modeling concerns the calculation of a risk measure VaR$_\gamma$ at a confidence level $\gamma = 99,9\%$ for a loss random variable L corresponding to the aggregate losses over a given period, usually one year [5]. As this distribution has no closed form, standard way of examining it is again a Monte Carlo approach. Therefore we generated K times, with K 10^5, sums $L = \sum L_k$ of $D = 100$ variables L_k having the mixed distribution derived and estimated in preceding parts. Table 2 shows a comparison of chosen right empirical quantiles of L obtained by random generation.

Quantile	0.995	0.996	0.997	0.998	0.999	0.9995
a) Estimated	18 730	23 671	31 638	46 747	97 711	202 301
b) ”True”	15 555	19 183	25 214	37 952	84 846	154 176

Table 3 Empirical quantiles of L based on a) model using the medians of estimated parameters $\mu = 2.0512, \sigma = 0.5416, \lambda = 0.903, a = 1.521, \alpha = 0.091$; b) the ”true” model with parameters $\mu_0 = 2, \sigma_0 = 0, 5, \lambda_0 = 1, a_0 = 2, \alpha_0 = 0.1$.

The quantiles based on estimated parameters exceed slightly the quantiles of true distribution of sums. It indicates that the method could be applicable without large danger of underestimation of real aggregate losses. Naturally, each analysis of this kind has to start from careful exploration of available real data.

5 Concluding remarks

The first aim of the study was to examine and compare performance of several estimators of distribution parameters in the case of fixed left truncated data. The data were generated randomly, the sense of examples was to simulate a set of losses of a financial institution encountered during certain period. Their distribution was modeled via the log-normal distribution contaminated by the Pareto one. The main objective was then the estimation of distribution of sums of losses over a given period..It means to summarize the values coming from (possibly contaminated) log-normal distribution and, moreover, not observed fully. Theoretically, the distribution could be approximated on the basis of the central limit theorem. However, there are many issues leading to doubts on its correctness and practical usefulness. The asymptotic behavior of the C.L.Th. on distribution tails is rather slow in general, not speaking about the fact that Pareto distribution of our choice does not fulfil theoretical requirement for the C.L.Th. validity.

That is why this part of analysis was also based on Monte Carlo approach and estimated parameters. We hope that such an approach is suitable also for practical use. As a rule, a sufficiently large database is available, usually omitting values under given threshold. Hence, the parameters of assumed type of baseline distribution can be estimated, e.g. using sufficiently robust Cramér-von Mises estimator. Then the model for heavy-tailed part of losses distribution can be identified, in this stage a less robust method is appropriate, we can recommend the Anderson-Darling method. Finally, random generation from obtained model helps to recover expected behavior of aggregated losses.

References

[1] Andersen, P. K., and Keiding, N.: *Survival and Event History Analysis*. John Wiley & Sons, New York, 2007.

[2] Duchesne, T., Rioux, J., and Luong, A.: Minimum Cramér-von Mises estimators and their influence function. *Actuarial Research Clearing House* **1** (1997), 349–361.

[3] Ergashev, B., Pavlikov, K., Uryasev, S., and Sekeris, E.: Estimation of truncated data samples in operational risk modeling. *The Journal of Risk and Insurance* **83** (2016), 613-640.

[4] Huber, P. J., and Ronchetti, E.: *Robust Statistics (2-nd Edition)*. John Wiley & Sons, New York, 2009.

[5] Nešlehová, J., Embrechts, P., and Chavez-Demoulin, V.: Infinite mean models and the LDA for operational risk. *Journal of Operational Risk* **1** (2006), 3–25.

[6] Opdyke, J. D., and Cavallo, A.: Estimating operational risk capital: the challenges of truncation, the hazards of maximum likelihood estimation, and the promise of robust statistics. *Journal of Operational Risk* **7** (2012), 3–90.

[7] Turnbull, B. C.: The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)* **38** (1976), 290–295.

35th International Conference

Mathematical Methods in Economics

MME 2017

Conference Proceedings

Hradec Králové, Czech Republic
September 13th – 15th, 2017

University of Hradec Králové

ISBN 978-80-7435-678-0