A Simple and Robust Approach for Expected Shortfall Estimation

Zhibin Pan*, Tao Pang† and Yang Zhao‡

April 14, 2020
(First Version: January 2, 2018)

Abstract

In risk management, estimating Expected Shortfall (ES), though important and indispensable, is difficult when the sample size is small. This paper makes efforts to create a recipe for such a challenge. A tail-based normal approximation with explicit formulas is derived by matching a specific quantile and the mean excess square of the sample observations. To enhance the estimation accuracy, we propose an adjusted tail-based normal approximation based on the sample's tail weight. The adjusted ES estimator is robust and efficient in the sense that it can be applied to various heavy-tailed distributions, such as student's t, lognormal, Gamma, Weibull, etc., and the errors are all small. Moreover, compared to two common ES estimators—the arithmetic average of excessive losses and extreme value theory estimator, the proposed estimator achieves smaller mean square errors for small samples, especially at high confidence levels. The properties of linear transformations on the ES estimator are also investigated to ensure its practicality.

Keywords: Expected Shortfall; Tail-based Normal Approximation; Conditional Skewness; Tail-Weight Adjustment; Heavy-Tailed Distribution, Small Sample

JEL Code: C51; C63; G32

Key Messages:

- Propose a tail-based normal approximation for ES estimation.
- Develop an adjusted tail-based normal ES estimation approach for heavy-tailed distributions.
- The approach is easy to implement and robust for various heavy-tailed distributions.
- The approach is substantially accurate and works well for small samples.

*Department of Finance, East China Normal University, Shanghai 200062, P. R. China. Email: zbpan@finance.ecnu.edu.cn

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205. Email: tpang@ncsu.edu Corresponding Author.

‡SAS Institute, 500 SAS Campus Dr, Cary, NC 27513. Email: Yang.Zhao@sas.com

1 Introduction

Expected Shortfall (ES), also known as Conditional Value-at-Risk (CVaR), is a coherent measure of risk which considers losses exceeding the corresponding Value-at-Risk (VaR). As ES remedies the tail risk and non-sub-additivity, problems VaR inherently suffers (**?**), it has been attracting more and more attention in the field of risk management.

The ES of a sample can be estimated by definition with a parametric distribution or by extreme value theory (EVT) approach (**?**). Besides the Generalized Pareto distribution (GPD) supported by EVT, other asymmetric distributions such as skewed normal (**?**), asymmetric t and exponential power distributions (**?**), and Laplace (**?**) have also been studied. In view of location-scale distributions, **?** propose various large-sample parametric confidence intervals for ES estimates. Moreover, **?** introduce a multiple-period ES estimation method within frameworks of random walks, autoregressive process or GARCH(1, 1) models with t distributed innovations. **?** also evaluate the ES forecasts using GARCH models with different estimation frequencies. For mixture distributions, normal and t mixture distributions and their closed-form ES estimates are investigated by **?**.

Compared to parametric methods, non-parametric ES estimators avoid distributional assumptions. A popular non-parametric estimator is defined by **?** as the arithmetic average of losses that are beyond a specific VaR estimator, which is similar to the natural estimator for expected losses in the worst case proposed by **?**. **?** introduce an extrapolation method to estimate ES. Applying empirical likelihood, an asymptotically valid confidence interval for ES is derived by **?**. Besides, **?** propose a kernel-smoothed ES estimate that worked well even in the dependent situations. The kernel smoothing produces a VaR estimator with less mean square errors, especially for small samples (**?**), while **?** shows that such smoothing cannot give a more accurate ES estimator. Taking advantage of its simplicity, the arithmetic average of exceedances is still one of the most popular non-parametric ES estimators.

Though many methods have been studied, selecting an accurate ES estimator is a challenge in practice. In light of the arithmetic average of exceedances, the given sample of losses, however, is not always large enough to give a robust estimator. Assuming the sample size is 250, the number of one year's observations, only the largest two losses are covered when estimating ES at the 99% confidence level and only the maximal one is valuable for ES at the 99.5% level. **?** indicate that ES estimator is quite unstable especially for a heavy-tailed[1] loss distribution where it is easily affected by whether infrequent losses would occur in the realized sample. Moreover, given such small samples, estimating GARCH models precisely is also difficult.

Normal distributions have many nice properties with the presence justified by Central Limit Theorem but they cannot properly capture the heavy-tailed behaviors that are common in the financial data. A model building approach with normal distributions usually suffers the 'underestimation' problem when estimating ES. Heavy-tailed distributions such as t and other stable distributions have also been considered but some inherent drawbacks impede their applications. For example, the sum of two t distributed r.v.'s generally no longer follows a t distribution. Moreover, even though the sum of two stable r.v.'s follows a stable distribution, there is usually no general explicit formula for its probability density function.

In this paper, we propose a new ES estimator based on the tail-based normal approximation. This tail-based feature proves to be effective to alleviate the 'underestimation' problem. To further improve the estimation accuracy, a regression-adjusted tail-based normal approximation is then

[1] In this paper, a heavy-tailed distribution refers to any distribution that has a heavier right tail than the normal distribution.

introduced where the sample tail weight is considered. The robust tests indicate the adjusted ES estimator works well for various heavy-tailed loss distributions. It also outperforms the widely-used arithmetic average and EVT ES estimators in terms of mean square error (MSE) for small loss samples simulated from heavy-tailed distributions. Moreover, the properties of the proposed ES estimator on linear transformation further facilitate its application to portfolio management.

The rest of paper is organized as follows. In Section 2, we propose a tail-based normal approximation and derive the explicit formula of its ES estimator. The accuracy analysis is also provided. In Section 3, we adjust the tail-based normal approximation using the sample's tail weight, and explores whether such adjustment leads to a more accurate ES estimation. In Section 4, a self-consistency test and some robust tests are carried out to validate the proposed ES estimator. We compare the proposed ES estimator to the arithmetic average and EVT estimators for small loss samples in Section 5. Moreover, the effects of linear transformation on the ES estimator are studied in Section 6. In Section 7, we conclude the paper and suggest some topics of the future work. All detailed derivations are included in the Appendix.

2 Tail-Based Normal Approximation for ES Estimation

For a loss (or negative return) random variable (r.v.) L, its ES at level $\beta \in (0,1)$ is defined as follows (see **?**):

$$\mathrm{ES}_\beta(L) = \frac{1}{1-\beta}\int_\beta^1 \mathrm{VaR}_\phi(L)d\phi, \tag{2.1}$$

where $\mathrm{VaR}_\phi(L)$ is the value-at-risk at level $\phi \in (0,1)$ which is defined by

$$\mathrm{VaR}_\phi(L) = \inf\{z \in R | Pr(L \le z) \ge \phi\}. \tag{2.2}$$

It can be shown through a variable transformation that if the loss r.v. L is continuously distributed with a PDF $f(\cdot)$, then Eq. (2.1) is equivalent to

$$\mathrm{ES}_\beta(L) = \mathbb{E}[L | L \ge \mathrm{VaR}_\beta(L)] = \frac{1}{1-\beta}\int_{\mathrm{VaR}_\beta(L)}^\infty x f(x)dx. \tag{2.3}$$

In traditional model building approaches, all available sample points are utilized to estimate the distribution parameters. For example, a normal approximation is usually obtained by matching its mean and variance to the sample mean and sample variance. This type of approach is global-based in the sense that all sample points are taken into consideration.

ES, however, is a statistic that mainly depends on tail behaviors. Therefore, global-based approaches may not give accurate ES estimations. This is one of the reasons why the global-based normal approximation typically underestimates the ES for real market data, especially at high confidence levels such as 99% and 99.5%.

In this section, we propose a tail-based approach that only considers the tail sample points. In particular, focusing on the excessive observations, we build a tail-based normal approximation by equating its specific quantile (e.g. 95%-quantile) and mean excess square to the counterparts of the objective sample. This approximation will be further improved in the next section through some adjustment factors related to the sample's tail weight.

We want to point out the idea of the tail-based approximation can be applied to distributions other than normal distributions, such as student's t, Gamma, etc. In this paper, we only consider normal distributions for the following reasons. Firstly, normal distributions are simple and have many nice properties. For example, the sum of two normal r.v.'s is still a normal r.v., which is useful in risk management when calculating n-day ES (or VaR) based on daily estimates. Secondly, the tail-based normal approximation gives sufficiently accurate results so it might not be necessary to explore other distributions.

2.1 Explicit Formulas of Tail-Based Normal Approximation

Let $\mathbf{Y} = \{y_n\}_{n=1}^N$ denote a sample of losses and our goal is to develop a model-based approach to estimate β-level ES estimate ES_β. β is usually close to 1, and two popular choices are 99% and 99.5%. Apparently, this approach should mainly depends on the right-tail behavior of the loss sample. Firstly, for any $\alpha \in (0, 1)$, we define A_α, the α-quantile of the sample $\mathbf{Y}$, as follows:

$$A_\alpha \equiv (\lfloor N\alpha \rfloor + 1 - N\alpha) y_{(\lfloor N\alpha \rfloor)} + (N\alpha - \lfloor N\alpha \rfloor) y_{(\lfloor N\alpha \rfloor + 1)}, \tag{2.4}$$

where $\lfloor N\alpha \rfloor$ represents the greatest integer that is less than or equal to $N\alpha$ (i.e. $\lfloor\cdot\rfloor$ is the floor function), and $y_{(1)}, y_{(2)}, \cdots, y_{(N)}$ are the ascending order statistics of the sample $\mathbf{Y}$. In case of possible misunderstandings, 'α-quantile' ($0 < \alpha < 1$) in this paper is equivalent to (100α)th percentile. For example, 0.75-quantile or 75%-quantile is equivalent to 75th percentile.

Next, we choose a threshold level α that is less than β (e.g. $\alpha = 0.95$ when $\beta = 0.99$) and define a normal r.v. $X \sim N(\mu, \sigma^2)$ to approximate the right tail of the sample $\mathbf{Y}$ beyond the α-quantile A_α; that is, we are going to find a tail-based normal approximation for the given sample. In particular, μ and σ^2 are solved such that the two tail statistics, the α-quantile and the 'conditional tail variance' (mean excess square), match the corresponding sample statistics:

$$Pr(X \leq A_\alpha) = \alpha, \qquad \mathbb{E}[(X - A_\alpha)^2 | X > A_\alpha] = \frac{\sum_{n=1}^N (y_n - A_\alpha)^2 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}, \tag{2.5}$$

where $\mathbb{1}_{\{\cdot\}}$ is the indicator function.

We can derive the unique solution (μ, σ^2) of the system of equations listed above. Define $Z \equiv \frac{X-\mu}{\sigma}$, and then Z follows a standard normal distribution. Thus the first equation in Eq. (2.5) can be transformed as follows:

$$Pr(X \leq A_\alpha) = Pr\left(\frac{X-\mu}{\sigma} \leq \frac{A_\alpha - \mu}{\sigma}\right) = Pr\left(Z \leq \frac{A_\alpha - \mu}{\sigma}\right) = \Phi\left(\frac{A_\alpha - \mu}{\sigma}\right) = \alpha, \tag{2.6}$$

$$\frac{A_\alpha - \mu}{\sigma} = \Phi^{-1}(\alpha) = z_\alpha. \tag{2.7}$$

where $\Phi(\cdot)$ and z_α denote the cumulative distribution function (CDF) and z-score of a standard normal distribution, respectively; that is, $z_\alpha = \Phi^{-1}(\alpha)$. The conditional first and second moments of X (See Appendix A for details) are calculated as

$$\mathbb{E}[X | X > A_\alpha] = \mathbb{E}\left[\sigma Z + \mu \middle| Z > \frac{A_\alpha - \mu}{\sigma}\right] = \mu + \frac{\sigma}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{(A_\alpha - \mu)^2}{2\sigma^2}}, \tag{2.8}$$

$$\begin{aligned} \mathbb{E}[X^2 | X > A_\alpha] &= \mathbb{E}\left[\mu^2 + \sigma^2 Z^2 + 2\mu\sigma Z \middle| Z > \frac{A_\alpha - \mu}{\sigma}\right] \\ &= \mu^2 + \sigma^2 + \frac{\sigma(A_\alpha + \mu)}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{(A_\alpha - \mu)^2}{2\sigma^2}}. \end{aligned} \tag{2.9}$$

By Eq. (2.7), Eq. (2.8) and Eq. (2.9):

$$\begin{aligned}\mathbb{E}[(X-A_\alpha)^2|X>A_\alpha] &= \mathbb{E}[X^2|X>A_\alpha]-2A_\alpha\mathbb{E}[X|X>A_\alpha]+A_\alpha^2\\ &= (\mu-A_\alpha)^2+\sigma^2+\frac{\sigma(\mu-A_\alpha)}{(1-\alpha)\sqrt{2\pi}}\exp\big(-\frac{(A_\alpha-\mu)^2}{2\sigma^2}\big)\\ &= \sigma^2\left[z_\alpha^2+1-\frac{z_\alpha}{(1-\alpha)\sqrt{2\pi}}e^{-\frac{1}{2}z_\alpha^2}\right].\end{aligned}\tag{2.10}$$

By Eq. (2.5), Eq. (2.7) and Eq. (2.10), we can get the explicit formulas for the parameters μ, σ of the tail-based normal approximation using the sample data set $\{y_n\}_{n=1}^N$ as follows:

$$\hat{\sigma}^2 = \frac{1}{\left[z_\alpha^2+1-\frac{z_\alpha}{(1-\alpha)\sqrt{2\pi}}e^{-\frac{1}{2}z_\alpha^2}\right]}\left[\frac{\sum_{n=1}^N(y_n-A_\alpha)^2\mathbb{1}_{\{y_n>A_\alpha\}}}{\sum_{n=1}^N\mathbb{1}_{\{y_n>A_\alpha\}}}\right],\tag{2.11}$$

$$\hat{\mu} = A_\alpha-\hat{\sigma}z_\alpha.\tag{2.12}$$

In practice, when calculating the β-level (e.g. $\beta = 99\%$ or 99.5%) ES estimate of the loss sample $\mathbf{Y} = \{y_n\}_{n=1}^N$, we first derive the tail-based normal approximation given by Eq. (2.11) and Eq. (2.12) with a threshold level α that is less than β (e.g. $\alpha = 95\%$). Once the normal distribution is obtained, it can be used to calculate the ES_β as:

$$\text{VaR}_\beta = \hat{\mu}+\hat{\sigma}z_\beta,\tag{2.13}$$

$$\text{ES}_\beta = \mathbb{E}[X|X>\text{VaR}_\beta] = \hat{\mu}+\frac{\hat{\sigma}}{(1-\beta)\sqrt{2\pi}}e^{-\frac{1}{2}z_\beta^2},\tag{2.14}$$

where $z_\beta = \Phi^{-1}(\beta)$ is the z-score of the standard normal distribution at confidence level β.

To test the accuracy, the tail-based normal approximation is also implemented for a loss r.v. with an explicit distribution function. As for a loss r.v. W with CDF $F_W(\cdot)$ and probability density function (PDF) $f_W(\cdot)$, its α-quantile A_α and mean excess square can be calculated by:

$$A_\alpha = F_W^{-1}(\alpha), \quad \mathbb{E}[(W-A_\alpha)^2|W>A_\alpha] = \frac{1}{1-\alpha}\int_{A_\alpha}^\infty (x-A_\alpha)^2 f_W(x)dx.\tag{2.15}$$

By Eq. (2.12), Eq. (2.11) and Eq. (2.15), the parameters, $(\hat{\mu}_w,\ \hat{\sigma}_w)$, of the tail-based normal approximation for the loss r.v. W can be obtained similarly as

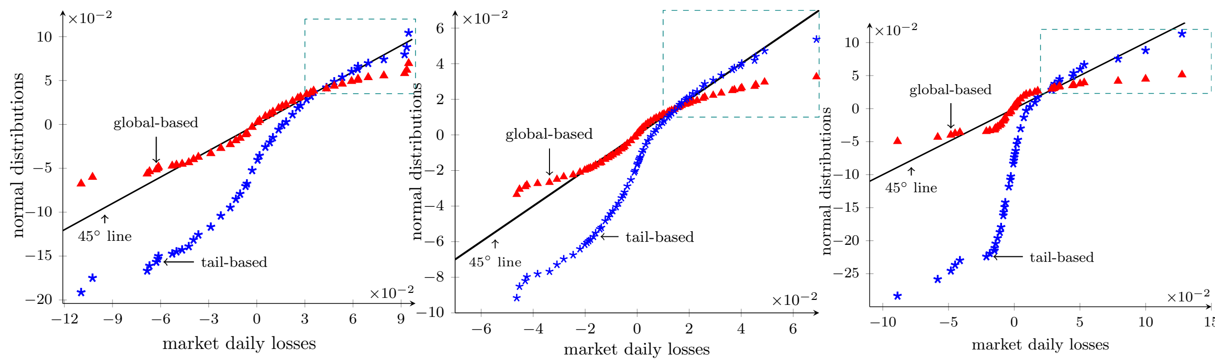
$$\hat{\sigma}_w^2 = \frac{1}{\left[z_\alpha^2+1-\frac{z_\alpha}{(1-\alpha)\sqrt{2\pi}}e^{-\frac{1}{2}z_\alpha^2}\right]}\mathbb{E}[(W-A_\alpha)^2|W>A_\alpha],\tag{2.16}$$

$$\hat{\mu}_w = A_\alpha-\hat{\sigma}z_\alpha.\tag{2.17}$$

2.2 Accuracy Analysis for Tail and Global Based Normal Approximations

In what follows, the accuracy of the tail-based normal approximation will be tested. Firstly, we compare the global-based and tail-based normal approximations using three samples of daily S&P 500 Index losses. The daily loss of Day m is calculated by $-\ln(P_m/P_{m-1})$, where P_m is the index value at Day m. As previously mentioned, the global-based normal approximation is obtained by matching its mean and variance to the sample mean and sample variance of the losses. The tail-based normal approximation is obtained by Eq. (2.11) and Eq. (2.12). The results are displayed in the form of Q-Q probability plots in Figure 1 where the tail-based and global-based normal approximations are marked by asterisks and triangles, respectively.

Concerning the right tail, Figure 1 shows the tail-based normal approximation is much closer to the given daily loss data than the global-based one. Since only the right tail is used for ES estimation, the tail-based normal approximation is expected to give a better ES estimate.



(a) Jul 2, 2007 - Jun 30, 2009 (b) Jan 4, 2010 - Jun 30, 2017 (c) Mar 21, 2019 - Mar 21, 2020

Figure 1: Q-Q Plot: global & tail-based normal approximations (y-axis) vs market daily losses (x-axis), $\alpha = 95\%$

To further test the accuracy of the tail-based normal approximation, we examine its performances for some common heavy-tailed loss distributions. The idea is that, for a r.v. W with a known distribution (such as t, Lognormal, etc.), we use formulas Eq. (2.16) and Eq. (2.17) to obtain its tail-based normal approximation. The formulas Eq. (2.13), Eq. (2.14) are then used to derive its VaR estimator, denoted by VaR^t_β, and ES estimator, denoted by ES^t_β, where the superscript t stands for tail-based. The theoretical values of VaR and ES estimates can be calculated by the PDF of the actual distribution and they are denoted by VaR_β and ES_β, respectively.

To quantitatively describe the estimation errors, the relative errors of the tail-based normal approximation are defined as follows:

$$\mathbf{e}^t_\beta(\mathrm{ES}) \equiv \frac{\mathrm{ES}_\beta - \mathrm{ES}^t_\beta}{\mathrm{ES}_\beta}, \quad \mathbf{e}^t_\beta(\mathrm{VaR}) \equiv \frac{\mathrm{VaR}_\beta - \mathrm{VaR}^t_\beta}{\mathrm{VaR}_\beta}. \tag{2.18}$$

From the above definitions, we can see that if the relative error is positive, there is an underestimation while if it is negative, there is an overestimation.

Additionally, for comparison reasons, we obtain the traditional global-based normal approximation by matching its mean and variance to the counterparts of W. Its VaR and ES estimators are denoted by VaR^g_β and ES^g_β respectively, where the superscript g stands for global-based. Similarly, the relative errors of the global-based normal approximation are defined by:

$$\mathbf{e}^g_\beta(\mathrm{ES}) \equiv \frac{\mathrm{ES}_\beta - \mathrm{ES}^g_\beta}{\mathrm{ES}_\beta}, \quad \mathbf{e}^g_\beta(\mathrm{VaR}) \equiv \frac{\mathrm{VaR}_\beta - \mathrm{VaR}^g_\beta}{\mathrm{VaR}_\beta}. \tag{2.19}$$

To investigate the performance of the tail-based normal approximation, we test it with heavy-tailed distributions including t, Gamma, Lognormal, GPD, and Weibull with different parameters. The location parameter is set to zero and the scale parameter is set to one if they exist. The distribution functions of each distribuiton are listed in Table 11 in Appendix B.

The relative errors for ES and VaR at $\beta = 99\%$ or 99.5% for tail-based and global-based normal distributions are given in Table 1, which shows that the tail-based normal approximation has much smaller ES estimation errors than the global-based one for all examined loss distributions.

W	β (%)	ES_β	$\mathbf{e}^g_\beta(\mathrm{ES})$ (%)	$\mathbf{e}^t_\beta(\mathrm{ES})$ (%)	VaR_β	$\mathbf{e}^g_\beta(\mathrm{VaR})$ (%)	$\mathbf{e}^t_\beta(\mathrm{VaR})$ (%)
t, df=3.5	99	5.895	30.940	-4.848	4.061	12.489	-19.838
	99.5	7.290	39.405	3.152	5.086	22.633	-14.716
t, df=5	99	4.452	22.721	-0.919	3.065	10.747	-9.075
	99.5	5.250	28.886	3.924	4.032	17.528	-6.054
t, df=8	99	3.591	14.296	0.121	2.897	7.258	-4.023
	99.5	4.083	18.222	2.770	3.355	11.357	-2.380
Gamma(5, 1)	99	13.001	15.699	0.225	11.605	12.088	-0.851
	99.5	13.956	17.837	0.977	12.594	14.565	-0.339
Gamma(3, 1)	99	9.639	20.981	0.303	8.406	16.376	-1.225
	99.5	10.485	23.618	1.332	9.274	19.542	-0.489
Gamma(0.3, 1)	99	3.494	49.627	0.819	2.639	40.358	-4.947
	99.5	4.092	53.956	3.954	3.221	46.878	-1.895
LogN(0, 1)	99	15.228	51.348	-2.409	10.241	34.803	-18.511
	99.5	18.971	58.364	5.598	13.142	45.096	-11.720
LogN(0, 0.9^2)	99	11.527	48.269	-1.316	8.115	33.510	-14.107
	99.5	14.059	54.884	5.417	10.158	42.769	-8.693
LogN(0, 0.3^2)	99	2.235	14.911	0.225	2.010	10.786	-1.179
	99.5	2.391	17.420	1.237	2.166	13.521	-0.564
GPD(0.3,1)	99	15.624	52.326	-7.747	9.937	32.743	-29.377
	99.5	20.006	60.208	2.547	13.004	44.274	-21.385
GPD(0.2,1)	99	10.699	48.118	-1.726	7.559	33.803	-14.187
	99.5	13.034	54.604	4.933	9.427	42.646	-9.154
GPD(0.1,1)	99	7.610	41.892	0.179	5.849	31.594	-6.576
	99.5	8.874	46.994	4.121	6.987	38.296	-3.584
Weibull(0.6,1)	99	17.990	52.448	0.339	12.747	39.923	-10.158
	99.5	21.773	57.955	5.711	16.103	48.344	-4.972
Weibull(0.9,1)	99	6.801	38.632	0.584	5.457	30.791	-3.347
	99.5	7.739	42.637	2.936	6.377	36.193	-1.386
Weibull(1.4,1)	99	3.415	21.839	0.262	2.977	17.831	-0.859
	99.5	3.714	24.100	1.005	3.290	20.660	-0.290

Table 1: Estimation errors of tail-based and global-based normal approximations, $\alpha = 95\%$

The improvements in ES estimation using the tail-based normal approximation are associated with overestimated VaR values (see the last column in Table 1). The overestimates of the VaR are needed to compensate for the originally underestimated ES estimates since the normal distribution has a lighter right tail than all the examined distributions.

Though the tail-based normal approximation gives more accurate ES estimates than the global-based one, its errors are not small enough and further improvements are needed. Moreover, it seems the errors have some dependence on the tail weight (shape) parameters of those distributions. Therefore, we consider an adjustment factor related to some tail weight statistics in the next section.

3 Adjusted Tail-Based Normal Approximation

The distributions of financial data are usually heavy-tailed so extreme losses are more frequent than the normal distributed samples. Table 1 indicates the estimation errors of the tail-based normal approximation depend on the tail weight of its actual distribution. Therefore, to further decrease the estimation errors, let us propose an adjusted tail-based normal approximation.

Suppose the theoretical β-level ES estimate of a loss r.v. W with a given distribution is denoted by $\mathrm{ES}_\beta(W)$. We pick a value of α such that $\alpha < \beta$ and the α-quantile of W is denoted by A_α. Assuming X, obtained by Eq. (2.16) and Eq. (2.17), is the r.v. of the tail-based normal approximation for W, a ratio $R_{\alpha,\beta}$ that measures estimation errors is defined by

$$R_{\alpha,\beta} \equiv \frac{\mathrm{ES}_\beta(W) - A_\alpha}{\mathrm{ES}_\beta(X) - A_\alpha}, \tag{3.1}$$

where $\mathrm{ES}_\beta(X)$ is obtained by Eq. (2.14) and A_α is the α-quantile for both X and W. W and X always have the same α-quantile, which is ensured by the derivation of the tail-based normal approximation: X is obtained by matching its α-quantile and mean excess square to the counterparts of W.

The reason why we subtract A_α from both the numerator and denominator in Eq. (3.1) is it ensures $R_{\alpha,\beta}$ stays unchanged when the loss r.v. is subject to a linear transformation (see Section 6 for details). Since the relative distribution shapes of X and W are not changed by a linear transformation, neither is $R_{\alpha,\beta}$. If the estimation error is close to 0, $R_{\alpha,\beta}$ is close to 1 and vice versa. Therefore, $R_{\alpha,\beta}$ can be used as a measure of the estimation error. Moreover, the results in Table 1 imply this ratio depends on the tail weight of W. To further navigate the relationship between the ratio $R_{\alpha,\beta}$ and the tail weight, we need a variable to quantitatively measure the tail weight generally. For a loss r.v. W whose α-quantile equals A_α, its conditional skewness is defined by

$$\gamma_\alpha \equiv \frac{\mathbb{E}[(W - A_\alpha)^3 | W > A_\alpha]}{(\mathbb{E}[(W - A_\alpha)^2 | W > A_\alpha])^{\frac{3}{2}}}. \tag{3.2}$$

A similarly defined conditional kurtosis can be considered, too. Since results with the conditional skewness are already outstanding, we stay with it in this paper. Our idea is to develop a regression model between $R_{\alpha,\beta}$ and γ_α such that

$$\hat{R}_{\alpha,\beta} = f_{\alpha,\beta}(\gamma_\alpha). \tag{3.3}$$

We can subsequently have a more accurate ES estimator by adjusting the tail-based normal approximation based on Eq. (3.1).

Remark 1 *From the formula of γ_α, it seems the regression-adjusted tail-based normal approximation only works for distributions with a finite third moment. This drawback may limit the application of the proposed method. However, the third moment of any given sample always exists (it might be very large), and the method is still feasible. We give the numerical results for samples generated from some distributions without finite third moments, such as student's t with the degree of freedom equals to 2.5 or 3, and GPD with $\xi = 0.35$ or 0.5. Please refer to the results in Table 10 and the discussion after that.*

Now we take the student's t distribution for the r.v. W as the training distribution. Choosing different degrees of freedom (df), we can obtain different W with the corresponding tail weights

	$\alpha = 95\%, \beta = 99\%$			$\alpha = 95\%, \beta = 99.5\%$		
Coefficient	Value	SE	p-value	Value	SE	p-value
b_0	0.8611	1.7×10^{-4}	0.00	0.9919	2.2×10^{-4}	0.00
b_1	0.5191	7.0×10^{-3}	0.00	0.6681	8.5×10^{-3}	0.00
b_2	0.9747	3.8×10^{-3}	0.00	0.9607	3.7×10^{-3}	0.00
b_3	0.6099	2.0×10^{-3}	0.00	0.6022	2.5×10^{-3}	0.00
b_4	-0.9413	6.9×10^{-5}	0.00	-1.4623	6.2×10^{-3}	0.00
	Adjusted $R^2 = 0.9999$			Adjusted $R^2 = 0.9998$		

Table 2: Summary for regression models with $\alpha = 95\%, \beta = 99\%$ and $\alpha = 95\%, \beta = 99.5\%$

γ_α. Then a regression model between $R_{\alpha,\beta}$ and γ_α is built. One thing we want to point out is that, although the regression model is developed based on student's t distribution, it can be applied to other heavy-tailed distributions (see Remark 2 and the robust test results in Subsection 4.2).

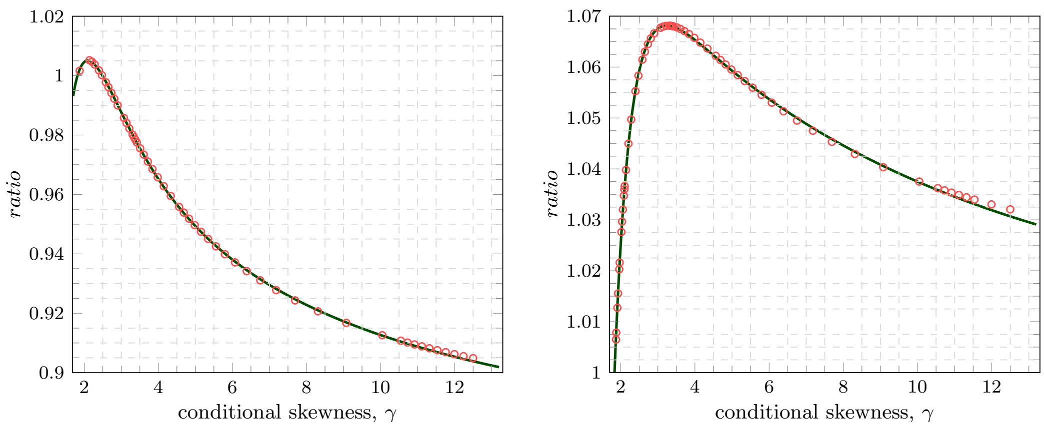


Figure 2: Scatter plot of $R_{\alpha,\beta}$ on γ_α with regression line; $\alpha = 95\%$, $\beta = 99\%$ (left), $\beta = 99.5\%$ (right)

To illustrate the relation between $R_{\alpha,\beta}$ and γ_α, we plot some values in Figure 2 (the small circles) for $\beta = 99\%$ (left) and $\beta = 99.5\%$ (right). A regression analysis is subsequently conducted to develop an equation that predicts $R_{\alpha,\beta}$ from the conditional skewness of the t distributed r.v. W. According to Figure 2, the scatter plots[2] of $R_{\alpha,\beta}$ on γ_α, the following regression lines are devised (see the solid lines in Figure 2) :

$$\hat{R}_{\alpha,\beta} = f_{\alpha,\beta}(\gamma_\alpha) = b_0 + b_1 e^{-b_2\gamma_\alpha} + b_3\gamma_\alpha^{-1} + b_4\gamma_\alpha^{-2}, \quad \alpha = 95\%,\ \beta = 99\% \text{ or } 99.5\%. \tag{3.4}$$

The coefficient estimates in Eq. (3.4) with p-values and the adjusted R^2 are reported in Table 2 which shows all variables are useful to predict $R_{\alpha,\beta}$. We want to mention that, in (3.4), the coefficient values depend on the values of α and β. The other thing we want to mention is that, the regression model is not unique. Any prediction formula works as long as it precisely describes relations between $R_{\alpha,\beta}$ and γ_α.

The adjustment factor $f_{\gamma,\beta}(\gamma_\alpha)$ depends on the level of α and β, but it is usually robust on the distribution of W. By Eq. (3.1) and Eq. (3.3), we can get a more accurate ES estimator based on the regression-adjusted tail-based normal approximation as

$$\widehat{\mathrm{ES}}_\beta(X) \equiv [\mathrm{ES}_\beta(X) - A_\alpha] f_{\alpha,\beta}(\gamma_\alpha) + A_\alpha. \tag{3.5}$$

[2]To ensure the regression line is shown clearly, only a few points are displayed in Figure 2 while regression analysis covers a sufficiently large number of points to guarantee the effectiveness.

Remark 2 *Though the two regression models ($\beta = 99\%$ or 99.5%) in Eq. (3.4) are built based on the student's t distributions, we can adopt them to adjust the tail-based normal approximation obtained from loss samples or loss r.v.'s given by other distributions. This adjustment is feasible because the corresponding γ_α defined by Eq. (3.6) or Eq. (3.2) is a general statistic. The accuracy of the ES estimator after the adjustment will be discussed in the next two sections.*

For a sample of losses $\mathbf{Y} = \{y_n\}_{n=1}^N$, its conditional skewness γ_α is calculated by

$$\gamma_\alpha \equiv \frac{\frac{\sum_{n=1}^N (y_n - A_\alpha)^3 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}}{\left[\frac{\sum_{n=1}^N (y_n - A_\alpha)^2 \mathbb{1}_{\{y_n > A_\alpha\}})}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}\right]^{\frac{3}{2}}}, \tag{3.6}$$

where A_α is the sample's α-quantile given by Eq. (2.4).

The γ_α of a loss r.v. W has been defined by Eq. (3.2) which will be used later for error estimation and robust tests. The calculation in Eq. (3.2) depends on the third moment of W. If W follows a distribution without a finite third moment such as GPD(0.35,1), then the adjust tail-based ES estimator of W is not available. This limitation is negligible in most cases since our estimator is designed for small samples whose γ_α can always be obtained by Eq. (3.6).

For a given loss sample $\mathbf{Y}$, the algorithm for obtaining our regression-adjusted ES estimator from the tail-based normal approximation can be summarized as follows:

Algorithm for the adjusted tail-based normal ES$_\beta$ estimation

Step 1. Choose a value for α (e.g. $\alpha = 95\%$), and calculate A_α using Eq. (2.4);

Step 2. Obtain the tail-based normal r.v. X by Eq. (2.11) and Eq. (2.12);

Step 3. Compute the sample's conditional skewness γ_α by Eq. (3.6);

Step 4. Compute the adjustment factor $f_{\alpha,\beta}(\gamma_\alpha)$ by Eq. (3.4);

Step 5. Obtain the adjusted tail-based normal ES estimator at β level by Eq. (3.5).

4 Consistency and Robust Tests

Although the adjustment factor $R_{\alpha,\beta}$ is developed based on the student's t distributions, it works well for many other heavy-tailed distributions, such as Gamma, Lognormal, GPD, Weibull, etc. Therefore, the adjusted tail-based normal approximation is a simple and robust approach for ES estimations. In this section, we present some results on the consistency test and robust test.

4.1 Consistency Test

Before examining the accuracy of the adjusted tail-based normal approximation, a self-consistency test is conducted. Suppose the loss r.v. W itself follows a normal distribution; that is, $W \sim N(\mu, \sigma^2)$. Under this circumstance, the corresponding tail-based normal r.v. X obviously also follows $N(\mu, \sigma^2)$. Therefore, the adjustment based on the tail weight is not necessary; that is, the adjustment factor $R_{\alpha,\beta}$ should be close to 1. Next, let us verify the result by considering an arbitrary normal distribution.

Define $z_\alpha \equiv \Phi^{-1}(\alpha)$, and $q_\alpha \equiv \frac{1}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{z_\alpha^2}{2}}$. Let A_α denote the α-quantile of W. We have

$$A_\alpha = \mu + \sigma z_\alpha. \tag{4.1}$$

Moreover, we have (see Appendix A for details)

$$\begin{aligned}
\mathbb{E}[W\ \ |W > A_\alpha] &= \mu + \sigma q_\alpha, & (4.2)\\
\mathbb{E}[W^2|W > A_\alpha] &= \mu^2 + \sigma^2 + \sigma(A_\alpha + \mu) q_\alpha, & (4.3)\\
\mathbb{E}[W^3|W > A_\alpha] &= \mu^3 + 3\mu\sigma^2 + [\sigma(A_\alpha^2 + \mu^2 + A_\alpha \mu) + 2\sigma^3] q_\alpha. & (4.4)
\end{aligned}$$

Define $s_2 \equiv \mathbb{E}[(W - A_\alpha)^2|W > A_\alpha]$, and $r_3 \equiv \mathbb{E}[(W - A_\alpha)^3|W > A_\alpha]$. By Eq. (4.1)-Eq. (4.4), s_2 and r_3 can be expanded as

$$\begin{aligned}
s_2 &= (\mu - A_\alpha)[\sigma q_\alpha + (\mu - A_\alpha)] + \sigma^2, & \frac{s_2}{\sigma^2} &= -z_\alpha(q_\alpha - z_\alpha) + 1,\\
r_3 &= (\mu - A_\alpha)^3 + 3(\mu - A_\alpha)\sigma^2 + [2\sigma^2 + (A_\alpha - \mu)^2]\sigma q_\alpha, & \frac{r_3}{\sigma^3} &= -z_\alpha^3 - 3z_\alpha + (2 + z_\alpha^2) q_\alpha.
\end{aligned}$$

It is proved that r_3/σ^3 and s_2/σ^2 only depend on α. Therefore, the square of the conditional skewness of W is derived by

$$\gamma_\alpha^2 = \frac{r_3^2}{s_2^3} = \frac{(r_3/\sigma^3)^2}{(s_2/\sigma^2)^3} = \frac{[-z_\alpha^3 - 3z_\alpha + (2 + z_\alpha^2) q_\alpha]^2}{[-z_\alpha(q_\alpha - z_\alpha) + 1]^3}. \tag{4.5}$$

Based on Eq. (4.5), γ_α, a constant that only depends on α, is independent of μ and σ^2.

Obviously, the original estimation error of the tail-based normal approximation is zero since X is a replication of W. For $\alpha = 95\%$, we have $\gamma_\alpha = 1.838$ by Eq. (4.5), $f_{\alpha,\beta_1}(\gamma_\alpha) = 1.0008$ and $f_{\alpha,\beta_2}(\gamma_\alpha) = 1.0009$ by Eq. (3.4) at $\beta_1 = 99\%$ and $\beta_2 = 99.5\%$ levels. Since $f_{\alpha,\beta}(\gamma_\alpha)$ is close to 1, the estimation error of the adjusted tail-based normal approximation is still close to zero. Therefore the adjusted tail-based normal ES estimation method is self-consistent.

4.2 Robust Test and Estimation Errors for Heavy-Tailed Loss Distributions

To test the accuracy of the proposed adjusted tail-based approximation, we calculate the ES estimates of some heavy-tailed distributions. The results will be compared with the accurate values derived from the distribution functions.

In particular, for a loss r.v. W with a known distribution, its corresponding tail-based normal r.v. X and conditional skewness, γ_α, can be determined respectively by Eq. (2.16)-Eq. (2.17) and Eq. (3.2). Apart from the original ES estimation error defined in Eq. (2.18), a measure of the regression-adjusted estimation error for our ES estimator at β level is now defined by

$$\widehat{\mathbf{e}}_\beta^t(\mathrm{ES}) \equiv \frac{\mathrm{ES}_\beta(W) - \widehat{\mathrm{ES}}_\beta(X)}{\mathrm{ES}_\beta(W)}, \tag{4.6}$$

where $\widehat{\mathrm{ES}}_\beta(X)$, the β-level adjusted tail-based normal approximation, is given in Eq. (3.5).

The results are summarized in Table 3, in which the original and adjusted errors are both reported for various heavy-tailed loss distributions. For each kind of distributions, three different parameter sets are chosen and the most heavy-tailed one is listed first.

W	γ_α	$\text{ES}_{\beta_1}(W)$	$\mathbf{e}^t_{\beta_1}(\text{ES})$ (%)	$\widehat{\mathbf{e}}^t_{\beta_1}(\text{ES})$ (%)	$\text{ES}_{\beta_2}(W)$	$\mathbf{e}^t_{\beta_2}(\text{ES})$ (%)	$\widehat{\mathbf{e}}^t_{\beta_2}(\text{ES})$ (%)
t, df=3.5	7.181	5.895	-4.848	-0.028	7.290	3.152	-0.036
t, df=5	3.165	4.452	-0.919	-0.003	5.250	3.924	-0.004
t, df=8	2.359	3.591	0.121	-0.001	4.083	2.770	-0.001
Gamma(5,1)	1.998	13.001	0.225	0.091	13.956	0.977	0.142
Gamma(3,1)	2.033	9.639	0.303	0.135	10.485	1.332	0.214
Gamma(0.3,1)	2.249	3.494	0.819	0.572	4.092	3.954	0.985
LogN(0,1)	3.902	15.228	-2.409	-0.161	18.971	5.598	1.178
LogN(0,0.9^2)	3.416	11.527	-1.316	0.104	14.059	5.417	1.116
LogN(0,0.3^2)	2.098	2.235	0.225	0.091	2.391	1.237	0.158
GPD(0.3, 1)	11.225	15.624	-7.747	-0.689	20.006	2.547	0.065
GPD(0.2, 1)	3.674	10.699	-1.726	0.062	13.034	4.933	0.672
GPD(0.1, 1)	2.571	7.610	0.179	0.274	8.874	4.121	0.652
Weibull(0.6, 1)	2.673	17.990	0.339	0.610	21.773	5.711	1.526
Weibull(0.9, 1)	2.192	6.801	0.584	0.352	7.739	2.936	0.612
Weibull(1.4, 1)	1.967	3.415	0.262	0.114	3.714	1.005	0.166
$\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 3: Regression-adjusted errors for various heavy-tailed loss distributions, $\alpha = 95\%$

Table 3 shows the adjusted tail-based normal approximation is much more accurate than the original one. The relative errors of the t distributions after adjustment are very small as expected since the adjustment factors are obtained from the t distribution.

Moreover, the adjusted tail-based normal approximations of other heavy-tailed distributions also have small relative errors ($-0.7\% \sim 1.6\%$), though the adjustment factors are initially developed from the t distribution. Therefore, the regression models work effectively for distributions other than the t distribution and the adjusted tail-based normal approximation is a robust ES estimation method for loss r.v.'s with various heavy-tailed distributions.

5 Expected Shortfall Estimators for Small Samples

To further test the accuracy of our method, the adjusted tail-based normal ES estimator is compared to the arithmetic average (AA) of excessive losses and the extreme value theory (EVT) ES estimator for small loss samples simulated from various heavy-tailed distributions.

5.1 Arithmetic Average of Excessive Losses

We adopt the AA estimator proposed by **?**: assuming $\tilde{Y}_{(1)}, \ \ldots, \ \tilde{Y}_{(N-1)}, \ \tilde{Y}_{(N)}$ is a sample of N losses sorted in the ascending order, its ES estimator at level β is defined as

$$\widetilde{\text{ES}}_\beta = \frac{\sum_{n=\lceil N\beta \rceil}^{N} \tilde{Y}_{(n)}}{N + 1 - \lceil N\beta \rceil}, \tag{5.1}$$

where $\lceil N\beta \rceil$ denotes the smallest integer greater than or equal to $N\beta$.

5.2 Extreme Value Theory Expected Shortfall Estimator

The EVT estimator of ES is also considered. Assuming $F(w)$ is the CDF for a loss r.v. W, the distribution function of excesses beyond a threshold v is defined by **?** as follows:

$$F_v(y) = Pr(W - v \leq y | W > v) = \frac{F(v+y) - F(v)}{1 - F(v)}. \tag{5.2}$$

According to **?**, as the threshold v increases, $F_v(y)$ converges to a GPD whose CDF and PDF are listed as below:

$$G_{\xi,\sigma}(y) = 1 - \left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi}, \quad g_{\xi,\sigma}(y) = \frac{1}{\sigma}\left(1 + \xi \frac{y}{\sigma}\right)^{-1/\xi - 1}, \quad \xi \neq 0, \ \sigma > 0, \tag{5.3}$$

where ξ is a shape parameter and σ is a scale parameter. Assuming there are n_v losses, $\{w_k\}_{k=1}^{n_v}$, greater than the threshold v, then the sequence $\{w_k - v\}_{k=1}^{n_v}$ will show Generalized Pareto behavior (**?**). If $\xi < 0$, an extra condition is $1 + \xi \frac{(w_i - v)}{\sigma} > 0$ for $i = 1, \ldots, n_v$. When $\xi = 0$, the GPD becomes the exponential distribution so $G_{\xi,\sigma}(y) = 1 - \exp\left(-\frac{y}{\sigma}\right)$ and $g_{\xi,\sigma}(y) = \frac{1}{\sigma}\exp\left(-\frac{y}{\sigma}\right)$, $\sigma > 0$. Accordingly, ξ and σ estimates in GPD are solutions that maximize its log-likelihood function:

$$(\xi^*, \sigma^*) = \arg\max_{\xi,\sigma} \sum_{k=1}^{n_v} \ln\left(g_{\xi,\sigma}(w_k - v)\right). \tag{5.4}$$

The estimation of GPD by MLE can be limited since for some samples the likelihood function appears to have no local maximum. **?** show that for $\xi < -1$, there never exists a local maximum and then the BIC or AIC criteria cannot be used to model selection.

In this paper, given a loss sample and the 95th percentile threshold, we calculate the maximum likelihood estimates of the GPD for all the three situations: $\xi < 0, \xi = 0$, and $\xi > 0$. Numerical algorithms[3] are used to obtain the optimal estimators in each situation. For $\xi > 0$, it is a unconstrained optimization problem and the Nelder-Mead simplex algorithm is applied. For $\xi = 0$, the GPD becomes the exponential distribution and the MLE has a analytical solution. For $\xi < 0$, the interior point algorithm is used with the strict constraint $\xi > -\frac{\sigma}{\max(w) - v}$ where v is the threshold and $\max(w)$ is the maximum value of the loss sample. Then the estimate that has the maximum likelihood in the three situations will be selected to fit the simulated loss sample.

Remark 3 *In practice, the threshold v is usually set to equal to the 95th percentile of the loss sample. We keep this selection so that the threshold v of the EVT estimator is same as A_α ($\alpha = 0.95$) in our adjusted tail-based ES estimator.*

$F_v(y)$ in Eq. (5.2) is a conditional probability if $W \leq v + y$ given $W > v$. The unconditional distribution of excesses beyond the threshold v is thus derived by

$$Pr(W \leq w) = (1 - F(v))G_{\xi,\sigma}(w - v) + F(v), \quad w > v. \tag{5.5}$$

When the parameters of the GPD are determined, the β-level ($\beta > F(v)$) EVT VaR and ES estimators of W can be derived based on Eq. (5.5). The EVT estimator for VaR at β level is

[3]The algorithms are implemented in MATLAB with 'fminsearch' and 'fmincon' functions. MATLAB also offers a function 'gpfit' to calculate the maximum likelihood estimates of the GPD with all situations considered. In practice, for all simulated samples, the estimates required by numerical algorithms are identical to the ones got by 'gpfit'.

calculated by solving $Pr(W \leq \mathrm{VaR}_\beta(W)) = \beta$ where the unconditional probability $Pr(W \leq \mathrm{VaR}_\beta(W))$ is given in Eq. (5.5). When $\xi \neq 0$, we have

$$(1 - F(v))\left(1 - \left[1 + \frac{\xi(\mathrm{VaR}_\beta(W) - v)}{\sigma}\right]^{-1/\xi}\right) + F(v) = \beta. \tag{5.6}$$

Thus $\mathrm{VaR}_\beta(W) = v + \frac{\sigma}{\xi}[(\frac{1-\beta}{1-F(v)})^{-\xi} - 1]$. Since $W - v \sim \mathrm{GPD}(\xi, \sigma)$, its ES estimate is given by

$$\begin{aligned}\mathrm{ES}_\beta(W) &= \mathbb{E}[W|W \geq \mathrm{VaR}_\beta(W)] = v + \mathbb{E}[W - v|W - v \geq \mathrm{VaR}_\beta(W) - v] \\ &= v + \mathrm{VaR}_\beta(W) - v + \frac{\sigma + \xi\mathrm{VaR}_\beta(W) - \xi v}{1-\xi} = \frac{\mathrm{VaR}_\beta(W) + \sigma - \xi v}{1-\xi},\end{aligned} \tag{5.7}$$

where Eq. (5.7) is obtained by GPD's ES closed-form formula which can be found in Table 11 (see Appendix B). When $\xi = 0$, we have

$$(1 - F(v))\left(1 - e^{-\frac{1}{\sigma}(\mathrm{VaR}_\beta(W) - v)}\right) + F(v) = \beta. \tag{5.8}$$

Thus $\mathrm{VaR}_\beta(W) = v - \sigma\ln(\frac{1-\beta}{1-F(v)})$. Now $W - v$ follows an exponential distribution with parameter σ and we have the following derivation based on PDF of the exponential distribution:

$$\begin{aligned}&\mathbb{E}[W - v|W - v \geq b] = \frac{1}{1 - G_\sigma(b)}\int_b^\infty \frac{y}{\sigma}e^{-\frac{y}{\sigma}}dy = \mathrm{VaR}_\beta(W) - v + \sigma, \\ &\mathrm{ES}_\beta(W) = v + \mathbb{E}[W - v|W - v \geq b] = \mathrm{VaR}_\beta(W) + \sigma.\end{aligned} \tag{5.9}$$

where $b = \mathrm{VaR}_\beta(W) - v$ and $G_\sigma(b) = Pr(W - v \leq b) = 1 - e^{-b/\sigma}$. In summary, given a loss sample, the EVT ES estimate is calculated as follows:

- Calculate the threshold v using the 95th percentile of the loss sample.
- Determine the parameters of the GPD by maximizing the likelihood function for all the three situations: $\xi < 0$, $\xi = 0$ and $\xi > 0$.
- Calculate the β-level ES estimate $\mathrm{ES}_\beta(\mathrm{EVT})$ based on Table 4. If the loss sample size is N, $1 - F(v)$ can be approximated by n_v/N (**?**).

	$\mathrm{VaR}_\beta(\mathrm{EVT})$	$\mathrm{ES}_\beta(\mathrm{EVT})$
$\xi \neq 0$	$v + \frac{\sigma}{\xi}\left[\left(\frac{1-\beta}{1-F(v)}\right)^{-\xi} - 1\right]$	$\frac{\mathrm{VaR}_\beta(\mathrm{EVT}) + \sigma - \xi v}{1-\xi}$
$\xi = 0$	$v - \sigma\ln\left(\frac{1-\beta}{1-F(v)}\right)$	$\mathrm{VaR}_\beta(\mathrm{EVT}) + \sigma$

Table 4: EVT estimators for VaR and ES at β-level

5.3 Comparison between the Three Expected Shortfall Estimators

To further validate the adjusted tail-based normal approximation, we implement the Monte Carlo simulation and compare our ES estimates to the ones of AA and EVT methods. 2500 random loss samples with equal sample size (250 or 500) are generated from each examined distribution. For each simulated sample, the three ES estimates at 99% and 99.5% levels are computed, respectively. The evaluation of the estimates are based on the following three metrics:

$$\text{MSE of ES}_{\beta,j} = \frac{1}{M}\sum_{i=1}^{M}\left(\text{ES}_{\beta,j}(i) - \text{true ES}_{\beta}\right)^2, \quad j = 1,2,3, \tag{5.10}$$

$$\text{Variance of ES}_{\beta,j} = \frac{1}{M}\sum_{i=1}^{M}\left[\text{ES}_{\beta,j}(i) - \frac{1}{M}\sum_{i=1}^{M}\text{ES}_{\beta,j}(i)\right]^2, \quad j = 1,2,3, \tag{5.11}$$

$$\text{Bias of ES}_{\beta,j} = \frac{1}{M}\sum_{i=1}^{M}\text{ES}_{\beta,j}(i) - \text{true ES}_{\beta}, \quad j = 1,2,3, \tag{5.12}$$

where 'true ES_β' is the theoretical β-level ES value of the underlying distribution, $j = 1$ stands for our adjusted tail-based normal approximation, $j = 2$ stands for the AA method, $j = 3$ stands for the EVT method and i is the sample index of all M samples. In practice, the extremely skewed sample that leads to a MLE EVT paramters with $\xi > 0.65$ is discarded so M is slightly less than 2500. The MSE of any estimator is equal to the sum of its variance and square of bias.

The results are summarized in Table 5 - Table 9, in which our adjusted tail-based normal approximation ES estimator is denoted by $\widehat{\text{ES}}_\beta$, and the best metric of each column for a given underlying distribution is highlighted with an underscore. The true ES values of the underlying distributions in Table 5 - Table 9 can be found in Table 3.

From those tables, we can see that $\widehat{\text{ES}}_\beta$ outperforms the other two in terms of MSE for $\beta = 99\%$ or 99.5% for almost all the underlying distributions. In particular, it performs the best for all scenarios with a smaller sample size ($n = 250$) or a higher confidence level (99.5%). For scenarios with sample size $n = 500$ and confidence level 99%, the proposed ES estimator $\widehat{\text{ES}}_\beta$ outperforms

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
t, df=3.5	$\widehat{\text{ES}}$	<u>2.514</u>	<u>5.243</u>	<u>2.280</u>	<u>3.897</u>	-0.484	-1.160
	AA	3.080	7.426	3.053	6.968	<u>-0.162</u>	<u>-0.676</u>
	EVT	2.908	7.358	2.809	6.879	-0.315	-0.692
t, df=5	$\widehat{\text{ES}}$	<u>0.949</u>	<u>1.821</u>	<u>0.901</u>	<u>1.521</u>	-0.219	-0.548
	AA	1.214	2.687	1.213	2.616	<u>-0.028</u>	-0.266
	EVT	1.071	2.562	1.060	2.492	-0.106	<u>-0.265</u>
t, df=8	$\widehat{\text{ES}}$	<u>0.356</u>	<u>0.645</u>	<u>0.340</u>	<u>0.552</u>	-0.126	-0.305
	AA	0.445	0.910	0.445	0.888	<u>-0.010</u>	-0.147
	EVT	0.367	0.791	0.366	0.779	-0.032	<u>-0.107</u>
size=500		β_1	β_2	β_1	β_2	β_1	β_2
t, df=3.5	$\widehat{\text{ES}}$	1.683	<u>3.561</u>	1.605	<u>2.970</u>	-0.279	-0.769
	AA	<u>1.511</u>	4.904	<u>1.449</u>	4.878	-0.248	<u>-0.161</u>
	EVT	1.795	4.758	1.770	4.674	<u>-0.157</u>	-0.289
t, df=5	$\widehat{\text{ES}}$	0.533	<u>1.068</u>	<u>0.517</u>	<u>0.946</u>	-0.129	-0.348
	AA	<u>0.530</u>	1.583	0.518	1.583	-0.108	<u>-0.002</u>
	EVT	0.589	1.485	0.585	1.473	<u>-0.063</u>	-0.109
t, df=8	$\widehat{\text{ES}}$	<u>0.193</u>	<u>0.366</u>	<u>0.187</u>	<u>0.331</u>	-0.077	-0.187
	AA	0.201	0.521	0.197	0.521	-0.060	<u>-0.001</u>
	EVT	0.194	0.434	0.194	0.434	<u>-0.008</u>	-0.006
loss samples from t distribution, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 5: Comparisons of three ES estimators, t distributed samples

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
Gamma(5, 1)	$\widehat{\text{ES}}$	1.247	2.164	1.155	1.785	-0.303	-0.616
	AA	1.507	2.722	1.500	2.587	-0.080	-0.367
	EVT	1.363	2.885	1.296	2.574	-0.260	-0.558
Gamma(3, 1)	$\widehat{\text{ES}}$	1.018	1.788	0.946	1.483	-0.268	-0.552
	AA	1.243	2.311	1.237	2.200	-0.076	-0.334
	EVT	1.157	2.567	1.110	2.342	-0.217	-0.475
Gamma(0.3, 1)	$\widehat{\text{ES}}$	0.510	0.912	0.485	0.780	-0.155	-0.364
	AA	0.641	1.256	0.641	1.223	-0.002	-0.179
	EVT	0.568	1.330	0.560	1.266	-0.089	-0.253
size=500		β_1	β_2	β_1	β_2	β_1	β_2
Gamma(5, 1)	$\widehat{\text{ES}}$	0.649	1.145	0.610	1.001	-0.197	-0.379
	AA	0.700	1.488	0.672	1.484	-0.166	-0.059
	EVT	0.733	1.568	0.699	1.479	-0.185	-0.299
Gamma(3, 1)	$\widehat{\text{ES}}$	0.525	0.939	0.491	0.812	-0.185	-0.356
	AA	0.568	1.207	0.543	1.201	-0.156	-0.076
	EVT	0.586	1.271	0.558	1.199	-0.168	-0.269
Gamma(0.3, 1)	$\widehat{\text{ES}}$	0.265	0.488	0.255	0.435	-0.101	-0.229
	AA	0.281	0.673	0.276	0.673	-0.069	0.010
	EVT	0.320	0.749	0.316	0.741	-0.061	-0.089
loss samples from Gamma distribution, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 6: Comparisons of three ES estimators, Gamma distributed samples

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
LogN(0, 1)	$\widehat{\text{ES}}$	18.295	37.418	16.619	28.010	-1.294	-3.067
	AA	22.731	51.119	22.570	48.008	-0.400	-1.764
	EVT	22.898	57.597	22.197	54.309	-0.837	-1.813
LogN(0, 0.9^2)	$\widehat{\text{ES}}$	8.444	16.871	7.762	13.003	-0.826	-1.967
	AA	10.649	23.047	10.608	21.858	-0.204	-1.091
	EVT	10.355	25.523	10.072	24.147	-0.532	-1.173
LogN(0, 0.3^2)	$\widehat{\text{ES}}$	0.035	0.062	0.034	0.054	-0.035	-0.086
	AA	0.045	0.084	0.045	0.083	0.003	-0.041
	EVT	0.042	0.084	0.042	0.082	-0.004	-0.046
size=500		β_1	β_2	β_1	β_2	β_1	β_2
LogN(0, 1)	$\widehat{\text{ES}}$	10.809	22.928	10.190	18.685	-0.787	-2.060
	AA	10.122	31.278	9.690	31.139	-0.657	-0.373
	EVT	12.792	34.110	12.682	33.862	-0.332	-0.498
LogN(0, 0.9^2)	$\widehat{\text{ES}}$	4.959	10.263	4.726	8.640	-0.483	-1.274
	AA	4.778	14.412	4.630	14.396	-0.384	-0.128
	EVT	5.992	15.840	5.955	15.761	-0.194	-0.281
LogN(0, 0.3^2)	$\widehat{\text{ES}}$	0.017	0.031	0.016	0.028	-0.024	-0.054
	AA	0.019	0.044	0.018	0.044	-0.018	0.003
	EVT	0.016	0.034	0.016	0.034	0.002	0.010
loss samples from Lognormal distribution, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 7: Comparisons of three ES estimators, Lognormal distributed samples

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
GPD(0.3, 1)	$\widehat{\text{ES}}$	21.634	49.003	17.271	29.052	-2.089	-4.467
	AA	24.628	59.532	23.204	50.409	-1.193	-3.021
	EVT	24.767	62.953	22.095	53.099	-1.635	-3.139
GPD(0.2, 1)	$\widehat{\text{ES}}$	7.246	14.638	6.518	10.995	-0.853	-1.909
	AA	8.947	20.088	8.857	18.984	-0.300	-1.050
	EVT	8.774	21.755	8.403	20.304	-0.609	-1.205
GPD(0.1, 1)	$\widehat{\text{ES}}$	2.207	4.183	2.037	3.368	-0.412	-0.903
	AA	2.772	5.752	2.762	5.542	-0.097	-0.458
	EVT	2.582	6.060	2.488	5.678	-0.307	-0.618
size=500		β_1	β_2	β_1	β_2	β_1	β_2
GPD(0.3, 1)	$\widehat{\text{ES}}$	15.195	34.398	13.513	24.882	-1.297	-3.085
	AA	13.469	42.283	11.991	40.732	-1.216	-1.245
	EVT	15.862	42.180	15.145	40.157	-0.847	-1.422
GPD(0.2, 1)	$\widehat{\text{ES}}$	4.306	8.990	4.010	7.302	-0.544	-1.299
	AA	4.124	12.175	3.910	12.099	-0.463	-0.275
	EVT	4.900	12.628	4.802	12.401	-0.312	-0.476
GPD(0.1, 1)	$\widehat{\text{ES}}$	1.264	2.471	1.196	2.135	-0.259	-0.580
	AA	1.286	3.499	1.244	3.497	-0.204	-0.044
	EVT	1.532	3.841	1.508	3.794	-0.154	-0.216
loss samples from GPD, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 8: Comparisons of three ES estimators, General Pareto distributed samples

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
Weibull(0.6, 1)	$\widehat{\text{ES}}$	19.342	37.144	17.652	29.131	-1.300	-2.831
	AA	24.114	50.042	23.998	47.898	-0.341	-1.464
	EVT	23.360	55.835	22.529	52.720	-0.912	-1.765
Weibull(0.9, 1)	$\widehat{\text{ES}}$	1.249	2.248	1.168	1.881	-0.285	-0.606
	AA	1.576	3.005	1.574	2.914	-0.049	-0.302
	EVT	1.453	3.281	1.409	3.094	-0.210	-0.432
Weibull(1.4, 1)	$\widehat{\text{ES}}$	0.131	0.226	0.124	0.195	-0.084	-0.176
	AA	0.164	0.297	0.164	0.289	-0.014	-0.091
	EVT	0.151	0.319	0.151	0.319	0.012	0.004
size=500		β_1	β_2	β_1	β_2	β_1	β_2
Weibull(0.6, 1)	$\widehat{\text{ES}}$	10.999	21.795	10.273	18.209	-0.852	-1.894
	AA	11.155	29.882	10.720	29.822	-0.659	-0.245
	EVT	13.468	34.001	13.277	33.707	-0.438	-0.543
Weibull(0.9, 1)	$\widehat{\text{ES}}$	0.660	1.221	0.621	1.063	-0.197	-0.396
	AA	0.702	1.675	0.678	1.674	-0.154	-0.032
	EVT	0.782	1.825	0.761	1.784	-0.143	-0.201
Weibull(1.4, 1)	$\widehat{\text{ES}}$	0.067	0.117	0.064	0.105	-0.058	-0.110
	AA	0.073	0.157	0.071	0.157	-0.049	-0.009
	EVT	0.070	0.144	0.069	0.139	0.029	0.077
loss samples from Weibull distribution, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 9: Comparisons of three ES estimators, Weibull distributed samples

the other two in terms of MSE for most of the scenarios. For scenarios where $\widehat{\text{ES}}_\beta$ is not the best in terms of MSE, its MSEs are comparable to the ones of the other two estimators.

To further investigate the performance, we also calculate the variances and biases for all three ES estimators and the results are given in Table 5 - Table 9. As we can tell from the tables, the superiority in the MSE of the proposed ES estimator $\widehat{\text{ES}}_\beta$ actually comes from the smaller variance. It indicates our ES estimator is more robust to the small samples than the other two estimators. In terms of bias which reveals the difference between the estimator's expected value and the true value, the AA estimator is preferred for most underlying distributions. However, if the sample size is small, the AA estimator is easily influenced by whether the infrequent data would occur or not in the sample. Therefore, the AA estimator for a small sample is unstable with a large variance. After all, it only uses a small percentage ($1-\beta$, with $\beta = 99\%$ or 99.5%) of all the sample points while our ES estimator always incorporates much more data—the largest 5% ($1-\alpha$, with $\alpha = 95\%$) of losses whatever β is.

The EVT estimator is similar to the AA estimator in terms of the variance and bias. Like the AA estimator, an extreme value in a small sample will have a big effect on the estimate. Therefore, it is very unstable which is reflected by the large variance in Table 5 - Table 9. Regarding the MSE, it is less robust than our ES estimator though both of them utilize the largest 5% of the sample points within the distribution fitting. Another disadvantage for the EVT ES estimator is that it is not easy to derive the MLE of a GPD for small samples when computing the EVT estimates.

Another thing we can observe from the tables is that, for the three estimators, their biases are almost all negative. When the sample size is small, the infrequent extreme values barely appear in the sample. Therefore, all three estimators tend to underestimate the ES, especially in the high confidence level ($\beta = 99.5\%$) case. This is a common phenomenon for sample-based ES estimators, especially for small samples or high confidence levels.

5.4 Comparison for Distributions without the Third Moment

As mentioned in Remark 1, one limitation of our proposed ES estimator is that the third moment of the underlying distribution is needed. Therefore, for distributions without the third moment, it seems the proposed method may not work. However, in reality, with a given data set, the third moment of the sample always exists, although it might be very large. In this subsection, we investigate the performance of the proposed method by comparing it with the performance of the AA ES estimator and the EVT ES estimator.

In particular, we consider the student's t distributions with degrees of freedom equal to 2.5 or 3, and the GPD with $\xi = 0.35$ or 0.5. For those distributions, their third moments do not exist. We draw samples from those distributions and apply the three ES estimators. The numerical results are given in Table 10. For reference purpose, we list their true ES values as follows: $\text{ES}_{99\%} = 9.091$, $\text{ES}_{99.5\%} = 12.067$ for t with df=2.5; $\text{ES}_{99\%} = 7.003$, $\text{ES}_{99.5\%} = 8.913$ for t with df=3; $\text{ES}_{99\%} = 38.000$, $\text{ES}_{99.5\%} = 54.569$ for GPD(0.5,1) and $\text{ES}_{99\%} = 19.173$, $\text{ES}_{99.5\%} = 25.222$ for GPD(0.35,1).

Table 10 shows the proposed ES estimator is still the best one in terms of MSE for most of the scenarios, especially for the smaller sample size (250) or the higher confidence level (99.5%) scenario. Therefore, even for distributions without skewness (the third moment), the proposed ES estimation method still gives decent results, compared to the AA and EVT ES estimators.

Another thing we need to mention is that, in reality, the skewness of the financial market data is not that large, and the proposed ES estimation method should work well.

		MSE		Var		Bias	
size=250		β_1	β_2	β_1	β_2	β_1	β_2
	$\widehat{\text{ES}}$	9.237	22.882	6.749	11.310	-1.577	-3.402
t, df=2.5	AA	10.199	25.784	9.125	18.964	-1.036	-2.611
	EVT	10.316	27.173	8.649	20.547	-1.291	-2.574
	$\widehat{\text{ES}}$	4.290	9.646	3.588	6.120	-0.838	-1.878
t, df = 3	AA	5.010	12.508	4.808	10.903	-0.449	-1.267
	EVT	4.962	12.942	4.567	11.288	-0.629	-1.286
	$\widehat{\text{ES}}$	259.758	753.222	132.640	223.358	-11.275	-23.019
GPD(0.5, 1)	AA	257.813	755.729	176.495	388.803	-9.018	-19.155
	EVT	269.263	771.402	173.239	422.384	-9.799	-18.682
	$\widehat{\text{ES}}$	38.150	92.495	27.710	46.481	-3.231	-6.783
GPD(0.35, 1)	AA	41.887	104.238	37.500	79.897	-2.095	-4.934
	EVT	42.385	109.786	35.550	84.892	-2.614	-4.989
size=500		β_1	β_2	β_1	β_2	β_1	β_2
	$\widehat{\text{ES}}$	6.570	16.347	5.414	9.980	-1.076	-2.523
t, df = 2.5	AA	5.664	18.168	4.565	16.193	-1.048	-1.405
	EVT	6.262	17.002	5.569	14.679	-0.832	-1.524
	$\widehat{\text{ES}}$	2.774	6.430	2.459	4.574	-0.561	-1.362
t, df = 3	AA	2.439	7.894	2.161	7.548	-0.528	-0.588
	EVT	2.828	7.667	2.672	7.163	-0.394	-0.710
	$\widehat{\text{ES}}$	180.212	550.677	98.948	181.718	-9.015	-19.208
GPD(0.5, 1)	AA	163.560	493.845	82.473	289.897	-9.005	-14.281
	EVT	165.789	482.412	107.032	284.871	-7.665	-14.055
	$\widehat{\text{ES}}$	29.115	69.017	24.920	46.044	-2.048	-4.793
GPD(0.35, 1)	AA	24.808	79.719	20.853	74.092	-1.989	-2.372
	EVT	28.258	76.334	26.118	69.869	-1.463	-2.543
loss samples from other distributions, $\beta_1 = 99\%$, $\beta_2 = 99.5\%$							

Table 10: Comparisons of three ES estimators, other distributed samples

6 Effects of Linear Transformations

In this section, we consider the effects of a linear transformation of the loss sample or loss r.v.'s. Linear transformations are very common in portfolio management and risk management, such as exchanging from one currency to another, or changing of measuring units, etc.

Assuming an underlying distribution is transformed linearly with a positive scale multiplier—m and a constant summand—c, then any random loss sample generated from it will be subject to the same transformation. In what follows, we use the superscript τ for the new variables after transformation. Assume that the transformation is in forms of $\mathbf{Y}^\tau = m\mathbf{Y} + c$, where $\mathbf{Y}$ represents the original loss sample (or the original loss r.v.).

The three β-level ES estimates for the i-th original sample are denoted by $\{\text{ES}_{\beta,j}(i)\}_{j=1}^3$ ($j = 1$ for the adjusted tail-based normal approximation, $j = 2$ for AA, and $j = 3$ for EVT). As for the corresponding i-th linearly-transformed sample, the new estimators are denoted by $\{\text{ES}^\tau_{\beta,j}(i)\}_{j=1}^3$. Apparently, the β-level theoretical ES estimate for the new sample is:

$$\text{true ES}^\tau_\beta = m(\text{true ES}_\beta) + c.$$

First, let us investigate the effects of the linear transformation on the adjusted tail-based normal estimator. We have the following result:

Proposition 1 *If a loss sample or loss r.v. is transformed linearly with a positive multiplier, the* *corresponding tail-based normal r.v. X and the regression-adjusted tail-based normal approximation ES estimator are subject to the same linear transformation. Moreover, the conditional skewness stays unchanged.*

Proof: Suppose the original loss sample is denoted by $\{y_n\}_{n=1}^N$ whose α-quantile is A_α and conditional skewness is γ_α. Undergoing a linear transformation, the loss sample becomes a new one denoted by $\{y_n^\tau\}_{n=1}^N$ such that $y_n^\tau = my_n + c$ for $n = 1, \dots, N$ where m, c are both constants and $m > 0$. The α-quantile for the new sample is then $mA_\alpha + c$ and the new conditional skewness is

$$\gamma_\alpha^\tau = \frac{\frac{\sum_{n=1}^N (y_n^\tau - mA_\alpha - c)^3 \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}}{\left[\frac{\sum_{n=1}^N (y_n^\tau - mA_\alpha - c)^2 \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}\right]^{\frac{3}{2}}} = \frac{m^3 \left[\frac{\sum_{n=1}^N (y_n - A_\alpha)^3 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}\right]}{m^3 \left[\frac{\sum_{n=1}^N (y_n - A_\alpha)^2 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}\right]^{\frac{3}{2}}} = \gamma_\alpha. \tag{6.1}$$

That is, after the linear transformation, γ_α^τ is equal to γ_α. Furthermore, by Eq. (2.11) and Eq. (2.12), the parameters of the original tail-based normal r.v. $X \sim N(\mu, \sigma^2)$ is solved by

$$\begin{aligned}\sigma &= \left([\Phi^{-1}(\alpha)]^2 + 1 - \frac{\Phi^{-1}(\alpha)}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\alpha))^2}\right)^{-\frac{1}{2}} \left[\frac{\sum_{n=1}^N (y_n - A_\alpha)^2 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}\right]^{\frac{1}{2}}, \\ \mu &= A_\alpha - \sigma\Phi^{-1}(\alpha).\end{aligned} \tag{6.2}$$

Assuming the tail-based normal r.v. for $\{y_n^\tau\}_{n=1}^N$ is $X^\tau \sim N(\mu^\tau, (\sigma^\tau)^2)$, we have

$$\begin{aligned}\sigma^\tau &= \left([\Phi^{-1}(\alpha)]^2 + 1 - \frac{\Phi^{-1}(\alpha)}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\alpha))^2}\right)^{-\frac{1}{2}} \left[\frac{\sum_{n=1}^N (y_n^\tau - mA_\alpha - c)^2 \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n^\tau > mA_\alpha + c\}}}\right]^{\frac{1}{2}} \\ &= \left([\Phi^{-1}(\alpha)]^2 + 1 - \frac{\Phi^{-1}(\alpha)}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{1}{2}(\Phi^{-1}(\alpha))^2}\right)^{-\frac{1}{2}} \left[\frac{m^2 \sum_{n=1}^N (y_n - A_\alpha)^2 \mathbb{1}_{\{y_n > A_\alpha\}}}{\sum_{n=1}^N \mathbb{1}_{\{y_n > A_\alpha\}}}\right]^{\frac{1}{2}}.\end{aligned}$$

Therefore, $\sigma^\tau = m\sigma$ and $\mu^\tau = mA_\alpha + c - m\sigma\Phi^{-1}(\alpha) = m\mu + c$. So we have $X^\tau \sim N(m\mu + c, m^2\sigma^2)$ and X^τ follows the same distribution as $mX + c$. By Eq. (3.5), the new β-level ES estimator after the transformation is derived as follows:

$$\begin{aligned}\widehat{\mathrm{ES}}_\beta(X^\tau) &= [\mathrm{ES}_\beta(X^\tau) - mA_\alpha - c] f_{\alpha,\beta}(\gamma_\alpha^\tau) + mA_\alpha + c \\ &= [m\mathrm{ES}_\beta(X) + c - mA_\alpha - c] f_{\alpha,\beta}(\gamma_\alpha) + mA_\alpha + c \\ &= m[\mathrm{ES}_\beta(X) - A_\alpha] f_{\alpha,\beta}(\gamma_\alpha) + mA_\alpha + c = m\widehat{\mathrm{ES}}_\beta(X) + c.\end{aligned} \tag{6.3}$$

where $\mathrm{ES}_\beta(X^\tau) = m\mathrm{ES}_\beta(X) + c$ comes from Eq. (2.14). The other scenario that a loss r.v. W is transformed linearly can be proved similarly. **Q.E.D.**

From Proposition 1, we have $\mathrm{ES}_{\beta,1}^\tau(i) = m\,\mathrm{ES}_{\beta,1}(i) + c$ for our proposed estimator. As for the AA estimator, $\mathrm{ES}_{\beta,2}^\tau(i)$ equals $m\,\mathrm{ES}_{\beta,2}(i) + c$ obviously. Next, let us consider the effects on the EVT estimator. In particular, we examine whether $\mathrm{ES}_{\beta,3}^\tau(i)$ equals $m\,\mathrm{ES}_{\beta,3}(i) + c$.

Suppose $\{w_k\}_{k=1}^{n_v}$ denote the n_v losses that are greater than the threshold v in the original sample. After the linear transformation, the new threshold is $m\,v + c$ and losses beyond it become $\{m\,w_k + c\}_{k=1}^{n_v}$. Assuming $(\hat{\xi}, \hat{\sigma})$ are the maximum likelihood estimation (MLE) GPD parameters of the i-th original sample, we have

$$(\hat{\xi}, \hat{\sigma}) = \arg\max_{\xi,\sigma} \sum_{k=1}^{n_v} -\ln\sigma - \left(\frac{1}{\xi} + 1\right) \ln\left(1 + \frac{\xi}{\sigma}(w_k - v)\right), \quad \xi \neq 0. \tag{6.4}$$

Suppose $(\hat{\xi}^\tau, \hat{\sigma}^\tau)$ are the MLE GPD parameters after the linear transformation. Then they satisfy:

$$
\begin{aligned}
(\hat{\xi}^\tau, \hat{\sigma}^\tau) &= \arg\max_{\xi,\sigma} \sum_{k=1}^{n_v} -\ln\sigma - \left(\frac{1}{\xi}+1\right)\ln\left(1+\frac{\xi}{\sigma}(m\,w_k + c - m\,v - c)\right), \quad \xi \neq 0 \\
&= \arg\max_{\xi,\sigma} \sum_{k=1}^{n_v} -\ln(\sigma/m) - \left(\frac{1}{\xi}+1\right)\ln\left(1+\frac{\xi}{\sigma/m}(w_k - v)\right), \quad \xi \neq 0.
\end{aligned} \tag{6.5}
$$

Comparing Eq. (6.5) to Eq. (6.4), we have $\hat{\xi}^\tau = \hat{\xi}$ and $\hat{\sigma}^\tau = m\,\hat{\sigma}$. According to Table 4, the EVT β-level ES estimator ($\xi \neq 0$) for the sample after that linear transformation is

$$
\begin{aligned}
\mathrm{ES}^\tau_{\beta,3}(i) &= \frac{m\,v + c + \frac{\hat{\sigma}^\tau}{\hat{\xi}^\tau}\left[(\frac{1-\beta}{n_v/N})^{-\hat{\xi}^\tau} - 1\right] + \hat{\sigma}^\tau - \hat{\xi}^\tau(m\,v + c)}{1-\hat{\xi}^\tau} \\
&= \frac{m\,v + m\frac{\hat{\sigma}}{\hat{\xi}}\left[(\frac{1-\beta}{n_v/N})^{-\hat{\xi}} - 1\right] + m\hat{\sigma} - m\hat{\xi}v + (1-\hat{\xi})c}{1-\hat{\xi}} = m\,\mathrm{ES}_{\beta,3}(i) + c.
\end{aligned} \tag{6.6}
$$

The other situation when $\xi = 0$ can be demonstrated similarly. Therefore, all three ES estimators are subject to the same linear transformation as the loss sample. Furthermore, we have

$$
\begin{aligned}
(\mathrm{ES}^\tau_{\beta,j}(i) - \text{true } \mathrm{ES}^\tau_\beta)^2 &= (m\mathrm{ES}_{\beta,j}(i) + c - m(\text{true } \mathrm{ES}_\beta) - c)^2 \\
&= m^2(\mathrm{ES}_{\beta,j}(i) - \text{true } \mathrm{ES}_\beta)^2, \quad j = 1,2,3, \quad i = 1,2,\ldots,M.
\end{aligned} \tag{6.7}
$$

Obviously, the MSE of the new sample are proportional to the original one with a constant multiple m^2 based on Eq. (6.7). We can prove this property also holds for the variance and the square of bias similarly. Therefore, the comparison results among the three estimators in terms of MSE, variance, and bias are maintained after a linear transformation ($m > 0$) is applied to the underlying distribution or the loss samples. For example, the advantage of our proposed ES estimator still exists after the linear transformation. Moreover, this property ensures it is effective in Section 5 to only consider the situation that the scale and locations are equal to 1 and 0, respectively.

7 Conclusion

In this paper, we propose a simple and robust ES estimation method based on the tail-based normal approximation. The regression model related to a sample's tail weight is also introduced to make the estimations more accurate. For various heavy-tailed loss distributions, the regression-adjusted ES estimation errors are all sufficiently small. Moreover, compared to the commonly-used arithmetic average or EVT estimators, our proposed ES estimator is preferred in terms of MSE at high levels such as 99% and 99.5% for small samples. It also shows that our method works well under linear transformations, which further adds to its practicality in the portfolio management.

Nonetheless, we only consider the scenario that $\beta = 99\%$ or 99.5% with $\alpha = 95\%$ and other combinations may have better performances. It is also possible that the proposed regression models would not perform as well for a sample with an extremely large conditional skewness ($\gamma_\alpha > 12$) though that situation is quite rare in practice. Instead of fitting the tailed data, the tail-based normal approximation matches the specific statistics of excessive losses. That normal approximation itself cannot describe the tail behaviors correctly and needs to work jointly with the regression model to give the ES estimate. Furthermore, MSE, variance, and bias are used to evaluate the ES estimator but the underestimation error should be paid more attention. Hence, an asymmetric loss function may be designed to better evaluate the estimators. These deficiencies are worth looking into in the future.

Acknowledgements and Declaration of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Appendix A Derivations

Suppose $X \sim N(\mu, \sigma^2)$ and $Pr(X \leq A_\alpha) = \alpha$. Let $Z \equiv \frac{X-\mu}{\sigma}$ and $a \equiv \frac{A_\alpha - \mu}{\sigma}$. Then $Z \sim N(0,1)$ and $\Phi(a) = \Phi\left(\frac{A_\alpha-\mu}{\sigma}\right) = Pr\left(Z \leq \frac{A_\alpha-\mu}{\sigma}\right) = Pr(X \leq A_\alpha) = \alpha$, where $\Phi(\cdot)$ is the CDF of the standard normal distribution. So we can get

$$
\begin{aligned}
\mathbb{E}[X|X > A_\alpha] &= \sigma \mathbb{E}\left[Z|Z > a\right] + \mu = \mu + \frac{\sigma}{1-\Phi(a)} \int_a^\infty z \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \mu + \frac{\sigma}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{a^2}{2}} = \mu + \frac{\sigma}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{(A_\alpha-\mu)^2}{2\sigma^2}}, \\
\mathbb{E}[X^2|X > A_\alpha] &= \frac{1}{Pr(X > A_\alpha)} \int_{A_\alpha}^\infty \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \frac{1}{1-\alpha} \int_a^\infty \frac{(z\sigma+\mu)^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \frac{1}{(1-\alpha)\sqrt{2\pi}} \int_a^\infty (z^2\sigma^2 + 2\mu\sigma z + \mu^2) e^{-\frac{z^2}{2}} dz \\
&= \mu^2 + \sigma^2 + \frac{\sigma(A_\alpha+\mu)}{(1-\alpha)\sqrt{2\pi}} e^{-\frac{(A_\alpha-\mu)^2}{2\sigma^2}},
\end{aligned}
$$

where the form $\int_a^\infty z^2 e^{-\frac{z^2}{2}} dz$ is evaluated using integration by parts as follows:

$$
\int_a^\infty z^2 e^{-\frac{z^2}{2}} dz = \int_a^\infty -z de^{-\frac{z^2}{2}} = \left(-ze^{-\frac{z^2}{2}}\right)\Big|_a^\infty - \int_a^\infty -e^{-\frac{z^2}{2}} dz = ae^{-\frac{a^2}{2}} + \sqrt{2\pi}(1-\Phi(a)).
$$

$\mathbb{E}[X^3|X > A_\alpha]$ can be derived using integration by parts in a similar way.

Appendix B Closed-form Formulas for Some Heavy-tailed Distributions

In what follows we summarize closed-form expressions of some necessary statistics used in this paper. Results are reported in Table 11, in which A is a constant, $\Gamma(\cdot)$ is the gamma function, $\gamma(\alpha;x)=\int_0^x z^{\alpha-1}e^{-z}dz$ is the lower incomplete gamma function and $\hat{\Gamma}(\alpha;x)=\int_x^\infty z^{\alpha-1}e^{-z}dz$ is the upper incomplete gamma function. Except t and GPD, the domains of PDF and CDF are $w>0$. In GPD, $\xi\neq 0$, $\sigma>0$ always hold and $1+\xi w/\sigma>0$ should be guaranteed when ξ is negative. Some non-analytical expressions need to be computed by numerical methods or software packages. For the ease of presentation, long equations are shown below separately:

$$m_1^t=\frac{1}{1-F_W(A)}\frac{\Gamma(\frac{v+1}{2})v}{\sqrt{v\pi}\Gamma(\frac{v}{2})(v-1)}\left(\frac{A^2}{v}+1\right)^{-\frac{v-1}{2}},\quad v>1;$$

$$m_2^{GPD}=A^2+2\left(1+\frac{\xi}{\sigma}A\right)\frac{\sigma(A+\sigma-A\xi)}{(\xi-1)(2\xi-1)},\quad \xi<\frac{1}{2};$$

$$m_3^{GPD}=A^3-\frac{3\sigma\left(1+\xi\frac{A}{\sigma}\right)}{\xi-1}A^2+\frac{6\sigma^2\left(1+\xi\frac{A}{\sigma}\right)^2}{(\xi-1)(2\xi-1)}A-\frac{6\sigma^3\left(1+\xi\frac{A}{\sigma}\right)^3}{(\xi-1)(2\xi-1)(3\xi-1)},\quad \xi<\frac{1}{3}.$$

W	t, df $=v$	Gamma(α,β)	LogN(μ,σ^2)	GPD(ξ,σ)	Weibull(k,λ)
PDF, $f_W(w)$	$\frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1+\frac{w^2}{v})^{-\frac{v+1}{2}}$	$\frac{w^{\alpha-1}e^{-\frac{w}{\beta}}}{\Gamma(\alpha)\beta^\alpha}$	$\frac{1}{\sqrt{2\pi}\sigma w}e^{-\frac{(\ln w-\mu)^2}{2\sigma^2}}$	$\frac{1}{\sigma}\left(1+\xi\frac{w}{\sigma}\right)^{-\frac{1}{\xi}-1}$	$\frac{k}{\lambda}(\frac{w}{\lambda})^{k-1}e^{-(\frac{w}{\lambda})^k}$
CDF, $F_W(w)$	$\int_{-\infty}^w f_W(x)dx$	$\frac{\gamma\left(\alpha;\frac{w}{\beta}\right)}{\Gamma(\alpha)}$	$\Phi(\frac{\ln w-\mu}{\sigma})$	$1-\left(1+\xi\frac{w}{\sigma}\right)^{-\frac{1}{\xi}}$	$1-e^{-(\frac{w}{\lambda})^k}$
$\mathbb{E}[W\|W\geq A]$	m_1^t	$\frac{\beta\hat{\Gamma}\left(\alpha+1;\frac{A}{\beta}\right)}{(1-F_W(A))\Gamma(\alpha)}$	$\frac{\Phi(\frac{\mu+\sigma^2-\ln A}{\sigma})}{1-F_W(A)}e^{\mu+\frac{\sigma^2}{2}}$	$A+\frac{\xi A+\sigma}{1-\xi},\xi<1$	$\frac{\lambda\hat{\Gamma}\left(\frac{1}{k}+1;\left(\frac{A}{\lambda}\right)^k\right)}{1-F_W(A)}$
$\mathbb{E}[W^2\|W\geq A]$	$\frac{\int_A^\infty x^2 f_W(x)dx}{1-F_W(A)},v>2$	$\frac{\beta^2\hat{\Gamma}\left(\alpha+2;\frac{A}{\beta}\right)}{(1-F_W(A))\Gamma(\alpha)}$	$\frac{\Phi(\frac{\mu+2\sigma^2-\ln A}{\sigma})}{1-F_W(A)}e^{2\mu+2\sigma^2}$	m_2^{GPD}	$\frac{\lambda^2\hat{\Gamma}\left(\frac{2}{k}+1;\left(\frac{A}{\lambda}\right)^k\right)}{1-F_W(A)}$
$\mathbb{E}[W^3\|W\geq A]$	$\frac{\int_A^\infty x^3 f_W(x)dx}{1-F_W(A)},v>3$	$\frac{\beta^3\hat{\Gamma}\left(\alpha+3;\frac{A}{\beta}\right)}{(1-F_W(A))\Gamma(\alpha)}$	$\frac{\Phi(\frac{\mu+3\sigma^2-\ln A}{\sigma})}{1-F_W(A)}e^{3\mu+\frac{9}{2}\sigma^2}$	m_3^{GPD}	$\frac{\lambda^3\hat{\Gamma}\left(\frac{3}{k}+1;\left(\frac{A}{\lambda}\right)^k\right)}{1-F_W(A)}$

Table 11: Closed-form formulas of 1st, 2nd & 3rd conditional moments.