Learning extreme Expected Shortfall and Conditional Tail Moments with neural networks. Application to cryptocurrency data

Michaël Allouche, Stéphane Girard, Emmanuel Gobet

Learning extreme Expected Shortfall and Conditional Tail Moments with neural networks. Application to cryptocurrency data

Michaël Allouche[a], Stéphane Girard[b], Emmanuel Gobet[c]

[a] *Kaiko - Quantitative Data, 2 rue de Choiseul, Paris, 75002, France*
[b] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France*
[c] *Centre de Mathématiques Appliquées (CMAP), CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, Route de Saclay, Palaiseau, 91128, France*

Abstract

We propose a neural networks method to estimate extreme Expected Shortfall, and even more generally, extreme conditional tail moments as functions of confidence levels, in heavy-tailed settings. The convergence rate of the uniform error between the log-conditional tail moment and its neural network approximation is established leveraging extreme-value theory (in particular the high-order condition on the distribution tails) and using critically two activation functions (eLU and ReLU) for neural networks. The finite sample performance of the neural network estimator is compared to bias-reduced extreme-value competitors using synthetic heavy-tailed data. The experiments reveal that our method largely outperforms others. In addition, the selection of the anchor point appears to be much easier and stabler than for other methods. Finally, the neural network estimator is tested on real data related to extreme loss returns in cryptocurrencies: here again, the accuracy obtained by cross-validation is excellent, and is much better compared with competitors.

Keywords: Extreme-value theory, heavy-tailed distribution, risk measure estimation, neural networks.

60G70, 62G32, 68T07.

1. Introduction

The 80's and 90's saw major financial crises, highlighting the need for risk monitoring indicators, for defensive control or active management purposes. In July 1993, the G-30

– a consultative group of top bankers, financiers, and academics from leading industrial nations – met to discuss and share best risk management practices and came up with the notion of Value at Risk (VaR), which was already being investigated at the bank J.P. Morgan [41]. The VaR summarizes the *worst loss over a target horizon that will not be exceeded with a given level of confidence*. The VaR is then used to determine the amount of capital to set aside for solvency purposes, see [35, Section 3.4]. Major criticisms on the VaR include (i) its failure to fulfill the subadditivity property [1], making it a non coherent risk measure [9], and (ii) its failure to account for the magnitude of losses beyond the given confidence level, since quantiles only depend on the frequency of tail losses and not on their values, see [23, Section 1.3.3]. For these reasons, it has been proposed as early as 2002 to change the method for assessing risks from the usual VaR to the alternative coherent Expected Shortfall (ES) [1, 46]. Unlike the VaR, ES satisfies all of the requirements to be a coherent risk measure, namely translation invariance, monotonicity, positive homogeneity, and subadditivity [9]. Moreover, the work of [2] took part in promoting the ES by introducing backtesting methodologies and by proving that the ES is jointly elicitable with the VaR. Because of the paramount importance of risk measures (such as VaR or ES) in financial institutions, it is essential to accurately estimate them: This point has been widely studied by both researchers and practitioners, see [23] for a review.

In this work, we focus on the ES and Conditional Tail Moments (that generalize the ES) and provide a new estimation methodology. Let us define the quantities of interest. The VaR is the quantile at a given confidence level $(1-\alpha)$:

$$\texttt{VaR}(1-\alpha) = q(1-\alpha) = \inf\{x \in \mathbb{R},\, F(x) \geq 1-\alpha\},$$

where $\alpha \in (0,1)$ and with F the cumulative distribution function of the non-negative random variable X of interest (X is usually the opposite of a loss and α is close to 0). Besides, the ES, also known as Conditional Value at Risk or Average Value at Risk, is defined as the average of the quantile function above the confidence level $(1-\alpha)$:

$$\texttt{ES}(1-\alpha) = \frac{1}{\alpha}\int_0^\alpha q(1-u)\,\mathrm{d}u.$$

When the distribution of X is continuous (no atoms), $\texttt{ES}(1-\alpha)$ coincides with the conditional expectation of X given that it exceeds the VaR $q(1-\alpha)$:

$$\texttt{ES}(1-\alpha) = \mathbb{E}\left(X \mid X > q(1-\alpha)\right),$$

and it is referred to as Conditional Tail Expectation (CTE). In this work, we consider more generally Conditional Tail Moments (CTMs) of order $p > 0$, defined in [25] as

$$\texttt{CTM}_p(1-\alpha) = \mathbb{E}\left(X^p \mid X > q(1-\alpha)\right) = \frac{1}{\alpha}\int_0^\alpha q^p(1-u)\,\mathrm{d}u \tag{1}$$

under the assumption of a continuous distribution. Clearly, $\texttt{CTM}_1 = \texttt{ES}$. Explicit derivations of the ES are available in some parametric families, namely elliptical [40], or phase-type distributions [15], opening the way to parametric estimation. See also Table 1 below for ES calculations associated with some heavy-tailed distributions deduced from [8, Table 8]. Derivations of the CTM for elliptical and skew-elliptical distributions are provided in [39] and [24] respectively. Nonparametric estimators have also been introduced in [13, 21] based on the empirical counterpart of (1):

$$\widehat{\texttt{CTM}_p}(1-\alpha) = \frac{1}{\lfloor n\alpha \rfloor}\sum_{j=1}^{\lfloor n\alpha \rfloor} X_{n-j+1,n}^p \text{ with } \widehat{\texttt{ES}}(1-\alpha) = \widehat{\texttt{CTM}_1}(1-\alpha), \tag{2}$$

and where $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics associated with the initial n-sample. See also [47] for a kernel version and [18] for the associated asymptotic properties.

We consider extreme risks, *i.e.* the situation where α is so small that the VaR is larger than the maximal observation: $q(1-\alpha) > X_{n,n}$. This is asymptotically the case when $n\alpha$ is small, and in this challenging situation, the naive estimator (2) cannot be used to estimate the ES or CTM out of the range of the data set. To overcome this problem, we propose in Section 2 an extrapolation formula dedicated to heavy-tailed distributions and leveraging extreme-value theory, linking the (out-of-range) CTM at an extreme level to the (in-range) CTM at a non-extreme one. Note that the non-Gaussian and heavy-tailed behaviors of traditional financial asset returns are well established since the last century [42], this conclusion being recently extended to cryptocurrency data [4] which often translates into a heavy-tailed loss X. We show that the linking

function can be approximated by a neural network (NN) with a controlled error. The performance of the proposed NN estimator is illustrated on simulated data in Section 3: It is compared to five other extrapolation techniques stemming from extreme-value theory on a large variety of heavy-tailed situations. An application to cryptocurrency data is provided in Section 4 and a short conclusion is drawn in Section 5. Proofs and technical details are postponed to Appendix B and Appendix A respectively.

2. Extrapolation principle for conditional tail moments

Let $X_1, \ldots, X_n$ be a sample from an unknown cumulative distribution function F assumed to be continuous. The associated order statistics are denoted by $X_{1,n} \leq \cdots \leq X_{n,n}$. We address the estimation of $\texttt{ES}(1-\alpha_n)$ at an extreme level α_n *i.e.* such that $n\alpha_n \to 0$ as $n \to \infty$. Let us assume for the sake of argument (and only in the following calculation) that $X_1, \ldots, X_n$ are independent. It follows that

$$\begin{aligned}\mathbb{P}(X_{n,n} \leq \texttt{ES}(1-\alpha_n)) \geq \mathbb{P}(X_{n,n} \leq q(1-\alpha_n)) = F^n(q(1-\alpha_n)) = (1-\alpha_n)^n \\ = \exp(-n\alpha_n(1+o(1))) \to 1 \text{ as } n \to \infty,\end{aligned}$$

thanks to a first order Taylor expansion. It appears that the ES of interest is out of the sample with probability tending to one. Dedicated extrapolation techniques are thus necessary since, in that challenging situation, the empirical estimator (2) cannot be used.

2.1. Theoretical framework

We focus on heavy-tailed distributions whose survival function $(1-F)$ is regularly-varying with index $-1/\gamma$, where $\gamma > 0$ is called the tail-index. Our main assumption can be written as

(A) $(1-F) \in \mathcal{RV}_{-1/\gamma}$ with $\gamma > 0$ *i.e.*

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-1/\gamma},$$

for all $x > 0$.

The index γ tunes the tail heaviness of F: the larger the index, the heavier the right tail. Examples include the (generalized) Pareto, Burr, Fisher, Inverse gamma and Student distributions. Equivalently, the tail quantile function

$$t > 1 \mapsto U(t) = q(1 - 1/t) \tag{3}$$

can be rewritten as

$$U(t) = t^{\gamma} L(t) \tag{4}$$

where L is a slowly-varying function, *i.e.* a regularly-varying function with index zero:

$$\lim_{t \to \infty} \frac{L(tz)}{L(t)} = 1, \tag{5}$$

for all $z > 0$. See [12] for more details on regular variation theory. The following result provides a sufficient condition for the existence of the CTM associated with an heavy-tailed distribution.

Proposition 1. *If* (**A**) *holds for some* p *such that* $p\gamma \in (0, 1)$, *then* $\texttt{CTM}_p(1-\alpha)$ *is finite for any* $\alpha \in (0, 1)$.

We also consider a J-th order condition ($J \geq 2$), introduced in [52], on the slowly-varying function to control the rate of convergence in (5):

(**B$_J$**) There exist for all $j \in \{2, \ldots, J\}$, functions A_j with $A_j(t) \to 0$ as $t \to \infty$ and an asymptotically constant sign, $|A_j| \in \mathcal{RV}_{\rho_j}$ and $\rho_j \leq 0$ such that

$$\log L(tz) - \log L(t) = \sum_{j=2}^{J} \prod_{\ell=2}^{j} A_\ell(t) R_j(z) + o\left(\prod_{j=2}^{J} A_j(t) \right),$$

as $t \to \infty$ for all $z > 0$, where:

$$R_j(z) := \int_1^z z_2^{\rho_2 - 1} \int_1^{z_2} z_3^{\rho_3 - 1} \cdots \int_1^{z_{j-1}} z_j^{\rho_j - 1} \, \mathrm{d}z_j \ldots \mathrm{d}z_3 \, \mathrm{d}z_2.$$

Such a condition refines the initial heavy-tail assumption and is therefore a classical device in extreme-value analysis for extrapolating beyond the observation range. For instance, when $J = 2$, we recover the classical second-order condition [30, Equation (13)]. Moreover, $J = 3$ and $J = 4$ yield back respectively the third-order [27, 44] and fourth-order conditions [29]. For most of heavy-tailed distributions, $A_1, \ldots, A_J$ are power functions, which will be our assumption in the sequel:

($\mathbf{C_J}$) $A_j(t) = c_j t^{\rho_j}$ where $c_j \neq 0$ and $\rho_j < 0$ for $j \in \{2, \ldots, J\}$.

When $J = 2$, ($\mathbf{C_J}$) amounts to supposing that the underlying distribution belongs to the Hall-Welsh class [32]. We give a list of examples of classical heavy-tailed distributions in Table 1 we shall work with in the simulation study (Section 3), with their respective values of γ and ρ_2. See [8] for derivation details associated with Inverse Generalized Gamma (IGG) and Burr-Fisher-Snedecor (BFS) families of distributions respectively. Finally, we shall also need the following assumption for the subsequent analysis.

($\mathbf{D}$) L is differentiable and $|(\log L)'| \in \mathcal{RV}_{\rho_2 - 1}$ with $\rho_2 < 0$.

See Lemma 6 in Appendix A for a discussion on the assumptions.

2.2. Basic properties

Assume ($\mathbf{A}$) holds for some p such that $p\gamma \in (0, 1)$, and consider the slowly-varying function $t \mapsto L(t) = t^{-\gamma} U(t)$. The following two functions are then introduced:

$$(y_1, y_2) \in [0, \infty)^2 \mapsto \varphi(y_1, y_2) = \log L(\exp(y_1 + y_2)) - \log L(\exp(y_2)), \tag{6}$$

$$u \in (0, 1] \mapsto \Psi_p(u) = \log \left(\int_0^1 w^{-p\gamma} \, \frac{L^p(1/(wu))}{L^p(1/u)} \, \mathrm{d}w \right). \tag{7}$$

As we will explain, the proposed approach consists in learning these functions φ and Ψ_p using neural network approximations. In [7], we have established in a similar statistical framework that the log-spacings between two quantiles can be written with φ as

$$\log q(1 - \alpha) - \log q(1 - \delta) = \gamma \log(\delta/\alpha) + \varphi\big(\log(\delta/\alpha), \log(1/\delta) \big). \tag{8}$$

The key observation is that the log-spacings associated with the CTM can also be rewritten in an appealing manner with φ and the additional function Ψ_p compared to the extreme quantile framework.

Proposition 2. *If* ($\mathbf{A}$) *holds for some p such that $p\gamma \in (0, 1)$, then, for all probability levels $(\alpha, \delta) \in (0, 1)^2$:*

$$\log \mathtt{CTM}_p(1 - \alpha) = p \log q(1 - \alpha) + \Psi_p(\alpha), \tag{9}$$

$$\begin{aligned} \log \mathtt{CTM}_p(1 - \alpha) - \log \mathtt{CTM}_p(1 - \delta) &= p\gamma \log(\delta/\alpha) + p\varphi\big(\log(\delta/\alpha), \log(1/\delta) \big) \\ &\quad + \Psi_p(\alpha) - \Psi_p(\delta) \\ &=: g_p(\alpha, \delta). \end{aligned} \tag{10}$$

Combining Lemma 7(i) in Appendix A and Proposition 2, Equation (9) yields an asymptotic link between the CTM and the Value-at-Risk:

$$\lim_{\alpha\to 0} \frac{\texttt{CTM}_p(1-\alpha)}{q^p(1-\alpha)} = \frac{1}{1-p\gamma}, \tag{11}$$

and we thus recover a result from [26, Proposition 1(ii)]. It is shown in particular that the limiting result (11) also holds true for light-tailed distributions ($\gamma = 0$) which is a heuristic indication that the proposed methods may also work in this framework. We also refer to [43] for first and second order asymptotic expansions of generalized shortfall risk measures. One can then use this first order approximation of the extreme CTM to build a crude *indirect* estimator based on an estimator of the extreme quantile $q(1-\alpha)$ and an estimator of the tail-index γ, see Paragraph 3.2 for further details.

Besides, thanks to (10) in Proposition 2, one can easily derive an extrapolation principle for the CTM. Indeed, the out-of-range CTM at an extreme level $1-\alpha_n$ (such that $n\alpha_n \to 0$ as $n \to \infty$) can be deduced from the in-range CTM at an intermediate level $1-\delta_n$ (such that $n\delta_n \to \infty$ as $n \to \infty$) since

$$\texttt{CTM}_p(1-\alpha_n) = \texttt{CTM}_p(1-\delta_n)(\delta_n/\alpha_n)^{p\gamma} \exp\left(p\varphi\big(\log(\delta_n/\alpha_n), \log(1/\delta_n)\big) + \Psi_p(\alpha_n) - \Psi_p(\delta_n)\right), \tag{12}$$

or equivalently

$$\texttt{CTM}_p(1-\alpha_n) = \texttt{CTM}_p(1-\delta_n)\exp(g_p(\alpha_n,\delta_n)).$$

This kind of extrapolation method is widely used in extreme-value theory, see for instance Weissman estimator [53] dedicated to extreme quantiles and [19, 20] for extremiles and expectiles respectively. The challenge is to design an extrapolation function g_p in a way that the approximation of $\texttt{CTM}_p(1-\alpha_n)$ is as stable as possible with respect to the anchor level δ_n. The implementation of this idea requires the estimation of two real quantities, namely $\texttt{CTM}_p(1-\delta_n)$ and γ as well as the two functions φ and Ψ_p. For the latter problem, the simplest choice is to approximate the slowly-varying function L by a constant, leading to $\varphi \simeq 0$. Moreover, Lemma 7(i) in Appendix A shows that $\Psi_p(\alpha_n) - \Psi_p(\delta_n) \to 0$ as $n \to 0$ suggesting the first-order approximation:

$$\texttt{CTM}_p(1-\alpha_n) \simeq \texttt{CTM}_p(1-\delta_n)(\delta_n/\alpha_n)^{p\gamma}. \tag{13}$$

This result is used in Paragraph 3.2 to build the so-called *direct* estimator of the extreme CTM. Note that we do not expect it to be accurate enough since this estimator accounts for a simplified behavior of the tails.

In the following, we focus on a nonparametric estimation of φ and Ψ_p using NNs whose first step consists in approximating the two unknown functions using expansions on well-chosen families of functions.

2.3. Approximation of the CTM function with neural networks

The Hölder properties of Ψ_p are established in Lemma 7(ii) in Appendix A, depending on the position of the second-order parameter ρ_2 compared to -1. We show in the next Proposition that Ψ_p can be approximated using rectified linear unit (ReLU) activation functions $x \in \mathbb{R} \mapsto \sigma^{\mathtt{R}}(x) := \max(0, x)$ with a controlled error. The choice of ReLU activation functions relies on the theoretical result that piecewise linear functions are universal approximators of Hölder functions, see Lemma 8 in Appendix A for technical details.

Proposition 3. *Let L be a slowly-varying function such that* **(D)** *holds. Let $p\gamma \in (0,1)$ and consider the function Ψ_p defined in (7). Then, for any $K_1 \geq 6$ and $\varepsilon \in (0, |\rho_2|)$, there exists a ReLU-NN with K_1 neurons, parameterized by $\theta_1 = (a^{(1)}_{1:K_1}, a^{(2)}_{1:K_1}, a^{(3)}_{1:K_1}) \in \Theta_1 := \mathbb{R}^{3K_1}$ and defined by*

$$u \in [0,1] \mapsto \tilde{\Psi}^{K_1}_{p,\theta_1}(u) = \sum_{k=1}^{K_1} a^{(1)}_k \sigma^{\mathtt{R}}\left(a^{(2)}_k u + a^{(3)}_k\right),$$

such that

$$\sup_{u\in[0,1]} \left|\Psi_p(u) - \tilde{\Psi}^{K_1}_{p,\theta_1}(u)\right| \leq 6p\, H^{\varepsilon}_{\Psi_1} (1/K_1)^{\rho_2^{\varepsilon}\wedge 1},$$

where $H^{\varepsilon}_{\Psi_1} > 0$ is the $(\rho_2^{\varepsilon} \wedge 1)$-Hölder constant associated with Ψ_1 and $\rho_2^{\varepsilon} = |\rho_2| - \varepsilon$.

Turning to φ, the next Proposition presents how, starting from the J-th order condition $(\mathbf{B_J})$ and $(\mathbf{C_J})$, a NN approximation of

$$(y_1, y_2) \in [0,\infty)^2 \mapsto \varphi(y_1, y_2) = \sum_{j=2}^{J} \bar{c}_j \exp(\bar{\rho}_j y_2) R_j(\exp y_1) + o(\exp(\bar{\rho}_J y_2)) \qquad (14)$$

can be built using exponential linear unit (eLU) activation functions $x \in \mathbb{R} \mapsto \sigma^{\mathtt{e}}(x) := x\mathbb{1}_{\{x\geq 0\}} + (\exp(x)-1)\mathbb{1}_{\{x<0\}}$. Here, and in the sequel, we let $\bar{\rho}_J = \rho_2 + \cdots + \rho_J$ and $\bar{c}_J = |c_2 \times \cdots \times c_J|$. Note that in (14), the exponential functions only involve negative arguments, and as such, they can be interpreted as eLU functions. This justifies the use of eLU activation functions in the approximation of φ.

Proposition 4. *Let L be a continuously differentiable slowly-varying function verifying the J-th order condition* $(\mathbf{B_J})$ *with* $(\mathbf{C_J})$ *and $J \geq 2$. Consider the function φ defined in (6). Then, one can design an eLU-NN with $K_2 = J(J-1)/2$ neurons, parameterized by $\theta_2 = (b^{(1)}_{1:K_2}, b^{(2)}_{1:K_2}, b^{(3)}_{1:K_2}, b^{(4)}_{1:K_2}) \in \Theta_2 := (\mathbb{R}\times\mathbb{R}^3_-)^{K_2}$ defined by*

$$(y_1,y_2) \in \mathbb{R}^2_+ \mapsto \tilde{\varphi}^{K_2}_{\theta_2}(y_1,y_2) = \sum_{k=1}^{K_2} b^{(1)}_k \left(\sigma^{\mathtt{e}}\left(b^{(2)}_k y_1 + b^{(3)}_k y_2\right) - \sigma^{\mathtt{e}}(b^{(4)}_k y_2)\right), \tag{15}$$

and, for all $\varepsilon > 0$, there exists $y_\varepsilon > 0$ such that, for all $y_1, y_2 \geq y_\varepsilon$,

$$\left|\varphi(y_1,y_2) - \tilde{\varphi}^{K_2}_{\theta_2}(y_1,y_2)\right| \leq \varepsilon \bar{c}_J \exp((\bar{\rho}_J+\varepsilon)(y_1+y_2) - y_2\varepsilon).$$

Following (12), we thus propose to approximate the CTM at level $(1-\alpha_n)$ by

$$\begin{aligned}\widetilde{\mathtt{CTM}}^K_{p,\theta}(1-\alpha_n;1-\delta_n) := \mathtt{CTM}_p(1-\delta_n)(\delta_n/\alpha_n)^{p\theta_0} \exp\left(p\tilde{\varphi}^{K_2}_{\theta_2}(\log(\delta_n/\alpha_n),\log(1/\delta_n))\right)\\ \times \exp\left(\tilde{\Psi}^{K_1}_{p,\theta_1}(\alpha_n) - \tilde{\Psi}^{K_1}_{p,\theta_1}(\delta_n)\right),\end{aligned} \tag{16}$$

where $\theta := (\theta_0,\theta_1,\theta_2) \in \Theta := \mathbb{R}_+ \times \Theta_1 \times \Theta_2$ and $K := K_1 + K_2$, or equivalently,

$$\widetilde{\mathtt{CTM}}^K_{p,\theta}(1-\alpha_n;1-\delta_n) = \mathtt{CTM}_p(1-\delta_n)\exp(\tilde{g}^K_{p,\theta}(\alpha_n,\delta_n)),$$

with $\tilde{g}^K_{p,\theta}(\alpha_n,\delta_n) = p\theta_0\log(\delta_n/\alpha_n) + p\tilde{\varphi}^{K_2}_{\theta_2}(\log(\delta_n/\alpha_n),\log(1/\delta_n)) + \tilde{\Psi}^{K_1}_{p,\theta_1}(\alpha_n) - \tilde{\Psi}^{K_1}_{p,\theta_1}(\delta_n)$. Combining the above two results, we have:

Theorem 5. *Let L be a slowly-varying function verifying* $(\mathbf{B_J})$, $(\mathbf{C_J})$ *with $J \geq 2$ and* $(\mathbf{D})$. *Let $p\gamma \in (0,1)$ and consider the NN approximation (16) of the CTM involving $K_1 \geq 6$ ReLU units and $K_2 = J(J-1)/2$ eLU units. Consider two probability level sequences (α_n) and (δ_n) such that $\delta_n \to 0$ and $\alpha_n/\delta_n \to 0$ as $n \to \infty$. Then, for all $\varepsilon \in (0, |\rho_2|)$,*

$$\frac{1}{p}\inf_{\theta\in\Theta}\left|\log \mathtt{CTM}_p(1-\alpha_n) - \log \widetilde{\mathtt{CTM}}^K_{p,\theta}(1-\alpha_n;1-\delta_n)\right| \leq 12 H^{\varepsilon}_{\Psi_1}(1/K_1)^{\rho_2^{\varepsilon}\wedge 1} + \bar{c}_J\varepsilon(\delta_n/\alpha_n)^{\varepsilon}\alpha_n^{-\bar{\rho}_J}.$$

The approximation rate does not depend on the tail heaviness γ, but rather on the parameters ρ_2 and $\bar{\rho}_J$ stemming from the J-th order condition. Since $\bar{\rho}_J$ is a decreasing function of J, choosing $J > 2$ may be of interest to reduce the approximation error while usual extreme-value estimators rely on first- or second-order approximations, see Paragraph 3.2 for examples. The approximation rate is also driven by the complexity of the NN which can be quantified by both K_1 the number of ReLU units and K_2 the number of eLU units via J. Finally, note that the approximation error is an increasing function of α_n: The further in the tail the risk measure is, the better its approximation.

2.4. Estimation of conditional tail moments with neural networks

Here, we let $\delta_n = k/n$ where $k = k_n$ is an intermediate sequence *i.e.* such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. The in-range CTM at the intermediate level δ_n is estimated by its empirical counterpart (2):

$$\widehat{\mathtt{CTM}}_p(1-\delta_n) = \widehat{\mathtt{CTM}}_p(1-k/n) = \frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n},$$

and from (16), the out-of-range CTM at the extreme level α_n is thus estimated by

$$\begin{aligned}\widehat{\mathtt{CTM}}^{\mathtt{NN}}_{p,\theta}(1-\alpha_n; 1-k/n) &:= \left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right)(k/(n\alpha_n))^{p\hat{\theta}_0} \\ &\quad \times \exp\left(p\tilde{\varphi}^{K_2}_{\hat{\theta}_2}(\log(k/(n\alpha_n)), \log(n/k)\right) \\ &\quad \times \exp\left(\tilde{\Psi}^{K_1}_{p,\hat{\theta}_1}(\alpha_n) - \tilde{\Psi}^{K_1}_{p,\hat{\theta}_1}(k/n)\right) \\ &= \left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right)\exp\left(\tilde{g}^{K}_{p,\hat{\theta}}(\alpha_n, k/n)\right),\end{aligned}$$

where $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ is computed by fitting the NN model of the log-spacings to the empirical ones. More specifically, Proposition 2 shows that, for all $k \in \{1, \dots, n\}$ and $i \in \{k+1, \dots, n\}$, one as

$$S_{i,k} := \log \mathtt{CTM}_p(1-i/n) - \log \mathtt{CTM}_p(1-k/n) = g_p(i/n, k/n).$$

The empirical estimator is deduced from (2) and given by

$$\hat{S}_{i,k} := \log\left(\frac{1}{i}\sum_{j=1}^{i} X^p_{n-j+1,n}\right) - \log\left(\frac{1}{k}\sum_{j=1}^{k} X^p_{n-j+1,n}\right), \tag{17}$$

while the NN model is $\tilde{g}^K_{p,\theta}(i/n, k/n)$. In practice, small indices i and k are discarded since the associated empirical estimates of the CTM suffer from a large variance. To this end, we consider $\zeta \in [0, 1)$ and minimize the following distance between the above two estimations of the $N = (n - 2 - [\zeta n])(n - 1 - [\zeta n]))/2$ log-spacings:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{k=2+[\zeta n]}^{n-1} \sum_{i=1+[\zeta n]}^{k-1} \left| \hat{S}_{i,k} - \tilde{g}^K_{p,\theta} \left(i/n, k/n \right) \right|. \tag{18}$$

Here, a L_1 distance is adopted but other convex choices could be considered without significantly modifying the implementation, for instance a L_2 distance or the Huber loss [34] providing a compromise between the latter two criteria. See Figure 1 (right panel) for an illustration of the NN estimation of the ES log-spacing function g_1 on Burr data, compared with the crude linear approximation. Implementation details are reported in the next Section.

Besides, in this work, we do not analyze the statistical error $(\widehat{\texttt{CTM}}^{\texttt{NN}}_{p,\theta} - \log \widetilde{\texttt{CTM}}^K_{p,\theta})$, our result (Theorem 5) only focuses on the approximation error $(\log \texttt{CTM}_p - \log \widetilde{\texttt{CTM}}^K_{p,\theta})$. Usually, the analysis of the statistical error is conducted under the assumption that the sample is i.i.d.; Some extensions to data satisfying mixing conditions are also available in the extreme context (see for instance [38, Theorem 9.5.1]).

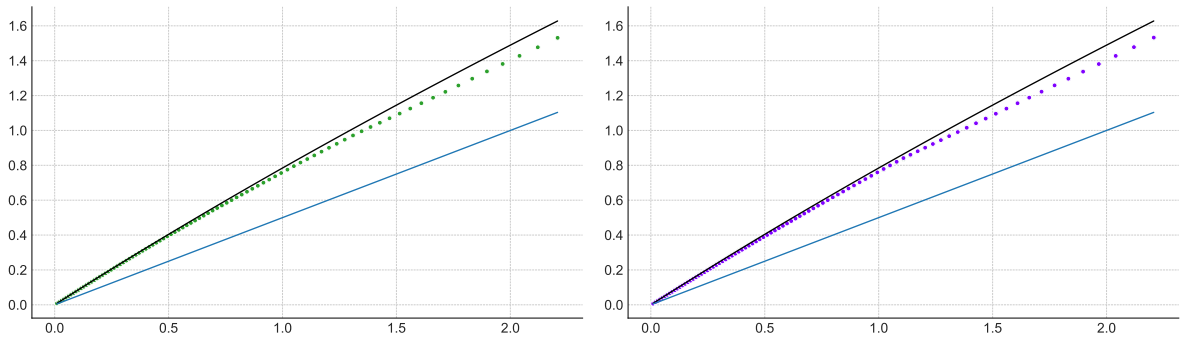


Figure 1: ES log-spacings associated with a Burr distribution ($\gamma = 1, \rho_2 = -1/2$), see Table 1 for the parameterization. Black curve: theoretical function $(x_1, g_1(x_1, \log(n/k)))$; blue line: first order approximation $(x_1, \gamma x_1)$ with $\varphi = 0$ and $\Psi_1(\cdot) = -\log(1-\gamma)$; green dots (left panel): empirical pointwise estimation $(\log(k/i), \hat{S}_{i,k})$; purple dots (right panel): NN estimation $(\log(k/i), \tilde{g}^K_{1,\hat{\theta}}(\log(k/i), \log(n/k)))$ with $i \in \{1, \ldots, k-1\}$, $k = 100$ and $n = 500$.

3. Validation on simulated data

The finite sample behavior of the NN estimator of the extreme CTM is illustrated on simulated data in the case where $p = 1$, *i.e.* focusing on the extreme ES. To this end, we first describe both the estimator implementation and the model selection technique. Then, some other bias-reduced estimators are briefly presented. Next, we list the heavy-tailed distributions as well as the performance criteria used to compare all considered estimators.

3.1. Implementation

All numerical experiments have been conducted on the Cholesky computing cluster from Ecole Polytechnique http://meso-ipp.gitlab.labos.polytechnique.fr/user_doc. It is composed by 4 nodes, where each one includes 2 CPU Intel Xeon Gold 6230 @ 2.1GHz, 20 cores and 4 Nvidia Tesla v100 graphics card. The code was implemented in Python 3.10.10 and using the library PyTorch 2.0.0.

The NN hyperparameters considered are the batch size set to 256, $K_1 = 10$ and $J \in \{2, 3, 4\}$ leading to $K \in \{11, 13, 16\}$ neurons. Note that the choice of K_1 has not been optimized, actually we have observed that numerical results (not reported here) are not very sensitive to that choice. We used the optimizer Adam [37] with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all tests performed during $M = 100$ iterations. Recall that one "iteration" in Adam algorithm corresponds to exploring the full data set using several mini-batches. The best iteration $m^\star \in \{1, \ldots, M\}$ is selected with [7, Algorithm 1] corresponding to the smallest median absolute deviation

$$\mathtt{MAD}\left(\left\{\widehat{\mathtt{ES}}^{\mathtt{NN}}_{\hat{\theta}_m}\left(1-\alpha_n; 1-\frac{k}{n}\right),\ k \in \{\lceil 3n/100\rceil, \ldots, \lceil 3n/4\rceil\}\right\}\right),$$

where, for any set $\mathcal{E} \subset \mathbb{R}$, the median absolute deviation is defined as

$$\mathtt{MAD}(\mathcal{E}) = \underset{\mathfrak{e}\in\mathcal{E}}{\mathrm{median}} \left|\mathfrak{e} - \mathrm{median}(\mathcal{E})\right|. \tag{19}$$

3.2. Competitors

Five extreme ES estimators arising from extreme-value theory are considered. They can be divided in two main families: the direct and the indirect ones. On the first

hand, plugging the empirical counterpart (2) of the intermediate ES in the first-order approximation (13) yields the direct extreme ES estimator

$$\widehat{\mathtt{ES}}^{\mathtt{D}}(1-\alpha_n;1-k/n)=\left(\frac{1}{k}\sum_{j=1}^{k}X_{n-j+1,n}\right)\left(\frac{k}{n\alpha_n}\right)^{\widehat{\gamma}_{\mathtt{H}}(k)}, \tag{20}$$

where

$$\widehat{\gamma}_{\mathtt{H}}(k)=\frac{1}{k}\sum_{i=1}^{k}\log(X_{n-i+1,n})-\log(X_{n-k,n})$$

is the Hill estimator [33] of the tail-index. A similar estimator is introduced in the conditional case [25] while this extrapolation principle is applied to multivariate versions of the ES in [16, 22]. On the other hand, the indirect estimator combines the first-order approximation (11) of the link between the extreme ES and the associated extreme quantile together with the Weissman estimator [53] to get:

$$\widehat{\mathtt{ES}}^{\mathtt{I}}(1-\alpha_n;1-k/n)=\frac{\widehat{q}_{\mathtt{W}}(1-\alpha_n;1-k/n)}{1-\widehat{\gamma}_{\mathtt{H}}(k)}=\frac{X_{n-k,n}}{1-\widehat{\gamma}_{\mathtt{H}}(k)}\left(\frac{k}{n\alpha_n}\right)^{\widehat{\gamma}_{\mathtt{H}}(k)}. \tag{21}$$

One can then propose reduced bias versions of (20) and (21) by replacing the Hill estimator by the Corrected-Hill estimator [14] denoted by $\widehat{\gamma}_{\mathtt{CH}}$ in the sequel. This gives rise respectively to $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$ and $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$. Additionally, replacing in (21) the Weissman estimator $\widehat{q}_{\mathtt{W}}$ by the Corrected Weissman estimator [31], which reduces both the extrapolation bias and the bias coming from the estimation of the tail-index in the extreme quantile estimator, gives $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$. Summarizing, the five competitors are $\widehat{\mathtt{ES}}^{\mathtt{D}}$, $\widehat{\mathtt{ES}}^{\mathtt{I}}$ (no bias correction), $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$, $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$ (one bias correction) and $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$ (two bias corrections).

3.3. Experimental design

The comparative study is achieved on seven heavy-tailed distributions which have a closed-form ES expression (see Table 1): Pareto, Student, Fréchet, Inverse gamma, Burr, generalized Pareto distribution (GPD), and Fisher-Snedecor. For all considered distributions, five values of the tail-index are considered: $\gamma \in \{0.1, 0.4, 0.5, 0.6, 0.9\}$. The Burr distribution has a free second-order parameter, four large values are investigated: $\rho_2 \in \{-1, -1/2, -1/4, -1/8\}$, while the ρ_2 parameter associated with the six remaining distributions is fixed, see Table 1 for details.

For each of these 50 considered configurations, $R = 500$ replicated data sets of size $n = 500$ are simulated with $\zeta = 0.02$ and the extreme ES of order $1 - \alpha_n = 1 - 1/(2n)$ is estimated using the NN ES estimator and the five estimators described in the above paragraph. Note that the generation of $R = 500$ replications is just used for evaluating the statistical fluctuations of the estimators. The performance is indeed assessed using the Relative median-squared error (RMedSE):

$$\text{RMedSE}\left(\widehat{\texttt{ES}}, \frac{1}{2n}\right) = \underset{r\in\{1,\dots,R\}}{\text{median}} \left[\left(\frac{\widehat{\texttt{ES}}^{(r)}\left(1-\frac{1}{2n}; 1-\frac{k^\star(r)}{n}\right)}{\texttt{ES}\left(1-\frac{1}{2n}\right)} - 1\right)^2\right],$$

where $\widehat{\texttt{ES}}^{(r)}\left(1-\frac{1}{2n}; 1-\frac{k^\star(r)}{n}\right)$ denotes an estimator of $\texttt{ES}\left(1-\frac{1}{2n}\right)$ (either the NN one or some of its competitors) computed with the anchor index $k^\star(r)$ selected using [5, Algorithm 1] with initial points $a^{(0)} = \lceil 3n/100 \rceil$ and $c^{(0)} = \lceil 3n/4 \rceil$ on the rth replication, $r \in \{1, \dots, R\}$. Other automated selection procedures could also be used, see for instance [10].

3.4. Results

It appears from the results in Table 2–Table 4 that the NN approach is an efficient tool for estimating extreme ES in difficult heavy-tailed situations where other estimators almost all fail. The NN estimator indeed yields the best results in 21 out of 50 times and always provides a valid result (RMedSE ≤ 1 – otherwise, it means that the estimator has an error larger than 100%, with probability at least 50%). As a comparison, $\widehat{\texttt{ES}}^{\texttt{D}}$, $\widehat{\texttt{ES}}^{\texttt{D,CH}}$, $\widehat{\texttt{ES}}^{\texttt{I}}$, $\widehat{\texttt{ES}}^{\texttt{I,CH}}$ and $\widehat{\texttt{ES}}^{\texttt{I,CW}}$ estimators give the best (and valid) results respectively only in 4(15), 3(31), 4(12), 7(33) and 11(28) out of 50 times. The second best estimator is thus $\widehat{\texttt{ES}}^{\texttt{I,CW}}$ which enjoys a double bias correction. Its performances are yet only half as good as the ones obtained with the NN approach. Note that in the Pareto case, the ES log-spacing function is linear and thus both $\widehat{\texttt{ES}}^{\texttt{D}}$ and $\widehat{\texttt{ES}}^{\texttt{I}}$ benefit from the perfect parametric form with γ being the only parameter to estimate. Unsurprisingly, these two methods yield the best results in this particular case, see Table 2. In contrast, they are outperformed by reduced biased approaches on all other distributions.

Note that the same experiments have been conducted in the case where $1 - \alpha_n = 1 - 1/(10n)$. The results provided in Table C.6–Table C.8 (see Appendix C) are very

similar to the ones discussed there with $1 - \alpha_n = 1 - 1/(2n)$.

Figure 2 and Figure 3 illustrate that the NN estimate features a nice stability in terms of median behavior and RMedSE for a wide range of k values on selected situations from Table 2–Table 4. This phenomenon may be highly appreciated either when the NN estimator is not ranked first on the RMedSE criteria basis (Figure 2) or for difficult heavy-tailed configurations where the NN outperforms all the other competitors (Figure 3). Note that this stability in terms of median behavior and RMedSE criteria with respect to the anchor point k comes from the fact that the empirical ES log-spacings function (see Figure 1) is smoother than the empirical quantile log-spacing function (see [7, Figure 2]) and so easier to approximate with a NN. Therefore, even with a low number of neurons, the NN is able to perfectly approximate the empirical log-spacings (17) for every k, so that the bias and variance mainly come from the statistical properties of the empirical log-spacings used as training data.

The RMedSE criterion may only provide an incomplete view of the estimators performance since it focuses on their distribution bulk. In order to have a complete picture of the estimators distributions, Figure 4 displays the boxplots associated with the six considered estimators computed on eight Burr datasets selected from Table 4. For each row (from the lowest to the largest second-order parameter ρ_2), two configurations of tail indices γ are represented as columns. It appears that all considered estimators (NN or extreme-value based) produce outliers when applied to heavy-tailed data. The NN method however provides the estimations the most concentrated (at least visually) around the true extreme expected-shortfall, compared to other competitors. This is in accordance with Table 4: The NN method is the best from the RMedSE point of view in situations (d-h) while $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$ and $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$ provide the best results in situations (a,b) and (c) respectively. Note that the boxplots associated with $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$ and $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$ are not displayed in situations (f-h): The bias corrections failed and yielded negative estimates of the expected shortfall.

As a conclusion, even though the NN method is numerically more expensive than its competitors, it provides a very effective estimator for all heavy-tailed situations. Here, the NNs are built with $K = 11$ to $K = 16$ neurons ($K_1 = 10$ and $K_2 \in \{1, 3, 6\}$)

which remains acceptable from the computational cost point of view (computing the NN estimate on 500 replications in multiprocessing with 40 cores took 3 hours with a batch size of 256).

4. Illustration on cryptocurrency data

While the cryptocurrency market is not regulated yet, the study of risk measures is of essential importance for financial actors such as investors, hedge funds, market makers, traders, and can provide solid foundations for future regulator policies. Some works dealing with risk estimation for cryptocurrency data started to emerge: ES estimation using GARCH models (see [3, 17] among others and [51] for a review), as well as expectile and Marginal Expected Shortfall estimation [49], but none of them focused on quantities at levels out of the range of the data set.

Here, the NN estimator is tested on the the absolute value of negative daily log-returns of BTC/USD (adverse outcomes are thus represented by large positive values and the value of one Bitcoin expressed in US dollars) during seven years between 2016 and 2022. Since Bitcoin is historically the first cryptocurrency, it is often used as a reference in crypto-market. The dataset provided by Kaiko is based on a daily robust aggregation of all trades between 12pm and 11:59pm. The variables of interest $X^{(j)}$ are the negative daily log-returns for each year composed respectively by $n^{(j)}$ data, $j \in \{1, \dots, 7\}$. More details on the data statistics are reported in Table 5, while the data distribution is illustrated in Figure 5a. The very large losses with daily negative log-returns between 1% and 3% are significantly higher than in traditional finance and are explained by a less mature and a more speculative market. Such a behavior suggests a heavy-tail behavior, even for the one of the most liquid pair in the cryptocurrency market. This is confirmed by the log quantile-quantile plot (Figure 5b) which is approximately linear for all years at level $\xi = 0.9$, providing a graphical evidence of the tail heaviness of each margin. The estimated tail-indices are reported Table 5, they vary between 0.25 and 0.6, which is in line with the conclusions of [4]. As mentioned in Section 2.4, for ensuring the statistical error to be well controlled (in our methodology or in those of competitors), we shall need independence or mixing properties within the

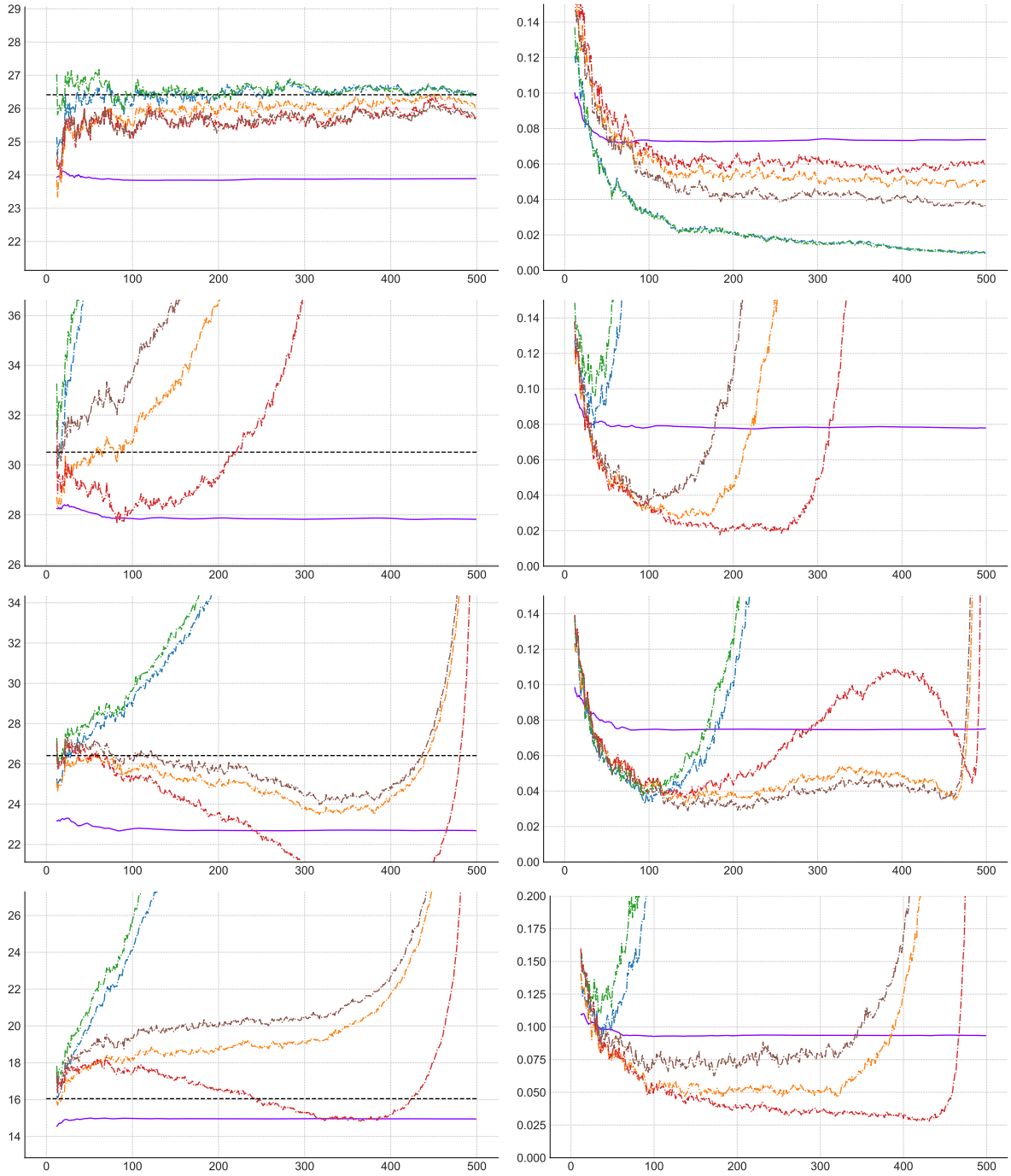


Figure 2: Illustration on simulated data sets of size $n = 500$ from Pareto, Student's t, Fréchet and Inverse gamma distributions (all with $\gamma = 0.4$) from top to bottom. Median of the estimators (left panel) of the extreme ES (black dashed line) at level $1 - \alpha_n = 1 - 1/(2n)$ and RMedSE (right panel), as functions of $k \in \{2 + [\zeta n], \ldots, n - 1\}$, computed on $R = 500$ replications, associated with $\widehat{\mathtt{ES}}^{\mathtt{D}}$ (blue), $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$ (orange), $\widehat{\mathtt{ES}}^{\mathtt{I}}$ (green), $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$ (red), $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$ (brown) and NN (purple) estimators.

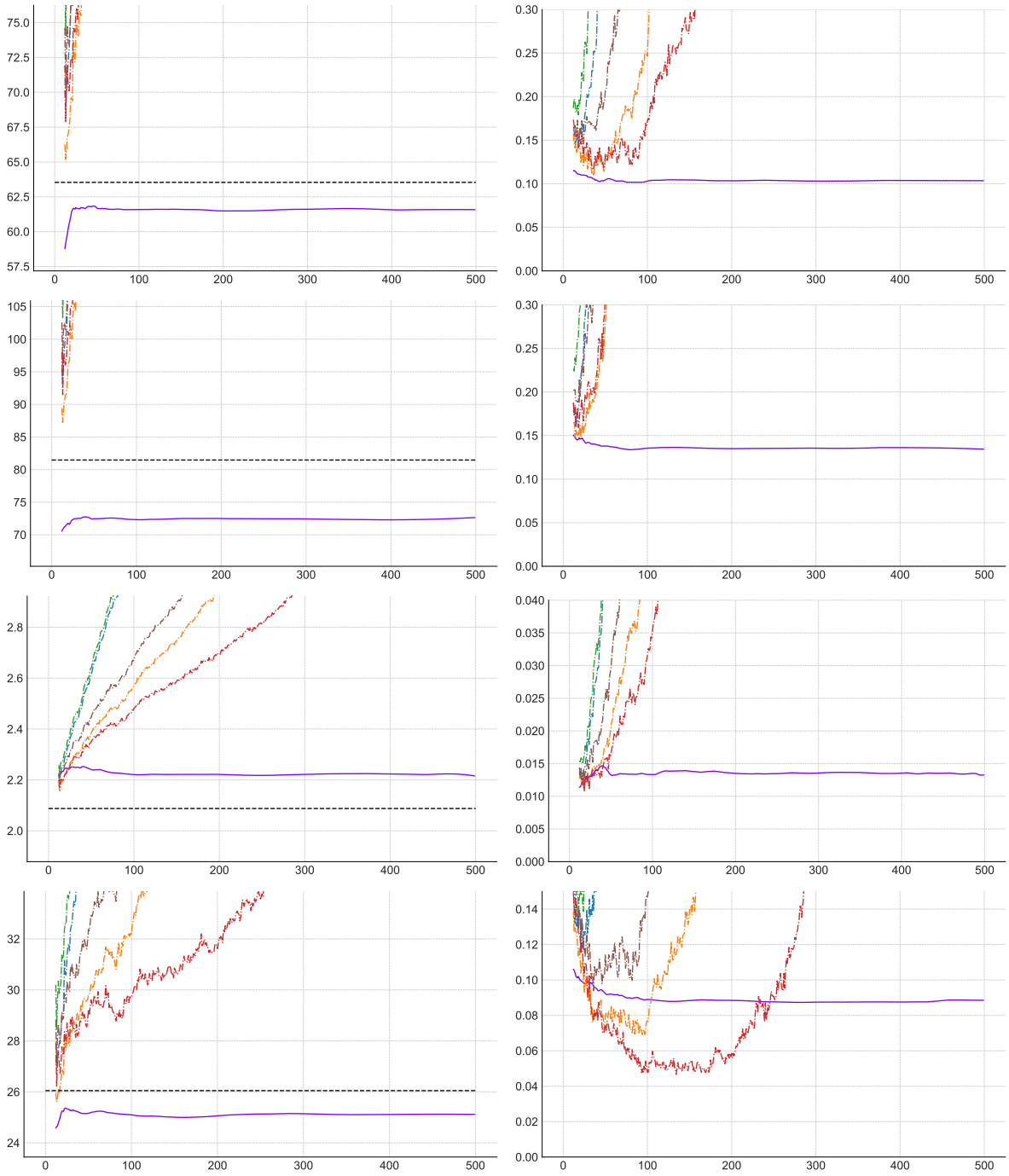
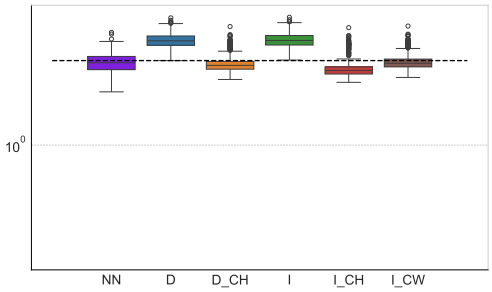
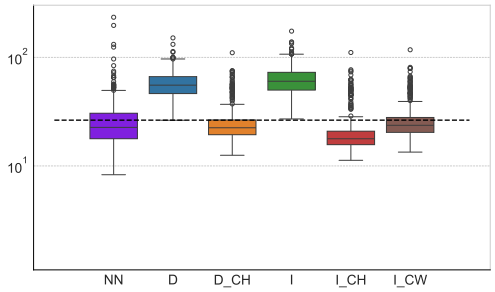


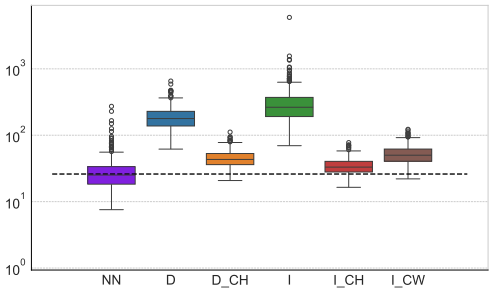
Figure 3: Illustration on simulated data sets of size $n = 500$ from GPD, Fisher-Snedecor (both with $\gamma = 0.4$) and Burr distributions with $(\gamma = 0.1, \rho_2 = -1/4)$ and $(\gamma = 0.4, \rho_2 = -1/2)$ from top to bottom. Median of the estimators (left panel) of the extreme ES (black dashed line) at level $1 - \alpha_n = 1 - 1/(2n)$ and RMedSE (right panel), as functions of $k \in \{2, \ldots, n-1\}$, computed on $R = 500$ replications, associated with $\widehat{\mathtt{ES}}^{\mathtt{D}}$(blue), $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$(orange), $\widehat{\mathtt{ES}}^{\mathtt{I}}$(green), $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$(red), $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$(brown) and NN (purple) estimators.



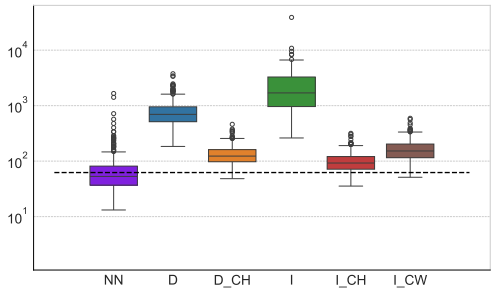
(a) $\gamma = 0.1$, $\rho_2 = -1$



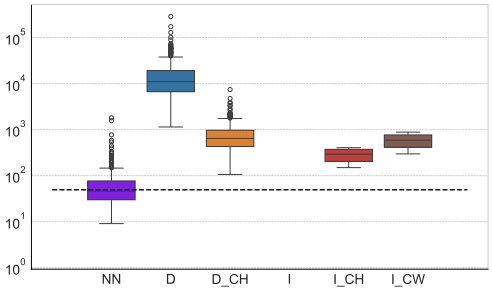
(b) $\gamma = 0.4$, $\rho_2 = -1$



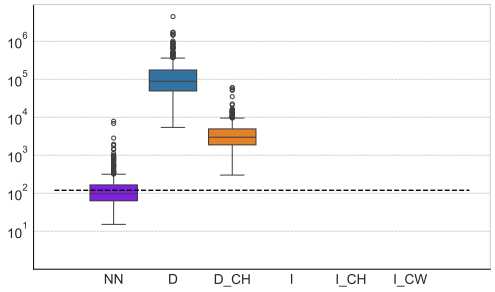
(c) $\gamma = 0.4$, $\rho_2 = -1/2$



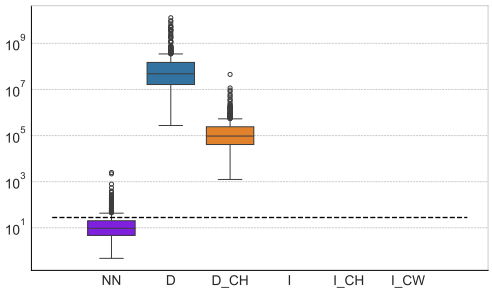
(d) $\gamma = 0.5$, $\rho_2 = -1/2$



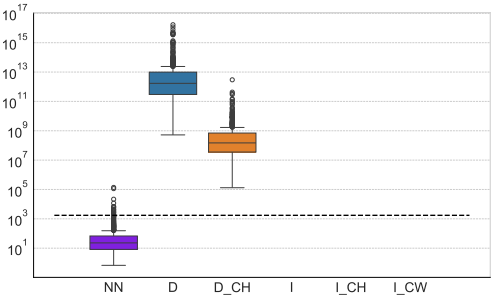
(e) $\gamma = 0.5$, $\rho_2 = -1/4$



(f) $\gamma = 0.6$, $\rho_2 = -1/4$



(g) $\gamma = 0.6$, $\rho_2 = -1/8$



(h) $\gamma = 0.9$, $\rho_2 = -1/8$

Figure 4: Boxplot (in a logarithmic scale) of the six estimators NN (purple), $\widehat{\mathtt{ES}}^{\mathtt{D}}$(blue), $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$(orange), $\widehat{\mathtt{ES}}^{\mathtt{I}}$(green), $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$(red) and $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$(brown) of the extreme ES at level $1 - \alpha_n = 1 - 1/(2n)$ on 500 replications of the Burr distribution. The real ES value is depicted by a horizontal black dashed line.

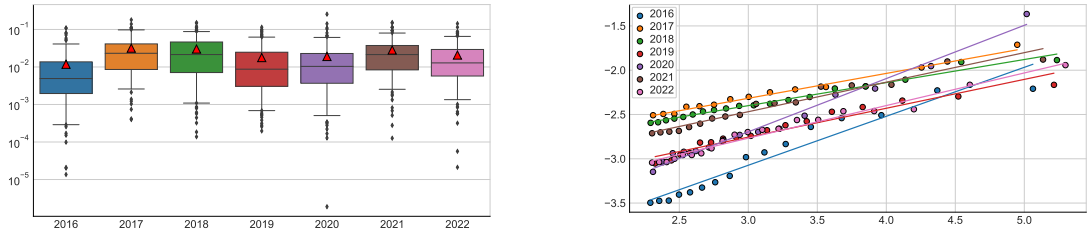
sample; this property holds at the daily scale, see the Pearson's autocorrelation plot in [48, Figure 1].

For each year $j \in \{1, \dots, 7\}$, denote by $\mathcal{X}_j^{\text{test}} = \left\{X_{n^{(j)}-i+1,n^{(j)}}, i = 1, \dots, \lceil n^{(j)}(1-\xi)\rceil\right\}$ the test set composed of the $\lceil n^{(j)}(1-\xi)\rceil$ largest order statistics with $\xi = 0.9$ and $\mathcal{X}_j^{\text{train}}$ the remaining data set with size $\lceil n^{(j)}\xi\rceil$. Clearly, $X_{n^{(j)}-\lceil n^{(j)}(1-\xi)\rceil+1,n^{(j)}}$ can be interpreted as the empirical estimator computed on $\mathcal{X}_j^{\text{test}}$ of the extreme quantile $q(1-1/\lceil n^{(j)}\xi\rceil)$. Similarly, our objective is, for each year $j \in \{1, \dots, 7\}$, to compute the extreme ES of level $(1-1/\lceil n^{(j)}\xi\rceil)$ on $\mathcal{X}_j^{\text{train}}$ using the six above competitors and to compare the estimations to the empirical estimation computed on $\mathcal{X}_j^{\text{test}}$ by the sample average. The estimation procedure (18) as well as the model selection (19) are similar to the ones used for the simulated data with hyperparameters $\zeta = 0$, $K_1 = 10$, $K_2 \in \{1, 3, 6\}$), a batch size of 32 and 500 epochs.

The nice stability of the NN estimator with respect to the anchor points is again confirmed on this real data set as illustrated in Figure 6a, and highly appreciated compared to the other estimators which have an estimation highly dependent on the anchor point. All the estimations are reported in Figure 6b for each year, and, it appears that the NN estimator is the closest to the empirical one. Here again, the results associated with the NN estimator are more satisfying than those obtained with other extreme ES estimators.

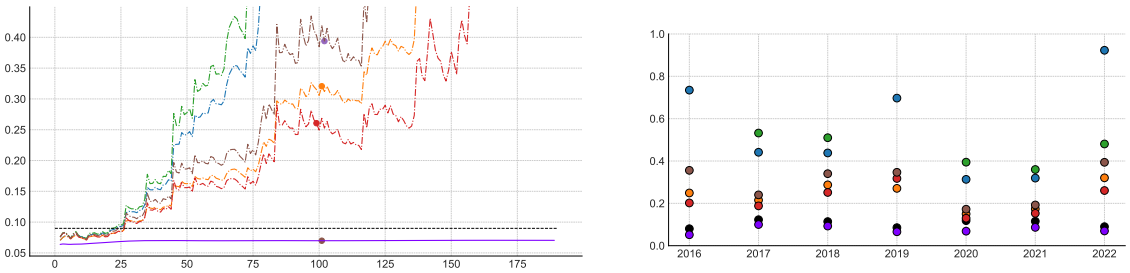
5. Conclusion

We have introduced, to the best of our knowledge, the first NN approach dedicated to extreme ES and CTM estimation. From the theoretical point of view, the uniform convergence rate of the approximations underpinning the estimator is established within an extreme-value framework. The nice accuracy properties of our ES and CTM estimator are partly due to the suitable combination of NN with both eLU and ReLU activation functions. From the practical point of view, our estimator has been tested on both simulated and real data; showing in the former case that the NN estimator outperforms most of usual estimators in challenging heavy-tailed situations. It has been noticed that the anchor point selection can be performed in much stabler way than



(a) (b)

Figure 5: Illustration on real data. (a): Boxplot of the dataset (in a logarithmic scale) for each considered year. The mean is represented by a red triangle and the whiskers represent the 5th and 95th percentile. (b): Log quantile-quantile plots $\log((n^{(j)}+1)/i) \mapsto \log X^{(j)}_{n^{(j)}-i+1,n^{(j)}}$, for $i \in \{1,\dots,\lceil(1-\xi)n^{(j)}\rceil\}$ on the selected years $j \in \{1,\dots,7\}$ at probability level $\xi = 0.90$.



(a) (b)

Figure 6: Illustration on real data: Estimation of the extreme ES associated with empirical (black), $\widehat{\mathtt{ES}}^{\mathtt{D}}$(blue), $\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$(orange), $\widehat{\mathtt{ES}}^{\mathtt{I}}$(green), $\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$(red), $\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$(brown) and NN (purple) estimators. (a): Estimated ES as functions of $k \in \{2,\dots,n-1\}$ for the year 2022. The empirical estimate is depicted with a horizontal black dashed line. (b): Estimated ES at the selected anchor point for all considered years. The vertical axis is clipped between 0 and 1.

what it is traditionally done. In the cryptocurrency data application, the NN estimator is the closest to the empirical one when investigating extreme loss returns.

Our further work will consist in adapting this framework to the estimation of conditional extreme Expected Shortfalls, taking inspiration from previous works dedicated to conditional extreme quantiles [7, 45]. Besides, since expectiles [28] and distorsion risk measures [11] are possible alternatives to the VaR, their estimation in distribution tails using NN can also be of interest, both in conditional and unconditional settings. Finally, this work could also be extended to other distribution tails (without necessarily positive tail indices) and compared to the associated bias reduction techniques [50].

Acknowledgments: The authors thank the referees and Associate Editor for their constructive remarks which improve the manuscript. This work is supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003). S. Girard and E. Gobet also acknowledge the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas.

	Distribution (parameters)	Density function f Expected Shortfall $\texttt{ES}(1-\alpha)$	γ	ρ_2
	Pareto	$f(t) = \theta t^{-\theta-1},\ t \geq 1$	$1/\theta$	$-\infty$
	$(\theta > 1)$	$\texttt{ES}(1-\alpha) = \dfrac{\theta \alpha^{-1/\theta}}{\theta - 1}$		
	Student's t	$f(t) = \dfrac{1}{\sqrt{\nu} B\left(\nu/2, 1/2\right)} \left(1 + \dfrac{t^2}{\nu}\right)^{-(\nu+1)/2},\ t \in \mathbb{R}$	$1/\nu$	$-2/\nu$
	$(\nu > 1)$	$\texttt{ES}(1-\alpha) = \dfrac{\nu + q^2(1-\alpha) f(q(1-\alpha))}{(\nu-1)\alpha}$		
IGG	Fréchet	$f(t) = \theta t^{-\theta-1} \exp\left(-t^{-\theta}\right),\ t > 0$	$1/\theta$	-1
	$(\theta > 1)$	$\texttt{ES}(1-\alpha) = \dfrac{1}{\alpha} \Gamma_\ell(1 - 1/\theta, -\log(1-\alpha))$		
	Inverse gamma	$f(t) = \dfrac{1}{\Gamma(\zeta)} t^{-\zeta-1} \exp(-1/t),\ t > 0$	$1/\zeta$	$-1/\zeta$
	$(\zeta > 1)$	$\texttt{ES}(1-\alpha) = \dfrac{1}{\alpha} \dfrac{\Gamma_\ell(\zeta - 1, 1/q(1-\alpha))}{\Gamma(\zeta)}$		
BFS	Burr*	$f(t) = \zeta\theta t^{\zeta-1} \left(1 + t^\zeta\right)^{-\theta-1},\ t > 0$	$1/(\zeta\theta)$	$-1/\theta$
	$(\zeta > 0, \zeta\theta > 1)$	$\texttt{ES}(1-\alpha) = \frac{\theta\zeta q(1-\alpha)}{(\theta\zeta - 1)} {}_2F_1\left(-\frac{1}{\zeta}, 1; \theta + 1 - 1/\zeta; \frac{1}{1-\alpha^{-1/\theta}}\right)$		
	GPD	$f(t) = (1 + \xi t)^{-1-1/\xi},\ t > 0$	ξ	$-\xi$
	$(0 < \xi < 1)$	$\texttt{ES}(1-\alpha) = \dfrac{\alpha^{-\xi} + \xi - 1}{\xi(1-\xi)}$		
	Fisher-Snedecor	$f(t) = \dfrac{(\nu_1/\nu_2)^{\nu_1/2}}{B(\nu_1/2, \nu_2/2)} t^{\nu_1/2-1} \left(1 + \dfrac{\nu_1}{\nu_2} t\right)^{-(\nu_1+\nu_2)/2},\ t > 0$	$2/\nu_2$	$-2/\nu_2$
	$(\nu_1 > 0, \nu_2 > 2)$	$\texttt{ES}(1-\alpha) = \frac{(\nu_1/\nu_2)^{-\nu_2/2} q(1-\alpha)^{1-\nu_2/2}}{\alpha B(\nu_1/2, \nu 2/2)(\nu_2/2-1)} {}_2F_1\left(\frac{\nu_1+\nu_2}{2}, \frac{\nu_2}{2} - 1; \frac{\nu_2}{2}; -\frac{\nu_2}{\nu_1 q(1-\alpha)}\right)$		

Table 1: Examples of heavy-tailed distributions (divided in four classes) satisfying the J-th order condition ($\mathbf{B_J}$) with the associated values of $\texttt{ES}(1-\alpha)$, γ and ρ_2. Here, $\Gamma(\cdot)$, $\Gamma_\ell(\cdot,\cdot)$, $B(\cdot,\cdot)$ and ${}_2F_1(\cdot,\cdot;\cdot;\cdot)$ respectively denote the gamma, lower incomplete gamma, beta and hypergeometric functions. See [8, Appendix B] for derivation details associated with IGG and BFS families of distributions respectively.
* Note that the expression of the ES associated with the Burr distribution provided in [36] is incorrect.

	$\widehat{\mathtt{ES}}^{\mathtt{NN}}_{\hat{\theta}}$	$\widehat{\mathtt{ES}}^{\mathtt{D}}$	$\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$
Pareto ($\rho_2 = -\infty$)						
$\gamma = 0.1$	0.0026	0.0009	0.0030	**0.0009**	0.0034	0.0022
$\gamma = 0.4$	0.0734	0.0167	0.0513	**0.0163**	0.0602	0.0423
$\gamma = 0.5$	0.1250	0.0278	0.0816	**0.0268**	0.0994	0.0716
$\gamma = 0.6$	0.1951	0.0459	0.1256	**0.0442**	0.1593	0.1139
$\gamma = 0.9$	0.6627	**0.2648**	0.3651	0.2654	0.6774	0.5994
Student's t ($\rho_2 = -2\gamma$)						
$\gamma = 0.1$	**0.0187**	-	-	-	-	-
$\gamma = 0.4$	0.0782	-	-	0.1081	**0.0247**	0.2327
$\gamma = 0.5$	0.1605	-	**0.0372**	-	0.0760	0.0430
$\gamma = 0.6$	0.2068	-	0.0844	-	0.2355	**0.0697**
$\gamma = 0.9$	0.6606	**0.2644**	0.6477	-	0.8147	0.6682

Table 2: RMedSE associated with six estimators of $\mathtt{ES}(1-\alpha_n = 1-1/(2n))$ on two heavy-tailed distributions. The best result is emphasized in bold. RMedSEs larger than 1 are not reported (as mentioned before, these large values correspond to errors likely to be larger than 100%).

	$\widehat{\mathtt{ES}}^{\mathtt{NN}}_{\hat{\theta}}$	$\widehat{\mathtt{ES}}^{\mathtt{D}}$	$\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$
Fréchet ($\rho_2 = -1$)						
$\gamma = 0.1$	0.0027	0.0095	0.0026	0.0101	0.0039	**0.0021**
$\gamma = 0.4$	0.0748	0.2041	0.0458	0.2540	0.0650	**0.0376**
$\gamma = 0.5$	0.1415	0.3404	0.0715	0.5052	0.1027	**0.0618**
$\gamma = 0.6$	0.2306	0.5266	0.1133	-	0.1540	**0.1037**
$\gamma = 0.9$	0.6755	**0.2446**	0.4083	-	0.5487	0.4478
Inverse gamma ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	0.0111	0.7545	0.0409	0.8272	**0.0105**	0.0622
$\gamma = 0.4$	0.0936	-	0.0479	-	**0.0374**	0.0723
$\gamma = 0.5$	0.1520	-	0.0665	-	**0.0647**	0.0685
$\gamma = 0.6$	0.2163	-	**0.0903**	-	0.1130	0.0920
$\gamma = 0.9$	0.6762	**0.3630**	0.3941	-	0.5481	0.4533

Table 3: RMedSE associated with six estimators of $\mathtt{ES}(1-\alpha_n = 1-1/(2n))$ on two heavy-tailed distributions included in the IGG family [8]. The best result is emphasized in bold. RMedSEs larger than 1 are not reported.

	$\widehat{\mathtt{ES}}^{\mathtt{NN}}_{\hat{\theta}}$	$\widehat{\mathtt{ES}}^{\mathtt{D}}$	$\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$
Burr ($\rho_2 = -1$)						
$\gamma = 0.1$	0.0038	0.0420	0.0036	0.0435	0.0100	**0.0021**
$\gamma = 0.4$	0.0831	-	0.0457	-	0.1264	**0.0351**
$\gamma = 0.5$	0.1386	-	0.0693	-	0.1828	**0.0560**
$\gamma = 0.6$	0.2170	-	0.0979	-	0.2579	**0.0879**
$\gamma = 0.9$	0.6805	-	0.4261	-	0.5718	**0.3878**
Burr ($\rho_2 = -1/2$)						
$\gamma = 0.1$	0.0062	0.3896	0.0128	0.4165	**0.0022**	0.0233
$\gamma = 0.4$	0.0874	-	0.4337	-	**0.0816**	0.8469
$\gamma = 0.5$	**0.1461**	-	0.9603	-	0.2350	-
$\gamma = 0.6$	**0.2046**	-	-	-	0.8951	-
$\gamma = 0.9$	**0.6691**	-	-	-	-	-
Burr ($\rho_2 = -1/4$)						
$\gamma = 0.1$	**0.0135**	-	0.2536	-	0.1203	0.3733
$\gamma = 0.4$	**0.1266**	-	-	-	-	-
$\gamma = 0.5$	**0.1896**	-	-	-	-	-
$\gamma = 0.6$	**0.2032**	-	-	-	-	-
$\gamma = 0.9$	**0.7341**	-	-	-	-	-
Burr ($\rho_2 = -1/8$)						
$\gamma = 0.1$	**0.0427**	-	-	-	-	-
$\gamma = 0.4$	**0.1822**	-	-	-	-	-
$\gamma = 0.5$	**0.2639**	-	-	-	-	-
$\gamma = 0.6$	**0.4219**	-	-	-	-	-
$\gamma = 0.9$	**0.9766**	-	-	-	-	-
GPD ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	**0.0464**	-	-	-	-	-
$\gamma = 0.4$	**0.1030**	-	-	-	0.6973	-
$\gamma = 0.5$	**0.1736**	-	0.9603	-	0.2350	-
$\gamma = 0.6$	0.2335	-	0.3596	-	**0.0975**	-
$\gamma = 0.9$	0.6775	-	0.3383	-	0.4861	**0.3132**
Fisher-Snedecor ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	**0.0654**	-	-	-	-	-
$\gamma = 0.4$	**0.1352**	-	-	-	-	-
$\gamma = 0.5$	**0.1828**	-	-	-	-	-
$\gamma = 0.6$	**0.2456**	-	-	-	-	-
$\gamma = 0.9$	0.6928	-	**0.1802**	-	-	-

Table 4: RMedSE associated with six estimators of $\mathtt{ES}(1 - \alpha_n = 1 - 1/(2n))$ on three heavy-tailed distributions included in the BFS family [8]. The best result is emphasized in bold. RMedSEs larger than 1 are not reported.

Period	$n^{(j)}$: nb. data	Mean	Std.	Largest return	Tail-index
2016	157	1.18	1.82	11	0.55
2017	140	3.17	3.15	18	0.28
2018	187	3.01	2.98	15	0.26
2019	183	1.78	2.16	11	0.33
2020	150	1.90	2.91	25	0.60
2021	169	2.83	2.77	15	0.33
2022	199	2.08	2.19	14	0.37

Table 5: Real data statistics: number of data, mean, standard deviation, largest negative daily log-return (last three scaled by 10^2) and estimated tail-index for each year.

Appendix A. Preliminary results

Lemma 6. *Let L be a slowly-varying function with Karamata representation*

$$L(t) = c(t) \exp\left(\int_{t_0}^{t} \frac{\eta(s)}{s} \, \mathrm{d}s \right),$$

where $t \geq t_0 > 0$, $c(t) \to c_0 > 0$ and $\eta(t) \to 0$ as $t \to \infty$.

*(i) If $|\eta| \in \mathcal{RV}_{\rho_2}$ with $\rho_2 < 0$ and L is normalized i.e. $c(t) = c_0$ for all $t \geq t_0$, then L satisfies (**D**).*

*(ii) Conversely, if L satisfies (**D**) and $(\log L)'$ has asymptotically constant sign, then L satisfies ($\mathbf{B_J}$) with $J = 2$.*

Proof. (i) Since L is normalized, one has for all $t \geq t_0$,

$$\log\left(\frac{L(t)}{L(t_0)} \right) = \int_{t_0}^{t} \frac{\eta(s)}{s} \, \mathrm{d}s.$$

It is thus clear that $t \geq t_0 \mapsto \log L(t)$ is differentiable and $(\log L)'(t) = \eta(t)/t$ which proves the result.

(ii) Assumption (**D**) implies that there exists $L_2 \in \mathcal{RV}_0$ such that $|(\log L)'(x)| = x^{\rho_2 - 1} L_2(x)$. Suppose for instance that $(\log L)'$ is eventually postive. Integrating, one obtains for $t, z > 0$,

$$\log\left(\frac{L(tz)}{L(t)} \right) = \int_{t}^{tz} x^{\rho_2 - 1} L_2(x) \, \mathrm{d}x = t^{\rho_2} L_2(t) \int_{1}^{z} y^{\rho_2 - 1} \frac{L_2(ty)}{L_2(t)} \, \mathrm{d}y,$$

with $L_2(ty)/L_2(t) \to 1$ as $t \to \infty$ for all $y > 0$. Combining Potter bounds [12, Theorem 1.5.6, p.25] with Lebesgue's dominated convergence theorem shows that

$$\lim_{t \to \infty} \frac{1}{A_2(t)} \log\left(\frac{L(tz)}{L(t)} \right) = \int_{1}^{z} y^{\rho_2 - 1} \, \mathrm{d}y = R_2(z),$$

where $A_2(t) := t^{\rho_2} L_2(t)$. This is assumption ($\mathbf{B_J}$) with $J = 2$, the result is thus proved. □

Let $I \subset \mathbb{R}$. In the following, $\mathcal{C}^{0,\xi}(I)$ denotes the set of Hölder continuous functions on I with exponent $\xi \in (0, 1]$. The next Lemma establishes the regularity properties of Ψ_p depending on the assumptions made on the slowly-varying function L.

Lemma 7. *Let L be a continuously differentiable slowly-varying function, $p\gamma \in (0,1)$ and consider the function Ψ_p defined in (7). Then,*

(i) Ψ_p is differentiable on $(0,1]$ and can be continuously extended to $[0,1]$ by setting $\Psi_p(0) = -\log(1-p\gamma)$.

(ii) If, moreover, (D) holds, then for all $\varepsilon \in (0, |\rho_2|)$, $\Psi_p \in \mathcal{C}^{0,\rho_2^\varepsilon \wedge 1}([0,1])$ where we have set $\rho_2^\varepsilon := -\rho_2 - \varepsilon = |\rho_2| - \varepsilon > 0$. The associated $(\rho_2^\varepsilon \wedge 1)$-Hölder constant can be written as $H^\varepsilon_{\Psi_p} = pH^\varepsilon_{\Psi_1}$.

Proof. (i) It is clear that

$$u \in (0,1] \mapsto \Psi_p(u) = \log\left(\int_0^1 w^{-p\gamma}\,\frac{L^p(1/(wu))}{L^p(1/u)}\,\mathrm{d}w\right)$$

is differentiable on $(0,1]$. By the slowly-varying property, one has $L(1/(wu))/L(1/u) \to 1$ as $u \to 0^+$ for all $w > 0$. From Potter bounds [12, Theorem 1.5.6, p.25], the dominated convergence theorem applies and then $\Psi_p(u) \to \log(\int_0^1 w^{-p\gamma}\,\mathrm{d}w)$ as $u \to 0^+$, the first part of the result follows.

(ii) Let $\varepsilon \in (0, |\rho_2|)$. From (D), $w \geq 1 \mapsto h(w) := w^{1-\rho_2-\varepsilon}|(\log L)'(w)|$ is regularly-varying with index $-\varepsilon < 0$ and thus $h(w) \to 0$ as $w \to \infty$. As a consequence, there exists $C_\varepsilon > 0$ such that $0 \leq h(w) \leq C_\varepsilon$ or equivalently $0 \leq |(\log L)'(w)| \leq C_\varepsilon w^{\rho_2 - 1 + \varepsilon}$ for all $w \geq 1$. Then, for $0 < u \leq v \leq 1$,

$$\begin{aligned}|\log L(1/v) - \log L(1/u)| &\leq \int_{1/v}^{1/u} |(\log L)'(w)|\,\mathrm{d}w \\ &\leq C_\varepsilon \int_{1/v}^{1/u} w^{\rho_2 - 1 + \varepsilon}\,\mathrm{d}w = \frac{C_\varepsilon}{|\rho_2| - \varepsilon}\left(v^{-\rho_2-\varepsilon} - u^{-\rho_2-\varepsilon}\right). \quad \text{(A.1)}\end{aligned}$$

Let us decompose the increments of Ψ_p as

$$\begin{aligned}\Psi_p(v) - \Psi_p(u) &= \log\left(\int_0^1 w^{-p\gamma}\,\frac{L^p(1/(wv))}{L^p(1/v)}\,\mathrm{d}w\right) - \log\left(\int_0^1 w^{-p\gamma}\,\frac{L^p(1/(wu))}{L^p(1/u)}\,\mathrm{d}w\right) \\ &= p(\log L(1/u) - \log L(1/v)) \quad \text{(A.2)} \\ &\quad + \log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wv))\,\mathrm{d}w\right) \\ &\quad - \log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wu))\,\mathrm{d}w\right). \quad \text{(A.3)}\end{aligned}$$

The first term (A.2) can be handled as follows. One has, for any $0 < u \leq v \leq 1$ and any $\beta > 0$,

$$0 \leq v^\beta - u^\beta \leq (\beta \vee 1)(v-u)^{\beta \wedge 1}. \tag{A.4}$$

Combining (A.1) and (A.4) yields

$$p\ |\log L(1/v) - \log L(1/u)| \leq p\,(\rho_2^\varepsilon \vee 1)\frac{C_\varepsilon}{\rho_2^\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1} =: p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1}. \tag{A.5}$$

We thus have proved that (A.2) satisfies a Hölder continuity condition with Hölder index $\rho_2^\varepsilon \wedge 1$. Let us now focus on (A.3). Similarly to (A.5), one has, for any $w \in (0,1)$ any $(u,v) \in (0,1]^2$,

$$p\ |\log L(1/(wv)) - \log L(1/(wu))| \leq p\,C_{2,\varepsilon}|wv - wu|^{\rho_2^\varepsilon \wedge 1} \leq p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1},$$

or equivalently,

$$\exp(-p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1}) \leq \frac{L^p(1/(wv))}{L^p(1/(wu))} \leq \exp(p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1}).$$

Therefore, (A.3) is controlled as

$$\begin{aligned}
&\log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wv))\,\mathrm{d}w\right) - \log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wu))\,\mathrm{d}w\right)\\
&\leq \log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wu))\exp(p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1})\,\mathrm{d}w\right) - \log\left(\int_0^1 w^{-p\gamma}\,L^p(1/(wu))\,\mathrm{d}w\right)\\
&\leq p\,C_{2,\varepsilon}|v-u|^{\rho_2^\varepsilon \wedge 1}.
\end{aligned}$$

A similar lower bound can be easily derived, and this proves that (A.3) also satisfies the Hölder continuity criterion with exponent ρ_2^ε. All in all, we have proved that

$$|\Psi_p(v) - \Psi_p(u)| \leq H_{\Psi_p}^\varepsilon |v-u|^{\rho_2^\varepsilon \wedge 1},$$

where $H_{\Psi_p}^\varepsilon = 2p\,C_{2,\varepsilon}$. □

Our next goal is to study the uniform approximation error of a $\mathcal{C}^{0,\xi}([0,1])$-function, $\xi \in (0,1]$ by a feedforward ReLU-NN. To this end, consider the triangular function

$\widehat{\sigma}^{\mathtt{R}} : \mathbb{R} \to [-1, 1]$ built using three shifted ReLU functions $x \in \mathbb{R} \mapsto \sigma^{\mathtt{R}}(x) := \max(0,\, x)$:

$$\widehat{\sigma}^{\mathtt{R}}(x) := \sigma^{\mathtt{R}}(x+1) - 2\sigma^{\mathtt{R}}(x) + \sigma^{\mathtt{R}}(x-1) = \begin{cases} 1+x, & \text{if } -1 \leq x \leq 0, \\ 1-x, & \text{if } 0 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

It is then possible to control the uniform error between the function f and its piece-wise linear approximation based on triangular functions, depending on the regularity of f.

Lemma 8. *Let $\widehat{\sigma}^{\mathtt{R}}$ be a triangular function, $\tau \in (0,1)$ and $f \in \mathcal{C}^{0,\xi}([0,\tau])$ with $\xi \in (0,1]$. For all $M \in \mathbb{N}\backslash\{0\}$, let $h = \tau/M$ and $u_j = jh$ for $j \in \{0, \ldots, M\}$. Then,*

$$\sup_{u\in[0,\tau]} \left| f(u) - \sum_{j=0}^{M} f(u_j)\widehat{\sigma}^{\mathtt{R}}\left(\frac{u-u_j}{h}\right)\right| \leq H_f\,(\tau/M)^{\xi},$$

where H_f is the Hölder constant associated with f.

Proof. The proof follows the same lines as the one of [6, Lemma 12]. Clearly,

$$\sup_{u\in[0,\tau]} \left| f(u) - \sum_{j=0}^{M} f(u_j)\widehat{\sigma}^{\mathtt{R}}\left(\frac{u-u_j}{h}\right)\right| =: \max_{i\in\{0,\ldots,M-1\}} \sup_{u\in[u_i,u_{i+1}]} |\Delta_i(u)|\,,$$

where

$$\Delta_i(u) := f(u) - \left(f(u_i)\left(\frac{u_{i+1}-u}{h}\right) + f(u_{i+1})\left(\frac{u-u_i}{h}\right)\right).$$

As a consequence, if $f \in \mathcal{C}^{0,\xi}([0,\tau])$, one has

$$|\Delta_i(u)| \leq |f(u) - f(u_i)|\left(\frac{u_{i+1}-u}{h}\right) + |f(u) - f(u_{i+1})|\left(\frac{u-u_i}{h}\right) \leq H_f\, h^{\xi},$$

and the conclusion follows. □

Note that the above construction involves $K = 3(M+1) \geq 6$ neurons.

Appendix B. Proofs of main results

Proof of Proposition 1. Assume $p\gamma \in (0,1)$. From (1), (3) and (4), one has

$$\mathtt{CTM}_p(1-\alpha) = \frac{1}{\alpha}\int_0^{\alpha} u^{-p\gamma} L^p(1/u)\,\mathrm{d}u = \frac{1}{\alpha}\int_{1/\alpha}^{\infty} v^{p\gamma-2} L(v)\,\mathrm{d}v,$$

where L is slowly-varying and [12, Proposition 1.5.10] proves the result. □

Proof of Proposition 2. Let $(\alpha, u) \in (0,1)^2$. Starting from (4), the regular variation property yields

$$U(1/u) = U(1/\alpha)(\alpha/u)^{\gamma}\frac{L(1/u)}{L(1/\alpha)},$$

or equivalently using (3), for all $p\gamma \in (0,1)$,

$$q^p(1-u) = q^p(1-\alpha)\,(\alpha/u)^{p\gamma}\frac{L^p(1/u)}{L^p(1/\alpha)}. \tag{B.1}$$

Replacing in (1) yields

$$\begin{aligned}\log \texttt{CTM}_p(1-\alpha) &= p\log q(1-\alpha) + \log\left(\frac{1}{\alpha}\int_0^{\alpha}\left(\frac{u}{\alpha}\right)^{-p\gamma}\frac{L^p(1/u)}{L^p(1/\alpha)}\,\mathrm{d}u\right)\\ &= p\log q(1-\alpha) + \log\left(\int_0^1 w^{-p\gamma}\frac{L^p(1/(w\alpha))}{L^p(1/\alpha)}\,\mathrm{d}w\right)\\ &= p\log q(1-\alpha) + \Psi_p(\alpha),\end{aligned}$$

and (9) is thus proved. As an immediate consequence, one has

$$\begin{aligned}\log \texttt{CTM}_p(1-\alpha) - \log \texttt{CTM}_p(1-\delta) &= p(\log q(1-\alpha) - \log q(1-\delta))\\ &\quad + \Psi_p(\alpha) - \Psi_p(\delta),\end{aligned} \tag{B.2}$$

while (6) and (B.1) entail

$$\log q(1-\alpha) - \log q(1-\delta) = \gamma\log(\delta/\alpha) + \varphi\big(\log(\delta/\alpha), \log(1/\delta)\big), \tag{B.3}$$

and we thus find back the expansion of the log-spacings (8). Combining (B.2) and (B.3) proves (10). □

Proof of Proposition 3. Lemma 7(ii) shows that $\Psi_p \in \mathcal{C}^{0,\rho_2^{\varepsilon}}([0,1])$ for all $\varepsilon \in (0, |\rho_2|)$. The result thus follows from Lemma 8: for any $K_1 \geq 6$,

$$\sup_{u\in[0,1]}\left|\Psi_p(u) - \tilde{\Psi}^{K_1}_{p,\theta_1}(u)\right| \leq pH^{\varepsilon}_{\Psi_1}(K_1/3-1)^{-(\rho_2^{\varepsilon}\wedge 1)} \leq pH^{\varepsilon}_{\Psi_1}(6/K_1)^{\rho_2^{\varepsilon}\wedge 1} \leq 6pH^{\varepsilon}_{\Psi_1}(1/K_1)^{\rho_2^{\varepsilon}\wedge 1},$$

which is the desired result. □

Proof of Proposition 4. Following [7, Lemma 6, Supplementary Material], there exists a NN defined as in (15) such that, for all $\varepsilon > 0$, there exists y_{ε} such that

$$\varphi(y_1, y_2) = \tilde{\varphi}^{K_2}_{\theta_2}(y_1, y_2) + \Delta(\exp(y_1), \exp(y_2))\prod_{j=2}^{J} A_j(\exp(y_2)),$$

with $|\Delta(\exp(y_1),\exp(y_2))| \leq \varepsilon \exp(y_1(\bar{\rho}_J+\varepsilon))$ for all $y_1, y_2 \geq y_\varepsilon$. It follows from $(\mathbf{C_J})$ that

$$\left|\varphi(y_1,y_2) - \tilde{\varphi}_{\theta_2}^{K_2}(y_1,y_2)\right| \leq \varepsilon \bar{c}_J \exp((\bar{\rho}_J+\varepsilon)(y_1+y_2) - y_2\varepsilon),$$

where we have introduced $\bar{c}_J = \prod_{j=2}^{J} |c_j|$. □

Proof of Theorem 5. Collecting (12) and (16), one has

$$\begin{aligned}
&\inf_{\theta\in\Theta} \left|\log \mathtt{CTM}_p(1-\alpha_n) - \log \widetilde{\mathtt{CTM}}_{p,\theta}^{K}(1-\alpha_n; 1-\delta_n)\right| \\
&\leq p \inf_{\theta_2\in\Theta_2} \left|\left(\varphi - \tilde{\varphi}_{\theta_2}^{K_2}\right)(\log(\delta_n/\alpha_n), \log(1/\delta_n))\right| \\
&+ \inf_{\theta_1\in\Theta_1} \left|\left(\Psi_p - \tilde{\Psi}_{p,\theta_1}^{K_1}\right)(\alpha_n)\right| + \left|\left(\Psi_p - \tilde{\Psi}_{p,\theta_1}^{K_1}\right)(\delta_n)\right|.
\end{aligned}$$

Let $\varepsilon > 0$. First, Proposition 3 entails that, for all $n \geq 1$,

$$\inf_{\theta_1\in\Theta_1} \left|\left(\Psi_p - \tilde{\Psi}_{p,\theta_1}^{K_1}\right)(\alpha_n)\right| + \left|\left(\Psi_p - \tilde{\Psi}_{p,\theta_1}^{K_1}\right)(\delta_n)\right| \leq 12pH_{\Psi_1}^{\varepsilon}(1/K_1)^{\rho_2^{\varepsilon}\wedge 1}.$$

Second, Proposition 4 implies that, for all $\varepsilon > 0$, there exists $n_\varepsilon \geq 1$ such that, for all $n \geq n_\varepsilon$,

$$\inf_{\theta_2\in\Theta_2} \left|\left(\varphi - \tilde{\varphi}_{\theta_2}^{K_2}\right)(\log(\delta_n/\alpha_n), \log(1/\delta_n))\right| \leq \bar{c}_J \varepsilon (\delta_n/\alpha_n)^{\varepsilon} \alpha_n^{-\bar{\rho}_J}.$$

The conclusion follows. □

Appendix C. Additional numerical results

The following Table C.6–Table C.8 collect the results obtained in the same setting as in Section 3 for a larger confidence level $1-\alpha_n = 1-1/(10n)$. Even though the obtained RMEdSE are slightly larger than in the case where $1-\alpha_n = 1-1/(2n)$, the conclusions are qualitatively the same, emphasizing the stability of the tested estimation methods.

	$\widehat{\texttt{ES}}^{\texttt{NN}}_{\hat{\theta}}$	$\widehat{\texttt{ES}}^{\texttt{D}}$	$\widehat{\texttt{ES}}^{\texttt{D,CH}}$	$\widehat{\texttt{ES}}^{\texttt{I}}$	$\widehat{\texttt{ES}}^{\texttt{I,CH}}$	$\widehat{\texttt{ES}}^{\texttt{I,CW}}$
Pareto ($\rho_2 = -\infty$)						
$\gamma = 0.1$	0.0037	0.0013	0.0045	**0.0013**	0.0050	0.0034
$\gamma = 0.4$	0.1115	0.0237	0.0753	**0.0233**	0.0860	0.0624
$\gamma = 0.5$	0.1902	**0.0389**	0.1169	0.0391	0.1390	0.1046
$\gamma = 0.6$	0.2730	0.0626	0.1662	**0.0620**	0.2137	0.1599
$\gamma = 0.9$	0.7487	0.3065	0.4371	**0.3151**	0.7515	0.6847
Student's t ($\rho_2 = -2\gamma$)						
$\gamma = 0.1$	**0.0719**	-	-	-	-	-
$\gamma = 0.4$	0.1221	0.0237	0.0753	**0.0233**	0.0860	0.0624
$\gamma = 0.5$	0.1981	-	**0.0520**	-	0.0709	0.0615
$\gamma = 0.6$	0.3099	-	0.1121	-	0.2525	**0.0941**
$\gamma = 0.9$	0.7763	**0.5384**	0.7223	-	0.8581	0.7576

Table C.6: RMedSE associated with six estimators of $\texttt{ES}(1 - \alpha_n = 1 - 1/(10n))$ on two heavy-tailed distributions. The best result is emphasized in bold. RMedSEs larger than 1 are not reported.

	$\widehat{\texttt{ES}}^{\texttt{NN}}_{\hat{\theta}}$	$\widehat{\texttt{ES}}^{\texttt{D}}$	$\widehat{\texttt{ES}}^{\texttt{D,CH}}$	$\widehat{\texttt{ES}}^{\texttt{I}}$	$\widehat{\texttt{ES}}^{\texttt{I,CH}}$	$\widehat{\texttt{ES}}^{\texttt{I,CW}}$
Fréchet ($\rho_2 = -1$)						
$\gamma = 0.1$	0.0042	0.0164	0.0040	0.0173	0.0054	**0.0034**
$\gamma = 0.4$	0.1339	0.3663	0.0616	0.4523	0.0880	**0.0549**
$\gamma = 0.5$	0.2241	0.6362	0.1021	0.8826	0.1364	**0.0903**
$\gamma = 0.6$	0.3259	-	0.1518	-	0.1959	**0.1481**
$\gamma = 0.9$	0.7713	0.5206	**0.4666**	-	0.5986	0.5072
Inverse gamma ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	**0.0226**	-	0.1172	-	0.0520	0.1567
$\gamma = 0.4$	0.1579	-	0.0997	-	**0.0497**	0.1512
$\gamma = 0.5$	0.2410	-	0.0945	-	**0.0852**	0.1104
$\gamma = 0.6$	0.3281	-	**0.1304**	-	0.1379	0.1327
$\gamma = 0.9$	0.7711	-	**0.4537**	-	0.5939	0.5160

Table C.7: RMedSE associated with six estimators of $\texttt{ES}(1 - \alpha_n = 1 - 1/(10n))$ on two heavy-tailed distributions included in the IGG family [8]. The best result is emphasized in bold. RMedSEs larger than 1 are not reported.

	$\widehat{\mathtt{ES}}^{\mathtt{NN}}_{\hat{\theta}}$	$\widehat{\mathtt{ES}}^{\mathtt{D}}$	$\widehat{\mathtt{ES}}^{\mathtt{D,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CH}}$	$\widehat{\mathtt{ES}}^{\mathtt{I,CW}}$
Burr ($\rho_2 = -1$)						
$\gamma = 0.1$	0.0063	0.0741	0.0045	0.0767	0.0111	**0.0029**
$\gamma = 0.4$	0.1231	-	0.0581	-	0.1435	**0.0489**
$\gamma = 0.5$	0.2091	-	0.0890	-	0.2078	**0.0783**
$\gamma = 0.6$	0.3018	-	0.1281	-	0.2769	**0.1176**
$\gamma = 0.9$	0.7733	-	0.4586	-	0.6038	**0.4424**
Burr ($\rho_2 = -1/2$)						
$\gamma = 0.1$	0.0133	0.8017	0.0314	0.8358	**0.0098**	0.0495
$\gamma = 0.4$	**0.1389**	-	-	-	0.3488	-
$\gamma = 0.5$	**0.2098**	-	-	-	0.9332	-
$\gamma = 0.6$	**0.2903**	-	-	-	-	-
$\gamma = 0.9$	**0.7450**	-	-	-	-	-
Burr ($\rho_2 = -1/4$)						
$\gamma = 0.1$	**0.0237**	-	0.6512	-	0.3858	0.8667
$\gamma = 0.4$	**0.2014**	-	-	-	-	-
$\gamma = 0.5$	**0.2691**	-	-	-	-	-
$\gamma = 0.6$	**0.2367**	-	-	-	-	-
$\gamma = 0.9$	**0.7494**	-	-	-	-	-
Burr ($\rho_2 = -1/8$)						
$\gamma = 0.1$	**0.1325**	-	-	-	-	-
$\gamma = 0.4$	**0.3532**	-	-	-	-	-
$\gamma = 0.5$	**0.2873**	-	-	-	-	-
$\gamma = 0.6$	**0.3893**	-	-	-	-	-
$\gamma = 0.9$	**0.9765**	-	-	-	-	-
GPD ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	**0.1746**	-	-	-	-	-
$\gamma = 0.4$	**0.1569**	-	-	-	-	-
$\gamma = 0.5$	**0.2083**	-	-	-	0.9332	-
$\gamma = 0.6$	0.2962	-	0.9232	-	**0.2345**	-
$\gamma = 0.9$	0.7730	-	0.3659	-	0.5142	**0.3586**
Fisher-Snedecor ($\rho_2 = -\gamma$)						
$\gamma = 0.1$	**0.1390**	-	-	-	-	-
$\gamma = 0.4$	**0.2033**	-	-	-	-	-
$\gamma = 0.5$	**0.2445**	-	-	-	-	-
$\gamma = 0.6$	**0.3126**	-	-	-	-	-
$\gamma = 0.9$	0.7735	-	**0.2278**	-	-	-

Table C.8: RMedSE associated with six estimators of $\mathtt{ES}(1-\alpha_n = 1-1/(10n))$ on three heavy-tailed distributions included in the BFS family [8]. The best result is emphasized in bold. RMedSEs larger than 1 are not reported.

References

[1] C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26(7):1505–1518, 2002.

[2] C. Acerbi and B. Szekely. Back-testing expected shortfall. *Risk*, 27(11):76–81, 2014.

[3] B. Acereda Serrano, A. M. León Valle, and J. Mora-López. Estimating the expected shortfall of cryptocurrencies: An evaluation based on backtesting. *Finance Research Letters*, 33:101181, 2020.

[4] M. Allouche, M. Echenim, E. Gobet, and A.-C. Maurice. Statistical error bounds for weighted mean and median, with application to robust aggregation of cryptocurrency data. *(tentatively accepted in) Mathematical Finance*, 2024.

[5] M. Allouche, J. El Methni, and S. Girard. A refined Weissman estimator for extreme quantiles. *Extremes*, 26(3):545–572, 2023.

[6] M. Allouche, S. Girard, and E. Gobet. EV-GAN: Simulation of extreme events with ReLU neural networks. *Journal of Machine Learning Research*, 23(150):1–39, 2022.

[7] M. Allouche, S. Girard, and E. Gobet. Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Statistics and Computing*, 34:12, 2024.

[8] M. Allouche, J. El Methni, and S. Girard. Reduced-bias estimation of the extreme conditional tail expectation for Box-Cox transforms of heavy-tailed distributions. *Journal of Statistical Planning and Inference*, 233, 106189, 2024.

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

[10] B. Bader, J. Yan, and X. Zhang. Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *The Annals of Applied Statistics*, 12(1), 310–329, 2018.

[11] R. Bairakdar, F. Godin, M. Mailhot and F. Yang. Estimation of generalized tail distortion risk measures with applications in reinsurance. *Available at SSRN*, 4826996, 2024.

[12] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987.

[13] V. Brazauskas, B. Jones, M. Puri, and R. Zitikis. Estimating conditional tail expectation with actuarial applications in view. *Journal of Statistical Planning and Inference*, 138(11):3590–3604, 2008.

[14] F. Caeiro, M. I. Gomes, and D. Pestana. Direct reduction of bias of the classical Hill estimator. *Revstat - Statistical Journal*, 3(2):113–136, 2005.

[15] J. Cai and H. Li. Conditional tail expectations for multivariate phase-type distributions. *Journal of Applied Probability*, 42(3):810–825, 2005.

[16] J.-J. Cai, J. Einmahl, L. de Haan, and C. Zhou. Estimation of the marginal expected shortfall: the mean when a related variable is extreme. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 77(2):417–442, 2015.

[17] G. M. Caporale and T. Zekokh. Modelling volatility of cryptocurrencies using Markov-switching GARCH models. *Research in International Business and Finance*, 48:143–155, 2019.

[18] S. X. Chen. Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6(1):87–107, 2008.

[19] A. Daouia, I. Gijbels, and G. Stupfler. Extremiles: A new perspective on asymmetric least squares. *Journal of the American Statistical Association*, 114(527):1366–1381, 2019.

[20] A. Daouia, S. Girard, and G. Stupfler. Estimation of tail risk based on extreme expectiles. *Journal of the Royal Statistical Society, Series B*, 80(2):263–292, 2018.

[21] E. Deme, S. Girard, and A. Guillou. Reduced-bias estimators of the conditional tail expectation for heavy-tailed distributions. In M. Hallin et al., editor, *Mathematical Statistics and Limit Theorems*, pages 105–123. Springer, 2015.

[22] E. Di Bernardino and C. Prieur. Estimation of the multivariate conditional tail expectation for extreme risk levels: Illustration on environmental data sets. *Environmetrics*, 29(7):e2510, 2018.

[23] K. Dowd. *Measuring market risk*. John Wiley & Sons, 2007.

[24] E. J. Eini and H. Khaloozadeh. Tail conditional moment for generalized skew-elliptical distributions. *Journal of Applied Statistics*, 48(13–15):2285–2305, 2021.

[25] J. El Methni, L. Gardes, and S. Girard. Non-parametric estimation of extreme risk measures from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41(4):988–1012, 2014.

[26] J. El Methni, L. Gardes, and S. Girard. Kernel estimation of extreme regression risk measures. *Electronic Journal of Statistics*, 12:359–398, 2018.

[27] I. Fraga Alves, L. de Haan, and T. Lin. Third order extended regular variation. *Publications de l'Institut Mathématique*, 80(94):109–120, 2006.

[28] S. Girard, G. Stupfler, and A. Usseglio-Carleve. Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *The Annals of Statistics*, 49(6):3358–3382, 2021.

[29] Y. Goegebeur and T. de Wet. Estimation of the third-order parameter in extreme value statistics. *Test*, 21(2):330–354, 2012.

[30] M. I. Gomes, L. de Haan, and L. Peng. Semi-parametric estimation of the second order parameter in statistics of extremes. *Extremes*, 5(4):387–414, 2002.

[31] M. I. Gomes and D. Pestana. A sturdy reduced-bias extreme quantile (VaR) estimator. *Journal of the American Statistical Association*, 102(477):280–292, 2007.

[32] P. Hall and A. W. Welsh. Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13:331–341, 1985.

[33] B. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

[34] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[35] P. Jorion. *Value at risk*. McGraw-Hill Professional Publishing, third edition, 2011.

[36] V. Khokhlov. Conditional value-at-risk for uncommon distributions. *Available at SSRN*, 2018.

[37] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[38] R. Kulik and P. Soulier. *Heavy-tailed time series.* Springer, 2020.

[39] Z. Landsman, U. Makov, and T. Shushi. Tail conditional moments for elliptical and log-elliptical distributions. *Insurance: Mathematics and Economics*, 71:179–188, 2016.

[40] Z. Landsman and E. Valdez. Tail conditional expectations for elliptical distributions. *North American Actuarial Journal*, 7(4):55–71, 2003.

[41] J. Longerstaey and M. Spencer. Riskmetrics – Technical Document. *Morgan Guaranty Trust Company and Reuters Ltd*, 51:54, 1996. https://www.msci.com/documents/10199/5915b101-4206-4ba0-aee2-3449d5c7e95a.

[42] B. Mandelbrot. *The variation of certain speculative prices*. Springer, 1997.

[43] T. Mao, G. Stupfler, and F. Yang. Asymptotic properties of generalized shortfall risk measures for heavy-tailed risks. *Insurance: Mathematics and Economics*, 111:173–192, 2023.

[44] C. Neves. From extended regular variation to regular variation with application in extreme value statistics. *Journal of Mathematical Analysis and Applications*, 355(1):216–230, 2009.

[45] O. Pasche and S. Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *arXiv preprint arXiv:2208.07590*, 2022.

[46] R. T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26(7):1443–1471, 2002.

[47] O. Scaillet. Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance*, 14(1):115–129, 2004.

[48] E. Tartakovsky, K. Plesovskikh, A. Sarmakeeva and A. Bibik. Autocorrelation of returns in major cryptocurrency markets. *arXiv preprint arXiv:2003.13517*, 2020.

[49] A. Teruzzi. Tail risk and systemic risk estimation of cryptocurrencies: an expectiles and marginal expected shortfall based approach. *arXiv preprint arXiv:2311.17239*, 2023.

[50] D. Troop, F. Godin, and J. Y. Yu. Bias-corrected peaks-over-threshold estimation of the CVaR. *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence*, PMLR 161:1809–1818, 2021.

[51] C. Trucíos and J. W. Taylor. A comparison of methods for forecasting value at risk and expected shortfall of cryptocurrencies. *Journal of Forecasting*, 42:989–1007, 2023.

[52] X. Q. Wang and S. H. Cheng. General regular variation of the n-th order and 2nd order Edgeworth expansions of the extreme value distribution. II. *Acta Mathematica Sinica (English Series)*, 22(1):27–40, 2006.

[53] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978.