Bias-Corrected Peaks-Over-Threshold Estimation of the CVaR

Dylan Troop\textsuperscript{1} Frédéric Godin\textsuperscript{2} Jia Yuan Yu\textsuperscript{3}

\textsuperscript{1}Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada
\textsuperscript{2}Department of Mathematics and Statistics, Concordia University, Montreal, Canada
\textsuperscript{3}Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada

Abstract

The conditional value-at-risk (CVaR) is a useful risk measure in fields such as machine learning, finance, insurance, energy, etc. When measuring very extreme risk, the commonly used CVaR estimation method of sample averaging does not work well due to limited data above the value-at-risk (VaR), the quantile corresponding to the CVaR level. To mitigate this problem, the CVaR can be estimated by extrapolating above a lower threshold than the VaR using a generalized Pareto distribution (GPD), which is often referred to as the peaks-over-threshold (POT) approach. This method often requires a very high threshold to fit well, leading to high variance in estimation, and can induce significant bias if the threshold is chosen too low. In this paper, we address this bias-variance tradeoff by deriving a new expression for the GPD approximation error of the CVaR, a bias term induced by the choice of threshold, as well as a bias correction method for the estimated GPD parameters. This leads to the derivation of a new CVaR estimator that is asymptotically unbiased and less sensitive to lower thresholds being used. An asymptotic confidence interval for the estimator is also constructed. In a practical setting, we show through experiments that our estimator provides a significant performance improvement compared with competing CVaR estimators in finite samples from heavy-tailed distributions.

1 INTRODUCTION

Traditional machine learning algorithms typically consider the expected value of a random variable as the target to optimize. In a risk-averse setting, the objective function needs to be adapted to consider the full distribution and account for severe outcomes. Recently, risk-averse machine learning has become an important area of study, especially in the context of multi-armed bandits and reinforcement learning, for example, Chow and Ghavamzadeh [2014], Tamar et al. [2015], Keramati et al. [2020], Torossian et al. [2019] and Hiraoka et al. [2019]. Most often, the risk measure of interest is the conditional value-at-risk (CVaR). Given a random variable $X$ representing losses (i.e., where larger values are less desirable), the CVaR at a confidence level $\alpha \in (0, 1)$ measures the expected value of $X$ given that $X$ exceeds the quantile at $\alpha$. This quantile is referred to as the value-at-risk (VaR). Compared to the VaR, the CVaR captures more information about the weight of a distribution’s tail, making it a more useful object of study in risk-averse decision making. In practice, the CVaR is usually estimated by averaging observations above the estimated VaR, which we call the sample average estimator of the CVaR. When $\alpha$ is close to 1, these observations can be very scarce in small samples leading to volatile estimates of the CVaR. This work is motivated by a lack of reliable estimators and performance guarantees for the CVaR at these extreme levels.

In this paper, we consider estimating the CVaR of heavy-tailed random variables, which are ubiquitous in areas such as finance, insurance, energy, and epidemiology, e.g., Manz and Mansmann [2020]. In this setting, extreme events correspond to very large observations (and hence severe losses), which is in contrast to the light- or short-tailed cases where similar low probability events are closer to the mean. Extreme value theory provides the tools to construct a new CVaR estimator that is appropriate for this setting. By selecting a threshold lower than the VaR, it is possible to approximate the tail distribution of a random variable by using a generalized Pareto distribution (GPD) and extrapolating beyond available observations. The estimation of quantities using this approximation is commonly referred to as the peaks-over-threshold (POT) approach. To the best of our knowledge, the only existing CVaR estimator based on the POT approach is given in, for example, McNeil et al. [2005 Section 7.2.3], where the CVaR is referred to as the...
expected shortfall. This estimator suffers from one of the main drawbacks of the POT approach, which is the difficult bias-variance tradeoff in selecting the threshold. Unless the threshold is chosen very high, the estimator will encounter two sources of potentially significant bias: the deviation between the GPD and the true tail distribution, and the bias associated with parameter estimation using the approximate GPD tail data. Perhaps even more significantly, the CVaR estimator of McNeil et al. [2005] comes with no performance guarantees unless one assumes exactness of the GPD approximation and of the empirical distribution function. Therefore, it has not been previously possible to determine whether our results.

Appendix in the supplementary material, along with code to reproduce our results. Additional details for experiments are provided in the appendix in the supplementary material, along with code to reproduce our results.

The remainder of this paper is organized as follows. In section 2, the VaR and CVaR are formally defined, and the sample average estimator of the CVaR is given. Needed background from extreme value theory and second-order regular variation is presented, along with the CVaR approximation based on the POT approach. Section 3 derives the GPD approximation error for the CVaR and its asymptotic behaviour. In section 4, bias-corrected maximum likelihood estimators for the GPD parameters are derived, as well as details on second-order parameter estimation, which play an important role in bias correction. Section 5 establishes an estimator for the GPD approximation error, and our results are consolidated to give the unbiased POT estimator for the CVaR. In section 6, simulations are shown to provide empirical evidence of the finite sample performance of our estimator on data, and Section 7 concludes. All proofs and additional details for experiments are provided in the appendix in the supplementary material, along with code to reproduce our results.

2 PRELIMINARIES

Let \( X \) denote a random variable and \( F \) its corresponding cumulative distribution function (cdf). In this paper, we adopt the convention that \( X \) represents a loss, so larger values of \( X \) are less desirable.

Definition 2.1 (Value-at-Risk). The value-at-risk of \( X \) at level \( \alpha \in (0, 1) \) is

\[
q_\alpha \triangleq \text{VaR}_\alpha(X) = \inf \{ x \in \mathbb{R} | F(x) \geq \alpha \}.
\]

VaR\(_\alpha\) (\( X \)) is equivalent to the quantile at level \( \alpha \) of \( F \). If the inverse of \( F \) exists, VaR\(_\alpha\) (\( X \)) = \( F^{-1}(\alpha) \). The VaR can be estimated in the same way as the standard empirical quantile. Let \( X_1, \ldots, X_n \) be i.i.d. random variables with common cdf \( F \). Let \( X_{(1,n)} \leq X_{(2,n)} \cdots \leq X_{(n,n)} \) denote the set of order statistics for the sample of size \( n \), i.e., the sample sorted in non-decreasing order. An estimator for the VaR is

\[
\hat{\text{VaR}}_{n,\alpha}(X) = \min \{ X_{(i,n)} | i = 1, \ldots, n; \hat{F}_n(X_{(i,n)}) \geq \alpha \} = X_{(m,n)},
\]

where \( \hat{F}_n \) denotes the empirical cdf and \( m = \lceil \alpha n \rceil \). We now define the CVaR as in Acerbi and Tasche [2002]. The current work will only consider continuous random variables.

Definition 2.2 (Conditional Value-at-Risk). The conditional value-at-risk of a continuous random variable \( X \) at level \( \alpha \in (0, 1) \) is

\[
c_\alpha \triangleq \text{CVaR}_\alpha(X) = \mathbb{E}[X | X \geq \text{VaR}_\alpha(X)]
\]

\[
= \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_\gamma(X) d\gamma.
\]

Typical values of \( \alpha \) are 0.95, 0.99, 0.999, etc. Besides the CVaR, expectile-based approaches to risk estimation can be considered, such as in Daouia et al. [2018], Girard et al. [2021]. Expectile-based risk estimation comes with the advantage of the elicitability property but is arguably more difficult to understand and implement than CVaR, making it less common outside of financial risk management. We leave such discussion out-of-scope in the current work.

The CVaR can be estimated by averaging observations above VaR\(_\alpha\) (\( X \)). This estimator is given by

\[
\tilde{\text{CVaR}}_{n,\alpha}(X) = \frac{\sum_{i=1}^{n} I_{\{X_i \geq \text{VaR}_{n,\alpha}(X)\}}}{\sum_{j=1}^{n} I_{\{X_j \geq \text{VaR}_{n,\alpha}(X)\}}}.
\]

The use of the eq. (3) can be problematic when the confidence level \( \alpha \) is high due to the scarcity of extreme observations. We now provide tools from extreme value theory to address this problem.

2.1 \textit{Expected shortfall}. This estimator suffers from one of the main drawbacks of the POT approach, which is the difficult bias-variance tradeoff in selecting the threshold. Unless the threshold is chosen very high, the estimator will encounter two sources of potentially significant bias: the deviation between the GPD and the true tail distribution, and the bias associated with parameter estimation using the approximate GPD tail data. Perhaps even more significantly, the CVaR estimator of McNeil et al. [2005] comes with no performance guarantees unless one assumes exactness of the GPD approximation and of the empirical distribution function. Therefore, it has not been previously possible to determine whether our results.

Appendix in the supplementary material, along with code to reproduce our results. Additional details for experiments are provided in the appendix in the supplementary material, along with code to reproduce our results.

The remainder of this paper is organized as follows. In section 2, the VaR and CVaR are formally defined, and the sample average estimator of the CVaR is given. Needed background from extreme value theory and second-order regular variation is presented, along with the CVaR approximation based on the POT approach. Section 3 derives the GPD approximation error for the CVaR and its asymptotic behaviour. In section 4, bias-corrected maximum likelihood estimators for the GPD parameters are derived, as well as details on second-order parameter estimation, which play an important role in bias correction. Section 5 establishes an estimator for the GPD approximation error, and our results are consolidated to give the unbiased POT estimator for the CVaR. In section 6, simulations are shown to provide empirical evidence of the finite sample performance of our estimator on data, and Section 7 concludes. All proofs and additional details for experiments are provided in the appendix in the supplementary material, along with code to reproduce our results.

2 PRELIMINARIES

Let \( X \) denote a random variable and \( F \) its corresponding cumulative distribution function (cdf). In this paper, we adopt the convention that \( X \) represents a loss, so larger values of \( X \) are less desirable.

Definition 2.1 (Value-at-Risk). The value-at-risk of \( X \) at level \( \alpha \in (0, 1) \) is

\[
q_\alpha \triangleq \text{VaR}_\alpha(X) = \inf \{ x \in \mathbb{R} | F(x) \geq \alpha \}.
\]

VaR\(_\alpha\) (\( X \)) is equivalent to the quantile at level \( \alpha \) of \( F \). If the inverse of \( F \) exists, VaR\(_\alpha\) (\( X \)) = \( F^{-1}(\alpha) \). The VaR can be estimated in the same way as the standard empirical quantile. Let \( X_1, \ldots, X_n \) be i.i.d. random variables with common cdf \( F \). Let \( X_{(1,n)} \leq X_{(2,n)} \cdots \leq X_{(n,n)} \) denote the set of order statistics for the sample of size \( n \), i.e., the sample sorted in non-decreasing order. An estimator for the VaR is

\[
\hat{\text{VaR}}_{n,\alpha}(X) = \min \{ X_{(i,n)} | i = 1, \ldots, n; \hat{F}_n(X_{(i,n)}) \geq \alpha \} = X_{(m,n)},
\]

where \( \hat{F}_n \) denotes the empirical cdf and \( m = \lceil \alpha n \rceil \). We now define the CVaR as in Acerbi and Tasche [2002]. The current work will only consider continuous random variables.

Definition 2.2 (Conditional Value-at-Risk). The conditional value-at-risk of a continuous random variable \( X \) at level \( \alpha \in (0, 1) \) is

\[
c_\alpha \triangleq \text{CVaR}_\alpha(X) = \mathbb{E}[X | X \geq \text{VaR}_\alpha(X)]
\]

\[
= \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_\gamma(X) d\gamma.
\]

Typical values of \( \alpha \) are 0.95, 0.99, 0.999, etc. Besides the CVaR, expectile-based approaches to risk estimation can be considered, such as in Daouia et al. [2018], Girard et al. [2021]. Expectile-based risk estimation comes with the advantage of the elicitability property but is arguably more difficult to understand and implement than CVaR, making it less common outside of financial risk management. We leave such discussion out-of-scope in the current work.

The CVaR can be estimated by averaging observations above VaR\(_\alpha\) (\( X \)). This estimator is given by

\[
\tilde{\text{CVaR}}_{n,\alpha}(X) = \frac{\sum_{i=1}^{n} I_{\{X_i \geq \text{VaR}_{n,\alpha}(X)\}}}{\sum_{j=1}^{n} I_{\{X_j \geq \text{VaR}_{n,\alpha}(X)\}}}.
\]

The use of the eq. (3) can be problematic when the confidence level \( \alpha \) is high due to the scarcity of extreme observations. We now provide tools from extreme value theory to address this problem.
Let $F^n(x) = \mathbb{P}(\max(X_1, \ldots, X_n) \leq x)$ denote the cdf of the sample maxima. Suppose there exists a sequence of real-valued constants $a_n > 0$ and $b_n$, $n = 1, 2, \ldots$, and a nondegenerate cdf $H$ such that $\lim_{n \to \infty} F^n(a_n x + b_n) = H(x)$ for all $x$, where nondegenerate refers to a distribution not concentrated at a single point. The class of distributions $F$ that satisfy this limit are said to be in the maximum domain of attraction of $H$, denoted $F \in \text{MDA}(H)$. The Fisher–Tippett–Gnedenko theorem (see De Haan and Ferreira [2006] Theorem 1.1.3) states that $H$ must then be a generalized extreme value distribution (GEVD)$^1$ i.e., if $F \in \text{MDA}(H)$, then there exists a unique $\xi \in \mathbb{R}$ such that $H = H_\xi$. It is important to note that essentially all common continuous distributions used in applications are in MDA$(H_\xi)$ for some value of $\xi$. When $\xi > 0$, $F$ is a heavy-tailed distribution, defined next.

**Definition 2.3 (Heavy-tailed random variable).** Let $X$ be a random variable with cdf $F$. Then $X$ (or $F$) is heavy-tailed if $F \in \text{MDA}(H_\xi)$ with $\xi > 0$.

If $F$ is heavy-tailed, then moments of order greater than or equal to $1/\xi$ do not exist. Otherwise, $F$ is light-tailed with a tail having exponential decay ($\xi = 0$) or the right endpoint of $F$ is finite ($\xi < 0$). If $\xi \geq 1$, then $F$ has infinite mean, and therefore the true CVaR, eq. (2), is also infinite. For the remainder of this paper, we assume the following condition is satisfied.

**Assumption 2.1.** $F$ is heavy-tailed with $\xi < 1$.

When $F \in \text{MDA}(H_\xi)$, there exists a useful approximation of the distribution of sample extremes above a threshold, and we define this distribution next. We denote the tail distribution $F = 1 - F$.

**Definition 2.4 (Excess distribution function).** For a given threshold $u \geq \text{ess inf } X$, the excess distribution function is defined as

$$F_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{F(y + u) - F(u)}{F(u)}, \quad y \geq 0.$$ 

Note that the domain of $F_u$ is $[0, \infty)$ under assumption 2.1. The $y$-values are referred to as threshold excesses. Given that $X$ has exceeded some high threshold $u$, this function represents the probability that $X$ exceeds the threshold by at most $y$. The Pickands-Balkema-de Haan theorem states that $F_u$ can be well-approximated by the GPD, which we give now.

**Theorem 2.1** (Pickands III [1975], Balkema and de Haan [1994]). Suppose assumption 2.1 is satisfied. Then, there exists a positive function $\sigma = \sigma(u)$ such that

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

where $G_{\xi, \sigma}$ is the generalized Pareto distribution, which for $\xi \neq 0$ has a cdf and density function given by, respectively,

$$G_{\xi, \sigma}(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \quad g_{\xi, \sigma}(y) = \frac{1}{\sigma} \left(1 + \frac{\xi y}{\sigma}\right)^{-1-1/\xi}.$$

Using theorem 2.1, it is quite straightforward to derive approximate formulas for the VaR and CVaR using the definition of the excess cdf and eqs. (1) and (2), for example, see McNeill et al. [2005, Section 7.2.3]. Before stating these formulas, we make precise the choice of function $\sigma(u)$ in theorem 2.1 which we give next after some needed definitions.

Let $U = (1/F)^{-1}$, the functional inverse of $1/F$. Assume such $U$ exists and is twice-differentiable. The following functions will become important tools for characterizing the tail behaviour of $F$.

**Definition 2.5.** The first- and second-order auxiliary functions are defined as, respectively,

$$a(t) = tU'(t), \quad A(t) = \frac{tU''(t)}{U'(t)} - \xi + 1.$$ 

For the remainder of this paper, let $\sigma(u) = a(1/F(u))$. It is proven in [2003, Corollary 1], with different notation, that eq. (4) achieves the optimal rate of convergence with $\sigma(u) = a(1/F(u))$ when the following condition on $A$ holds, which we assume to be true for the rest of this paper.

**Assumption 2.2.** If $F \in \text{MDA}(H_\xi)$, the second-order auxiliary function $A$ exists and satisfies the following conditions: (1) $\lim_{t \to \infty} A(t) = 0$, (2) $A$ is of constant sign in a neighborhood of $\infty$, (3) $|A| \leq 0$ such that $|A| \in \text{RV}_R$.

While assumption 2.2 may seem restrictive at first glance, it is in fact a very general condition, satisfied for all common distributions that belong to a maximum domain of attraction [Drees et al., 2004]. Counterexamples are fairly contrived and rarely seen in practice, e.g., De Haan and Ferreira [2006, Exercise 2.7 on p. 61].

Now, with a precise definition of $\sigma(u)$, we state the approximations for the VaR and CVaR which follow from theorem 2.1. For the rest of this paper, we shall denote $s_{u, \alpha} = F(u)/(1 - \alpha)$.

**Definition 2.6 (POT approximations).** Suppose that assumption 2.2 and assumption 2.2 are satisfied. Fix $u \in \mathbb{R}$ and let $\rho > 0$. A positive, measurable function $f$ is regularly varying with unique index $\rho$, denoted $f \in \text{RV}_\rho$, if $\lim_{x \to \infty} f(tx)/f(x) = t^\rho$ for all $t > 0$. 

1For $\xi > 0$, the GEVD has cdf $H_\xi(x) = \exp\left(-\left(1+\xi x\right)^{-1/\xi}\right)$ over its support, which is $[-1/\xi, \infty)$. 

\[ \sigma = a(1/\bar{F}(u)). \] Then, the POT approximations for the VaR and CVaR are given by, respectively,

\[ q_{u,\alpha} = u + \frac{\sigma^2(\xi)}{\xi} (k_{u,\alpha} - 1), \quad c_{u,\alpha} = u + \frac{\sigma}{1 - \xi} \left( 1 + k_{u,\alpha} - 1 \right). \] (7)

The accuracy of the POT approximations depends on how high of a threshold is used. When these approximations are used in statistical estimation, a lower threshold is preferable to make use of as much data as possible, but this can introduce a significant bias. To estimate this bias, explicit expressions are required for the approximation error when using eq. (7). In the next section, we derive these expressions.

3 GPD APPROXIMATION ERROR

When applying the POT approximation for the CVaR, there is a deviation between \( c_{u,\alpha} \) and \( c_\alpha \) that can be quantified asymptotically. We define this deviation as follows.

**Definition 3.1.** The GPD approximation error (of the CVaR) at level \( \alpha \) and threshold \( u \) is defined as

\[ \epsilon_{u,\alpha} = \frac{\Delta c_{u,\alpha}}{a(\tau_u)} A(\tau_u) K_{\xi,\rho}(\beta). \]

Note that when we do not consider parameter estimation, \( \epsilon_{u,\alpha} \) is a deterministic quantity. In this section, the asymptotic behaviour of \( \epsilon_{u,\alpha} \) as \( u \to \infty \) is derived, which leads to a useful approximation for finite \( u \). For the rest of this paper, we shall denote \( \tau_u = 1/\bar{F}(u) \).

**Theorem 3.1.** Suppose assumption 2.1 and assumption 2.2 hold. Let \( \alpha = \alpha_u = 1 - \bar{F}(u)/\beta \), where \( \beta > 1 \) is a constant not depending on \( u \). Then,

\[ \epsilon_{u,\alpha} a(\tau_u) A(\tau_u) K_{\xi,\rho}(\beta) \to 1 \quad \text{as} \quad u \to \infty, \]

where

\[ K_{\xi,\rho}(\beta) = \left\{ \begin{array}{ll} \frac{1}{\rho} \left( \frac{\beta}{\xi(1-\xi)} - \frac{1}{\xi + \rho} \left( \frac{\beta}{1-\xi} + \frac{1}{\xi} \right) \right), & \rho < 0, \xi + \rho \neq 0, \\
\frac{1}{\rho} \left( \frac{\beta}{\xi(1-\xi)} - \log \beta + \frac{\xi - 1}{\xi} \right), & \rho > 0, \xi + \rho = 0, \\
\frac{2}{\xi(1-\xi)} - \log \beta + \frac{1}{\xi}, & \rho = 0. \end{array} \right. \] (8)

In practice, we would typically be interested in the CVaR at a fixed value of \( \alpha \), so it may appear unsatisfactory that \( \alpha \to 1 \) in theorem 3.1. However, a useful approximation in the non-asymptotic setting which holds for large \( u \) is \( \epsilon_{u,\alpha} \approx a(\tau_u) A(\tau_u) K_{\xi,\rho}(\epsilon_{u,\alpha}) \), which is valid as long \( \alpha > F(u) \).

In subsequent sections, we derive estimators for all needed quantities to estimate \( \epsilon_{u,\alpha} \) and \( \epsilon_{u,\alpha} \) (and thus \( c_\alpha \)) from data, namely the parameters \( \xi, \sigma, \rho, \) and function \( A \), leading to an asymptotically unbiased estimator of \( c_\alpha \).

4 PARAMETER ESTIMATION

In this section, we discuss the estimation of \( \xi, \sigma, \rho, \) and \( A \). The starting point is to first select a threshold \( u \), and then estimate \( \xi \) and \( \sigma \) using maximum likelihood with the threshold excesses above \( u \). Let \( X_{(1,n)} \leq X_{(2,n)} \ldots \leq X_{(n,n)} \) denote the order statistics for a sample of size \( n \). Let \( u = X_{(n-k,n)} \) for some value of \( k = k_n < n \). Then, the threshold excesses \( Y_i = X_{(n-k+i,n)} - u \), \( i = 1, \ldots, k \) are i.i.d. [De Haan and Ferreira 2006, Section 3.4] and approximately distributed by a GPD (theorem 2.1). Maximum likelihood estimators (MLEs) are obtained by maximizing the approximate log-likelihood function with respect to \( \xi \) and \( \sigma \).

\[ (\hat{\xi}_{n,MLE}, \hat{\sigma}_{n,MLE}) = \arg \max_{\xi,\sigma} \sum_{i=1}^{k} \log g_{\xi,\sigma}(Y_i). \] (9)

Based on partial derivatives of the log-pdf with respect to parameters, the resulting maximum likelihood first-order conditions when \( \xi > 0 \) are given by

\[ \frac{1}{k} \sum_{i=1}^{k} \log \left( 1 + \frac{\xi Y_i}{\sigma} \right) = \xi, \quad \frac{1}{k} \sum_{i=1}^{k} \frac{Y_i}{\sigma + \xi Y_i} = \frac{1}{\xi + 1}. \] (10)

A closed-form solution to eq. (10) does not exist, but the MLEs can be obtained numerically through standard software packages. See, for example, [Grimshaw 1993] for an overview of the commonly implemented algorithm.

While the usual asymptotic theory of maximum likelihood does not apply in the approximate GPD model, the following theorem establishes the fact that the MLEs are asymptotically normal with a biased mean as long as the number of threshold excesses is chosen suitably. We will include a correction for the asymptotic bias in an estimator for the CVaR subsequently. The following theorem is given in [De Haan and Ferreira 2006 Theorem 3.4.2].

**Theorem 4.1.** Suppose that assumption 2.1 and assumption 2.2 hold. Then for \( k = k_n \to \infty \) and \( k/n \to 0 \) as \( n \to \infty \), if \( \lim_{n \to \infty} \sqrt{k} A(n/k) = \lambda < \infty \), then the MLEs satisfy

\[ \sqrt{k}(\hat{\xi}_{n,MLE} - \xi, \hat{\sigma}_{n,MLE}/a(n/k) - 1) \overset{d}{\to} N(\lambda b_{\xi,\rho}, \Sigma), \]

where \( N \) denotes the normal distribution and

\[ b_{\xi,\rho} = \left( b_{\xi,\rho}^{(1)}, b_{\xi,\rho}^{(2)} \right) \quad \text{where} \quad \begin{pmatrix} \xi + 1, -\rho \\ (1-\rho)(1+\xi-\rho) \end{pmatrix}, \]

\[ \Sigma = \begin{pmatrix} (1+\xi)^2 & -(1+\xi) \\ -1(1+\xi) & 1+1(1+\xi)^2 \end{pmatrix}. \] (11)

For the remaining theory sections of this paper, let \( u_{n} = X_{(n-k,n)} \). In the assumption of theorem 4.1, it does not seem possible to give conditions to guarantee \( \sqrt{k} A(n/k) \to \lambda < \infty \) in full generality, but a common
approach when working with heavy-tailed distributions is to assume that they belong to the Hall class \[ \text{Hall}[1982] \], which nests those most often seen in practice, for example, the Burr, Fréchet, Cauchy, Pareto, F, stable etc. The Hall class satisfies assumption \[ 2.2 \] with \( A(t) = ct^p \) for some constant \( c \in \mathbb{R} \), and so to ensure convergence we only require that \( k = O(n^{-2p}/(1-2p)) \).

To obtain an asymptotically unbiased estimator of the CVaR, we will first correct the asymptotic bias in theorem 4.1, which nests those most often seen in practice, for example, the estimator \( \hat{\xi} \).

To obtain an asymptotically unbiased estimator of the CVaR, we will first correct the asymptotic bias in theorem 4.1, which nests those most often seen in practice, for example, the estimator \( \hat{\xi} \).

\[ \hat{\xi} \]

\[ \hat{T}_n^{(\tau)}(m) = \frac{(M_n^{(1)}(m))^\tau - (M_n^{(1)}(m)/2)^{\tau/2}}{(M_n^{(2)}(m)/2)^{\tau/2} - (M_n^{(3)}(m)/6)^{\tau/3}}, \quad \tau \in \mathbb{R}, \]

with the notation \( a^{br} = b \log a \) if \( \tau = 0 \). Then, an estimator for \( \rho \) is given by [Fraga Alves et al. 2003] Equation 2.18,

\[ \hat{\rho}_n = \frac{3(T_n^{(\tau)}(m) - 1)}{T_n^{(\tau)}(m) - 3}. \]

The number of upper order statistics chosen to estimate \( \rho \) is usually much larger than the choice used to estimate \( (\xi, \sigma) \), i.e., \( m > k \). It is shown in [Fraga Alves et al. 2003] that \( \hat{\rho}_n \) is consistent, i.e., \( \hat{\rho}_n \xrightarrow{p} \rho \), under certain mild conditions. The estimator \( \hat{\rho}_n \) has an asymptotic bias, and the reduction of this bias is dependant on the choice of \( m \) as well as the tuning parameter \( \tau \). Fortunately, the adaptive algorithm given in [Caeiro and Gomes 2015] Section 4.1] provides an effective method of bias correction by choosing \( m \) and \( \tau \) via the most stable sample path of \( \hat{\rho}_n \). Details of the full estimation procedure are given in appendix B.1.

**Estimation of \( A(n/k) \).** As part of a secondary contribution of this paper, we derive an estimator for \( A(n/k) \) in order to estimate \( \hat{\xi} \) from i.i.d. samples. Following the formulation of [Haouas et al. 2018], we adapt their estimator for \( A_0(n/k) \) to non-truncated data. Then, using the relation between \( A_0 \) and \( A \) in [De Haan and Ferreira 2006] Table 3.1, an estimator for \( A(n/k) \) is

\[ \hat{A}_n = \left( \hat{\xi}_{n,k}^{(1)} + \hat{\rho}_n \right) \left( 1 - \hat{\rho}_n \right)^2 M_n^{(2)} - 2 \left( M_n^{(1)} \right)^2 \right) \right)^2, \]

where we define \( M_n^{(j)} = M_n^{(j)}(k) \). The proof that \( \hat{A}_n \) is consistent in the sense that \( \hat{A}_n / A(n/k) \xrightarrow{p} 1 \) is given in appendix A.5.

**Estimation of \( b_{\xi,\rho} \).** To obtain a consistent estimator for \( b_{\xi,\rho} \), it suffices to plug in any consistent estimators for \( \xi \) and \( \rho \) into eq. (11), which follows from the continuous mapping theorem (see, for example, [Vaart 1998] Theorem 2.3). Since \( \hat{\xi}_{n,k}^{(\tau)} \xrightarrow{p} \xi \) by theorem 4.1, we set

\[ \hat{b}_n = (\hat{b}_n^{(1)}, \hat{b}_n^{(2)}) \triangleq \frac{[\hat{\xi}_{n,k}^{(1)} + 1, \hat{\rho}_n]}{1 - \hat{\rho}_n}(1 + \hat{\xi}_{n,k}^{(1)} - \hat{\rho}_n) \]

as an estimator for \( b_{\xi,\rho} \), where \( \hat{b}_n \xrightarrow{d} b_{\xi,\rho} \). We now give bias-corrected estimates of the GPD parameters, which we define by

\[ \hat{\xi}_n \triangleq \hat{\xi}_{n,k}^{(1)} - \hat{A}_n \hat{b}_n^{(1)}, \quad \hat{\sigma}_n \triangleq \hat{\sigma}_{n,k}^{(1)}(1 - \hat{A}_n \hat{b}_n^{(2)}). \]

The following theorem shows that \( \hat{\xi}_n \) and \( \hat{\sigma}_n \) are asymptotically normal and centered with the same asymptotic variance \( \Sigma \) as in eq. (11).

**Theorem 4.2.** Suppose that the assumptions of theorem 4.1 hold. Then

\[ \sqrt{n} \left( \hat{\xi}_n - \xi, \hat{\sigma}_n / (n/k) - 1 \right) \xrightarrow{d} N(0, \Sigma). \]

Having now established estimators for all required distributional parameters, in the next section we introduce estimators for \( c_{u,\alpha} \) and \( c_{u,\alpha} \). Using this result, we derive an asymptotically unbiased estimator and confidence interval for \( c_{\alpha} \).

**5 UNBIASED POT ESTIMATOR**

Using theorem 4.2, a new estimator for \( c_{u,\alpha} \) can be constructed from eq. (7), which we then show is asymptotically normal and centered. The only missing requirement is an estimate for \( F(u) \), which, with \( u = X_{(n-k,n)} \), can be obtained using the empirical distribution function, i.e., \( F_n(u) = 1 - k/n \).

**Definition 5.1 (POT estimator).** Suppose that \( (\hat{\xi}_n, \hat{\sigma}_n) \) are obtained from \( k \) threshold excesses with \( \hat{\xi}_n < 1 \). Then, an estimator for \( c_{u,\alpha} \) at level \( \alpha > 1 - k/n \) is

\[ \hat{c}_\alpha(n) \triangleq \frac{\hat{\sigma}_n}{1 - \hat{\xi}_n} \left( 1 - \frac{1}{\hat{\xi}_n} \left( \frac{k}{n(1-\alpha)} \hat{\xi}_n - 1 \right) \right) + X_{(n-k,n)}. \]

Typically, when the CVaR is estimated using the POT approach in the literature, e.g., [McNeil et al. 2005], eq. (16) is used with \( (\hat{c}_{n,k}^{(1)}, \hat{c}_{n,k}^{(2)}) \) in place of our estimators \( (\hat{\xi}_n, \hat{\sigma}_n) \). Hence, the typical approach introduces two sources of bias with respect to the true CVaR: the bias from the MLEs and the bias from the misspecification of the threshold excesses by the GPD (which can be corrected using the GPD approximation error). The next theorem shows that \( \hat{c}_{\alpha}(n) \) is asymptotically unbiased with respect to \( c_{u,\alpha} \).

**Theorem 5.1.** Suppose that the assumptions of theorem 4.1 hold. Let \( \alpha = \alpha_n = 1 - (1/\beta)k/n \) where \( \beta > 1 \) is a constant not depending on \( n \). Let

\[ d_\beta(x, y) = \frac{y}{1 - x} \left( 1 + \frac{\beta^2 - 1}{x} \right). \]
If the contribution of variance by the random variable $\hat{F}_n(u)$ is ignored (it is approximately 0), then

$$\frac{\sqrt{k}}{a(n/k)} \left( c_{\alpha}^{(n)} - c_{\alpha,a} \right) \overset{d}{\to} N(0,V),$$

where $V = \nabla d_\beta(\xi,1)^\top \nabla d_\beta(\xi,1)$ and $\nabla d_\beta(\xi,1)$ denotes the gradient of $d_\beta$ evaluated at $(\xi,1)$, given at the end of appendix A.3.

**Remark 5.1.** In the proof of theorem 5.1 we show that the remainder term involving $\hat{F}_n(u)$ would only add 1 to $V$ if $\hat{F}_n(u)$ and $(\hat{\xi}_n, \hat{\sigma}_n)$ are asymptotically independent.

**Remark 5.2.** The conditions of theorem 5.1 imply that $\alpha \to 1$, however, this is not very restrictive in a practical setting since finite sample approximations will be valid for any fixed choice of $\alpha$ as long as $\alpha > 1 - \frac{k}{n}$, since $\beta$ is arbitrary.

While $\hat{c}_\alpha^{(n)}$ is asymptotically unbiased with respect to $c_{\alpha,a}$, we still need to include the GPD approximation error to correct the remaining deviation induced by the GPD model. Using theorem 5.1, we can derive an estimator for the GPD approximation error, given by

$$\hat{c}_\alpha^{(n)} \triangleq \hat{\sigma}_n \hat{A}_n \hat{K}_n,$$

where $\hat{K}_n = K_{\hat{\xi}_n} \hat{p}_n(k/(n(1-\alpha)))$, defined in eq. (8) with known values replaced by their respective estimators. We can now define the following estimator for the CVaR.

**Definition 5.2** (Unbiased POT estimator). The unbiased POT estimator is an estimator for the CVaR at level $\alpha > 1 - \frac{k}{n}$, which is defined for $\hat{\xi}_n < 1$, and is given by

$$\hat{c}_\alpha^{(n)} \equiv \hat{c}_\alpha^{(n)} - \hat{\sigma}_n.$$

Note that $\hat{c}_\alpha^{(n)}$ is asymptotically unbiased with respect to $c_{\alpha}$, a statement which is made precise in the following theorem.

**Theorem 5.2.** Suppose that the assumptions of theorem 5.1 hold. Then,

$$\frac{\sqrt{k}(\hat{c}_\alpha^{(n)} - c_{\alpha})}{\hat{\sigma}_n \sqrt{\hat{V}_n}} \overset{d}{\to} N(0,1),$$

where $\hat{V}_n$ denotes a consistent estimator of $V$, which can be obtained by plugging in $\hat{\xi}_n$ into the expression for $V$ given in theorem 5.1.

**Corollary 5.1.** Based on the above limit, an asymptotic confidence interval with level $1 - \delta$ for $c_{\alpha}$ is

$$C^{n}_{\delta} = \left( \hat{c}_\alpha^{(n)} \pm z_{\delta/2} \hat{\sigma}_n \sqrt{\hat{V}_n/k} , \hat{c}_\alpha^{(n)} \pm z_{\delta/2} \hat{\sigma}_n \sqrt{\hat{V}_n/k} \right),$$

where $z_{\delta/2}$ satisfies $P(Z > z_{\delta/2}) = \delta/2$ with $Z \sim N(0,1)$. Equation (21) has asymptotically correct coverage probability, i.e., $P(\hat{c}_\alpha^{(n)} \in C^{n}_{\delta}) \to 1 - \delta$ as $n \to \infty$.

## 6 NUMERICAL EXPERIMENTS

In this section, we investigate the finite sample performance of $c_{\alpha,a}^{(n)}$ (denoted UPOT in this section) compared with the sample average estimator (eq. (3)), and POT estimator with no bias correction, i.e., eq. (14) with $(\hat{\xi}_n, \hat{\sigma}_n)$ replaced by $(\hat{\xi}_{\text{MLE}}, \hat{\sigma}_{\text{MLE}})$. Denote these estimators as SA and BPOT, respectively. First, in the theoretical setting, we compare the exact values of the asymptotic variance of UPOT and SA at different values of $\alpha$ and sample sizes on the Fréchet distribution. This analysis provides justification for the cases where UPOT is expected to perform better than SA on data. Next, we assess the statistical accuracy of the three estimation methods at different sample sizes among several classes of heavy-tailed distributions. Finally, we assess the accuracy of the asymptotic confidence interval given in eq. (21) on finite samples by using the empirical coverage probability.

### 6.1 COMPARISON OF ASYMPOTIC VARIANCE

In this section, the magnitude of the asymptotic variance (AVAR) of UPOT and SA are compared. Since both estimators are asymptotically unbiased and assuming they are both efficient, the mean squared error of each estimator approaches the AVAR in large samples (by the Cramér-Rao lower bound). Hence, this comparison gives evidence of the distributional properties and level of $\alpha$ where UPOT results in lower error than SA. The comparison is made on the Fréchet distribution with single parameter $\gamma$, which has $\xi = 1/\gamma$, $\rho = -1$ (see appendix C.2). We compute $V/k$ (given in theorem 5.1) with $n = 10000, 20000, \ldots, 100000$ and set $k = \lceil n^{2/3} \rceil$ to satisfy the assumption of theorem 5.1. An expression for the AVAR of SA is given in, for example, [Trindade et al. (2007), and we provide the details of this calculation for the Fréchet distribution in appendix C.2.1. To the best of our knowledge, the AVAR of SA can only be derived for distributions with a bounded second moment, which corresponds to distributions with $\xi < 1/2$ (or $\gamma > 2$ in the Fréchet case). The AVAR of SA and UPOT is compared for the Fréchet distribution with $\gamma = 2.25, 2.5, 3$ and $\alpha = 0.99, 0.999$ in fig. [I]. The results indicate that UPOT is preferable for high values of $\alpha$ and low values of $\gamma$. Increasing $\alpha$ would lead to lower sample availability in SA, and thus higher variance, while UPOT is unaffected. Decreasing $\gamma$ is equivalent to increasing $\xi$ and thus increasing tail thickness. This increases the AVAR of SA since extreme observations are much further from the mean but not readily observed. Based on evidence from the Fréchet distribution, it is reasonable to extrapolate that UPOT should always perform better than SA on heavy-tailed distributions with $\xi \geq 1/2$ at high values of $\alpha$. 

1814
6.2 ERROR ANALYSIS OF CVAR ESTIMATORS

In the experiments that follow, samples are generated from the Burr, Fréchet, and half-

$t$ distributions, which provide a good characterization of heavy-tailed phenomena with

finite mean. Relevant details for each distribution class are provided in appendix C. The estimation performance of SA, BPOT, and UPOT are compared via the root-mean-

square error (RMSE) on five examples from each distribution class, shown in fig. 1. Similar plots of the absolute bias are provided in appendix D. We fix $\alpha = 0.998$ as an example of an extreme risk level. Experiments are conducted as follows. Generate $N = 1000$ random samples of size $50000$ from each distribution. For each sample, the CVaR is estimated using the three methods at sub-sample sizes $n = 5000, 10000, \ldots, 50000$. In practice, it can be difficult to choose the number of threshold exceedances $k$, and so we apply the ordered goodness-of-fits tests of Bader et al. [2018] to choose the optimal threshold. This threshold selection procedure, which we employ in both BPOT and UPOT, is given in detail in appendix B.2. The average threshold selected (in terms of the percentile of a given sample) was between 0.80 and 0.96 in all simulations performed. The complete algorithm for UPOT is summarized in appendix B.3. The chosen Burr distributions allow us to investigate the effect of varying $\rho$ while keeping a fixed $\xi$. In this case, we set $\xi = 2/3$ while $\rho = -0.25, -0.33, -0.44, -1.33, -2.22$ in the respective Burr distributions. In general, when $\rho$ approaches 0, the distribution’s tail deviates more severely from a strict Pareto model, and therefore we see the largest bias and RMSE occur in BPOT in the Burr(0.38, 4) and Burr(0.5, 3) models, while the bias-correction of UPOT leads to the most substantial performance gain. As a non-parametric estimator, SA is less affected by changes in the value of $\rho$, outperforming the POT estimators in terms of bias on some Burr distributions. However, as alluded to in section 6.1, high values of $\xi$ leads to high variance in observations, typically causing poor performance in SA in terms of RMSE. This effect is similarly observed in the Fréchet simulations, where SA has relatively low bias. The Fréchet distribution always has $\rho = -1$, a property shared with the GPD, giving its tail a similar shape. Therefore, the bias-correction of UPOT is less significant, but still provides a noticeable performance gain over BPOT. The results of the half-t simulations are similar to the Fréchet, but we note a larger bias in BPOT due to the fact that the half-t distribution has a $\rho$ value that varies with its parameter. Like in the Burr simulations, SA is unaffected by different values of $\rho$ and obtains good performance in terms of bias in the half-t simulations, except when $\xi$ is largest in the half-t(1.5) model. Finally, we note that UPOT consistently had the lowest RMSE in all simulations except in a few cases at a sample size of 5000. Next, the finite sample performance of the UPOT confidence interval is investigated.

6.3 COVERAGE PROBABILITY OF THE

ASYMPTOTIC CONFIDENCE INTERVAL

The accuracy of the confidence interval given in eq. (21) is assessed by its empirical coverage probability for each distribution using the same simulated data from section 6.2. Let $C_{i, \delta}$ denote the confidence interval computed for a sample of size $n$ for sample $i$, $i = 1, \ldots, N$. Then, the empirical coverage probability is defined as

$$\hat{P}_n^\alpha(N) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{c_i \in C_{i, \delta}\}}.$$

Plots of the coverage probability at each sample size for each distribution are shown in fig. 2. We set $\delta = 0.05$ and compute the coverage probability at sample sizes $n = 5000, 10000, \ldots, 50000$. The final value of each distribution’s coverage probability at $n = 50000$ is reported in appendix D. Most of the distributions tested achieve nearly the correct coverage of 0.95, sometimes surpassing it in some cases, and this is due to the estimated confidence interval being wider than its true asymptotic counterpart. The coverage is worst in the Burr(0.38, 4) distribution, achieving a final coverage probability of just 0.73. The small magnitude of $\rho$ in this distribution causes slow convergence of the tail to the GPD, and hence a relatively high average threshold percentile of 0.96 was chosen by the threshold selection procedure. This high threshold increases the variance of parameter estimation which explains the poor coverage.
We have studied the asymptotic properties of a new CVaR estimator based on the peaks-over-threshold approach. Using extreme value theory and second-order regular variation, we derived estimators for the bias induced by the approximate GPD model of the threshold excesses and the bias from maximum likelihood estimators of the GPD parameters. Using these results, we proved that our estimator is asymptotically normal and unbiased (up to some technical conditions). This convergence result allowed us to derive confidence intervals for the CVaR, enabling us to measure the level of uncertainty in our estimator. We compared the magnitudes of the asymptotic variance of our CVaR estimator with that of the sample average CVaR estimator, demonstrating a significant improvement in asymptotic performance for some cases. An empirical study showed that our CVaR estimator can lead to a significant performance improvement in heavy-tailed distributions when compared to the sample average estimator and the existing peaks-over-threshold estimator. Finally, we investigated the finite-sample performance of the asymptotic confidence interval, and found that good coverage probability is achieved in reasonable sample sizes. While our evidence suggests that our CVaR estimator is most effective in the heavy-tailed domain, it would also be instructive to perform the same theoretical analysis for

**Figure 2:** RMSE of estimating CVaR$_{0.998}$ using UPOT (black), BPOT (red), and SA (blue).

**Figure 3:** Coverage probabilities with $\alpha = 0.998, \delta = 0.05$. The solid line indicates the theoretical coverage, i.e., $1 - \delta = 0.95$.

### 7 CONCLUSION

We have studied the asymptotic properties of a new CVaR estimator based on the peaks-over-threshold approach. Using extreme value theory and second-order regular variation, we derived estimators for the bias induced by the approximate GPD model of the threshold excesses and the bias from maximum likelihood estimators of the GPD parameters. Using these results, we proved that our estimator is asymptotically normal and unbiased (up to some technical conditions). This convergence result allowed us to derive confidence intervals for the CVaR, enabling us to measure the level of uncertainty in our estimator. We compared the magnitudes of the asymptotic variance of our CVaR estimator with that of the sample average CVaR estimator, demonstrating a significant improvement in asymptotic performance for some cases. An empirical study showed that our CVaR estimator can lead to a significant performance improvement in heavy-tailed distributions when compared to the sample average estimator and the existing peaks-over-threshold estimator. Finally, we investigated the finite-sample performance of the asymptotic confidence interval, and found that good coverage probability is achieved in reasonable sample sizes. While our evidence suggests that our CVaR estimator is most effective in the heavy-tailed domain, it would also be instructive to perform the same theoretical analysis for
light-tailed distributions. Doing so would allow our CVaR estimator to be robust to situations where it is not possible to make any assumptions about the underlying data distribution.

Acknowledgements

Financial support from Natural Sciences and Engineering Research Council of Canada (NSERC, Godin: RGPIN-2017-06837, Yu: RGPIN-2018-05096) is gratefully acknowledged. We would like to thank Mélina Mailhot for her valuable feedback.

References


Stéphane Girard, Gilles Stupfler, and Antoine Usseglio-Carleve. On automatic bias reduction for extreme expectile estimation. working paper or preprint, January 2021. URL https://hal.archives-ouvertes.fr/hal-03086048.


Takuya Hiraoka, Takahisa Imagawa, Tatsuya Moris, Takashi Onishi, and Yoshimasa Tsuruoka. Learning


