A Least Squared approach on estimating the Conditional Tail Expectation for Heavy Tailed distributions

El hadji Deme

LERSTAD, Université Gaston Berger de Saint-Louis, Sénégal

1 Introduction

Risk measures are thought of as mappings from a set of real-valued random variables to the real numbers. There exist several types of risk measures in the literature. We refer to Govaerts *et al.* (1984) for various examples and properties of such Risk measures. One of the most popular measures in hydrology and climate is undoubtedly the return period. A frequency analysis in hydrology focuses on the estimation of quantities (e.g., flows or annual rainfall) corresponding to a certain return period. It is closely related to the notion of quantile which has therefore been extensively studied. For a real value random variable X with $\mathbb{E}[X] < \infty$, that represents the magnitude of an event that occurs at a given time and at a given site, the quantile of order $1 - \frac{1}{T}$ expresses the magnitude of the event which is exceeded with a probability equal to $\frac{1}{T}$. T is then called the return period. In the acturial financial litterature, or more generally in the risk theory the quantile is known as the Value-at Risk (VaR) and it is defined by

$$Q(\alpha) = \inf\{x \in \mathbb{R}_+ : \ F(x) \geq \alpha\}, \quad \text{for } \alpha \in (0,1),$$

with F the distribution function of event X. A second important risk measure, based on the quantile notion, is the Conditional-Tail-Expectation (CTE) defined by

$$CTE_\alpha[X] = \mathbb{E}(X|X > Q(\alpha)), \quad \text{for } \alpha \in (0,1).$$

Since the distribution function F is continuous, we easily check that $CTE_\alpha[X]$ is equal to

$$\mathbb{C}_\alpha[X] = \frac{1}{1-\alpha}\int_\alpha^1 Q_n(s)ds.$$

Hence, from now on we work with $\mathbb{C}_\alpha[X]$ and call it the CTE for short. Naturally, the CTE is unknown since the cdf F is unknown. Hence, it is desirable to establish appropriate statistical inferential results such as confidence intervals for $\mathbb{C}_\alpha[X]$ with specified confidence levels and margins of error.

Namely, suppose that we have $(X_1, ..., X_n)$ a sample of independent and identically distributed random variables from F and let $X_{1,n} \leq ... \leq X_{n,n}$ denote its order statistics.

A natural estimator for $\mathbb{C}_\alpha[X]$ can be obtained by

$$\widehat{\mathbb{C}}_{n,\alpha}[X] = \frac{1}{1-\alpha} \int_\alpha^1 Q_n(s)ds. \tag{1}$$

where $Q_n(s)$ is the empirical quantile function, which is equal to the ith order statistic $X_{i,n}$ for all $s \in ((i-1)/n, i/n]$, and for all $i = 1, ..., n$. The asymptotic behavior of the estimator $\widehat{\mathbb{C}}_{n,\alpha}[X]$ has been studied by Brazauskas *et al.* (2008), when $\mathbb{E}[X^2] < \infty$.

This paper deals with the estimation problem of the CTE within the class of heavy-tailed distribution, i.e. we assume that

$$\overline{F}(x) = x^{-1/\gamma} \ell_F(x) \tag{2}$$

where $\gamma > 0$ is the extreme value index and ℓ_F is a slowly varying function at infinity satisfying $\ell_F(\lambda x)/\ell_F(x) \to 1$ as $x \to \infty$ for all $\lambda > 0$. Moreover we focus our paper on the case $\gamma \in \left(\frac{1}{2}, 1\right)$ in order to ensure that the $\mathbb{C}_\alpha[X]$ is finite for every $\alpha \in (0,1)$ and since in that case the results of Brazauskas *et al.* (2008) cannot be applied, the second moment of X being infinite.

The estimation of γ has been extensively studied in the literature and the most famous estimator is the Hill (1975) estimator defined as:

$$\widehat{\gamma}^H_{n,k} = \frac{1}{k} \sum_{j=1}^{k} j \left(\log X_{n-j+1,n} - \log X_{n-j,n}\right) \tag{3}$$

for an intermediate sequence $k = k(n)$, i.e. a sequence such that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Note that the $\mathbb{C}_\alpha[X]$ can be rewriten by tranformation into

$$\begin{aligned} \mathbb{C}_\alpha[X] &= \frac{1}{1-\alpha} \int_\alpha^{1-k/n} Q(s)ds + \frac{1}{1-\alpha} \int_0^{k/n} Q(1-s)ds. \\ &=: \mathbb{C}_\alpha^{(1)}[X] + \mathbb{C}_\alpha^{(2)}[X]. \end{aligned}$$

By taking into account different asymptotic properties of moderate an high quantiles in the case of heavy-tailed distributions, we obtain the following alternative estimator of the CTE

$$\begin{aligned} \widetilde{\mathbb{C}}_{n,\alpha}[X] &= \frac{1}{1-\alpha} \int_\alpha^{1-k/n} Q_n(s)ds + \frac{k/n}{(1-\alpha)(1-\widehat{\gamma}^H_{n,k})} X_{n-k,n}. \\ &=: \widetilde{\mathbb{C}}^{(1)}_{n,\alpha}[X] + \widetilde{\mathbb{C}}^{(2)}_{n,\alpha}[X]. \end{aligned} \tag{4}$$

We estimate $\widetilde{\mathbb{C}}^{(1)}_{n,\alpha}[X]$ by using the same trick as for (1), whereas for $\widetilde{\mathbb{C}}^{(2)}_{n,\alpha}[X]$ we use a Weissman estimator for Q: $\widehat{Q}(1-s) := X_{n-k,n} \left(\frac{k}{n}\right)^{\widehat{\gamma}^H_{n,k}} s^{-\widehat{\gamma}^H_{n,k}}$, $s \to 0$ (see Weissman, 1978).

It is easy to check that $\widetilde{\mathbb{C}}^{(1)}_{n,\alpha}[X]$ can be rewritten as

$$\widetilde{\mathbb{C}}^{(1)}_{n,\alpha}[X] = \frac{1}{1-\alpha}\sum_{j=1}^{n-k}\left(\left(\frac{j}{n}-\alpha\right)_+ - \left(\frac{j-1}{n}-\alpha\right)_+\right)X_{j,n},$$

where $(s-\alpha)_+$ is the classical notation for the positive part of $(s-\alpha)$. In this paper we deal with the problem of bias of the estimator $\widetilde{\mathbb{C}}_{n,\alpha}[X]$. Asymptotic normality for $\widetilde{\mathbb{C}}_{n,\alpha}[X]$ is obviously related to the one of $\widehat{\gamma}^H_{n,k}$. As usual in the extreme value framework, to prove such type of results, we need a second-order condition on the function $\mathbb{U}(x) = Q(1-1/x)$ such as the following:

Condition $(\mathcal{R}_{\mathbb{U}})$. *There exist a function $A(x) \to 0$ as $x \to \infty$ of constant sign for large values of x and a second order parameter $\rho \leq 0$ such that, for every $x > 0$,*

$$\lim_{t\to\infty}\frac{\log\mathbb{U}(tx) - \log\mathbb{U}(t) - \gamma\log x}{A(t)} = \frac{x^\rho - 1}{\rho}, \tag{5}$$

when $\rho = 0$, then the ratio on the right-hand side of equation (5) should be interpreted as $\log x$.

Note that condition $(\mathcal{R}_{\mathbb{U}})$ implies that $|A|$ is regularly varying with index ρ (see, e.g. Geluk and de Haan, 1987). It is satisfied for most of the classical distribution functions such as the Pareto, Burr and Fréchet ones.

2 Main results

We start to give in Theorem 1, an approximation of $\widetilde{\mathbb{C}}_{n,\alpha}[X]$ in terms of Brownian bridges, which leads to its asymptotic normality stated in Corollary 1. As it exhibits some bias, we propose a reduced-bias estimator.

2.1 Asymptotic results for the CTE estimator

Theorem 1. *Assume that F satisfies $(\mathcal{R}_{\mathbb{U}})$ with $\gamma \in (1/2, 1)$. They for any sequence of integer $k = k(n)$ satisfies $k \to \infty$, $k/n \to 0$ and $\sqrt{k}A(n/k) = O(1)$ as $n \to \infty$, we have*

$$\frac{\sqrt{n}(1-\alpha)}{(k/n)^{1/2}\mathbb{U}(n/k)}\left(\widetilde{\mathbb{C}}_{n,\alpha}[X] - \mathbb{C}_\alpha[X]\right) \overset{\mathcal{D}}{=} \sqrt{k}A\left(\frac{n}{k}\right)\mathcal{AB}(\gamma,\rho) + \mathbb{W}_{n,1} + \mathbb{W}_{n,2} + \mathbb{W}_{n,3} + o_{\mathbb{P}}(1)$$

where

$$\mathcal{AB}(\gamma,\rho) := \frac{\gamma\rho}{(1-\rho)(\gamma+\rho-1)(1-\gamma)^2}$$

and

$$\begin{cases} \mathbb{W}_{n,1} := -\dfrac{\int_0^{1-k/n}\mathbb{B}_n(s)dQ(s)}{(k/n)^{1/2}Q(1-k/n)} \\[2ex] \mathbb{W}_{n,2} := -\dfrac{\gamma}{(1-\gamma)}\sqrt{\dfrac{n}{k}}\,\mathbb{B}_n(1-k/n) \\[2ex] \mathbb{W}_{n,3} := \dfrac{\gamma}{(1-\gamma)^2}\sqrt{\dfrac{n}{k}}\displaystyle\int_0^1 s^{-1}\mathbb{B}_n(1-sk/n)d(s\underline{K}(s)). \end{cases}$$

with $\underline{K}(s) = \mathbb{1}_{0<s<1)}$.

Corollary 1. *Under the assumptions of Theorem 1, if* $\sqrt{k}A(n/k) \to \lambda \in \mathbb{R}$*, we have*

$$\frac{\sqrt{n}(1-\alpha)}{(k/n)^{1/2}\mathbb{U}(n/k)}\left(\widetilde{\mathbb{C}}_{n,\alpha}[X] - \mathbb{C}_\alpha[X]\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\lambda \mathcal{AB}(\gamma,\rho), \mathcal{AV}(\gamma)\right).$$

where $\mathcal{AB}(\gamma,\rho)$ *is as above and*

$$\mathcal{AV}(\gamma) = \frac{\gamma^4}{(2\gamma-1)(1-\gamma)^4}.$$

The goal of the next section is to propose a reduced-bias estimator of $\mathbb{C}_\alpha[X]$.

2.2 Estimating the CTE with the Least Squared approach

In this paper, we use the bias-reduced estimator of the high quantile $Q(1-s)$ proposed by Feureverger and Hall, (1999), Beirlant et al. (2002).
Using $(\mathcal{R}_{\mathbb{U}})$, Feuerverger and Hall (1999) and Beirlant *et al* (1999, 2002) proposed the following exponential regression model for the log-spacings of order statistics:

$$Z_{j,k} \sim \left(\gamma + A(n/k)\left(\frac{j}{k+1}\right)^{-\rho}\right) + \varepsilon_{j,k},\ 1 \le j \le k, \tag{6}$$

where $\varepsilon_{j,k}$ are zero-centered error terms. If we ignore the term $A(n/k)$ in (6), we retrieve the Hill-type estimator $\widehat{\gamma}^{H}_{n,k}$ by taking the mean of the left-hand side of (6). By using a least-squares approach, (6) can be further exploited to propose a reduced-bias estimator for γ in which ρ is substituted by a consistent estimator $\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\rho}}_{n,k}$ (see for instance Beirlant *et al*, 2002) or by a canonical choice, such as $\boldsymbol{\rho} = -1$ (see e.g. Feuerverger and Hall (1999) or Beirlant *et al* (1999)). The least squares estimators for γ and $A(n/k)$ are then given by

$$\begin{cases} \widehat{\gamma}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}}) = \dfrac{1}{k}\displaystyle\sum_{j=1}^{k} Z_{j,k} - \dfrac{\widehat{A}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}})}{1-\widehat{\boldsymbol{\rho}}}, \\ \\ \hat{A}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}}) = \dfrac{(1-2\widehat{\boldsymbol{\rho}})(1-\widehat{\boldsymbol{\rho}})^2}{\widehat{\boldsymbol{\rho}}^2}\dfrac{1}{k}\displaystyle\sum_{j=1}^{k}\left(\left(\dfrac{j}{k+1}\right)^{-\widehat{\boldsymbol{\rho}}} - \dfrac{1}{1-\widehat{\boldsymbol{\rho}}}\right) Z_{j,k}. \end{cases} \tag{7}$$

The asymptotic normality of $\hat{\gamma}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}})$ and $\hat{A}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}})$ is stablised in Beirlant et al. (2002, Theorem 3.2 . Note that $\widehat{\gamma}^{\boldsymbol{LS}}_{n,k}(\rho)$ can be viewed as a kernel estimator

$$\widehat{\gamma}^{\boldsymbol{LS}}_{n,k}(\widehat{\boldsymbol{\rho}}) = \frac{1}{k}\sum_{j=1}^{k} K_{\widehat{\boldsymbol{\rho}}}\left(\frac{j}{k+1}\right) Z_{j,k},$$

where for $0 < u \le 1$:

$$K_\rho(u) = \frac{1-\rho}{\rho}\underline{K}(u) + \left(1 - \frac{1-\rho}{\rho}\right)\underline{K}_\rho(u)$$

with $\underline{K}(u) = \mathbb{1}_{\{0<u<1\}}$ and $\underline{K}_\rho(u) = ((1-\rho)/\rho)(u^{-\rho}-1)\mathbb{1}_{\{0<u<1\}}$.

Now, we are going to propose an adaptive unbiased estimation procedure for $\mathbb{C}_\alpha[X]$ that is based on the above estimators. Considering the following unbiased Weissman's estimator of the extreme quantile base on the second order rafinements,

$$\widehat{Q}^{\boldsymbol{LS},\widehat{\boldsymbol{\rho}}}(1-s) = (ns/k)^{-\hat{\gamma}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}})} X_{n-k,n} \left(1 - \widehat{\boldsymbol{\rho}}^{-1} \hat{A}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}}) \left(1 - (ns/k)^{-\widehat{\boldsymbol{\rho}}}\right)\right), \tag{8}$$

where $\widehat{\boldsymbol{\rho}}$, $\hat{\gamma}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}})$ and $\hat{A}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}})$ denote the corresponding estimators of ρ, γ and $A(n/k)$ outlined above based on the exponnential regression model. By using the same argument in (4), we arrive at the the following unbiased estimator of $\mathbb{C}_\alpha[X]$

$$\begin{aligned}\widetilde{\mathbb{C}}_{n,\alpha}^{\boldsymbol{LS},\widehat{\boldsymbol{\rho}}}[X] &= \frac{1}{1-\alpha}\sum_{j=1}^{n-k}\left(\left(\frac{j}{n}-\alpha\right)_+ - \left(\frac{j-1}{n}-\alpha\right)_+\right) X_{j,n} \\ &\quad + \frac{k/n}{(1-\alpha)(1-\hat{\gamma}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}}))}\left(1 - \frac{\hat{A}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}})}{\hat{\gamma}_{n,k}^{\boldsymbol{LS}}(\widehat{\boldsymbol{\rho}}) + \widehat{\boldsymbol{\rho}} - 1}\right) X_{n-k,n}.\end{aligned} \tag{9}$$

Our next goal is to establish, under suitable asumptions, the asymptotic normality of $\widetilde{\mathbb{C}}_{n,\alpha}^{\boldsymbol{LS},\widehat{\boldsymbol{\rho}}}[X]$ and we provide simulations which aim at studying the practical behavior of the new estimator $\widetilde{\mathbb{C}}_{n,\alpha}^{\boldsymbol{LS},\widehat{\boldsymbol{\rho}}}[X]$, as far as to compare its performances to the biased estimator $\widetilde{\mathbb{C}}_{n,\alpha}[X]$. A real case in environmental framework is also analyzed.

References

[1] Beirlant, J., Dierckx, G., Goegebeur, M., Matthys, G. (1999). Tail index estimation and an exponential regression model, *Extremes*, **2**, 177-200.

[2] Beirlant, J., Dierckx, G., Guillou, A., Starica, C. (2002). On exponential representations of log-spacings of extreme order statistics, *Extremes*, **5**, 157-180.

[3] Feuerverger, A., Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *Annals of Statistics*, **27**, 760-781.

[4] Geluk, J.L., de Haan, L. (1987). *Regular variation, extensions and Tauberian theorems*, CWI tract 40, Center for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

[5] Goovaerts, M.J., de Vlyder, F., Haezendonck, J. (1984). *Insurance premiums, theory and applications*, North Holland, Amsterdam.

[6] Hill, B. M. (1975). A simple approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1136–1174.

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