REVSTAT – Statistical Journal
Volume 17, Number 1, January 2019, 91–108

USING SHRINKAGE ESTIMATORS TO REDUCE BIAS AND MSE IN ESTIMATION OF HEAVY TAILS

Authors: Jan Beirlant
– Department of Mathematics, KU Leuven, Belgium, and Department of Mathematical Statistics and Actuarial Science, University of the Free State South Africa
jan.beirlant@kuleuven.be

Gaonyalelwe Maribe
– Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa
maribeg@ufs.ac.za

Andréhette Verster
– Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa
VersterA@ufs.ac.za

Received: June 2016 Revised: December 2016 Accepted: February 2017

Abstract:

- Bias reduction in tail estimation has received considerable interest in extreme value analysis. Estimation methods that minimize the bias while keeping the mean squared error (MSE) under control, are especially useful when applying classical methods such as the Hill (1975) estimator. In the case of heavy tailed distributions, Caeiro *et al.* (2005) proposed minimum variance reduced bias estimators of the extreme value index, where the bias is reduced without increasing the variance with respect to the Hill estimator. This method is based on adequate external estimation of a pair of parameters of second order slow variation under a third order condition. Here we revisit this problem exploiting the mathematical fact that the bias tends to 0 with increasing threshold. This leads to shrinkage estimation for the extreme value index, which allows for a penalized likelihood and a Bayesian implementation. This new approach is applied starting from the approximation to excesses over a high threshold using the extended Pareto distribution, as developed in Beirlant *et al.* (2009). We present asymptotic results for the resulting shrinkage penalized likelihood estimator of the extreme value index. Finite sample simulation results are proposed both for the penalized likelihood and Bayesian implementation. We then compare with the minimum variance reduced bias estimators.

Key-Words:

- *extreme value index; tail estimation; extended Pareto distribution; shrinkage estimators.*

AMS Subject Classification:

- 62G32, 62F10, 62F15, 62J07.

1. INTRODUCTION

In this paper we consider the estimation of the extreme value index ξ and tail probabilities $P(X > x)$ for x large, on the basis of independent and identically distributed observations $X_1, X_2, ..., X_n$ which follow a Pareto-type distribution with right tail function (RTF) given by

$$\bar{F}(x) = 1 - F(x) = P(X > x) = x^{-1/\xi}\ell(x) \tag{1.1}$$

where ℓ is a slowly varying function at infinity, i.e.

$$\frac{\ell(ty)}{\ell(t)} \to 1, \quad \text{as } t \to \infty, \quad \text{for every } y > 1.$$

The most famous estimator of ξ was first derived by Hill (1975) as a maximum likelihood (ML) estimator approximating the RTF of the excesses $\frac{X}{t}|X > t$ over a large threshold t by a simple Pareto distribution with RTF $y^{-1/\xi}$:

$$\bar{F}(ty)/\bar{F}(t) \approx y^{-1/\xi}, \quad t \text{ large.} \tag{1.2}$$

When setting $t = X_{n-k,n}$ where $X_{1,n} \leq X_{2,n} \leq \cdots \leq X_{n,n}$ the ML estimator is given by

$$H_{k,n} = \frac{1}{k}\sum_{j=1}^{k} \log \frac{X_{n-j+1,n}}{X_{n-k,n}}. \tag{1.3}$$

A simple estimator of a tail probability $P(X > x)$ with x large, introduced in Weissman (1978), is then obtained from (1.2) setting $ty = x$ and estimating $P(X > t)$ by the empirical proportion k/n:

$$\hat{p}_{x,k} = \frac{k}{n}\left(\frac{x}{X_{n-k,n}}\right)^{-1/H_{k,n}}. \tag{1.4}$$

In practice, a way to verify the validity of model (1.1) is to check whether the Hill estimates are stable as a function of k. However in most cases the stability is not visible, which can be explained by slow convergence in (1.2). For this reason bias reduced estimators have been proposed which lead to plots that are much more horizontal in k which facilitates the analysis of a practical case to a great extent. Here we can refer to Peng (1998), Beirlant *et al.* (1999, 2008), Feuerverger and Hall (1999), Caeiro *et al.* (2005, 2009) and Gomes *et al.* (2000, 2007) for bias-reduced estimators based on functions of the top k order statistics. Several of these methods focus on the distribution of log-spacings of high order statistics.

Beirlant *et al.* (2009) proposed a more flexible model capable of capturing the deviation between the true excess RTF $\bar{F}(ty)/\bar{F}(t)$ and the asymptotic Pareto model. For a heavy tailed distribution (1.1), this deviation can be parametrized using a power series expansion (Hall, 1982), or more generally via second-order slow variation (Bingham *et al.*, 1987). More specifically in Beirlant *et al.* (2009) the subclass $\mathcal{F}(\xi,\tau)$ of the Pareto-type tails (1.1) was considered satisfying

$$\bar{F}(x) = Cx^{-1/\xi}\left(1+\xi^{-1}\delta(x)\right), \tag{1.5}$$

with $\delta(x)$ eventually nonzero and of constant sign such that $|\delta(x)| = x^{\tau}\ell_{\delta}(x)$ with $\tau < 0$ and ℓ_{δ} slowly varying. It was shown that under $\mathcal{F}(\xi,\tau)$ as $t\to\infty$

$$\sup_{y\geq 1}\left|\frac{\bar{F}(ty)}{\bar{F}(t)} - \bar{G}_{\xi,\delta,\tau}(y)\right| = o\left(|\delta(t)|\right)$$

with $\bar{G}_{\xi,\delta,\tau}$ the RTF of the extended Pareto distribution (EPD)

$$\bar{G}_{\xi,\delta,\tau}(y) = \{y(1+\delta-\delta y^{\tau})\}^{-1/\xi}, \quad y>1, \tag{1.6}$$

with $\tau < 0 < \xi$ and $\delta > \max(-1, 1/\tau)$. This shows that the EPD improves the approximation (1.2) with an order of magnitude. Then ML estimation of the parameters (ξ,δ) based on a set of excesses $(Y_{j,k} := X_{n-j+1,n}/X_{n-k,n},\ j=1,...,k)$ was used to obtain a bias reduced estimator $\hat{\xi}^{ML}_{k,n}$ of ξ. Bias reduction of the Weissman estimator of tail probabilities can analogously be obtained using

$$\hat{p}^{EP}_{x,k} = \frac{k}{n}\bar{G}_{\hat{\xi}_k,\hat{\delta}_k,\hat{\tau}}\left(\frac{x}{X_{n-k,n}}\right), \tag{1.7}$$

where $(\hat{\xi}_k,\hat{\delta}_k)$ denote the ML estimators based on the EPD model, and where $\hat{\tau}$ is a consistent estimator of τ, to be specified below, which was shown not to affect the asymptotic distribution of (ξ,δ).

If F satisfies $\mathcal{F}(\xi,\tau)$, it is shown in Beirlant *et al.* (2009) that $U(x) := Q(1-x^{-1})$ $(x>1)$, with $Q(p) = \inf\{x : F(x)\geq p\}$ $(p\in(0,1))$, satisfies

$$U(x) = C^{\xi}x^{\xi}\left(1+a(x)\right) \tag{1.8}$$

with $a(x) = \delta(Q(1-x^{-1}))\{1+o(1)\} = \delta(C^{\xi}x^{\xi})\{1+o(1)\}$ as $x\to\infty$. In particular a is eventually nonzero and of constant sign and $|a(x)| = x^{\rho}\ell_a(x)$ with ℓ_a slowly varying and $\rho = \xi\tau$. Here we assume $|\ell_a(x)| = C_a(1+o(1))$ as $x\to\infty$ for some constant $C_a>0$.

The following asymptotic results have been derived for $H_{k,n}$ and $\hat{\xi}^{ML}_{k,n}$ assuming that F satisfies $\mathcal{F}(\xi,\tau)$, and $\sqrt{k}a(n/k)\to\lambda\in\mathbb{R}$ and $\hat{\rho}_{k,n} = \rho + o_p(1)$ as $k,n\to\infty$ and $k/n\to 0$:

$$\sqrt{k}\left(H_{k,n}-\xi\right) \to_d \mathcal{N}\left(\lambda\frac{\rho}{1-\rho}, \xi^2\right), \tag{1.9}$$

$$\sqrt{k}\left(\hat{\xi}^{ML}_{k,n}-\xi\right) \to_d \mathcal{N}\left(0, \xi^2\left(\frac{1-\rho}{\rho}\right)^2\right). \tag{1.10}$$

An estimator $\hat{\rho}_{k,n}$ of ρ can be taken from Fraga Alves *et al.* (2003) using $k = k_1 = \lfloor n^{1-\epsilon} \rfloor$ for some $\epsilon > 0$. The required consistency for $\hat{\rho}_{k,n}$ was obtained under (1.8).

Asymptotic results of the type (1.9) and (1.10) are typical for bias reduced estimators when both ξ and $a(n/k)$ or δ are jointly estimated at every k value: for larger values of k corresponding to $\sqrt{k}a(n/k) \to \lambda \neq 0$, bias reduced estimators still have asymptotic bias 0 in contrast to the Hill estimator, but their variance is increased by a factor $((1-\rho)/\rho)^2$ compared to $H_{k,n}$. In a pioneering paper, Caeiro *et al.* (2005) proposed to estimate $(n/k)^{-\rho}a(n/k)$ at a high level $k = k_1 = \lfloor n^{1-\epsilon} \rfloor$, leading to a corrected Hill estimator (denoted below by $CH_{k,n}$) with asymptotic variance ξ^2 and excellent bias and MSE characteristics. To obtain the normal asymptotic behaviour of such minimum variance reduced bias estimators one needs a third-order slow variation condition which is more restrictive than (1.8) or condition $\mathcal{F}(\xi, \tau)$.

Up to now, to the best of our knowledge, the fact that $\delta(t) \to 0$ as $t \to \infty$, or $a(n/k) \to 0$ as $n/k \to \infty$ has not been exploited in the literature. However, this calls for shrinkage estimators. Such shrinkage approach can be implemented by putting a penalty on δ in an ML procedure, leading to penalized ML. Alternatively a penalty on δ can be naturally introduced in a Bayesian approach putting an appropriate prior on this parameter. Here we investigate the use of shrinkage estimation when modelling the distribution of the vector of excesses $\mathbf{Y}_k := (Y_{j,k}, j = 1, ..., k)$ with an EPD. In section 2 we show that a quadratic penalty, or equivalently a normal prior, on δ with zero mean and variance $\sigma^2_{k,n}$, depending in an appropriate way on k and n, leads to interesting asymptotic MSE results for ξ. In section 3 we consider the finite sample behaviour of the penalized likelihood and Bayes approach, and make a comparison with the minimum variance reduced bias estimator, and consider a practical case.

2. SHRINKAGE ESTIMATORS OF THE EPD PARAMETERS

2.1. Penalized likelihood and Bayesian interpretation

ML estimation of the EPD parameters (ξ, δ), given a value of τ, follows by maximizing the log-likelihood

$$\frac{1}{k} l_{EP}(\xi, \delta | \mathbf{y}) = -\log \xi - \left(\frac{1}{\xi} + 1\right) \frac{1}{k} \sum_{j=1}^{k} \left[\log y_{j,k} + \log(1 + \delta\{1 - y_{j,k}^{\tau}\})\right]$$

$$+ \frac{1}{k} \sum_{j=1}^{k} \log\left(1 + \delta\{1 - (1+\tau) y_{j,k}^{\tau}\}\right). \tag{2.1}$$

Shrinkage estimators are then obtained by putting a penalty on δ. Below it will be shown that a quadratic penalty is appropriate in view of the asymptotic results for the penalized maximum likelihood (PML) estimators $(\hat{\xi}^P_{k,n}, \hat{\delta}^P_{k,n})$. These estimators are then obtained by optimizing the log-likelihood

$$\frac{1}{k} l_{pen}(\xi, \delta|\mathbf{y}) = \frac{1}{k} l_{EP}(\xi, \delta|\mathbf{y}) - \omega \frac{\delta^2}{2k\sigma^2_{k,n}}, \tag{2.2}$$

where $\omega > 0$ serves as a tuning constant regulating the amount of penalty, and $\sigma^2_{k,n}$ indicating the penalty rate as a function of k. From the asymptotic analysis below, it follows that $\sigma^2_{k,n} = (k/n)^{-2\rho}$ is appropriate.

Alternatively, from a Bayesian perspective, a shrinkage estimator is obtained by considering the posterior mode estimators $(\hat{\xi}^B_k, \hat{\delta}^B_k)$ of the log-posterior

$$\frac{1}{k} \log p(\xi, \delta|\mathbf{y}) = \frac{1}{k} l_{EP}(\xi, \delta|\mathbf{y}) + \frac{1}{k} \log \pi(\xi, \delta), \tag{2.3}$$

where $\pi(\xi, \delta)$ denotes the prior density on (ξ, δ). Following a objective Bayesian point of view, we assign a maximal data information (MDI) prior to ξ, which for a general parameter θ is defined as $\pi(\theta) \propto \exp(E(\log f(\mathbf{Y}|\theta)))$. The concept of MDI priors was introduced in Zellner (1971) in order to maximize the information contributed by the data density, relative to that of the prior density. Beirlant *et al.* (2004) derived that the MDI for a Pareto distribution is given by

$$\pi(\xi) \propto \frac{e^{-\xi}}{\xi}. \tag{2.4}$$

Next, in correspondance with the choice for the penalized log-likelihood (2.2), we here choose a normal prior on δ with mean 0 and variance $\sigma^2_{k,n}$. We also truncate it from the left in order to comply with the restriction $\delta > \max(-1, 1/\tau)$:

$$\pi(\delta) = \frac{1}{\sqrt{2\pi}\sigma_{k,n}} e^{-\frac{1}{2}\frac{\delta^2}{\sigma^2_{k,n}}} \Big/ \left(1 - \Phi(\max(-1, \tau^{-1})/\sigma_{k,n})\right). \tag{2.5}$$

2.2. Asymptotic results for the penalized ML estimator $\hat{\xi}^P_k$

In the Appendix we derive that the first order approximations $(\hat{\xi}^P_k, \hat{\delta}^P_k)$ of the penalized ML estimators are given by

$$\begin{aligned} \hat{\xi}^P_k &= H_{k,n} + \hat{\delta}^P_k \left(1 - E_{k,n}(\tau)\right), \\ \hat{\delta}^P_k &= \frac{1 - H_{k,n}\tau}{D^P_{k,n}} \left(E_{k,n}(\tau) - \frac{1}{H_{k,n}\tau}\right) \end{aligned}$$

where

$$E_{k,n}(s) = \frac{1}{k} \sum_{j=1}^{k} Y^s_{j,k}, \quad s < 0$$

and

$$D^P_{k,n} = \frac{\omega \hat{\xi}^P_k}{k\sigma^2_{k,n}} - \Big(1 \;-\; 2(1-\hat{\xi}^P_k\tau)E_{k,n}(\tau) + (1-2\hat{\xi}^P_k\tau - \hat{\xi}^P_k\tau^2)E_{k,n}(2\tau)$$
$$\qquad - \; \tau(1-E_{k,n}(\tau))E_{k,n}(\tau)\Big).$$

These expressions are identical to the asymptotic EPD-ML estimators derived in Beirlant *et al.* (2009) except for the extra term $\frac{\hat{\xi}^P_k}{k\sigma^2_{k,n}}$ in the expression of $D^P_{k,n}$. As an external estimator of τ we use $\hat{\tau} = \hat{\rho}_{k,n}/H_{k,n}$ with $\hat{\rho}_{k,n}$ taken from Fraga Alves *et al.* (2003). Moreover we set $\zeta = \xi^2(1-2\rho)(1-\rho)^2$. The following result is derived in the Appendix.

Theorem. *Let $F \in \mathcal{F}(\xi,\tau)$ with $|a(x)| = x^\rho C_a(1+o(1))$ as $x \to \infty$. Assume that $\sqrt{k}a(n/k) \to \lambda$ as $k,n \to \infty$, $k/n \to 0$. Setting $\sigma^2_{k,n} = (k/n)^{-2\rho}$, it follows that $\Xi_{k,n} := \sqrt{k}\left(\hat{\xi}^P_k - \xi\right)$ is asymptotically normal with asymptotic mean and variance given by*

$$E_\infty(\Xi_{k,n}) = \frac{\lambda\rho}{1-\rho}\frac{\zeta C_a^2\omega}{\zeta C_a^2\omega + \rho^4\lambda^2}, \tag{2.6}$$

$$Var_\infty(\Xi_{k,n}) = \frac{\xi^2\rho^8\lambda^4}{(\rho^4\lambda^2+\zeta C_a^2\omega)^2}\left(\left(\frac{1-\rho}{\rho}\right)^2 + \frac{\zeta^2C_a^4\omega^2}{\rho^8\lambda^4} + 2\frac{\zeta C_a^2\omega}{\rho^4\lambda^2}\right). \tag{2.7}$$

Minimizing $MSE_\infty(\Xi_{k,n}) = E^2_\infty(\Xi_{k,n}) + Var_\infty(\Xi_{k,n})$ with respect to ω, after some lengthy calculations, leads to the asymptotically optimal value

$$\omega_{opt} = C_a^{-2}.$$

One then obtains from (2.6) and (2.7) that

$$E^{opt}_\infty(\Xi_{k,n}) = \frac{\lambda\rho}{1-\rho}\frac{\zeta}{\zeta+\lambda^2\rho^4},$$
$$Var^{opt}_\infty(\Xi_{k,n}) = \frac{\xi^2}{(\lambda^2\rho^4+\zeta)^2}\left\{(1-\rho)^2\rho^6\lambda^4 + \zeta^2 + 2\zeta\rho^4\lambda^2\right\},$$

from which

$$MSE^{opt}_\infty(\Xi_{k,n}) = \xi^2 + \frac{\lambda^2\rho^2\xi^2(1-2\rho)}{\xi^2(1-2\rho)(1-\rho)^2+\rho^4\lambda^2}. \tag{2.8}$$

Since the right hand side of (2.8) is an increasing function in λ^2 it follows that

$$MSE^{opt}_\infty(\Xi_{k,n}) \le \lim_{\lambda\to\infty} MSE^{opt}_\infty(\Xi_{k,n}) = MSE_\infty\left(\sqrt{k}(\hat{\xi}^{ML}_{k,n}-\xi)\right) = \xi^2\left(\frac{1-\rho}{\rho}\right)^2.$$

Also, expanding the right hand side of (2.8) for $\lambda^2 \to 0$ leads to

$$MSE^{opt}_\infty(\Xi_{k,n}) = \xi^2 + \lambda^2\frac{\rho^2}{(1-\rho)^2}(1+o(1)).$$

We can conclude that the asymptotic MSE of the optimal penalized estimator is uniformly smaller than the MSE of the EPD-ML estimator as given in (1.10), while for smaller λ this asymptotic MSE follows the asymptotic MSE of the Hill estimator, given in (1.9), up to terms of order λ^2. Hence with the penalty $\omega/\sigma^2_{k,n} = C_a^{-2}(k/n)^{2\rho} = a^{-2}(n/k)$ in (2.2), the penalized ML estimator asymptotically follows the better of the two existing estimators as a function of λ or k.

Replacing $(\hat{\xi}_k, \hat{\delta}_k)$ by $(\hat{\xi}^P_k, \hat{\delta}^P_k)$ in $\hat{p}^{EP}_{x,k}$, it follows from the proof of Theorem 5.2 in Beirlant *et al.* (2009) that the resulting tail probability estimator $\hat{p}^P_{x,k}$ satisfies the following asymptotic result under the conditions of the Theorem:

When $p_n = P(X > x_n)$ satisfies $np_n/k \to 0$ and $\log(np_n)/\sqrt{k} \to 0$, then

$$\frac{\sqrt{k}}{\log(k/(np_n))}\,\xi\left(\frac{\hat{p}^P_{x_n,k}}{p_n} - 1\right)$$

is asymptotically normal with the same limit distribution as in the Theorem.

Hence the asymptotic MSE behaviour for the tail probability estimator has the same characteristics as the tail index estimator.

From the simulations it will follow that the choice $\omega = 1$ and the use of estimator of ρ taken from Fraga Alves (2003) yields good results. However, in order to alleviate the problem of choosing the number of top order statistics k that are used in the estimation procedure, one can choose ω adaptively with each sample aiming for a plot of $\hat{\xi}^P_k$ as a function of k which is as horizontal as possible. Setting $\hat{\xi}^P_k = \hat{\xi}^P_k(\omega)$ in order to emphasize the dependence of the penalized ML estimator on ω, a possible choice of ω is obtained by minimizing the variance of the resulting estimators for $k = 1, ..., n$:

$$\omega_{mv} = \text{argmin}_\omega s^2_n\left(\hat{\xi}^P(\omega)\right), \tag{2.9}$$

with $s^2_n(\hat{\xi}^P(\omega)) = \frac{1}{n-1}\sum_{k=1}^n\left(\hat{\xi}^P_k(\omega) - \bar{\hat{\xi}}^P\right)^2$.

3. SIMULATIONS AND PRACTICAL CASE STUDIES

Both the Bayes maximum a posteriori probability estimator and the penalized maximum likelihood estimator are implemented in **R** using the general **optim** function with default parameters.

We performed a simulation study, taking 1000 repetitions of samples of size $n = 200, 500, 1000$ studying the finite sample behaviour of $\hat{\xi}^P_{k,n}(\omega)$ for different distributions. The bias and RMSE are plotted as a function of k.

The following distributions are used:

- *The extreme value distribution* (EV) with $F(x) = \exp(-(1+\xi x)^{-1/\xi})$ $(1+\xi x > 0)$ taking $\xi = 0.25$ in which case $\rho = -0.25$ and $C_a = 1$.
- *The Fréchet distribution* with $\bar{F}(x) = 1 - \exp(-x^{-1/\xi})$ taking $\xi = 0.5$ in which case $\rho = -1$ and $C_a = 0.25$.
- *The Burr distribution* with $\bar{F}(x) = (1+x)^{-4/3}$ so that $\xi = 0.75$ and $\rho = -0.75$ and $C_a = 1$.
- *The loggamma distribution* with $\bar{F}(x) \sim constant \times x^{-2}(\log x)^3$ so that $\xi = 0.5$, which does not belong to the class $\mathcal{F}(\xi, \tau)$.

First, in Figures 1-4 we plotted the bias and the RMSE of the Hill estimator H_k, the EPD-ML estimator $\hat{\xi}_k^{ML}$, the penalized ML estimator $\hat{\xi}_k^P(1)$ with $\omega = 1$, the Bayesian estimator $\hat{\xi}_k^B(1)$ with $\omega = 1$, and the minimum variance reduced bias estimator CH_k from Caeiro *et al.* (2005) given by

$$CH_k = H_{k,n}\left(1 - \frac{\hat{\beta}_{k_1}(\hat{\rho}_{k_1})}{1-\hat{\rho}_{k_1}}\left(\frac{n}{k}\right)^{\hat{\rho}_{k_1}}\right),$$

with

$$\hat{\beta}_k(\rho) = \frac{\left(\frac{k}{n}\right)^{\rho}\left\{\left(\frac{1}{k}\sum_{j=1}^{k}(\frac{j}{k})^{-\rho}\right)\left(\frac{1}{k}\sum_{j=1}^{k}Z_j\right) - \left(\frac{1}{k}\sum_{j=1}^{k}(\frac{j}{k})^{-\rho}Z_j\right)\right\}}{\left(\frac{1}{k}\sum_{j=1}^{k}(\frac{j}{k})^{-\rho}\right)\left(\frac{1}{k}\sum_{j=1}^{k}(\frac{j}{k})^{-\rho}Z_j\right) - \left(\frac{1}{k}\sum_{j=1}^{k}(\frac{j}{k})^{-2\rho}Z_j\right)},$$

where $Z_j := j(\log X_{n-j+1,n} - \log X_{n-j,n})$ $(j = 1, 2, ...)$, and $k_1 = \lfloor n^{0.99} \rfloor$.

In Figure 5 we briefly report on the effect of the choice of ω using $\omega = 1$ and $\omega = \omega_{mv}$ and compare these with the optimal asymptotic RMSE expression from (2.8).

We conclude from the simulations that the finite sample behaviour of the proposed estimators follows the characteristics predicted by the asymptotic analysis to a great extent: for small k the shrinkage estimators ξ_k^P and ξ_k^B show a similar behaviour as the Hill estimator, while for larger k the proposed estimators tend to follow the characteristics of the bias reduced EPD-ML estimator. In between these two k-regions the shrinkage estimators make a transition from the EPD-ML to the Hill RMSE curve. Only in the Fréchet case the Hill estimator shows a smaller RMSE than the shrinkage estimators for small k, while the shrinkage estimators then still show a much smaller RMSE than the EPD-ML estimator.

The Bayesian implementation shows a smaller RMSE than the penalized ML estimator, except for the Fréchet distribution where both RMSEs are comparable. In the latter case $\hat{\xi}_k^B$ shows a negative bias. Also note that the difference between both the Bayesian and penalized likelihood implementation decreases as n increases.

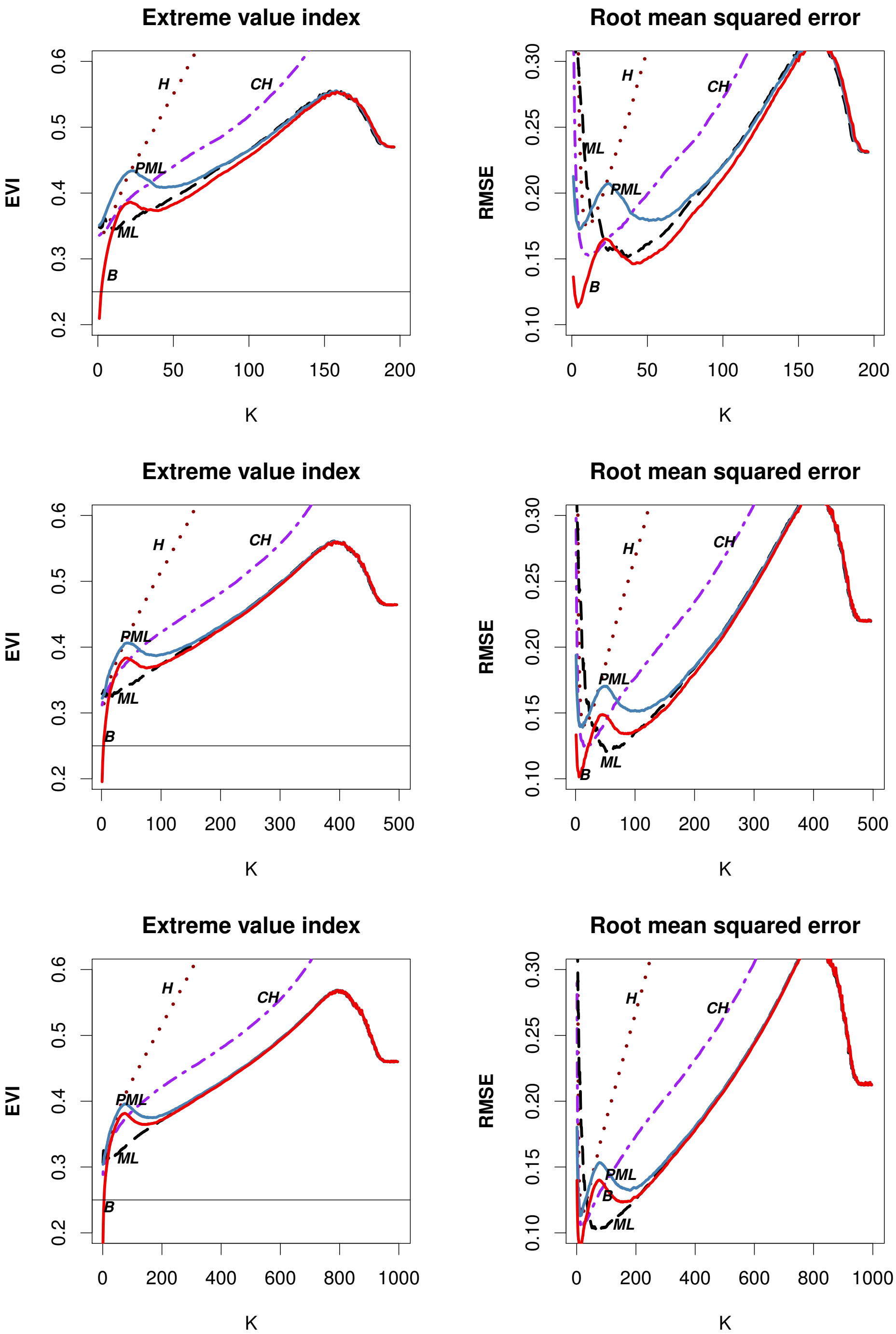

Figure 1: Bias (left) and root mean squared error (right) in case of the **EV distribution** with $\xi = 0.25$ for sample sizes $n = 200$ (top), $n = 500$ (middle) and $n = 1000$ (bottom) for the Hill estimator (H), the EPD-ML estimator $\hat{\xi}_k^{ML}$ (ML), the penalized ML estimator $\hat{\xi}_k^P(1)$ with $\omega = 1$ (PML), the Bayesian estimator $\hat{\xi}_k^B(1)$ with $\omega = 1$ (B), and the minimum variance reduced bias estimator CH_k (CH).

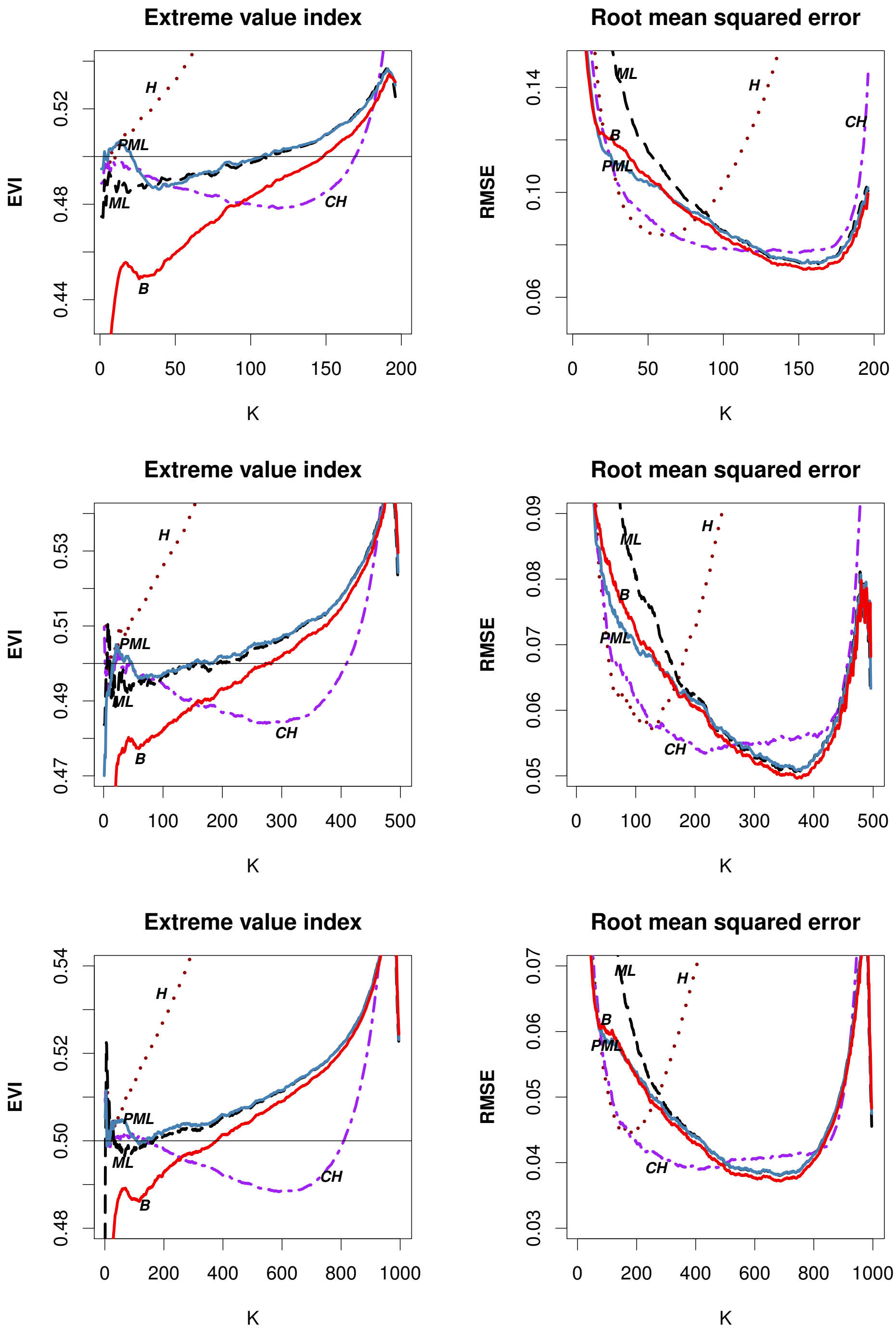

Figure 2: Bias (left) and root mean squared error (right) in case of the **Fréchet distribution** with $\xi = 0.5$ for sample sizes $n = 200$ (top), $n = 500$ (middle) and $n = 1000$ (bottom) for the Hill estimator (H), the EPD-ML estimator $\hat{\xi}_k^{ML}$ (ML), the penalized ML estimator $\hat{\xi}_k^{P}(1)$ with $\omega = 1$ (PML), the Bayesian estimator $\hat{\xi}_k^{B}(1)$ with $\omega = 1$ (B), and the minimum variance reduced bias estimator CH_k (CH).

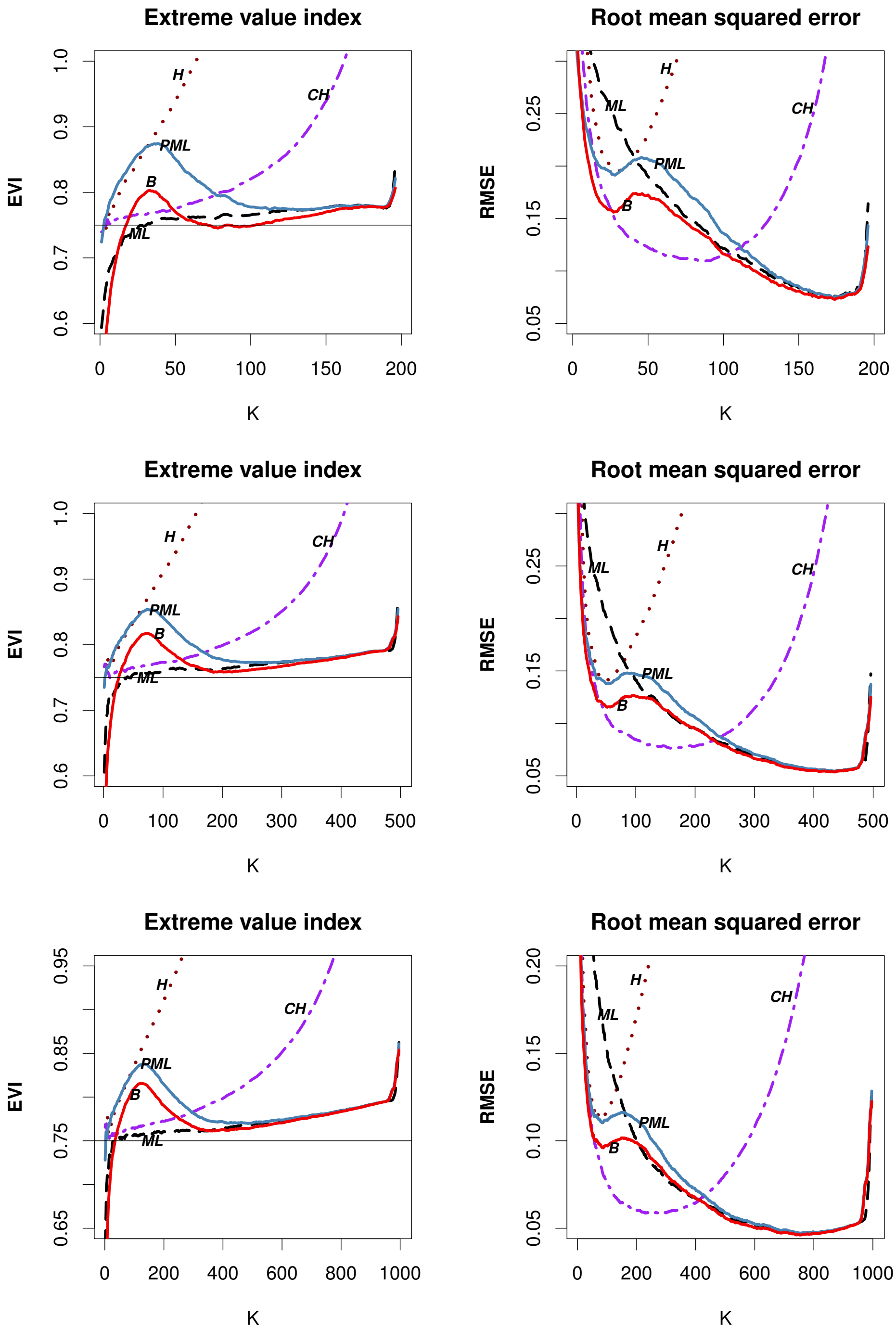

Figure 3: Bias (left) and root mean squared error (right) in case of the **Burr distribution** with $\xi = 0.75$ for sample sizes $n = 200$ (top), $n = 500$ (middle) and $n = 1000$ (bottom) for the Hill estimator (H), the EPD-ML estimator $\hat{\xi}_k^{ML}$ (ML), the penalized ML estimator $\hat{\xi}_k^{P}(1)$ with $\omega = 1$ (PML), the Bayesian estimator $\hat{\xi}_k^{B}(1)$ with $\omega = 1$ (B), and the minimum variance reduced bias estimator CH_k (CH).

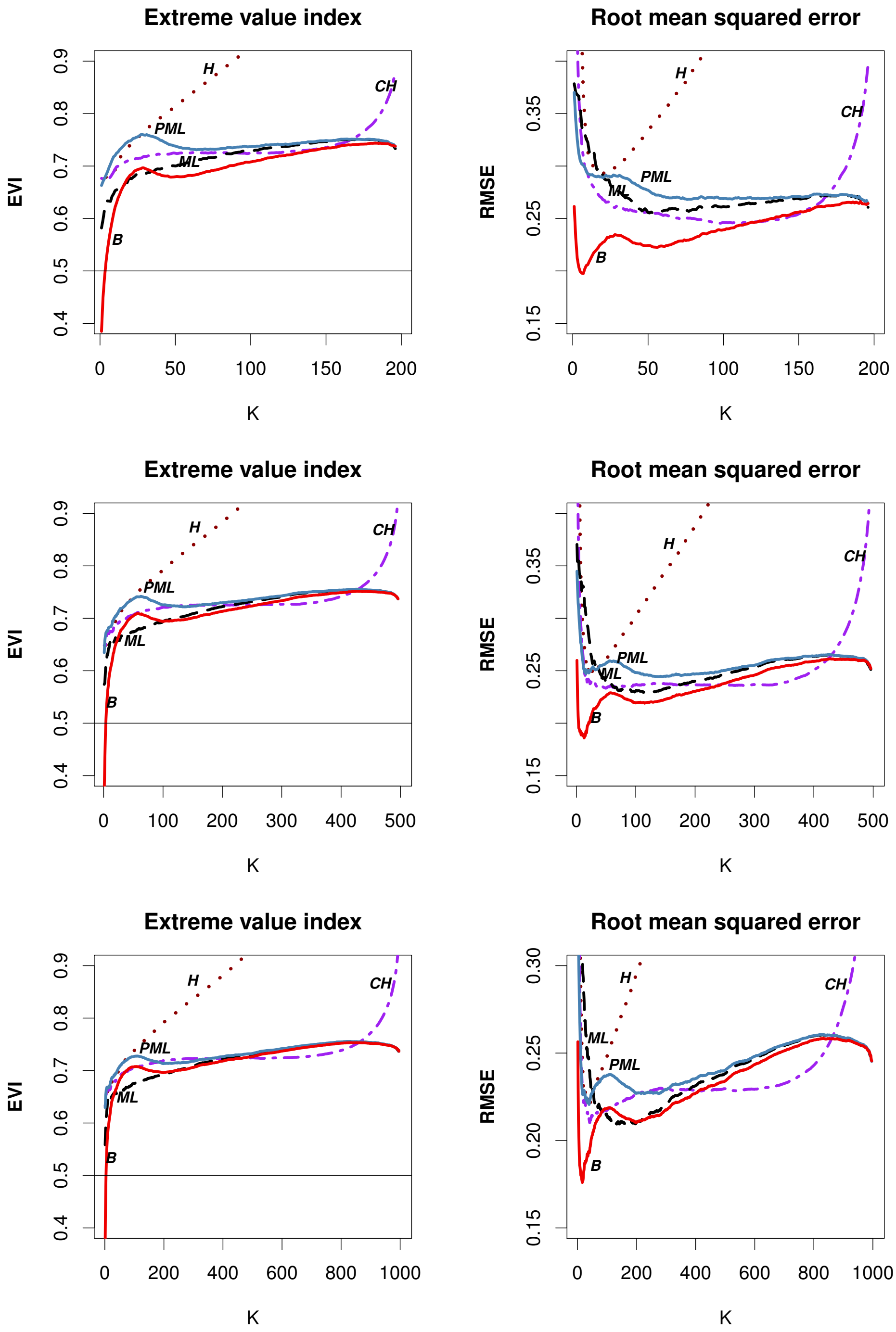

Figure 4: Bias (left) and root mean squared error (right) in case of the **loggamma distribution** with $\xi = 0.5$ for sample sizes $n = 200$ (top), $n = 500$ (middle) and $n = 1000$ (bottom) for the Hill estimator (H), the EPD-ML estimator $\hat{\xi}_k^{ML}$ (ML), the penalized ML estimator $\hat{\xi}_k^{P}(1)$ with $\omega = 1$ (PML), the Bayesian estimator $\hat{\xi}_k^{B}(1)$ with $\omega = 1$ (B), and the minimum variance reduced bias estimator CH_k (CH).

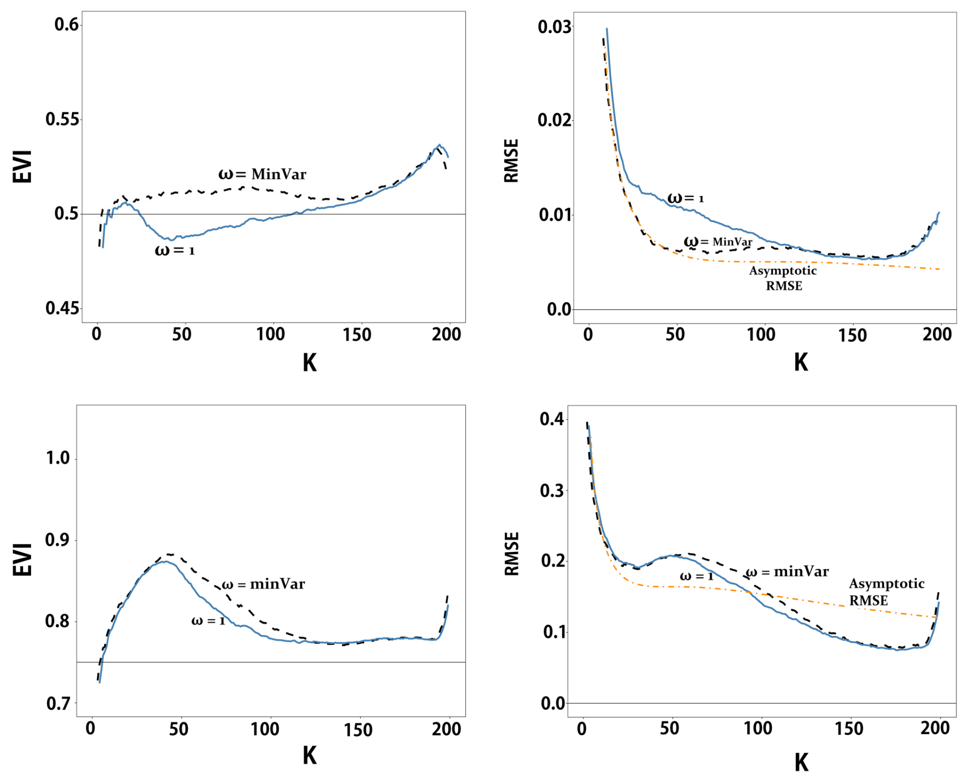

Figure 5: Bias (left) and root mean squared error (right) in case of the **Fréchet distribution** with $\xi = 0.5$ (top) and Burr distribution with $\xi = 0.75$ (bottom) for sample size $n = 200$ comparing the penalized ML estimator $\hat{\xi}_k^P(1)$ with $\omega = 1$, $\omega = \omega_{mv}$ from (2.9), and the optimal asymptotic RMSE from (2.8) replacing λ by $C_a\sqrt{k}(k/n)^{-\rho}$.

The results in case of the loggamma distribution are quite good. Hence it appears that the proposed method exhibits some robustness against deviations from the underlying model.

When the plots of the shrinkage estimators are not systematically increasing with increasing k as in the case of the Fréchet and the Burr distribution, it is useful to use the choice $\omega = \omega_{mv}$ when using the penalized ML estimator. In the case of the Fréchet distribution with $\omega_{opt} = 16$, this adaptive choice of ω leads to a clear RMSE improvement in the transition zone (in k) between the Hill and EPD-ML RMSE behaviour (see Figure 5, top). In the Burr case (see Figure 5, bottom) where $C_a = 1$ and hence $\omega_{opt} = 1$ the choice $\omega = 1$ is best, but the adaptive minimum variance choice $\omega = \omega_{mv}$ is almost as good in RMSE behaviour.

Overall, the proposed shrinkage estimators are competitive with respect to the minimum variance reduced bias estimator CH_k.

In order to illustrate the use of the proposed method we consider the Secura Belgian Re data introduced in section 6.2 in Beirlant *et al.* (2004). For $k \leq 100$ the penalized ML estimator $\hat{\xi}_k^P(1)$ is quite constant and follows the Hill estimator quite closely. This is in contrast with the EPD-ML estimates which vary a lot in that region. The Bayesian estimates $\hat{\xi}_k^B(1)$ and CH estimates show somewhat lower estimates. Beirlant *et al.* (2004) concluded that the Hill estimate in this k-region is an appropriate choice and the adaptive choice $\hat{k} = 98$ was proposed as one of the largest k-values in this region. This proposal is also supported by the present analysis, leading to an estimate $\hat{\xi}^P(1) = 0.28$.

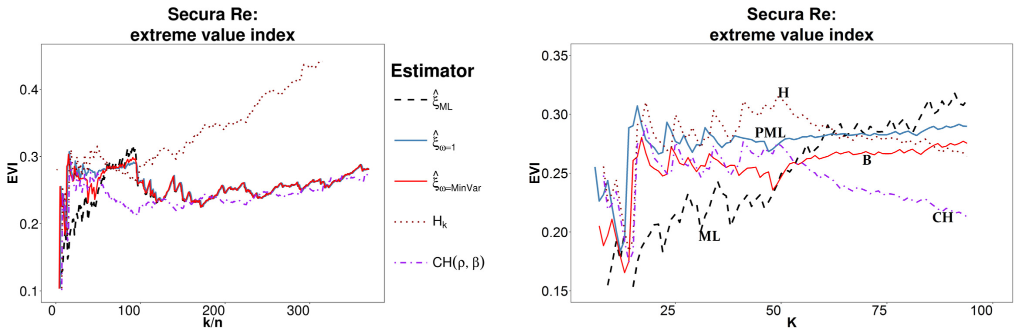

Figure 6: Estimates of ξ for Secura Belgian Re data set: results for the Hill estimator (H), the EPD-ML estimator $\hat{\xi}_k^{ML}$ (ML), the penalized ML estimator $\hat{\xi}_k^P(1)$ with $\omega = 1$ (PML), the Bayesian estimator $\hat{\xi}_k^B(1)$ with $\omega = 1$ (B), and the minimum variance reduced bias estimator CH_k (CH) (left), focused plot for $k = 1, ..., 100$ (right).

4. CONCLUSION

We introduced the use of shrinkage estimators in tail estimation, in order to obtain bias reduction jointly with good MSE behaviour. Shrinkage estimators can be obtained through a penalized ML approach, or through a Bayesian implementation. For larger thresholds the proposed estimators follow the behaviour of the classical Hill estimator with small bias and minimal variance, while the new estimators are never worse than the corresponding bias reduced ML estimators without penalization. The simulated MSE results are competitive with those of other bias reduced estimators. In contrast to existing minimum variance bias reduced estimators we only use second order slow variation conditions.

APPENDIX

Derivation of the expressions of $(\hat{\xi}_k^P, \hat{\delta}_k^P)$. First consider the asymptotic approximations of the penalize ML estimator of ξ based on maximization of (2.2). From (2.1)–(2.2) using expansions in $\delta \to 0$ we obtain

$$\begin{aligned}\frac{1}{k}\log l_{pen}(\xi,\delta|\mathbf{y}) &= -(1+\frac{1}{k})\log\xi - \frac{1}{k}(1+\xi) - (\frac{1}{\xi}+1)\frac{1}{k}\sum_{j=1}^{k}\log y_{j,k} \\ &\quad -\frac{\delta}{1+\xi}\frac{1}{k}\sum_{j=1}^{k}(1-y_{j,k}^{\tau}) + \delta\frac{1}{k}\sum_{j=1}^{k}(1-(1+\tau)y_{j,k}^{\tau}) \\ &\quad -\frac{\omega\delta^2}{2k\sigma_{k,n}^2} + \frac{\delta^2}{2(1+\xi)}\frac{1}{k}\sum_{j=1}^{k}(1-y_{j,k}^{\tau})^2 - \frac{\delta^2}{2}\frac{1}{k}\sum_{j=1}^{k}(1-(1+\tau)y_{j,k}^{\tau})^2 \\ &\quad + O(\delta^3) + c,\end{aligned}$$

where c is a constant only depending on $\sigma_{k,n}^2$ and τ. Note that $\frac{1}{k}\sum_{j=1}^{k}\log y_{j,k} = H_{k,n}$. Then the score functions admit the following expansions in $\delta \downarrow 0$ for $j = 1, ..., k$:

$$\begin{aligned}\frac{\partial}{\partial\xi}\log l_{pen}(\xi,\delta|y_{j,k}) &= -\frac{1}{\xi} + \frac{1}{\xi^2}\log y_{j,k} + \frac{\delta}{\xi^2}(1-y_{j,k}^{\tau}) + O(\delta^2), \\ \frac{\partial}{\partial\delta}\log l_{pen}(\xi,\delta|y_{j,k}) &= -\frac{1}{\xi}\left(1-(1-\xi\tau)y_{j,k}^{\tau}\right) - \frac{\omega\delta}{k\sigma_{k,n}^2} \\ &\quad +\frac{\delta}{\xi}\left(1-2(1-\xi\tau)y_{j,k}^{\tau} + (1-2\xi\tau-\xi\tau^2)y_{j,k}^{2\tau}\right) + O(\delta^2).\end{aligned}$$

Derivation of Theorem. Note that as $k, n \to \infty$, $k/n \to 0$ and $\sqrt{k}\,a(n/k) \to \lambda$, we also have $k\sigma_{k,n}^2 \to \lambda^2 C_a^{-2}$. Also as $\sqrt{k}a(n/k) \to \lambda$ we find using $E_{k,n}(s) \to 1/(1-\xi s)$ (see Theorem A.1 in Beirlant *et al.*, 2009) that

$$D_{k,n}^P = -\frac{\xi C_a^2}{\lambda^2} + \frac{\rho^4}{\xi(1-2\rho)(1-\rho)^2} + o_p(1).$$

Then, proceeding as in the proof of Theorem 3.1 in Beirlant *et al.* (2009), we obtain with $\Gamma_{k,n} = \sqrt{k}(H_{k,n} - \xi)$, $\mathbb{E}_{k,n}(s) = \sqrt{k}(E_{k,n}(s) - \frac{1}{1-\xi s})$ $(s < 0)$, that

$$\begin{aligned}\sqrt{k}\left(\hat{\xi}_k^P - \xi\right) &= \sqrt{k}\left(H_{k,n} - \xi - \hat{\delta}_k^P\frac{\rho}{1-\rho}\right) \\ &= \Gamma_{k,n} - \frac{\rho}{1-\rho}\sqrt{k}\hat{\delta}_k^P \\ &= \Gamma_{k,n}\left(1 + \frac{\rho^2}{\xi(1-\rho^2)}\frac{1}{\xi C_a^2/\lambda^2 + \rho^4/\xi(1-2\rho)(1-\rho)^2}\right) \\ &\quad -\frac{\rho}{\xi C_a^2/\lambda^2 + \rho^4/\xi(1-2\rho)(1-\rho)^2}\mathbb{E}_{k,n}(\hat{\tau}) + o_p(1) \\ &= \Gamma_{k,n}\left(1+\frac{\rho^2(1-2\rho)}{\zeta+\rho^4}\right) + \mathbb{E}_{k,n}(\hat{\tau})\left(\frac{(-\rho)\xi(1-2\rho)(1-\rho)^2}{\rho^4+\zeta}\right) + o_p(1).\end{aligned}$$

Using Theorem A.1 in Beirlant *et al.* (2009), (2.6) and (2.7) follow under $\sqrt{k}a(n/k) \to \lambda$.

ACKNOWLEDGMENTS

This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant Number 102628). The Grantholder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research is that of the author(s), and that the NRF accepts no liability whatsoever in this regard.

The authors take pleasure in thanking S. van der Merwe for his valuable advice concerning the Bayesian implementations.

Thanks to the referee for the constructive comments which improved the presentation of the paper significantly.

REFERENCES

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

[2] Beirlant, J.; Goegebeur, Y.; Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*, Wiley, Chichester.

[3] Beirlant, J.; Figueiredo, F.; Gomes, M.I. and Vandewalle, B. (2008). Improved reduced-bias tail index and quantile estimators, *Journal of Statistical Planning and Inference*, **138**(6), 1851–1870.

[4] Beirlant, J.; Joossens, E. and Segers, J. (2009). Second-order refined peaks-over-threshold modelling for heavy-tailed distribution, *Journal of Statistical Planning and Inference*, **139**, 2800–2815.

[5] Bingham, N.H.; Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation*, Cambridge University Press.

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

[7] Caeiro, F.; Gomes, M.I. and Rodrigues, L.H. (2009). Reduced-bias tail index estimators under a third-order framework, *Communications in Statistics, Theory and Methods*, **38**(7), 1019–1040.

[8] Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution, *The Annals of Statistics*, **27**(2), 760–781.

[9] Fraga Alves, M.I.; Gomes, M.I. and de Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter, *Portugaliae Mathematica*, **60**(2), 193–214.

[10] Gomes, M.I.; Martins, M.J. and Neves, M. (2000). Alternatives to a semi-parametric estimator of parameters of rare events — The jackknife methodology, *Extremes*, **3**(3), 207–229.

[11] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *Journal of the Royal Statistical Society. Series B (Methodological)*, 37–42.

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

[13] Peng, L. (1998). Asymptotically unbiased estimators for the extreme-value index, *Statistics & Probability Letters*, **38**(2), 107–115.

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

[15] Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley.