1 Interpretation of Second-Order Asymptotic Results

1.1 Block maxima method: MLE

The result of Dombry and Ferreira [4] may be summarized as follows.

Suppose we have n IID observations from a distribution function $F(\cdot)$ in the domain of attraction of $G_\xi(x) = \exp\left\{-(1+\xi x)_+^{-1/\xi}\right\}$ where the true value of ξ is denoted $\xi_0 > -\frac{1}{2}$. Suppose the second-order condition (2.45) holds. We suppose that the n observations are grouped into k_n blocks each of length m_n.

In reality there may not be a convenient factorization so that $n = k_n m_n$ exactly. However, if we just fix one of them, say k_n, and define m_n to be the integer part of $\frac{n}{k_n}$, then we can still have k_n blocks of length either m_n or $m_n + 1$ that cover all n observations. Since the difference between m_n and $m_n + 1$ in the following theory will not affect any of the asynptotic results, this is effectively the same as assuming $n = k_n m_n$ exactly. However, some convention of this nature might be worth fixing for the purpose of conducting simulations.

Following [4], we focus just on the estimation of ξ, though the same principles may be applied for estimating any function of the GEV parameters, of which tail probabilities and extreme quantiles are the most common applications

The Fisher information matrix is denoted $\mathcal{I}_\xi$, so let $v_1(\xi)$ be the (3,3) entry of $\mathcal{I}_\xi^{-1}$. Note that the GEV approximation in [4] is normalized so that the true values of μ and ψ are 0 and 1 respectively, so $\mathcal{I}_\xi$ and hence $v(\xi)$ are functions of ξ alone. This is an *explicit* function of ξ: $\mathcal{I}_\xi$ is calculated by substituting $\mu = 0,\ \psi = 1$ in formula (1.23), then invert the matrix. We are saving space and the likelihood of typographical error by not trying to write out $v_1(\xi)$ explicitly.

With these conventions, and writing ξ_0 for the true value of ξ, we then have

$$Var(\hat{\xi}_n) \quad \approx \quad \frac{v(\xi_0)}{k_n}.$$

Similarly, the function $\mathbf{b}(\xi, \rho)$ as defined in [4] is an *explicit* function of ξ and ρ; it involves and integral, and we assume that in practice it is evaluated by numerical quadrature, but it is explicit in the sense that it does not involve any additional parameters beyond ξ and ρ. Let $w_1(\xi, \rho)$ be the third entry of the three-dimensional vector $\mathcal{I}_\xi^{-1}\mathbf{b}(\xi, \rho)$. This depends just on ξ and ρ, though we have to use numerical integration and numerical matrix inversion to evaluate it.

The main result of [4], limited to just the $\hat{\xi}_n$ component, shows that

$$\sqrt{k_n}(\hat{\xi}_n - \xi_0) \quad \xrightarrow{d} \quad \mathcal{N}(\lambda w_1(\xi_0, \rho), v_1(\xi_0),$$

and we also have asymptotically $\lambda \approx \sqrt{k_n} A(m_n)$, so the asymptotic bias of $\hat{\xi}_n$ is

$$\frac{\lambda w_1(\xi_0, \rho)}{\sqrt{k_n}} \quad \approx \quad A(m_n) w_1(\xi_0, \rho)$$

Hence the mean squared error (MSE) of $\hat{\xi}_n$ is asymptotically of the form

$$\{A(m_n) w_1(\xi_0, \rho)\}^2 + \frac{v_1(\xi_0)}{k_n}.$$

Recall that $\rho \leq 0$ and $A(m)$ is regularly varying of index ρ, and if $\rho = 0$ it is necessary to impost the additional condition $A(m) \to 0$ as $m \to \infty$. The theory is simplified if we assume $\rho < 0$ and $A(\rho) \sim Dm^\rho$ for some $D \in (0, \infty)$ as $m \to \infty$.

With those simplifications, we have

$$MSE \approx D^2 w_1^2 \left(\frac{n}{k_n}\right)^{2\rho} + \frac{v_1}{k_n}$$

where we have omitted the fixed constants ξ_0 and ρ from the notations for w_1 and v_1.

By simple calculus, we then find that the k_n to minimize MSE is given by

$$k_n = \left\{\frac{v_1}{D^2 w_1^2 n^{2\rho} |2\rho|}\right\}^{1/(1-2\rho)}$$

and the MSE is then

$$MSE = \left\{D^2 w_1^2 n^{2\rho}\right\}^{1/(1-2\rho)} v_1^{-2\rho/(1-2\rho)} |2\rho|^{2\rho/(1-2\rho)} (1 + |2\rho|).$$

1.2 Threshold exceedance method: MLE

The block maxima method analyzed by Dombry and Ferreira [4] may be considered an alternative to the threshold exceedance estimator based on the generalized Pareto distribution (GPD). For the equivalent asymptotic theory for GPD estimators we follow de Haan and Ferreira [10], specifically Section 3.4. A summary of their treatment is as follows.

Suppose again we have n IID observations from a distribution F and define some k_n such that $k_n \to \infty$, $\frac{k_n}{n} \to 0$ as $n \to \infty$. Assume that (2.33) holds with the general form of $H(x)$ given by (2.37). Assume the observations are ordered as $X_1 > X_2 > \ldots > X_n$ and define estimators $(\hat{\sigma}_n, \hat{\xi}_n)$ as the values of (σ, ξ) that minimize

$$k \log \sigma + \left(\frac{1}{\xi} + 1\right) \log\left\{1 + \xi \frac{X_i - X_{k+1}}{\sigma}\right\}$$

They define an expression $\mathbf{b}(\xi, \rho)$ by

$$\mathbf{b}(\xi, \rho) = \begin{pmatrix} \frac{\xi+1}{(1-\rho)(1+\xi-\rho)} & \frac{-\rho}{(1-\rho)(1+\xi-\rho)} \end{pmatrix}$$

valid for $\rho < 0$, or just $\begin{pmatrix} 1 & 0 \end{pmatrix}$ when $\rho = 0$.

$$\mathbf{b}(\xi, \rho) = \begin{cases} \begin{pmatrix} \frac{\xi+1}{(1-\rho)(1+\xi-\rho)} & \frac{-\rho}{(1-\rho)(1+\xi-\rho)} \end{pmatrix} & \text{if } \rho < 0, \\ \begin{pmatrix} 1 & 0 \end{pmatrix} & \text{if } \rho = 0. \end{cases}$$

valid for $\rho < 0$, or just $\begin{pmatrix} 1 & 0 \end{pmatrix}$ when $\rho = 0$.

They again define $\lambda = \lim_{n\to\infty} \sqrt{k_n} A\left(\frac{n}{k_n}\right)$ and then show

$$\sqrt{k_n} \begin{pmatrix} \hat{\xi}_n - \xi_0 & \frac{\hat{\sigma}_n}{a(n/k_n)} - 1 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\lambda \mathbf{b}(\xi, \rho), \Sigma\right)$$

where the matrix Σ is given by

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

1.2.1 Heuristic derivation

Assume a sequence of thresholds $u_n \to \omega_F$ such that $1 - F(u_n) \approx \frac{k_n}{n}$; this is similar to the way we defined the block maxima model with block sizes of order $m_n \approx \frac{n}{k_n}$ with k_n a sequence of sample sizes that will satisfy $k_n \to \infty$, $\frac{k_n}{n} \to 0$ as $n \to \infty$. In terms of $U = \left(\frac{1}{1-F}\right)^{\leftarrow}$, we define $u_n = U\left(\frac{n}{k_n}\right)$.

Writing u in place of u_n we now define $Y_u = U\left(\frac{n}{Sk_n}\right) - U\left(\frac{n}{k_n}\right)$ with S having a uniform distribution on $(0, 1)$. Then

$$\begin{aligned}
\Pr\{Y_u \geq u + y\} &= \Pr\left\{U\left(\frac{n}{k_n S}\right) > y + u\right\} \\
&= \Pr\left\{\frac{n}{k_n S} > \frac{1}{1 - F(y+u)}\right\} \\
&= \frac{n}{k_n}\{1 - F(y+u)\} \\
&= \frac{1 - F(y+u)}{1 - F(u)}.
\end{aligned}$$

In other words, the distribution of Y_u is indeed that of $X - u$, condition on on $X > u$, where X is a random variable from the distribution F.

Under the model (2.33), the first-order approximation to $Y_u = U\left(\frac{n}{Sk_n}\right) - U\left(\frac{n}{k_n}\right)$ is $a(u) \cdot \frac{S^{-1/\xi} - 1}{\xi}$ which is indeed that of a GPD with scale parameter $\sigma = a(u)$ and shape parameters ξ.

The log likelihood for a single observation Y_u is

$$\ell(\sigma, \xi) = \log\sigma + \left(\frac{1}{\xi} + 1\right)\log\left(1 + \xi\frac{Y_u}{\sigma}\right).$$

form which we deduce

$$\begin{aligned}
\sigma\frac{\partial\ell}{\partial\sigma} &= -\frac{1}{\xi} + \left(\frac{1}{\xi} + 1\right)\left(1 + \xi\frac{Y_u}{\sigma}\right)^{-1}, \\
\frac{\partial\ell}{\partial\xi} &= -\frac{1}{\xi^2}\log\left(1 + \xi\frac{Y_u}{\sigma}\right) + \frac{1}{\xi}\left(\frac{1}{\xi} + 1\right)\left\{1 - \left(1 + \xi\frac{Y_u}{\sigma}\right)^{-1}\right\}.
\end{aligned}$$

To calculate the bias vector $\mathbf{b}$, we need to find expressions for the expected values of these terms.

Rewriting in terms of S and including the remainder term from (2.37),

$$\begin{aligned}
\sigma\frac{\partial\ell}{\partial\sigma} &= -\frac{1}{\xi} + \left(\frac{1}{\xi} + 1\right)\left\{(1-S)^{-\xi} + \xi A(u) H_{\xi,\rho}\left(\frac{1}{1-S}\right) + o(A(u))\right\}^{-1} \\
&= -\frac{1}{\xi} + \left(\frac{1}{\xi} + 1\right)\left\{(1-S)^{\xi} - \xi A(u)\,(1-S)^{2\xi}\, H_{\xi,\rho}\left(\frac{1}{1-S}\right) + o(A(u))\right\} \\
&= -\frac{1}{\xi} + \left(\frac{1}{\xi} + 1\right)\left\{(1-S)^{\xi} - \frac{\xi}{\rho}A(u)\left[\frac{(1-S)^{\xi-\rho}}{\xi+\rho} - \frac{(1-S)^{\xi}}{\xi}\right] + o(A(u))\right\}
\end{aligned}$$

and hence

$$\begin{aligned} E\left\{\sigma\frac{\partial\ell}{\partial\sigma}\right\} &= -\frac{1}{\xi}+\left(\frac{1}{\xi}+1\right)\left\{\frac{1}{1+\xi}-\frac{\xi}{\rho}A(u)\left[\frac{1}{(\xi+\rho)(1+\xi-\rho)}-\frac{1}{\xi(1+\xi)}\right]+o(A(u))\right\} \\ &= -\frac{1+\xi}{\rho}A(u)\left[\frac{\rho(\rho-1)}{\xi(\xi+\rho)(1+\xi)(1+\xi-\rho)}\right]+o(A(u)) \\ &= -A(u)\cdot\frac{\rho-1}{\xi(\xi+\rho)(1+\xi-\rho)}+o(A(u)). \end{aligned}$$

Similarly for $\frac{\partial\ell}{\partial\xi}$: write this as $B_1+B_2+B_3$ where

$$\begin{aligned} B_1 &= -\frac{1}{\xi^2}\log\left(1+\frac{\xi Y_u}{\sigma}\right) \\ &= -\frac{1}{\xi^2}\log\left\{(1-S)^{-\xi}\right\}+\frac{1}{\xi}A(u)(1-S)^{\xi}H_{\xi,\rho}\left(\frac{1}{1-S}\right)+o(A(u)), \\ &= \frac{1}{\xi}\log(1-S)+\frac{A(u)}{\xi\rho}\left[\frac{(1-S)^{-\rho}-(1-S)^{-\xi}}{\xi+\rho}-\frac{1-(1-S)^{-\xi}}{\xi}\right]+o(A(u)), \\ B_2 &= \frac{1+\xi}{\xi^2}, \\ B_3 &= -\left(\frac{1+\xi}{\xi^2}\right)\left(1+\xi\frac{Y_u}{\sigma}\right)^{-1} \\ &= -\frac{\sigma}{\xi}\frac{\partial\ell}{\partial\sigma}-\frac{1}{\xi^2}. \end{aligned}$$

Taking expectations in turn,

$$\begin{aligned} E\{B_1\} &= -\frac{1}{\xi}+\frac{A(u)}{\xi\rho}\left[\frac{1}{\xi+\rho}\left(\frac{1}{1-\rho}-\frac{1}{1-\xi}\right)-\frac{1}{\xi}\left(1-\frac{1}{1-\xi}\right)\right]+o(A(u)) \\ &= -\frac{1}{\xi}+A(u)\cdot\frac{2-\rho}{\xi(\xi+\rho)(1-\rho)(1-\xi)}+o(A(u)) \\ E\{B_2\} &= \frac{1+\xi}{\xi^2}, \\ E\{B_3\} &= -\frac{1}{\xi^2}+A(u)\cdot\frac{\rho-1}{\xi^2(\xi+\rho)(1+\xi-\rho)}+o(A(u)). \end{aligned}$$

Adding together the three expressions, then terms that do not involve $A(u)$ sum to 0 (as they must), so we are left with

$$E\left\{\frac{\partial\ell}{\partial\xi}\right\} = A(u)\cdot\left\{\frac{2-\rho}{\xi(\xi+\rho)(1-\rho)(1-\xi)}+\frac{\rho-1}{\xi^2(\xi+\rho)(1+\xi-\rho)}\right\}+o(A(u)).$$

We also have (see Chapter 1) for the Fisher information matrix with $\sigma=1$,

$$\mathcal{I}^{-1} = (1+\xi)\begin{pmatrix} 2 & -1 \\ -1 & 1+\xi \end{pmatrix}.$$

Hence the asymptotic expectation of $\begin{pmatrix} \frac{\hat{\sigma}_n}{\sigma_n} - 1 \\ \hat{\xi}_n - \xi_0 \end{pmatrix}$ is

$$A(u)(1+\xi)\begin{pmatrix} 2 & -1 \\ -1 & 1+\xi \end{pmatrix}\begin{pmatrix} E\left\{\sigma\frac{\partial \ell}{\partial \sigma}\right\} \\ E\left\{\frac{\partial \ell}{\partial \xi}\right\} \end{pmatrix}$$

This should correspond to the formula for $b_{\gamma,\rho}$ on page 92 of de Haan and Ferreira [10] (with ξ replacing γ), however our $\mathcal{I}^{-1}$ does not correspond to their Σ. It seems that a explanation of this discrepancy was given in a paper by Drees, Ferreira and de Haan [5] but I will need to check up on this.

1.2.2 Interpretation in terms of bias and mean squared error

If we focus specifically on $\hat{\xi}_n$, we can say,

$$\text{Bias of } \hat{\xi}_n \approx A\left(\frac{k_n}{n}\right)\frac{\xi+1}{(1-\rho)(1+\xi-\rho)}, \qquad \text{Variance of } \hat{\xi}_n \approx \frac{1+\xi^2}{k_n},$$

If we again assume $A(m) \sim Dm^\rho$ we then get

$$\text{MSE of } \hat{\xi}_n \approx D^2\left(\frac{n}{k_n}\right)^{2\rho} w_2^2 + \frac{v_2}{k_n^2}$$

with

$$w_2 = \frac{\xi+1}{(1-\rho)(1+\xi-\rho)},\ v_2 = (1+\xi^2).$$

This is of exactly the same structure as the MSE for the GEV estimator, but with $w_2,\ v_2$ replacing $w_1,\ v_1$. The expressions for the optimal k_n and the resulting optimal MSE therefore follow by the same arguments. The ratio of the optimal MSEs is

$$\left(\frac{w_1}{w_2}\right)^{2/(1-2\rho)} \cdot \left(\frac{v_1}{v_2}\right)^{-2\rho/(1-2\rho)}.$$

1.3 Other estimators

The method of *probability-weighted moments* (PWMs) has been proposed by hydrologists. The original method for the GEV was proposed in [17], and for the GPD in [15]. A book length treatment of the application of these methods in hydrology is due to Hosking and Wallis [16].

The motivation behind these methods is a claim (supported by simulations) that although the variances of PWMs is greater than that of MLEs, the bias is nevertheless smaller, suggesting that the MSE may be smaller for PWMs. This claim was disputed by Coles and Dixon [3], but the preceding sections suggest an asymptotic analysis that would rogirously compare the MSEs of the two approaches. This has been done in Section 3.6 of [10] for the GPD, and in [8] for the GEV.

2 Automated selection of the threshold

References: [21, 11, 12, 14, 1, 20, 18, 19, 7, 9, 2, 6, 22]

For the simple case of a Pareto tail, Hall and Welsh (1984) [11] shows that the optimal rates of convergence developed by Hall (1982) [13] cannot be improved, and Hall and Welsh (1985) [12] proposed an adaptive estimator that achieves the optimal mean squared error, with probability tending to 1 as sample size $n \to \infty$.

Their actual formulation assumed the model

$$F(x) \quad = \quad Cx^{\alpha}\left[1 + Dx^{\beta} + o(x^{\beta})\right]$$

which is equivalent to our assumption (2.55) if we replace x by $\frac{1}{x}$ everywhere (which obviously makes no difference to the substance of the problem). In this case they study Hill's estimator $(\hat{\alpha}_r, \hat{C}_r)$ based on the r smallest order statistics and they repeat the argument of [13] that shows the optimal value of r to be $\lambda n^{2\beta/(2\beta+\alpha}$ for some $lambda \in (0, \infty)$. They also write $\rho = \beta/\alpha$ (note that this is $-\rho$ in the notation of [10]).

From these starting points, Hall and Welsh propose:

1. Assume $\rho \in (\rho_1, \rho_2)$ for known ρ_1, ρ_2 and choose constants $\sigma,\ \tau_1,\ \tau_2$ such that $0 < \sigma < \frac{2\rho_1}{2\rho_1+1}$, $\frac{2\rho_2}{2\rho_2+1} < \tau_1 < \tau_2 < 1$, such that $2\rho_2(1-\tau_1) < \sigma$. Set $s = [n^{\sigma}]$ (integer part of n^{σ}), $t_1 = [n^{\tau_1}]$, $t_2 = [n^{\tau_2}]$.

2. Define

$$\begin{aligned}
\hat{\rho} &= \left|\log\left|\frac{\hat{\alpha}_{t_1} - \hat{\alpha}_s}{\hat{\alpha}_{t_2} - \hat{\alpha}_s}\right| \Big/ \log\left(\frac{t_1}{t_2}\right)\right| \\
\hat{\lambda}_0 &= \left|\hat{\alpha}_s \Big/ (2\hat{\rho})^{1/2}(n/t_1)^{\hat{\rho}}(\hat{\alpha}_{t_1} - \hat{\alpha}_s)\right|^{2/(2\hat{\rho}+1)}, \\
\hat{r}_0 &= \left[\hat{\lambda} n^{2\hat{\rho}/(2\hat{\rho}+1)}\right].
\end{aligned}$$

3. Then $\frac{\hat{r}_0}{r_0} \xrightarrow{p} 1$ (where r_0 is the true optimal value), and hence the estimator based on $\hat{r}_0$ order statistics achieves the optimal MSE.

References

[1] Shihong Cheng and Liang Peng. Confidence intervals for the tail index. *Bernoulli*, 7 (5):751–760, 2001.

[2] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51 (4):661–703, 2009.

[3] Stuart G. Coles and Mark J. Dixon. Likelihood-based inference for extreme value models. *Extremes*, 2 (1):5–23, 1999.

[4] C. Dombry and A. Ferreira. Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25:1690–1723, 2019.

[5] H. Drees, A. Ferreira, and L. de Haan. On maximum likelihood estimation of the extreme value index. *Annals of Applied Probability*, 14:1179–1201, 2004.

[6] Holger Drees, Anja Janssen, Sidney I. Resnick, and Tiandong Wang. On a minimum distance procedure for threshold selection in tail analysis. *SIAM J. Math. Data Sci.*, 2 (1):75–102, 2020.

[7] Ana Ferreira. Optimal asymptotic estimation of small exceedance probabilities. *Journal of Statistical Planning and Inference*, 104:83–102, 2002.

[8] Ana Ferreira and Laurens de Haan. On the block maxima method in extreme value theory: PWM estimators. *Annals of Statistics*, 43:276–298, 2015.

[9] Ana Ferreira, Laurens de Haan, and Liang Peng. Adaptive estimators for the endpoint and high quantities of a probability distribution. *Report Eurandom; Vol. 99042*, https://research.tue.nl/en/publications/adaptive-estimators-for-the-endpoint-and-high-quantities-of-a-pro, 1999.

[10] L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

[11] P. Hall and A.H. Welsh. Best attainable rates of convergence for estimates of parameters of regular variation. *Annals of Statistics*, 12:1079–1084, 1984.

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

[13] Peter Hall. On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B*, 44:37–42, 1982.

[14] Peter Hall and Ishay Weissman. On the estimation of extreme tail probabilities. *The Annals of Statistics*, 25 (3):1311–1326, 1997.

[15] J.R.M. Hosking and J.R. Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29:339–349, 1987.

[16] J.R.M. Hosking and J.R. Wallis. *Regional Frequency Analysis*. Cambridge University Press, Online ISBN 9780511529443, https://doi.org/10.1017/CBO9780511529443, 1997.

[17] J.R.M. Hosking, J.R. Wallis, and E.F. Wood. Estimation of the generalized extreme value distribution by the method of probability-weighted moments. *Technometrics*, 27:251–261, 1985.

[18] Johanna Mager. Automatic threshold selection of the peaks over threshold method. *Technische Universität München Master's Thesis*, https://mediatum.ub.tum.de/doc/1254349/document.pdf, 2015.

[19] Taku Moriyama. On tail inference in iid settings with nonnegative extreme value index. *https://arxiv.org/abs/2409.00906v1*, 2024.

[20] Liang Peng and Yongcheng Qi. Estimating the first- and second-order parameters of a heavy-tailed distribution. *Aust. N. Z. J. Stat.*, 46 (2):305–312, 2001.

[21] J. Pickands III. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.

[22] Laura Fee Schneider, Andrea Krajina, and Tatyana Krivobokova. Threshold selection in univariate extreme value analysis. *Extremes*, 24:881–913, 2021.