STOR 834: Spring 2025: Class of 2025-02-27

Here I essentially covered Section 2.5 of the text, though without filling in all the details. Since this section is likely to change in subsequent versions, I am attaching here the current version of this material.

It can be readily checked that this implies

$$y \;=\; (ct)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}t^{-\beta/\alpha}+o(t^{-\beta/\alpha})\right\}. \tag{2.38}$$

Therefore, $U(t)$ satisfies the right hand side of (2.38).

Hence,

$$\begin{aligned} U(tx)-U(t) &= (cxt)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}(xt)^{-\beta/\alpha}\right\}-(ct)^{1/\alpha}\left\{1+\frac{d}{\alpha}c^{-1-\beta/\alpha}t^{-\beta/\alpha}\right\}+o(t^{1/\alpha-\beta/\alpha}) \\ &= (ct)^{1/\alpha}(x^{1/\alpha}-1)+\frac{d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}(x^{1/\alpha-\beta/\alpha}-1)+o(t^{1/\alpha-\beta/\alpha}). \end{aligned} \tag{2.39}$$

If we define $a(t)=\alpha^{-1}(ct)^{1/\alpha}$, we get

$$\frac{U(tx)-U(t)}{a(t)}-\frac{x^{1/\alpha}-1}{1/\alpha} \;=\; dc^{-1-\beta/\alpha}t^{-\beta/\alpha}(x^{1/\alpha-\beta/\alpha}-1)+o(t^{-\beta/\alpha}),$$

which, however, does not give the form of limit function we are aiming at.

Therefore, we return to (2.39) and rewrite

$$\begin{aligned} U(tx)-U(t) &= \left\{(ct)^{1/\alpha}+\frac{(1-\beta)d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\right\}(x^{1/\alpha}-1) \\ &+\frac{(1-\beta)d}{\alpha^2}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\cdot\frac{\alpha}{1-\beta}\left\{x^{1/\alpha-\beta/\alpha}-1-(1-\beta)(x^{1/\alpha}-1)\right\}+o(t^{1/\alpha-\beta/\alpha}). \end{aligned}$$

Now define $a(t)=\alpha^{-1}\left\{(ct)^{1/\alpha}+\frac{\beta d}{\alpha}c^{1/\alpha-1-\beta/\alpha}t^{1/\alpha-\beta/\alpha}\right\}$, $A(t)=-\frac{(1-\beta)d}{\beta}c^{-1-\beta/\alpha}t^{-\beta/\alpha}$, then

$$\lim_{t\to\infty}\frac{\frac{U(tx)-U(t)}{a(t)}-\frac{x^{1/\alpha}-1}{(1/\alpha)}}{A(t)} \;=\; -\frac{\alpha}{\beta}\left(\frac{x^{1/\alpha-\beta/\alpha}-1}{1/\alpha-\beta/\alpha}-\frac{x^{1/\alpha}-1}{1/\alpha}\right).$$

This is precisely of the form (2.37) with $\xi=\frac{1}{\alpha}$, $\rho=-\beta/\alpha$.

2.5 Estimation theory based on second-order asymptotics

We focus here on a paper by Dombry and Ferreira [60], but this is just one of a series of papers going back to the 1980s [234, 63, 58, 79, 59, 182].

Consider an IID random sequence $\{X_i,\ i=1,2,\ldots\}$ where the common distribution function is F. Suppose the observations are grouped into blocks of length m, and let $M_{k,m}=\max\{X_i:(k-1)m+1,\ldots,km\}$ be the maximum of the k'th block. We assume F is in the domain of attraction of the GEV, so that

$$\Pr\left\{\frac{M_{k,m}-b_m}{a_m}\le x\right\} \;=\; F^m(a_mx+b_m) \;\to\; G_{\xi_0}(x) \;=\; \exp\left\{-(1+\xi_0x))_+^{-1/\xi_0}\right\}. \tag{2.40}$$

for some "true value" ξ_0 which we write that way to distinguish it from the unknown parameter ξ in the following likelihood analysis. We define $g_{\xi_0}(x) = \frac{dG_{\xi_0}(x)}{dx} = (1+\xi_0 x))^{-1/\xi_0-1} \exp\left\{-(1+\xi_0 x))^{-1/\xi_0}\right\}$ defined whenever $1+\xi_0 x > 0$ to be the density of G_{ξ_0} and let

$$\ell(\mu,\psi,\xi;x) \quad = \quad \log\psi + \log g_\xi\left(\frac{x-\mu}{\psi}\right) \tag{2.41}$$

be the log density for arbitrary ξ when the distribution is extended to include a location and scale parameter. The idea is that we treat the block maxima $M_{i,m}$ for $1 \le i \le k$ as if their exact distribution was GEV with parameters $\boldsymbol{\theta} = (\mu,\psi,\xi)$ though we know that for finite m this is only an approximation. Define the log likelihood

$$L_{k,m}(\boldsymbol{\theta}) \quad = \quad \sum_{i=1}^{k} \ell(\boldsymbol{\theta}, M_{i,m}) \tag{2.42}$$

In the following, we shall consider a sequence of sample sizes and block lengths $k_n,\ m_n$ where both k_n and M_n are indexed by n. We define $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\psi}_n, \hat{\xi}_n)$ to be a local maximizer of the log likelihood function, or just the MLE for short, if it satisfies the likelihood equations

$$\frac{\partial L_{k,m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad = \quad 0 \tag{2.43}$$

and if the hessian matrix $\frac{\partial^2 L_{k,m}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ is positive definite at $\hat{\boldsymbol{\theta}}_n$.

Dombry and Ferreira differ slightly from the notation of the previous section by defining $V = (-1/\log F)^{\leftarrow}$ (instead of $U = (1/(1-F))^{\leftarrow}$ as previously, though in most cases the two definitions will lead to the same asymptotics). In that context they assume, first, that there exists a_m such that

$$\lim_{m\to\infty} \frac{V(mx)-V(m)}{a_m} \quad = \quad \frac{x^{\xi_0}-1}{\xi_0} \tag{2.44}$$

and, second, that for some positive function $a(t)$ as $t\to\infty$ and some positive or negative function $A(t)$ as $t\to\infty$ with $\lim_{t\to\infty} A(t) = 0$,

$$\lim_{t\to\infty} \frac{\frac{V(tx)-V(t)}{a(t)} - \frac{x^{\xi_0}-1}{\xi_0}}{A(t)} \quad = \quad \int_1^x \int_1^s s^{\xi_0-1} u^{\rho-1} du ds \;=\; H_{\xi_0,\rho}(x),\ x>0, \tag{2.45}$$

where $\xi_0 > -\frac{1}{2},\ \rho \le 0$, the function A is regularly varying with index ρ, and $H_{\xi_0,\rho}$ is given by (2.37) with $\xi = \xi_0$. As noted previously, in any case where a limit of the form (2.45) exists, we can without loss of generality, redefining the functions $a(t)$ and $A(t)$ is necessary, assume that the right hand side is $H_{\xi_0,\rho}(x)$ for suitable $\rho \le 0$.

Dombry and Ferreira consider limiting cases as $k = k_n \to \infty,\ m = m_n \to \infty$ where

$$\lim_{n\to\infty} \sqrt{k_n} A(m_n) \quad = \quad \lambda \in \mathbb{R}. \tag{2.46}$$

They define $\boldsymbol{\theta}_0 = (0,1,\xi_0)$ and then

$$\begin{aligned} Q_{\xi_0}(s) &= \frac{(-\log s)^{-\xi_0}-1}{\xi_0},\ s\in(0,1) \\ \mathbf{b}(\xi_0,\rho) &= \int_0^1 \frac{\partial^2 \ell}{\partial x \partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0, Q_{\xi_0}(s)) H_{\xi_0,\rho}\left(\frac{1}{-\log s}\right) ds, \\ I_{\xi_0} &= -\int_0^1 \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}_0, Q_{\xi_0}(s)) ds. \end{aligned}$$

Note that I_{ξ_0} is the Fisher information for the GEV evaluated at $\boldsymbol{\theta}_0$; this is the same matrix as was shown in Chapter 1 following [194].

With these preliminaries, Theorem 2.2 of [60] states:

(a) There exists a sequence of estimators $\hat{\boldsymbol{\theta}}_n = \hat{\mu}_n, \hat{\psi}_n, \hat{\xi}_n$ such that

$$\begin{aligned} \lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n \text{ is a MLE}\right\} &= 1, \\ \sqrt{k_n}\left(\frac{\hat{\mu}_n - b_{m_n}}{a_{m_n}}, \frac{\hat{\psi}_n}{a_{m_n}} - 1, \hat{\xi}_n - \xi_0\right) &\xrightarrow{d} \mathcal{N}\left(\lambda I_{\xi_0}^{-1}\mathbf{b}, I_{\xi_0}^{-1}\right). \end{aligned}$$

(b) If $\hat{\boldsymbol{\theta}}_n^i = (\hat{\mu}_n^i, \hat{\psi}_n^i, \hat{\xi}_n^i),\ i = 1,2$ are two sequences of estimators satisfying

$$\begin{aligned} \lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n^i \text{ is a MLE}\right\} &= 1, \\ \lim_{n\to\infty} \Pr\left\{\sqrt{k_n}\left(\frac{\hat{\mu}_n^i - b_{m_n}}{a_{m_n}}, \frac{\hat{\psi}_n^i}{a_{m_n}} - 1, \hat{\xi}_n^i - \xi_0\right) \in H_n\right\} &= 1, \end{aligned}$$

where H_n is a ball in $\mathbb{R}^3$ of center 0 and radius r_n, where $r_n = O(k_n^\delta),\ 0 < \delta < \min\left(\frac{1}{2}, \xi_0 + \frac{1}{2}\right)$ as $n \to \infty$, then

$$\lim_{n\to\infty} \Pr\left\{\hat{\boldsymbol{\theta}}_n^1 = \hat{\boldsymbol{\theta}}_n^2\right\} = 1.$$

2.5.1 Side Section 1: A heuristic on biased estimation

Suppose we have a sequence of experiments indexed by n, where in the nth experiment there are k_n observations $X_1, \ldots, X_{k_n}$ whose true joint density is g_n, but for reasons of convenience or because we don't know how to exactly calculate g_n, we replace g_n by a known joint density f_n indexed by a parameter vector $\boldsymbol{\theta}_n$. The examples of interest to us include the X_i's being either block maxima or exceedances over a threshold and their density f_n being approximated by a GEV or GPD density. We will always want $f_n - g_n \to 0$ under some suitable metric (e.g. total variation norm or Hellinger distance) but we won't worry about precise modes of convergence for the moment — that can come later.

Suppose we estimate $\boldsymbol{\theta}$ by defining a set of equations

$$\sum_{i=1}^{k_n} \mathbf{T}(X_i; \boldsymbol{\theta}) = 0$$

where $\mathbf{T}(X_i;\boldsymbol{\theta})$ is a vector of the same length as $\boldsymbol{\theta}$ that form a set of *unbiased estimating equations* in the sense that

$$\mathrm{E}\{\mathbf{T}(X_i;\boldsymbol{\theta})\} \quad = \quad 0 \text{ when } X_i \sim f_n(\cdot\,;\,\boldsymbol{\theta}).$$

The classical case is when $\mathbf{T}$ is the vector of first-order derivatives of the log likelihood but we are writing the formula in this alternative format to allow for other possible estimators (in particular, in the case of extreme value theory, probability weighted moments estimators or PWMs, which are a popular alternative to maximum likelihood estimation).

We also define a matrix $W(X_i)$ with entries $w_{rs}(X_i) = \frac{\partial T_r(X_i)}{\partial \theta_s}$ where T_r is the rth component of T and θ_s is the sth component of $\boldsymbol{\theta}$. In standard maximum likelihood theory, W is the hessian matrix of the log likelihood function (for a single observation), also known as the observed information matrix, and the expectation of W when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is I_0, the Fisher information matrix assuming the model f_n is correct with parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Assuming suitable regularity conditions,

$$\begin{aligned} \mathbf{0} &= \sum_{i=1}^{k_n} \mathbf{T}(X_i;\hat{\boldsymbol{\theta}}_n) \\ &\approx \sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0) + W(X_i;\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned}$$

and hence

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \quad \approx \quad -\left\{\sum_{i=1}^{k_n} W(X_i;\boldsymbol{\theta}_0)\right\}^{-1} \left\{\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)\right\}. \tag{2.47}$$

If we assume

(i) The mean of $W(X_i;\boldsymbol{\theta}_0)$ is J_0 for each i,

(ii) The covariance matrix of $\mathbf{T}(X_i;\boldsymbol{\theta}_0)$ is C_0 for each i,

and assume f_n is the true density, we will have

$$\sqrt{k_n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \quad \xrightarrow{d} \quad \mathcal{N}(\mathbf{0}, J_0^{-1} C_0 J_0^{-1}). \tag{2.48}$$

Formula (2.48) is widely known as the *information sandwich formula*. When estimation is by maximum likelihood, J_0 and C_0 both reduce to I_0, the Fisher information matrix, and (2.48) is the standard asymptotic distribution for maximum likelihood estimators.

Now, however, suppose the true density is g_n rather than f_n. *Typically*, the following is true: the covariance matrix of $\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ and the mean of $\sum_{i=1}^{k_n} W(X_i;\boldsymbol{\theta}_0)$ are still asymptotic to $k_n C_0(\boldsymbol{\theta})$ and $k_n J_0(\boldsymbol{\theta})$ respectively, but the mean of $\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ is non-zero. To be precise the mean is $\mathbf{b}_n$. In that case, the CLT for

$\sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0)$ takes the form

$$k_n{}^{-1/2} \sum_{i=1}^{k_n} \mathbf{T}(X_i;\boldsymbol{\theta}_0) \quad \sim \quad \mathcal{N}[k_n{}^{-1/2}\mathbf{b}_n, C_0(\boldsymbol{\theta}_0)](1+o_p(1))$$

and the final result for $\hat{\boldsymbol{\theta}}_n$ becomes

$$\sqrt{k_n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + k_n{}^{-1/2} J_0^{-1}\mathbf{b}_n \quad \xrightarrow{d} \quad \mathcal{N}[0, J_0^{-1}C_0J_0^{-1}]. \tag{2.49}$$

Note that there are different special cases of this result depending on the asymptotic behavior of $k_n{}^{-1/2}\mathbf{b}_n$. If $k_n{}^{-1/2}\mathbf{b}_n \to 0$ then the asymptotic bias of $\hat{\boldsymbol{\theta}}_n$ is negligible compared with its statistical variability as represented by the Fisher information matrix. In effect, this means we can ignore the discrepancy between f_n and g_n. Conversely, if $n^{-1/2}\mathbf{b}_n \to \infty$ in at least one component, the bias dominates the variance, which has the practical interpretation that we can't really use the standard results in this case. However if $k_n{}^{-1/2}\mathbf{b}_n \to \mathbf{c}$ for some vector $\mathbf{c}$ whose components are finite and not all zero, we can rewrite the result (2.49) as

$$\sqrt{k_n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad \xrightarrow{d} \quad \mathcal{N}[-J_0^{-1}\mathbf{c}, J_0^{-1}C_0J_0^{-1}]. \tag{2.50}$$

This is a true case of "bias-variance tradeoff" which can be the basis for various decision processes, such as the choice of a threshold in a peaks over threshold analysis (the ultimate objective of [234]).

2.5.2 Side section 2: Asymptotics of the Hill-Weissman Estimator

In this section we consider the special case of extreme value theory based on the Type I or Fréchet limit. Gnedenko [92] showed that a limit of the form

$$F^n(a_n x) \quad \to \quad \Phi_\alpha(x) \;=\; \exp\left(-cx^{-\alpha}\right),\; x \geq 0,\; \alpha > 0,\; c > 0, \tag{2.51}$$

holds if $1-F(x)$ is regularly varying with index α, and in that case a_n may without loss of generality be taken as the solution of $F(a_n) = 1 - 1/n$, and $c = 1$. Note that in this case, there is no location parameter to the distribution ($b_n = 0$), but for statistical purposes, it makes sense to retain c as well as α as an unknown parameter.

In this case, Weissman's representation [265, 267] for the asymptotic joint distribution of the k largest order statistics $m_1 \geq m_2 \geq \ldots m_k$ reduces to

$$L(\alpha, c \mid m_1, ..., m_k) \quad = \quad \prod_{i=1}^{k} \left(c\alpha m_i^{-\alpha-1}\right) \cdot \exp\left(-cm_k^{-\alpha}\right). \tag{2.52}$$

where the notation is intended to indicate that we are thinking of (2.52) as a likelihood function for the parameters α and c. The dependence on $m_1, ..., m_k$ will be omitted in many of the formulas. Taking logarithms, we want to minimize

$$\ell(\alpha, c) \;=\; -\log L(\alpha, c) \;=\; -k\log\alpha - k\log c + (\alpha-1)\sum_{i=1}^{k} \log m_i + cm_k^{-\alpha}.$$

It is quickly established that this expression is minimized when $\alpha = \hat{\alpha},\ c = \hat{c}$ where

$$\hat{\alpha} \ = \ \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{m_i}{m_k} \right)^{-1}, \ \hat{c} \ = \ k m_k^{\hat{\alpha}}. \tag{2.53}$$

Note, in particular, the simple direct formula for the estimator of α. The derivation is the same as that in [265], but that paper did it for the equivalent case where the limit distribution is Gumbel (the Fréchet model is turned into the Gumbel model by taking logarithms of the observations).

An alternative, even simpler, derivation of an equivalent result was given by Hill [121]. Hill assumed, in effect, that the relationship $1-F(x) = cx^{-\alpha}$ is exact for $x \geq u$, for some known threshold u, but that $F(x)$ is unspecified for $x < u$. If data $X_1, ..., X_n$ are ordered so that $X_1 \geq X_2 \geq \ldots X_k > u \geq X_{k+1} \geq \ldots X_n$ then the likelihood function is

$$L(\alpha, c \mid X_1, ..., X_n) \quad = \quad \prod_{i=1}^{k} \left(\alpha c X_i^{-\alpha-1} \right) \cdot \left(1 - c u^{-\alpha} \right)^{n-k}$$

Taking logarithms and minimizing with respect to first c and then α leads to

$$\hat{\alpha} \ = \ \left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{X_i}{u} \right)^{-1}, \ \hat{c} \ = \ \frac{k}{n} u^{\hat{\alpha}}. \tag{2.54}$$

Note, in particular, the similarity of the two estimators of α: in effect, the role of the threshold u in (2.54) is replaced by the kth largest order statistic in (2.53). (The different estimators of c arise because of different definitions: (2.53) uses the limit distribution for sample maxima whereas (2.54) assumes the same functional form directly for the individual observations. The two definitions differ by a factor of n, which is reflected in the estimates.)

The estimator $\hat{\alpha}$ in (2.54) is widely known as *Hill's estimator* but in the present section, to emphasize the close similarity with Weissman's [267] result, we shall call it the *Hill-Weissman estimator*.

In order to develop some asymptotics for this estimator, we assume an expansion of the form

$$1 - F(x) \quad = \quad c x^{-\alpha} \left\{ 1 + d x^{-\beta} + o(x^{-\beta}) \right\}, \ x \to \infty. \tag{2.55}$$

In general, the assumption (2.55) may be replaced by an assumption of *second-order regular variation* which allows the terms with $x^{-\alpha}$ and $x^{-\beta}$ to be replaced by general regularly varying functions; see in particular [94] for a survey of this theory and its applications (including the present one). This, in turn, is a special case of the general sceond-order regular variation theory of [110], For the present discussion, we mke the simpler assumption (2.55) which is sufficient for most practical applications, and easier to manipulate.

Our focus will be on the condition distribution of X given $X > u$, for some high

threshold u. Let $Y_u = X/u$. Then the conditional probability $P\{Y_u > y \mid Y_u > 1\}$ is represented as

$$\frac{1-F(uy)}{1-F(u)} \quad = \quad y^{-\alpha}\left\{1+du^{-\beta}(y^{-\beta}-1)+o(u^{-\beta})\right\}$$

so, assuming it is valid to differentiate term by term, we calculate the density as

$$f_{Y_u}(y) \quad = \quad \alpha y^{-\alpha-1}+du^{-\beta}\left\{(\alpha+\beta)y^{-\alpha-\beta-1}-\alpha y^{-\alpha-1}\right\}+o(u^{-\beta}).$$

We note integrals of the form

$$\int_1^\infty (\log y)^k y^{-\alpha-1}dy \quad = \quad \alpha^{-k-1}k!$$

where we shall mainly be interested in the cases $k=1$ and 2 but for non-integer k the same formula holds with $k!$ replaced by $\Gamma(k+1)$. We therefore deduce

$$\mathrm{E}(\log Y_u)^k \quad = \quad \alpha^{-k}k!+du^{-\beta}k!\left\{(\alpha+\beta)^{-k}-\alpha^{-k}\right\}=o(u^{-\beta}). \qquad (2.56)$$

Now let's consider the bias and variance of $\frac{1}{\hat{\alpha}} = \frac{1}{k}\sum_{i=1}^k \log\frac{X_i}{u}$ as an estimator of $\frac{1}{\alpha}$, where k is the number of exceedances of u. Since $\mathrm{E}(\log Y_u) = \frac{1}{\alpha} - du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)} + o(u^{-\beta})$, we deduce

$$\text{Bias of } \frac{1}{\hat{\alpha}} \quad \approx \quad -du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)}.$$

However, we also have from the $k=1$ and $k=2$ cases of (2.56) that $\mathrm{Var}(\log Y_u) \to \frac{1}{\alpha^2}$ as $u\to\infty$ and hence the variance of $\frac{1}{\alpha}$ is asymptotically $\frac{1}{k\alpha^2}$. However if the whole sample is of size n, and k is the random number of exceedances of u, we have $k \sim ncu^{-\alpha}$. Therefore, in large samples we have

$$\text{Variance of } \frac{1}{\hat{\alpha}} \quad \approx \quad \frac{1}{\alpha^2 ncu^{-\alpha}}.$$

Combining the espressions for bias and variance, and writing mean squared error (MSE) for the sum of squared bias and variance, we deduce

$$\text{MSE of } \frac{1}{\hat{\alpha}} \quad \approx \quad \frac{Au^\alpha}{n}+B^2u^{-2\beta}$$

where $A = \frac{1}{\alpha^2 c}$ and $B = \frac{d\beta}{\alpha(\alpha+\beta)}$.

This asymptotic MSE is minimized with

$$u \quad = \quad \left(\frac{2\beta B^2 n}{\alpha A}\right)^{1/(\alpha+2\beta)}$$

which in turn leads to an asymptotic MSE of

$$MSE \quad = \quad \frac{B^2(\alpha+2\beta)}{\alpha}\left(\frac{2\beta B^2 n}{\alpha A}\right)^{-2\beta/(\alpha+2\beta)}.$$

The most important consequence of this is that the MSE is of $O\left(n^{-2\beta/(\alpha+2\beta)}\right)$ as $n \to \infty$, which could be arbitrarily slow for very small β but is of $O(n^{-1})$ as $\beta \to \infty$ — this makes sense, because in thaty limit the $cx^{-\alpha}$ result is exact and we are back in the original case considered by Hill.

2.5.2.1 Extension to the GPD

The above calculation was relatively straightforward because of the explicit closed form of the estimator. In most cases of interest (for example, estimating the two-parameter GPD or the three-parameter GEV distribution), there is no closed form estimator and the MLE is obtained by solving the likelihood equations. In such case, we may in principle proceed as follows. Suppose the negative log likelihood function based on n observations is $\ell_n(\theta)$ for some multidimensional parameter θ whose true value we shall write θ_0. Also write $\hat{\theta}_n$ for the MLE. The Taylor expansion

$$\nabla\ell_n(\hat{\theta}_n) - \nabla\ell_n(\theta_0) \quad \approx \quad \left(\hat{\theta}_n - \theta_0\right)^T \nabla^2\ell_n(\theta_0)$$

leads to the approximation

$$\hat{\theta}_n - \theta_0 \quad \approx \quad -\left(\nabla^2\ell_n(\theta_0)\right)^{-1}\nabla\ell_n(\theta_0).$$

Now suppose that as $n \to \infty$, $n^{-1}\nabla^2\ell_n(\theta_0) \xrightarrow{p} J$ (the Fisher information matrix) and $n^{-1}\nabla\ell_n(\theta_0) \xrightarrow{p} \mathbf{b}$ (bias due to model misspecification; if the model is correctly specified, $\mathbf{b} = \mathbf{0}$). Then for $\hat{\theta}_n$ we have, for large n,

$$\text{Bias } \approx J^{-1}\mathbf{b}, \text{ Covariance Matrix } \approx n^{-1}J^{-1}. \tag{2.57}$$

Now let's apply this to the case of the GPD, again under the assumption that the true distribution satisfies (2.55). Note that in the case where $1-F(x) = cx^{-\alpha}$ is exact, we have

$$\frac{1-F(u+y)}{1-F(u)} \quad = \quad \left(1+\frac{y}{u}\right)^{-\alpha} \quad = \quad \left(1+\xi\frac{y}{\sigma}\right)^{-1/\xi}$$

so the two forms are identical if $\sigma = \frac{u}{\alpha}$, $\xi = \frac{1}{\alpha}$. From now on, we treat these as the "true" GPD parameter values in this case.

In this model, the Fisher information matrix [234] is

$$J \quad = \quad \begin{pmatrix} \frac{1}{\sigma^2(1+2\xi)} & \frac{1}{\sigma(1+\xi)(1+2\xi)} \\ \frac{1}{\sigma(1+\xi)(1+2\xi)} & \frac{2}{(1+\xi)(1+2\xi)} \end{pmatrix}$$

provided $1+2\xi > 0$, and hence

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

Now let's compute the **b** term in (2.57). The log likelihood for a single observation is

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

Hence,

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

To calculate **b**, we need to find expressions for the expected values of these terms.

To recast in the notation of Section 2.5.2, we first make the substitutions $\sigma = \frac{u}{\alpha}$, $\xi = \frac{1}{\alpha}$, and also that if y denotes the excess over the threshold u, then $y = u(Y_u - 1)$ and so $1+\xi\frac{y}{\sigma} = Y_u$. Also, by the same reasoning as led to (2.56)

$$\mathrm{E}\left(Y_u^{-1}\right) = \frac{\alpha}{\alpha+1} + du^{-\beta}\cdot\frac{\beta}{(\alpha+1)(\alpha+\beta+1)} + o(u^{-\beta}).$$

We now calculate the expectations of $\sigma\frac{\partial\ell}{\partial\sigma}$ and $\frac{\partial\ell}{\partial\xi}$, respectively, to be

$$-\alpha+(\alpha+1)\left\{\frac{\alpha}{\alpha+1}+du^{-\beta}\cdot\frac{\beta}{(\alpha+1)(\alpha+\beta+1)}+o(u^{-\beta})\right\} = du^{-\beta}\cdot\frac{\beta}{\alpha+\beta+1}+o(u^{-\beta})$$

and

$$\begin{aligned} &-\alpha^2\left\{\frac{1}{\alpha}-du^{-\beta}\frac{\beta}{\alpha(\alpha+\beta)}\right\}+\alpha(\alpha+1)\left\{\frac{1}{\alpha+1}-du^{-\beta}\frac{\beta}{(\alpha+1)(\alpha+\beta+1)}\right\}+o(u^{-\beta}) \\ = \;& du^{-\beta}\cdot\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}+o(u^{-\beta}). \end{aligned}$$

Therefore, we conclude

$$\begin{aligned} \mathbf{b} &\sim du^{-\beta}\begin{pmatrix} \frac{1}{\sigma}\frac{\beta}{\alpha+\beta+1} \\ \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} \end{pmatrix}, \\ J^{-1}\mathbf{b} &\sim du^{-\beta}\frac{(\alpha+1)\beta}{\alpha(\alpha+\beta)(\alpha+\beta+1)}\begin{pmatrix} \sigma(\alpha+2\beta) \\ 1-\beta \end{pmatrix}. \end{aligned}$$

Focussing on the second entries in these vectors, we deduce that $\hat{\xi}$ has asymptotic bias

$$du^{-\beta}\frac{(\alpha+1)\beta(1-\beta)}{\alpha(\alpha+\beta)(\alpha+\beta+1)}$$

and asymptotic variance (based on $k \approx ncu^{-\alpha}$ exceedances of the threshold

$$\frac{1}{k}\left(\frac{\alpha+1}{\alpha}\right)^2 \sim \frac{(\alpha+1)^2}{\alpha^2 ncu^{-\alpha}}.$$

2.5.2.2 Comparisons with the Hill-Weissman Estimator

For the Hill-Weissman estimator, we deduced that the bias was asymptotically $Bu^{-\beta}$, variance Au^{α}/n, with $B = -d\beta/(\alpha(\alpha+\beta))$, $A = 1/(\alpha^2 c)$.

For the GPD estimator, we get asymptotic bias $B'u^{-\beta}$, asymptotic variance $A'u^{\alpha}/n$, where $B' = d\beta(1-\beta)(\alpha+1)/(\alpha(\alpha+\beta)(\alpha+\beta+1))$.

The optimal MSE is proportional to

$$|B|^{2\alpha/(\alpha+2\beta)} A^{2\beta/(\alpha+2\beta)}$$

Therefore, the ratio of the optimal MSE for the GPD estimator to that of the Hill-Weissman estimator is

$$\left|\frac{B'}{B}\right|^{2\alpha/(\alpha+2\beta)} \left|\frac{A'}{A}\right|^{2\beta/(\alpha+2\beta)} = \left|\frac{(1-\beta)(\alpha+1)}{\alpha(\alpha+\beta)(\alpha+\beta+1)}\right|^{2\alpha/(\alpha+2\beta)} |\alpha+1|^{4\beta/(\alpha+2\beta)}$$

See Figure 5.1.

2.5.2.3 Background References

The Hill estimator was introduced in [121] and the Weissman estimator, in its original form, in [267]. Asymptotic properties of the Hill estimator were obtained by [116, 112, 94] Optimality of the derived rate of convergence was proved by [114], and an adaptive estimator to achieve the optimal threshold was given by [115]. Many variants on the method exists, for example, [46] used a kernel-weighted version. The comparison of the two estimators was first derived in [234]. Many other authors have contributed to the theory and a more complete bibliography will be given later.

2.5.3 Outline Derivation of Dombry-Ferreira result

Health warning: This is not the proof. For that, we refer to the original paper [60]. The intention here is to motivate the result, and to show how it follows logically from the asymptotic approximations we have been developing in this chapter.

First, let us assume that the relationship (2.44) is exact, i.e. the left and right hand sides are identical for every m. Since $M_{i,m}$ has the distribution function F^m, by the probability integral transformation we can write $F^m(M_{i,m}) = S$ where S is uniform on (0,1). In that case $-\frac{1}{\log F(M_{i,m})} = \frac{m}{-\log S}$. But $-\frac{1}{\log F(\cdot)}$ was defined to be the inverse of V, so $M_{i,m} = V\left(\frac{m}{-\log S}\right)$. We also define $b_m = V(m),\ a_m = a(m)$. If we assume (2.44 is exact, then

$$\frac{M_{i,m} - b_m}{a_m} = \frac{\left(-\frac{1}{\log S}\right)^{\xi_0} - 1}{\xi_0}.$$

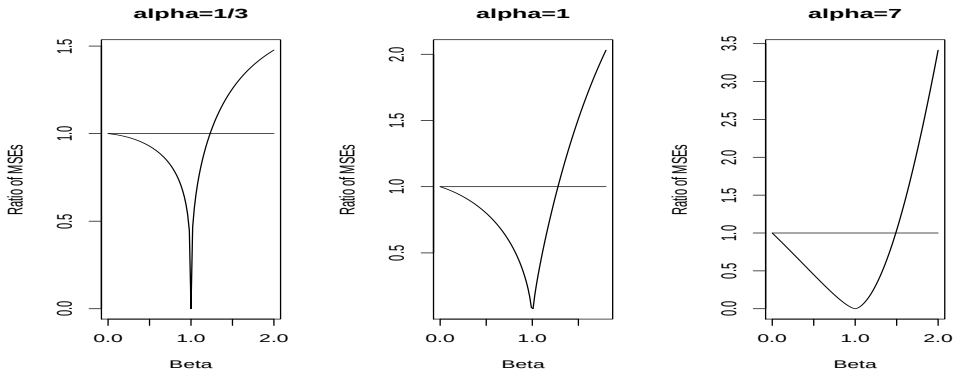

Figure 2.1 *Ratio of optimal mean squared error for the GPD estimator to that of the Hill-Weissman estimator, for a variety of values of* α *and* β.

But the right hand side has the GEV distribution:

$$\Pr\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\le y\right\} \;=\; \Pr\left\{S\le e^{-(1+\xi y)^{-1/\xi}}\right\} \;=\; e^{-(1+\xi y)^{-1/\xi}} \text{(provided } 1+\xi y>0).$$

Now, however, suppose (2.45) holds instead of (2.44) being exact. In that case, we can write

$$\frac{M_{i,m}-b_m}{a_m} \;=\; \frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}+A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log S}\right)+o_p(A(b_m)).$$

Suppose we want to find the expectation of $h\left(\frac{M_{i,m}-b_m}{a_m}\right)$, where h is some nonlinear continuously differentiable function. We proceed formally, assuming limiting operations are valid without rigorous proof. By Taylor expansion, we write

$$h\left(\frac{M_{i,m}-b_m}{a_m}\right) \;=\; h\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\right\}+A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log S}\right)h'\left\{\frac{\left(-\frac{1}{\log S}\right)^{\xi_0}-1}{\xi_0}\right\}+o_p(A(b_m)).$$

Taking expectations term by term

$$\mathrm{E}\left\{h\left(\frac{M_{i,m}-b_m}{a_m}\right)\right\} \;=\; \int_0^1 h\left\{\frac{\left(-\frac{1}{\log s}\right)^{\xi_0}-1}{\xi_0}\right\}ds+\int_0^1 A(b_m)H_{\xi_0,\rho}\left(\frac{1}{-\log s}\right)h'\left\{\frac{\left(-\frac{1}{\log s}\right)^{\xi_0}-1}{\xi_0}\right\}ds$$
$$+o_p(A(b_m)).$$

Now suppose the function h is any of $\frac{d\ell}{d\mu}$, $\frac{d\ell}{d\psi}$, $\frac{d\ell}{d\xi}$, where ℓ is given by (2.41). Because

h is a derivative of the log likelihood of the GEV model, $\int_0^1 h \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds = 0$ and we are left with

$$\mathrm{E}\left\{ h\left(\frac{M_{i,m} - b_m}{a_m} \right) \right\} \quad \sim \quad A(b_m) \int_0^1 H_{\xi_0,\rho}\left(\frac{1}{-\log s} \right) \frac{\partial h}{\partial x} \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds.$$

Representing $\ell(\mu, \psi, \xi; x) = \ell(\boldsymbol{\theta}, x)$ where $\boldsymbol{\theta} = (\ \theta_1 \quad \theta_2 \quad \theta_3$ and $\theta_1 = \mu,\ \theta_2 = \psi,\ \theta_3 = \xi$, we therefore have

$$\mathrm{E}\left\{ \frac{\partial \ell}{\partial \boldsymbol{\theta}} \left(\frac{M_{i,m} - b_m}{a_m} \right) \right\} \quad \sim \quad A(b_m) \int_0^1 H_{\xi_0,\rho}\left(\frac{1}{-\log s} \right) \frac{\partial^2 \ell}{\partial x \partial \boldsymbol{\theta}} \left\{ \frac{\left(-\frac{1}{\log s}\right)^{\xi_0} - 1}{\xi_0} \right\} ds. \tag{2.58}$$

The right hand side of (2.58) is $A(m_n)$ multiplied by $\mathbf{b}(\xi_0, \rho)$ in the notation of Dombry–Ferreira.

Equation (2.58) applies to just a single value of the likelihood function, whereas the formula (2.42) represents the sum of $k = k_n$ similar terms. In the notation of Section 2.5.1, we have $\mathbf{b}_n = k_n A(m_n)\mathbf{b}(\xi_0, \rho)$ and hence $k_n^{-1/2}\mathbf{b}_n = \sqrt{k_n} A(m_n)\mathbf{b}(\xi_0, \rho) \to \lambda \mathbf{b}(\xi_0, \rho)$. Since this case the method of estimation under the GEV model is maximum likelihood, in this case the matrices J_0 and C_0 of Section 2.5.1 are both I_0, the Fisher information matrix for the limiting GEV distribution. Thus, (2.50) implies

$$\sqrt{k_n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \quad \xrightarrow{d} \quad \mathcal{N}\left\{ \lambda I_0^{-1} \mathbf{b}(\xi_0, \rho), I_0^{-1} \right\}$$

The Dombry-Ferreira result differs from this because it assumes the GEV maximum likelihood estimation procedure is applied directly to the block maxima $M_{i,m}$, rather than the normalized maxima $\frac{M_{i,m} - b_m}{a_m}$ as we have written here. Nevertheless, this argument should serve to motivate their result and to define a general context to derive similar results under different variations of the basic model and estimation procedure.

2.6 Other topics to be added

2.6.1 Method of probability weighted moments

An alternative to the maximum likelihood method that achieved popularity after a famous paper of Hosking, Wallis and Wood [124], but theoretically do not perform as well as maximum likelihood estimators [58, 76].

2.6.2 Corresponding results for threshold estimators

Cite paper of Smith [234]; show how results may be reinterpreted in terms of the de Haan-Stadtmüller representation