CHAPTER 4

SEMI-PARAMETRIC APPROACH TO MODELLING EXTREMES

4.1 Introduction

Even from the previous chapter, we have raised some doubts about the applicability of the pure parametric approach to modelling extremes. Indeed, the assumption of an exact generalized extreme-value d.f. is a strict one. Moreover, in the majority of practical cases where the extreme-value analysis is called for, the main interest is not to fully describe the data at the expense of a very strict and unrealistic assumption. Instead, our main concern is to describe the 'behaviour' of extreme values (either maxima or minima). For example, in hydrology scientists are only concerned to describe the behaviour of the height of the largest waves, so as to build accordingly high the sea-dykes. The same holds for engineers constructing windmills (they are interested in the maximum speed of the wind so as to make the windmills strong enough to resist to them). Actuaries are also interested in the behaviour of largest claims since, if these are excessively large, they can even destroy a portfolio. Teletraffic is another field where extreme-value analysis has recently gained ground. Indeed, in order to evaluate and design properly the capacity of a system of telecommunications, it is essential, for the engineers, to be aware of the size of largest demands that can be required and the corresponding probabilities of occurrence. The previously described situations are only a few characteristic examples of the wider class of fields where extreme-value analysis is asked for.

One could argue that the method of block maxima (belonging to the previous parametric approach) is an appropriate method to answer these questions that practitioners have. Still, a problem with this method is that, sometimes, the grouping of data into epochs is somewhat arbitrary. Moreover, by using only the block maxima, we may loose important information, since some blocks may contain several among the largest observations, while other blocks may contain none. Moreover, in the case that we have few data, block maxima cannot be actually implemented.

The present approach is often referred to as 'Maximum Domain of Attraction Approach' (Embrechts et al., 1997), or Non-Parametric. We prefer the term 'semi-parametric' since this term reflects the fact that we make only partly assumptions about the unknown d.f. F.

The setting of the problems we are facing here, can be simply described as follows:

> "We have a data-set (i.i.d. sample) $X_1, X_2, ..., X_n$ (let $X_{1:n} \geq X_{2:n} \geq ... \geq X_{n:n}$ be the corresponding descending order statistics) from an unknown d.f. F and we are interested in the tail (upper or lower) behaviour of the d.f. F, i.e. in the behaviour of the extreme values"

So, essentially, we are interested in the distribution of the maximum (or minimum) value. Here is the point where extreme-value theory gets involved. As we have already discussed in chapter 2, the limiting d.f. of the (normalized) maximum value (if that exists) is the generalized extreme-value d.f. $H_\theta = H_{\gamma;\mu,\sigma}$. So, without making any assumptions about the unknown d.f. F (apart from some continuity conditions which ensure the existence of the limiting d.f.), extreme-value theory provides us with a fairly sufficient tool for describing the behaviour of extremes of the distribution that the data in hand stem from (i.e. a limiting d.f. for the maximum). The only thing that remains to be resolved is the estimation of the parameters of the generalized extreme-value d.f. $\theta = (\gamma, \mu, \sigma)$. Of these parameters, the *shape parameter* γ (also called *tail index* or *extreme-value index*) is the one that attracts most of the attention, since this is the parameter that determines, in general terms, the behaviour of extremes. More particularly,

- if $\gamma > 0$, the limiting d.f. of the maxima is of Fréchet type, i.e. the maximum value is unbounded to the right,
- if $\gamma = 0$, the limiting d.f. is of Gumbel type and is also unbounded to the right but tends faster to 0, and
- if $\gamma < 0$, the limiting d.f. is bounded to the right (Weibull type), meaning that the extreme values cannot increase indefinitely.

Furthermore, since the model $H_{\gamma;\mu,\sigma}$ includes location and scale parameters, it follows that the distribution function of the maximum $X_{1:n}$ itself, for large enough n, can be approximated by a member of the generalized extreme-value family. With the same reasoning, properly normalized maxima would follow the standard d.f. $H_{\gamma;0,1}$, which also stresses the dominance of the parameter γ among the three parameters of the generalized extreme-value d.f.

A natural question that follows is how to estimate the parameters θ. Extreme-value theory claims that these are the parameters of the generalized extreme-value d.f. that the maximum value follows asymptotically. Of course, in reality, we only have a finite sample and, in any case, we cannot use only the largest observation for inference. So, the procedure followed in practice is that we assume that the asymptotic approximation is achieved for the largest k observations (where k is large but not as large as the sample size n), which we subsequently use for the estimation of the parameter θ that characterizes the tail behaviour of our data. Notice, however, that the choice of k is not an easy task. On the contrary, it is a very controversial issue. Many authors have suggested several solutions but none of them has been universally adopted.

Summarizing, in the semi-parametric approach :

We have an i.i.d. sample $X_1, X_2, ..., X_n$ from an unknown d.f. F

We are interested in the tail behaviour of F

We take into account only the k largest observations

We assume, according to extreme-value theory, that these k largest observations come from a generalized extreme-value d.f. $H_{\gamma;\mu,\sigma}$

We estimate the parameters of generalized extreme-value d.f. based on these k largest observations

We proceed with any other inference of interest (e.g. large quantile estimation)

In the remaining of this chapter, we give the most prominent answers to the above question of parameter estimation. Of course, it would be unrealistic to claim that we can cover the whole literature on these issues, since the literature is indeed vast. We describe the most well-known proposals, ranging from the first contributions, of 1975, in the area to very recent modifications and new developments. We mainly concentrate to the estimation of the shape parameter γ due to its (already stressed) importance.

4.2 Pickands Estimator

4.2.1 Derivation

The first to suggest a so-called tail-index estimator (essentially an estimator of the parameter $\gamma \in \Re$ of generalized extreme-value d.f.) was Pickands (Pickands, 1975). Suppose that we have a sample of i.i.d r.v.'s $X_1, X_2, ..., X_n$ from an unknown *continuous* d.f. F. According to extreme-value theory, the normalized maximum of such a sample follows asymptotically a generalized extreme-value d.f. $H_{\gamma;\mu,\sigma}$, i.e. $F \in MDA(H_{\gamma;\mu,\sigma})$.

Pickands proved that the above statement is equivalent to

$$\lim_{u \to x_F} \inf_{0<\sigma<\infty} \sup_{0 \le x < \infty} \left| \frac{1-F(u+x)}{1-F(u)} - \exp\left\{ -\int_0^{x/\sigma} \frac{1}{(1+\gamma t)_+} dt \right\} \right| = 0 \qquad (4.1).$$

For any u and x, $\frac{1-F(u+x)}{1-F(u)}$ is the conditional probability that an observation is greater than u+x given that it is greater than u. This relation essentially means that, if u is large, the conditional distribution of X given that $X \ge u$ is very nearly of the form

$$G(x) = 1 - \exp\left\{ -\int_0^{x/\sigma} \frac{1}{(1+\gamma t)_+} dt \right\} \qquad (4.2),$$

for some σ, γ, $0<\sigma<\infty$, $-\infty<\gamma<\infty$.

Now, let M be an integer much smaller than n. Intuitively, the 4M (=k) largest observations contain information about the upper tail of F. By treating the values $\{X_{m:n} - X_{4M:n}\}$, m=1, ..., 4M-1 as (descending) order statistics from a sample of size 4M-1 from (4.2) and adopting a percentile estimation method (in particular by equating the theoretical 50- and 75- percentiles with their empirical counterparts) we end-up with the well-known Pickands estimator for the parameter γ :

$$\boxed{\hat{\gamma}_P = \frac{1}{\ln 2} \ln\left(\frac{X_{M:n} - X_{2M:n}}{X_{2M:n} - X_{4M:n}} \right)}.$$

Moreover, Pickands method provides us with an estimate of the scale parameter σ. The location parameter μ is not directly estimated (but as we will see in the sequel, this is not even necessary for making inference).

A more formal justification of Pickands estimator is provided by Embrechts et al. (1997). According to them, the basic idea behind the estimator is to find a condition equivalent to $F \in MDA(H_{\gamma;\mu,\sigma})$, involving the parameter γ in a straightforward way. The key-point is the characterization II property of $MDA(H_{\gamma;\mu,\sigma})$, as mentioned in chapter 2.

Another particular characteristic of Pickands estimator is the fact that the largest observation is not explicitly used in the estimation. One can argue that this makes sense since the largest observation may add too much uncertainty.

4.2.2 Choice of k

The following conditions on k (4M or equivalently on M) arise very naturally, not only for Pickands estimator, but for all estimators of γ that are based on the k upper order statistics. So, in order the Pickands estimator to have some 'good' properties, k, which can also be written as k(n) to stress its dependence on the sample size n, should satisfy the following conditions:

- $\lim_{n\to\infty} k(n) = \infty$, so that we actually take advantage of an increasing sample size, but
- $\lim_{n\to\infty} \frac{k(n)}{n} = 0$, i.e. the approximation at the population upper tail will not improve indefinitely.

In simple words, "k should be large enough, but not too large". This kind of development of k will often be referred to as '*moderate increase*'. For practical purposes, Pickands suggested the following criterion for the appropriate choice of k (in this and the following formulae k refers to M) :

$$k = \arg \min_{1 \le l \le [n/4]} d_l ,$$

where $d_l = \sup_{0 \le x < \infty} \left| \hat{F}_l(x) - \hat{G}_l(x) \right|$,

$\hat{F}_l$ is the empirical d.f. based on the $\{X_{m:n} - X_{4l:n}\}$, m=1, ..., 4l-1 order statistics, and

$\hat{G}_l$ is of the form (4.2) where γ and σ are estimated based on the 4l upper order statistics (M=l).

A more handy way that is used in practice for the choice of k (M) is based on the Pickands-plot (Embrechts, et al., 1997). For each value of k=1, ..., [n/4], we calculate the Pickands estimator and plot it against k. The range of k's that correspond to a 'plateau' in the plot (i.e. those values of k that give very similar values of $\hat{\gamma}_P$) are considered more appropriate, that is, *we choose* $\hat{\gamma}_P$ *from such a k-region where the plot is roughly horizontal*.

4.2.3 Properties of Pickands estimator

The properties of Pickands estimator were mainly explored by Dekkers and de Haan (1989). More precisely, they proved the followings :

- ***Weak Consistency***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $k(n) \to \infty$, $\frac{k(n)}{n} \to 0$,

then $\hat{\gamma}_P \to \gamma$, in probability, as $n \to \infty$.

This result was also shown by Pickands (1975).

- ***Strong Consistency***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $\frac{k(n)}{n} \to 0$, $\frac{k(n)}{\ln\ln(n)} \to \infty$,

then $\hat{\gamma}_P \to \gamma$, almost surely, as $n \to \infty$.

- ***Asymptotic Normality***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$,

$\lim_{t\to\infty} \frac{(tx)^{1-\gamma}U'(tx) - t^{1-\gamma}U'(t)}{a(t)} = \pm \ln x$ for x>0 (with either choice of sign),

where $U \equiv \left(1/(1-F)\right)^{\leftarrow}$ with positive derivative and a is a positive function,

$k(n) \to \infty$, $k(n) = o\left(n / g^{\leftarrow}(n)\right)$, where $g(t) \equiv t^{3-2\gamma}\left\{U'(t)/a(t)\right\}^2$

then $\sqrt{k}\left(\hat{\gamma}_P - \gamma\right) \to N(0, v(\gamma))$, in distribution as $n \to \infty$,

where $v(\gamma) = \gamma^2\left(2^{2\gamma+1}+1\right) / \left\{2\left(2^{\gamma}-1\right)\ln 2\right\}^2$.

Note that, while consistency depends only on the behaviour of k, asymptotic normality of Pickands estimator requires more delicate conditions (2^{nd} order conditions) on the underlying d.f. F. These conditions are generally difficult to verify in practice for an unknown d.f. Still, Dekkers and de Haan (1989) have shown that these conditions hold for various known and widely-used d.f.'s (normal, gamma, generalized extreme-value, exponential, uniform, cauchy) and found the precise rate of growth of k(n) such that $\hat{\gamma}_P$ is unbiased and asymptotically normal. In any case, based on the asymptotic normality of $\hat{\gamma}_P$, one can derive confidence intervals (and probably hypothesis testing) not only for the parameter γ, but also for higher quantile estimates as well as end-point estimates (for γ<0), since these estimates are essentially functions of $\hat{\gamma}_P$.

4.2.4 Quantile estimation

Irrespectively of the method used to estimate parameter γ (or μ, σ) the (large) quantiles can be estimated by direct substitution of the parameter estimates to the quantile function (generalized inverse function) of the generalized extreme-value d.f. Hence, for large p

$$\hat{x}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\gamma}}\left[1 - \left(-\ln p\right)^{-\hat{\gamma}}\right] \quad (4.3).$$

Note that the same formula was also used for quantile estimation in the parametric approach (chapter 3). The difference between these two formulae lies in the fact that in the parametric case (4.3) hold for every $p \in (0,1)$, while now it holds only for large p ($p \to 1$).

Moreover, apart from the general expression (4.3), there are several other quantile estimation formulae, depending on the estimator of γ. The quantile estimates (for large p) that are based on Pickands estimator of γ, have been derived by Dekkers and de Haan (1989) and are

$$\hat{Q}(p) = \hat{x}_{p,n} = \left[(n(1-p)/M)^{-\hat{\gamma}_P} - 1\right]\left(1 - 2^{-\hat{\gamma}_P}\right)^{-1}\left(X_{M:n} - X_{2M:n}\right) + X_{M:n}.$$

Notice that only $\hat{\gamma}_P$ is involved in the estimation of the quantile. Neither the scale σ nor the location parameter μ are required. It is worth mentioning the metaphor that the authors use for the above formula : '*In the absence of more observations (that would have allowed us to simply use the inverse empirical distribution function), one uses observed spacings (modulo a multiplicative constant) for the missing spacings, like a surgeon who uses a piece of skin from elsewhere to cover a wound'*.

In the case $\gamma < 0$, the limiting d.f. of maxima is bounded to the right. So, in such a case, the main interest is focused on the estimation of the right endpoint of the distribution instead of large quantiles. In the context of Pickands estimation of γ, Dekkers and de Haan (1989) proposed the following estimator for the right endpoint of the limiting distribution (when $\gamma < 0$)

$$\hat{x}_F = \left(2 - 1^{-\hat{\gamma}_P}\right)^{-1}\left(X_{M:n} - X_{2M:n}\right) + X_{M:n}.$$

The authors that proposed these estimators have also derived asymptotic distributions for these quantities, that can be used for the construction of asymptotic confidence intervals.

4.2.5 Modifications-Developments on Pickands Estimator

As we saw previously, in order to estimate consistently γ, we have to take a 'moderately increasing' sequence of k order statistics. Still this cannot be easily achieved in practice. When the number of upper order statistics used is too small, the variance of the estimator will be large. On the other hand, if k is too large, then bias is introduced in the estimation (since we are using observations which do not actually converge to the hypothesized limiting d.f.). This problem is not characteristic only of Pickands estimator, but of all estimators of the semi-parametric approach that utilize only the k largest observations.

So, the choice of k is an issue of balance between bias and variance. In the literature many efforts have been made to find the optimal k, i.e. that k which reconciles bias and variance (for references on this issue see Peng, 1998). From another standpoint, there have been efforts to make adjustments for bias, so that even when a large number k of upper order statistics is used, the resulting estimator to be unbiased. Unfortunately, these adjustments require some assumptions about the 2nd order behaviour of the unknown d.f. F, which usually are not even verifiable. Two of these attempts are presented in the sequel.

1st Modification

If we assume that there exist functions $a(t) > 0$, $A(t) \to 0$ (with constant sign near infinity), and constant $\rho < 0$ such that

$$\lim_{t \to \infty} \frac{\left(U(tx) - U(t)\right)/a(t) - \left(x^{\gamma} - 1\right)/\gamma}{A(t)} = \frac{1}{\rho}\left[\frac{x^{\gamma+\rho} - 1}{\gamma + \rho} - \frac{x^{\gamma} - 1}{\gamma}\right]$$

then $\tilde{\gamma}_P = \hat{\gamma}_P(k) - \dfrac{\hat{\gamma}_P(k) - \hat{\gamma}_P(k/4)}{1 - 4^{\rho_n}}$ (for $k \to \infty$, $k/n \to 0$, and $\sqrt{k}\, A(n/k) \to \lambda$

is an unbiased estimator of the shape parameter γ,

where $\rho_n = \left(\ln 2\right)^{-1} \ln \dfrac{\hat{\gamma}_P\left(n/(2\ln n)\right) - \hat{\gamma}_P\left(n/(4\ln n)\right)}{\hat{\gamma}_P\left(n/\ln n\right) - \hat{\gamma}_P\left(n/(2\ln n)\right)}$, and

$\hat{\gamma}_P(c)$ is the Pickands estimator based on the 4M=c upper order statistics (for proof see Peng, 1998).

Unfortunately, though this estimator is unbiased (even if we use a large number of upper order statistics), it has higher variance than the original Pickands estimator.

This problem of large variance is addressed by another class of Pickands-type estimators. Yun (2000) proposes a class of Pickands-type estimators which stem fronm the idea of reconstructing the original Pickands estimator itself in such a way that the upper-order statistics lying in the interval $\left[X_{n-4m+1:n}, X_{n-m+1:n}\right]$ are re-allocated efficiently. He proves strong consistency and asymptotic normality of the new estimator as well as its relative efficiency over standard Pickands estimator.

2nd Modification

To overcome the sensitivity of Pickands estimator to the choice of k, Drees (1996) suggested taking mixtures of Pickands estimators :

$$\hat{\gamma}_{P,v} = \int \hat{\gamma}_P(|4m_n t|)v(dt)$$
,

where m_n is an intermediate sequence, v is a probability measure on the Borel-σ-field B(0,1] and |x| denotes the smaller integer greater than or equal to x.

Under delicate 2nd order conditions on F, he calculates the bias of $\hat{\gamma}_{P,v}$, which consequently leads him to an unbiased estimator, even if we use a large number of upper order statistics. He also shows that his unbiased estimator is robust to deviations of the 2nd order assumptions (theoretically) and to an unsuitable choice of m_n (via simulation).

4.3 Hill Estimator

A few months after the publication of Pickands estimator, Hill proposed another estimator restricted, however, to the Fréchet case $\gamma > 0$ (Hill, 1975), i.e. this estimator is applicable only to regular varying d.f's. It is the most popular tail index estimator, the well-known Hill estimator.

4.3.1 Derivation

The *original derivation* of Hill estimator relied on the notion of conditional maximum likelihood estimation method.

Let's assume that our data $X_1, X_2, ..., X_n$ come from an unknown d.f. F, for which it holds that

$$1 - F(x) \sim cx^{-a} \qquad (4.4),$$

i.e. $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $\gamma = a^{-1} > 0$

and assume that relation (4.4) holds for $x \geq d$ (d fixed). Then our inference on a can be based on the conditional likelihood of $X_{i:n} \geq d$, i=1,...,k (where $X_{k:n} \geq d$).

By the Renyi representation theorem

$$X_{i:n} = F^{-1}\left[\exp\left\{-\left(\frac{e_1}{n} + \frac{e_2}{n-1} + \ldots + \frac{e_i}{n-i+1}\right)\right\}\right], \text{ i=1,...,n,}$$

where e_i are standard i.i.d. exponential r.v.'s.

Then, conditional on $X_{k:n} \geq d$, we have that

$e_i = (n-i+1)\left[\ln\left(1 - cX_{(i-1):n}^{-a}\right) - \ln\left(1 - cX_{i:n}^{-a}\right)\right]$, i=2,...,k, and

$e_1 = -n\ln\left(1 - cX_{1:n}^{-a}\right)$.

So, the conditional likelihood function of a is

$$L(a) = |J| \exp\left\{ n\ln\left(1 - cX_{1:n}^{-a}\right) - \sum_{i=1}^{k}(n-i)\ln\left(\frac{1 - cX_{i:n}^{-a}}{1 - cX_{(i-1):n}^{-a}}\right)\right\},$$

where the Jacobian J is proportional to $\prod_{i=1}^{k+1} \frac{d\ln\left(1 - cX_{i:n}^{-a}\right)}{dX_{i:n}}$.

Straightforward calculations yield the conditional MLE

$$\hat{a} = k\left(\sum_{i=1}^{k}\ln X_{i:n} - k\ln X_{k+1:n}\right)^{-1}.$$

In the context of generalized extreme-value distribution, the Hill estimator for the shape parameter γ>0, is obtained by inversion of the above relation :

$$\boxed{\hat{\gamma}_H = \frac{1}{k}\sum_{i=1}^{k}\ln X_{i:n} - \ln X_{k+1:n}}.$$

The above estimator will also be denoted as $\hat{\gamma}_H(k)$ whenever we want to stress the number of upper order statistics (k) used in the estimation.

An interesting aspect of this approach is that the likelihood function given above can be used, apart from obtaining conditional MLE's in the classical framework, in conjunction with a prior distribution for a so as to derive a conditional posterior distribution for a (in the Bayesian framework).

One of the appealing features of Hill estimator is that it can be derived even if we start from very different motivation points. Apart from the likelihood approach described above, the Hill estimator can be derived via a regular variation approach (Embrechts, et al., 1997) as well as via a graphical approach using mean excess function or QQ plots (Beirlant, et al., 1996).

The *regular variation approach* is in the same spirit as the construction of Pickands estimator, i.e. we base the inference of γ on a reformulation of $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, γ>0. According to the characterization property of MDA(Frechét), $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, γ>0 if and only if the tail of F (1-F) is regularly varying with index $-a$, that means that

$$\lim_{t\to\infty}\frac{1-F(tx)}{1-F(t)} = x^{-a}.$$

Partial integration of the above and use of Karamata's theorem (see chapter 2) leads to the relationship

$$\frac{1}{1-F(t)}\int_t^{\infty}\left(\ln x - \ln t\right)dF(x) \to \frac{1}{a}, \text{ as } t \to \infty.$$

If, now, we replace the unknown F by the empirical distribution function $F_n(\cdot)$, and t by an appropriate high, data-dependent level, such as $X_{k+1:n}$, then the form of Hill estimator arises as an estimator for a (or $1/a$ respectively) (see Embrechts et al., 1997).

The *graphical justification* of Hill estimator can be based either on the Pareto QQ plot or on the mean excess function. In the first case, we utilize the fact that if 1-F is regularly varying with index $-a$, the tail 1-F is of Pareto type for large x, i.e. $1-F(x) \sim x^{-a}$, for $x \to \infty$. That implies that the corresponding Pareto QQ plot should be linear for $x \to \infty$ (ultimately linear as it is often referred to). It can be shown (Beirlant et al., 1996) that the slope of this line is equal to $1/a$, while an estimator of this slope for the upper part of the plot (for values of x larger that $X_{k+1:n}$), where linearity can be regarded valid, is the known Hill estimator. The same reasoning holds for the mean excess plot.

4.3.2 Choice of k

The discussion that was made about the number of upper order statistics used in the calculation of Pickands estimator, holds for the Hill estimator as well. In order for the Hill estimator to have its claimed good properties (mentioned later on), the number k of upper order statistics used for its calculation must again have a moderate increase (i.e. the conditions $\lim_{n\to\infty} k(n) = \infty$, and $\lim_{n\to\infty} \frac{k(n)}{n} = 0$ should again hold). The choice of k seems to be an art, no 'hard and dry' solution exists.

A simple, and sometimes effective, way to choose k is again through the Hill plot of the pairs (k, $\hat{\gamma}_H(k)$), where $\hat{\gamma}_H(k)$ is the Hill estimator calculated on the basis of k upper order statistics. An area of k where the graph is almost horizontal indicates a proper range of k-values to choose from. Still, as Embrechts et al. (1997) present, there are sometimes where the results of Hill-plots are totally misleading. They refer to these plots as Hill-horror plots. Actually, as Drees et al. (2000) prove, traditional Hill-plot is most effective only when the underlying distribution is Pareto or very close to Pareto. For the Pareto distribution, one expects the Hill plot to be close to extreme-value index γ in the right side of the plot (since Hill is the MLE). On the other hand, in the case of regularly varying d.f.'s, i.e. when Pareto distribution is only approximated in the tail, Hill is only an approximate MLE based on 'large' observations and it is less clear what portion of the plot is most accurate.

For such cases, another way to plot the values of the Hill estimator is the *alternative Hill-plot*, proposed by Drees et al. (2000). Alternative Hill-plot is constructed by plotting the points $\left(\theta, \hat{\gamma}_H\left(\left[n^{\theta}\right]\right), 0 \le \theta < 1\right)$, that is one uses a logarithmic scale for the k-axis (horizontal axis). This has the effect of 'stretching' the left part of the Hill plot, giving more display-space to smaller values of k (larger observations). Clearly, this will not be beneficial when the underlying distribution is Pareto, but is much helpful in many other cases of regularly varying distributions. Drees et al. (2000) prove that proposition by quantifying superiority in terms of the occupation time of the plots in the neighbourhood of the true value γ. The figures that follow graphically illustrate the above statements.

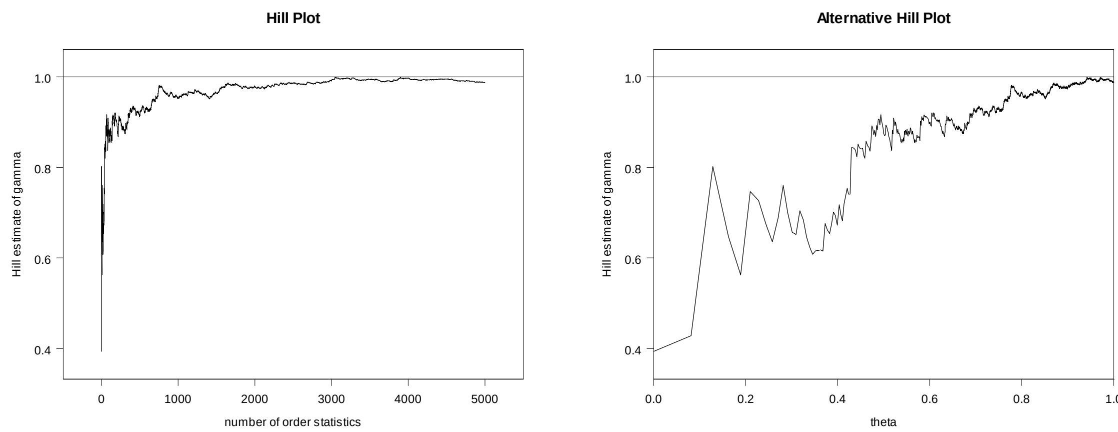

Figure 4.1. Hill Plot and Alternative Hill Plot of 5000 Pareto observations, with γ=1.

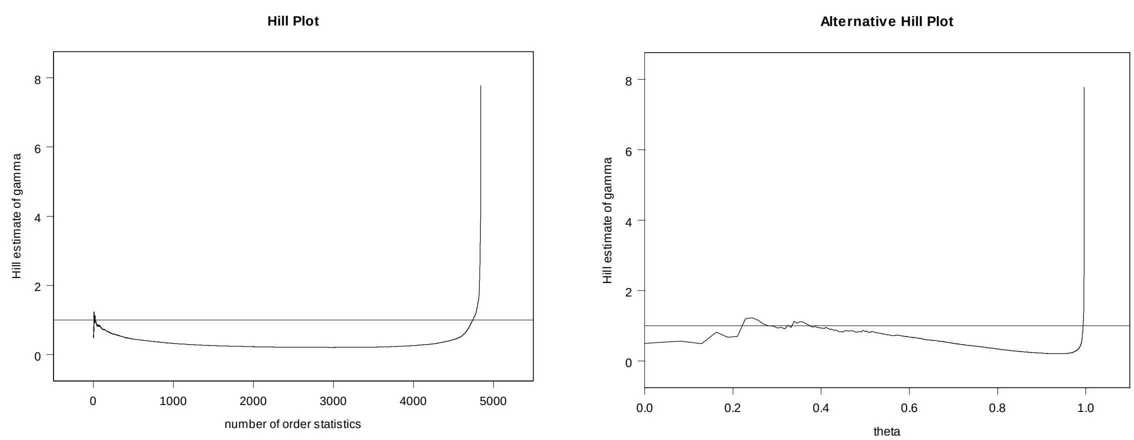

Figure 4.2. Hill Plot and Alternative Hill Plot of 5000 Cauchy observations, with γ=1 (μ=10,σ=1).

As we mentioned above, due to the high volatility in the Hill plots, it is difficult in practice to know exactly how to choose k in order to optimize the estimation of γ. A more 'objective' method to choose k has been proposed by Beirlant et al. (1996), in a regression setting. This method is described in the last section of the next chapter.

Another procedure for choosing k has been proposed by Hsieh (1999). His technique is based on the idea of spacing-statistics and proven (via simulation) to be relatively robust with respect to the true value of extreme-value index and to the underlying distribution.

4.3.3 Properties of Hill Estimator

The statistical behaviour and properties of Hill estimator have been studied by many authors separately, and under diverse conditions. In the sequel we present the main statistical properties of Hill estimator under the assumption of i.i.d. data as these are summarized by Embrechts et al. (1997). References for the proofs of these properties can be found therein. Similar (or slightly modified) results have been derived for data with several types of dependence or some other specific structures.

- ***Weak Consistency***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$ with $\gamma > 0$, $k(n) \to \infty$, $\dfrac{k(n)}{n} \to 0$,

then $\hat{\gamma}_H \to \gamma$, in probability, as $n \to \infty$.

- ***Strong Consistency***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$ with $\gamma > 0$, $\dfrac{k(n)}{n} \to 0$, $\dfrac{k(n)}{\ln\ln(n)} \to \infty$,

then $\hat{\gamma}_H \to \gamma$, almost surely, as $n \to \infty$.

- ***Asymptotic Normality***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, with $\gamma > 0$, under certain 2nd order conditions on F,

$k(n) \to \infty$, $\dfrac{k(n)}{n} \to 0$, and an additional restriction on the sequence k(n) depending on the 2nd order condition,

then $\sqrt{k}\left(\hat{\gamma}_H - \gamma\right) \to N(0,\gamma^2)$, in distribution as $n \to \infty$.

Note that, the conditions on k that ensure the consistency of Hill estimator are the same as the conditions imposed on k for the consistency of Pickands estimator. Moreover, as in the case of Pickands estimator, while consistency of Hill estimator depends only on the behaviour of k, asymptotic normality requires more delicate conditions (2nd order conditions) on the underlying d.f. F. Such conditions have been discussed by many authors, such as Davis and Resnick (1984), Haeusler and Teugels (1985), de Haan and Resnick (1998). These conditions are, again, difficult to verify in practice for an unknown d.f.

Using the asymptotic normality of $\hat{\gamma}_H$, Beirlant et al. (1996) derived asymptotic confidence intervals for $\hat{\gamma}_H(k_{opt})$, where k_{opt} is the optimal value of k according to their algorithm. Still, as they mention, such a confidence interval is expected to be conservative since the variability in the choice of k is ignored as should not be the case. A more precise method suggested, for confidence interval construction, is a bootstrap one. In the context of extreme-value index estimation, the bootstrap idea is utilized as follows:

– We sample with replacement (from the empirical distribution function $\hat{F}_n$) B bootstrap samples $X_i^* = (X_{i1}^*, X_{i2}^*, \ldots, X_{in}^*)$, i=1,...,B

– For each bootstrap sample i we compute $\hat{\gamma}_{H_i}$ (there should be a predetermined procedure for choosing k)

– $\{\hat{\gamma}_{H_i}$, i=1,...,B} constitute an estimate of the distribution of $\hat{\gamma}_H$, which can be used to estimate bootstrapped standard deviation of the estimator, confidence intervals, and so on.

4.3.4 Quantile estimation

Almost every author that was involved in the estimation of γ, was also led to an estimator for large quantiles based on the same rationale as for the estimation of γ.

In the context of Hill estimator, Beirlant et al. (1996), based on linear regression in a Pareto QQ plot, proposed the following quantile estimator :

$$\hat{x}_p = X_{(k+1):n}\left(\frac{k+1}{(n+1)p}\right)^{\hat{\gamma}_H(k)}.$$

4.3.5 Modifications-Developments on Hill Estimator

Looking at the properties of the Hill estimator and the associated conditions on k, one easily notices, that the Hill estimator is a consistent estimator only if $\hat{\gamma}_H(k)$ is based on a 'moderately increasing' (with respect to sample size n) sequence of k upper order statistics. The relationship between variance and bias that exists, depending on the rate of increase of k(n), has already been discussed in the context of Pickands estimator. The dependence of Hill's bias and mean squared error on k has been studied by Martins et al. (1999) for a number of heavy-tailed underlying models. The authors, also, introduce convex combinations of Hill estimators, which provide an, admittedly small, improvement over the standard Hill.

One of the methods to render the problem is to find the bias of the estimator (Hill estimator, in this case) and, consequently formulae for unbiased estimators, when a large (larger than moderate) number k of upper order statistics is used in the estimation. Of course, such a calculation requires additional assumptions (2^{nd} order conditions) imposed on the d.f. F. Such an attempt has been made by Peng (1998) and is summarized below.

If we assume that

(i) $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$ with $\gamma > 0$, and

(ii) there exist function $A(t) \to 0$ (with constant sign near infinity), and constant $\rho < 0$ such that

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^{\gamma}}{A(t)} = x^{\gamma}\,\frac{x^{\rho}-1}{\rho}, \text{ for } x > 0$$

then $\tilde{\gamma}_H(k) = \hat{\gamma}_H(k) - \dfrac{M_2(k) - 2\left(\hat{\gamma}_H(k)\right)^2}{2\hat{\gamma}_H(k)\hat{\rho}}\left(1-\hat{\rho}\right)$

(for $k \to \infty$, $k/n \to 0$, and $\sqrt{k}\,A(n/k) \to \lambda$)

is an unbiased (and normally distributed) estimator of the shape parameter γ>0,

where $\hat{\rho} = \left(\ln 2\right)^{-1} \ln \dfrac{M_2\left(n/(2\ln n)\right) - 2\left(\hat{\gamma}_H\left(n/(2\ln n)\right)\right)^2}{M_2\left(n/\ln n\right) - 2\left(\hat{\gamma}_H\left(n/(\ln n)\right)\right)^2}$, and

$M_2(k) = \frac{1}{k}\sum_{i=1}^{k}\left(\ln X_{i:n} - \ln X_{(k+1):n}\right)^2$. For proof see Peng (1998).

Another result for the above unbiased estimator, shown theoretically as well as via simulation, is that this modified estimator performs better (in terms of mean square error) than the classical Hill estimator as k becomes large. For small k, classical Hill estimator, actually, outperforms modified Hill estimator.

4.3.6 Asymptotic Behaviour of Hill Estimator Based on Dependent Data

Though not properly emphasized, all the above results-properties (as well as the results for the other estimators) hold in case that the estimator is based on a sample of i.i.d. data $\{X_1, X_2, \ldots, X_n\}$. However, since many real life applications provide one with dependent, stationary data rather than i.i.d. data, it is important to understand the behaviour of Hill estimator under more general assumptions such as stationarity of the observed sequence or, even more generally, only for a common marginal distribution. Several recent papers support the belief that Hill estimator performs well even under these weaker assumptions.

Hsing (1991) generalizes certain results on Hill estimator in the i.i.d. case by dropping independence. He studies some asymptotics of Hill estimator without requiring $\{X_1, X_2, \ldots, X_n\}$ to be independent. Actually, he proves weak consistency of $\hat{\gamma}_H$ by making no assumption on the $X_i's$ other than that they have a common marginal distribution F ($F \in MDA(H_\gamma)$, $\gamma > 0$). By adding some 2nd order assumptions on F (but still not assuming independence) he derives the asymptotic distribution of $\hat{\gamma}_H$, which, in this case, is more complicated than the well-known normal d.f.

Resnick and Stărică (1995) prove the consistency of Hill estimator for an infinite order moving average sequence whose marginal distribution is regularly varying. They also consider in detail the special case when the observations $X_i's$ come from a p-th order autoregressive process (AR(p)-process) whose residuals have regularly varying tail probabilities of index $-1/\gamma$. Since both the stationary sequence $\{X_1, X_2, ..., X_n\}$ and the residuals have distributions with regularly varying tails of index $-1/\gamma$, for estimating γ one could either

1. apply Hill estimator to the observed time series $\{X_1, X_2, ..., X_n\}$, or
2. assuming the order of autoregression p is known, fit coefficients of the autoregression and use this to estimate residuals. Then estimate γ by applying Hill estimator to the estimated residuals.

Both of these methods are proved to be consistent.

Later on, Resnick and Stărică (1996) compare the efficiencies of these two methods of estimation. Under quite general conditions on the innovations of the AR-process and on the asymptotic behaviour of the estimators for the coefficients of the autoregression, they prove that the second method based on estimated residuals, is a more efficient procedure. The asymptotic variance of the Hill estimator is always smaller when the second method is used. Actually, the asymptotic variance of Hill estimator applied to the estimated residuals does not depend on the coefficients of the AR process and is essentially equal to the asymptotic variance of the Hill estimator for independent data. Hence, they prove that the procedure of applying the Hill estimator directly to an autoregressive process is inferior to the procedure of first estimating autoregressive coefficients and then estimating γ using estimated residuals. Moreover, using a 'tail empirical' approach they prove Hill's asymptotic normality (when estimated from residuals) under 2^{nd} order regular variation conditions and a restriction on sequence k(n). More precisely, they associate the tail empirical process to the sequence of estimated residuals, show the weak convergence of the normalized tail empirical process to a process closely related to Brownian motion and deduce from that the asymptotic behaviour of Hill estimator.

More generally, Resnick and Stărică (1998) discuss consistency of Hill estimator when it is applied to certain classes of heavy-tailed stationary processes. The authors concentrate on non-linear models, since, as they say, while in the traditional setting of a stationary time series with finite variance every purely non-deterministic process can be expressed as a linear process driven by an uncorrelated input sequence, the situation is totally different when the stationary sequences has heavy tails and perhaps infinite variance. In this case we have no such confidence that heavy tailed linear models are sufficiently flexible and rich enough for modelling purposes. Actually, linear models do not seem to describe adequately the underlying random mechanism when heavy tails are present. A popular non-linear alternative to the linear model is the bilinear process. Other worthy non-linear models which the authors consider are two classes of random coefficient models, one of which includes the important example of the ARCH process and hidden semi-Markov models or r.v.'s defined on a semi-Markov chain. Such models have recently been used to fit times between packet transmissions at a terminal. In particular, the authors (using again the 'tail empirical' approach) prove weak consistency of Hill estimator when it is applied to the following three classes of heavy-tailed stationary processes :

(i) processes approximated by sequences of m-dependent r.v.'s (such as infinite moving average processes, simple stationary bilinear models, solutions of stochastic difference equations of specific form)

(ii) models which are solutions of stochastic difference equations (here is included the ARCH process), and

(iii) hidden semi-Markov models.

4.4 Adapted Hill Estimator

Though the Hill estimator has the apparent disadvantage that is restricted to the case $\gamma>0$, it has been widely used in practice and extensively studied by statisticians. Its popularity is partly due to its simplicity and partly to the fact that in most of the cases where extreme-value analysis is called for, we have long-tailed d.f.'s (i.e. $\gamma>0$). Still, it is a tempting problem to try to extend the Hill estimator (with its simplicity and good properties) to the general case $\gamma \in \Re$.

Such an attempt, led Beirlant et al. (1996) to the so-called adapted Hill estimator, which is applicable for any γ in the range of real numbers. The main theoretical result that stands behind adapted Hill estimator is the following proposition (Beirlant et al., 1996).

Proposition

If F belongs to the maximum domain of attraction of H_γ (F $\in$ MDA(H_γ)), then the function UH : $x \to U(x)e_{\ln X}\left(\ln U(x)\right) = U(x)E\left(\ln X - \ln(U(x)) \middle| X > U(x)\right)$ is regularly varying with index γ, i.e.

$$UH \in RV_\gamma.$$

Thus, following the same motivation as for the Hill estimator (under the QQ plot approach), where now UH is regularly varying instead of (1-F), an estimator of γ turns out to be the quantity :

$$\boxed{\hat{\gamma}_{adH} = \frac{1}{k}\sum_{i=1}^{k} \ln(UH_i) - \ln(UH_{k+1})},$$

where $UH_i = X_{(i+1):n}\left(\frac{1}{i}\sum_{j=1}^{i} \ln X_{j:n} - \ln X_{(i+1):n}\right)$ is the empirical counterpart of UH at x=n/i.

The original derivation of this estimator was based on an effort to extend the 'QQ plot' motivation of Hill (for γ>0) to the more general case $\gamma \in \Re$. This can be achieved by the introduction of the function UH (see proposition above). More precisely, from the theory of regular variation, $\ln\left(UH(x)\right) \xrightarrow{x\to\infty} \gamma \ln x$.

From this follows, that the generalized QQ plot $\left(-\ln\frac{j}{n}, \ln\left(UH_j\right)\right)$ (j=1,..., n-1) will be *ultimately linear* with slope γ. The same thinking that has been followed in the Hill estimator can also be adopted here, leading to the previously defined estimator.

The basic asymptotic result for the adapted Hill estimator is summarized in the following theorem (Beirlant et al., 1996).

Theorem

When UH is a (normalized) regularly varying function with index γ (UH∈ RV_γ)

such that $\sqrt{k}\int_0^1 \delta\left(\frac{n}{kt}\right)dt \to 0$, as $n \to \infty$, where $\delta(x) \to 0$ is the function involved in the characterization property of regular varying functions (see chapter 2), then

$$\sqrt{k}[\hat{\gamma}_{adH} - \gamma] \to \begin{cases} Normal\left(0,(1+\gamma)^2\right), & \gamma \geq 0 \\ Normal\left(0, \dfrac{\left(1-\gamma\left(1+\gamma+2\gamma^2\right)\right)}{(1-2\gamma)}\right), & \gamma < 0 \end{cases}$$ in distribution as $n \to \infty$,

For the choice of k (number of upper order statistics used in the calculation of the estimator) an iterative procedure, as the one described for Hill estimator, can also be applied here.

As far as the estimation of large quantiles is concerned, the corresponding estimation formula is

$$\hat{x}_p = X_{(k+1):n} + \frac{\left(n(1-p)/k\right)^{-\hat{\gamma}_{adH}} - 1}{\hat{\gamma}_{adH}} X_{(k+1):n} \cdot \hat{\gamma}_H .$$

Of course, such an estimate is meaningful only for $\gamma \geq 0$ when F has an infinite upper endpoint. In the case $\gamma < 0$, the interest is transferred to the estimation of the finite upper endpoint $x_F = \sup_x\{x: F(x) < \infty\} < \infty$. This quantity can be estimated as

$$\hat{x}_F = X_{(k+1):n} + \left(1 - \frac{1}{\hat{\gamma}_{adH}}\right)E_k ,$$

where $E_k = \frac{1}{k}\sum_{j=1}^{k} X_{j:n} - X_{(k+1):n}$ is the empirical counterpart of the mean excess function evaluated at $x = X_{(k+1):n}$.

4.5 Moment Estimator

Another estimator that can be considered as an adaptation of Hill estimator, in order to obtain consistency for all $\gamma \in \Re$, has been proposed by Dekkers et al. (1989). This is the so-called moment estimator, given by

$$\boxed{\hat{\gamma}_M = M_1 + 1 - \frac{1}{2}\left(1 - \frac{(M_1)^2}{M_2}\right)^{-1}},$$

where $M_j \equiv \frac{1}{k}\sum_{i=1}^{k}\left(\ln X_{i:n} - \ln X_{(k+1):n}\right)^j$, j=1, 2.

Note that the above quantities are functions of k (number of upper order statistics used in the estimation), so they are often denoted as $\hat{\gamma}_M(k)$ and $M_i(k)$. Moreover, it is easy to see that M_1 is essentially the well-known Hill estimator.

4.5.1 Derivation

As was the case for Pickands and Hill estimator, a reasonable approach, when trying to estimate extreme-value parameter γ, is to try to reformulate a characterization property of MDA(H_γ), which could lead to a simpler relationship with respect to γ. In particular, the key-point here is that MDA(H_γ) can be related with some extended form of regular variation.

Still, a more intuitive background for $\hat{\gamma}_M(k)$ is provided by Dekkers et al. (1989).

It is well known that the convergence of the Hill estimator for γ>0 is the sample analogue of the following relation, which constitutes a characterization property of MDA(H_γ), for γ>0 :

$$\gamma = \int_1^\infty u^{-1/\gamma}\frac{du}{u} \to \int_1^\infty \frac{1-F(tu)}{1-F(t)}\frac{du}{u} = \frac{\int_t^\infty (1-F(u))/(du/u)}{1-F(t)} = \frac{\int_t^\infty (\ln x - \ln t)dF(x)}{1-F(t)}$$
$$= E\left(\ln X - \ln t \middle| X > t\right), \text{ as } t \to \infty.$$

So, the reason for using the ln(.) of order statistics instead of the order statistics themselves is that otherwise the first integral may diverge. This forces one to use logarithms of order statistics instead of the order statistics themselves in the definition of

M_1. This is not possible when the random variables are negative. In order to avoid this problem (which comes up only for $\gamma \le 0$) we have to impose the extra condition $x_F > 0$. This does not cause any difficulty in applications, since it can always be achieved by a simple shift.

Moreover, an analogue of the above relation, for $\gamma=0$ is :

$F \in$ MDA(H_0) if and only if

$$\lim_{t \to x_F^-} \frac{E\left(\{X-t\}^2 \middle| X>t\right)}{\left\{E\left(X-t \middle| X>t\right)\right\}^2} = \frac{\int_0^\infty x^2 d\left(1-e^{-x}\right)}{\left\{\int_0^\infty x d\left(1-e^{-x}\right)\right\}^2} = 2.$$

These two considerations led the authors to consider the quotient $M_2 / (M_1)^2$. However, it is not clear that this quotient discriminates sufficiently, since taking logarithms transforms r.v.'s belonging to MDA(H_γ), $\gamma \ge 0$ into r.v.'s of MDA(H_0). Fortunately, M_1 itself also converges for any γ and discriminates the range of values of γ not covered by $M_2 / (M_1)^2$.

The denomination 'moment estimator' stems from the fact that the quantities M_1 and M_2 can be interpreted as empirical moments.

4.5.2 Properties of Moment Estimator

In the beginning of this section we have mentioned that the main reason for introducing the present estimator, was to extend the Hill estimator and its nice properties. So, weak and strong consistency, as well as asymptotic normality of the moment estimator have been proven by its creators Dekkers et al. (1989). The necessary and sufficient conditions for these desirable properties to hold are mentioned below.

- ***Weak Consistency***

If $F \in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $x_F > 0$, $k(n) \to \infty$, $\frac{k(n)}{n} \to 0$,

then $\hat{\gamma}_M \to \gamma$, in probability, as $n \to \infty$.

- ***Strong Consistency***

If $F \in MDA(H_{\gamma;\mu,\sigma})$, $x_F > 0$, $\frac{k(n)}{n} \to 0$, $\frac{k(n)}{(\ln(n))^{\delta}} \to \infty$, for some $\delta > 0$,

then $\hat{\gamma}_M \to \gamma$, almost surely, as $n \to \infty$.

- ***Asymptotic Normality***

If $F \in MDA(H_{\gamma;\mu,\sigma})$, and

for $\gamma > 0$, $\lim_{t\to\infty} \frac{(tx)^{-\gamma} U(tx) - t^{-\gamma} U(t)}{b_1(t)} = \pm \ln x$, x>0,

for $\gamma = 0$, $\lim_{t\to\infty} \frac{\ln(U(tx)) - \ln(U(t)) - b_2 \ln x}{b_3(t)} = \pm \frac{(\ln x)^2}{2}$, x>0, and

for $\gamma < 0$, $\lim_{t\to\infty} \frac{(tx)^{-\gamma} \{U(\infty) - U(tx)\} - t^{-\gamma} \{U(\infty) - U(t)\}}{b_4(t)} = \pm \ln x$, x>0,

where $U \equiv (1/(1-F))^{\leftarrow}$, and b_i, i=1,2,3,4 are positive function,

$k(n) \to \infty$, $k(n) = o(n/g^{\leftarrow}(n))$, where $g(t) \equiv t^{1-2\gamma} \{U(t)/b_1(t)\}^2$, for $\gamma > 0$,

$g(t) \equiv t^{1-2\gamma} \{(\ln U(\infty) - \ln U(t))/b_4(t)\}^2$, for $\gamma < 0$, and

$g(t) \equiv t\{U(t)/a(t)\}^2$, for $\gamma = 0$,

then $\sqrt{k}(\hat{\gamma}_M - \gamma) \to N(0, v(\gamma))$, in distribution as $n \to \infty$,

where $v(\gamma) = \begin{cases} 1+\gamma^2 & \gamma \ge 0 \\ (1-\gamma)^2(1-2\gamma)\left(4 - 8\frac{1-2\gamma}{1-3\gamma} + \frac{(5-11\gamma)(1-2\gamma)}{(1-3\gamma)(1-4\gamma)}\right) & \gamma < 0 \end{cases}$.

The above 2nd order conditions of F (or equivalently U) could be simplified substantially (in notation) if we used the notion of Π-varying functions. For a definition of a function see Dekkers et al. (1989). As far as the unbiasness (or better biasness) of moment estimator is concerned, if $k(n) \sim cn / g^{\leftarrow}(n)$ for some positive constant c, as

$n \to \infty$, then $\sqrt{k}\left(\hat{\gamma}_M - \gamma\right)$ has asymptotically a normal distribution with the same variance but with mean $\pm\sqrt{c}$.

4.5.3 Quantile estimation

Dekkers and de Haan (1989), use *differences* of large order statistics as building blocks for an estimator of γ (Pickands) and for estimating large quantiles. In the context of moment estimator, we can construct a similar estimate for large quantiles by using *sums* of large order statistics. Dekkers et al. (1989) propose to estimate $x_p = F^{\leftarrow}(p)$ for large p (close enough to 1) as follows :

$$\hat{x}_p = \frac{a_n^{\hat{\gamma}_M} - 1}{\hat{\gamma}_M} \cdot \frac{X_{(k+1):n} M_1}{\rho_1(\hat{\gamma}_M)} + X_{(k+1):n},$$

where $a_n = \dfrac{k}{n(1-p)}$, $\rho_1(\gamma) = \begin{cases} 1, & \gamma \geq 0 \\ (1-\gamma)^{-1}, & \gamma < 0 \end{cases}$.

Note that though this large-quantile estimator has been derived in the framework of moment estimator, the authors suggest that in the above formula any consistent estimator of γ could be used. The authors, also, explore an asymptotic d.f. for $\hat{x}_p$, based on which asymptotic confidence intervals can be constructed.

Furthermore, in the case $\gamma < 0$, the estimation of the right endpoint is of the major concern. A proposed estimator for this is :

$$\hat{x}_F = X_{(k+1):n} M_1 \left(1 - \frac{1}{\hat{\gamma}_M}\right) + X_{(k+1):n}.$$

Again here, instead of $\hat{\gamma}_M$ we could use any consistent estimator of γ.

4.6 Other Semi-Parametric Estimation Methods

Up to now, we have described analytically four semi-parametric methods of estimation of parameter γ (extreme-value index) of $H_{\gamma;\mu,\sigma}$. Still, there is a vast literature on other estimation alternatives. The applicability of extreme-value analysis on a variety of different fields led scientists with different background to work on this subject and consequently derive many and different estimators. Pickands, Hill and lately Moment estimators, continue to be the basis. If not anything else, most of the other proposed estimators constitute efforts to render some of the disadvantages of these three basic estimators, while others are efforts to generalize the framework of these. In the sequel, we present a number of such methods. Basically, we tried to include estimators diverging in motivation and derivation, as well as some of the latest developments. We present their rationale and their main properties, compare them with the 'basic' estimators, and point out their pros and cons.

The general framework we are working in, is that we have an i.i.d. sample (X_1, X_2, ..., X_n) from an unknown d.f. F ∈ $MDA\left(H_{\gamma;\mu,\sigma}\right)$ and our aim is to estimate γ.

4.6.1 Moments Ratio Estimator

Concentrating on cases where $\gamma > 0$, i.e. on regularly varying d.f.'s, the main disadvantage of Hill estimator is that it can be severely biased, depending on the 2nd order behaviour of the underlying d.f. F.

For $\beta \neq 0$, the expansion

$$F(x) = 1 - \mathrm{a}x^{-\alpha}\left[1 + bx^{-\beta} + o\left(x^{-\beta}\right)\right] \qquad (4.6),$$

holds as $x \to \infty$, and where $\alpha, \beta > 0$, $\mathrm{a} > 0$ and $b \in \Re$.

From the above formula, it becomes apparent that, for $b = 0$ Hill estimator is unbiased and consistent, but for $b \neq 0$ the Hill estimator is biased and may be inconsistent depending on the rate at which $k(n) \to \infty$ as $n \to \infty$. By taking into account the second order term of (4.6), one gets the bias of Hill estimator, and by assuming that (4.6) is an exact relation, Danielsson et al. (1996) propose the moments ratio estimator :

$$\boxed{\hat{\gamma}_{MR} = \frac{1}{2} \cdot \frac{M_2}{M_1}},$$

as an alternative to Hill estimator $\hat{\gamma}_H$. The sample (conditional) moments M_1, M_2 based on the k upper order statistics have been defined in the context of Moment estimator. They prove that $\hat{\gamma}_{MR}$ has lower asymptotic square bias than Hill (when evaluated at the same threshold, i.e. for the same k), though the convergence rates are the same. Additionally, Danielsson's et al. (1996) simulations reveal that the choice of k is also very important for $\hat{\gamma}_{MR}$ as it is for the Hill, i.e., $\hat{\gamma}_{MR}$ is sensitive to the choice of k.

4.6.2 Kernel Estimators

Extreme-value theory dictates that if F $\in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $\gamma > 0$, then it holds that $F^{\leftarrow}\left(1-x\right) \in RV_{-\gamma}$, where $F^{\leftarrow}(\ .)$ is the generalized inverse (quantile) function corresponding to d.f. F. Csörgő et al. (1985) show that for 'suitable' kernel functions K, it holds that

$$\int_0^{1/\lambda} \left\{\ln F^{\leftarrow}\left(1-u\lambda\right) d\left\{uK(u)\right\}\right\} \to \gamma \text{, as } \lambda \to 0.$$

Substituting $F^{\leftarrow}$ with its empirical counterpart $F_n^{\leftarrow}$ (which is a consistent estimator of $F^{\leftarrow}$), they propose

$$\boxed{\hat{\gamma}_{Kernel} = \left(\int_0^{1/\lambda} K(u)du\right)^{-1} \left(\sum_{j=1}^{n} \frac{j}{n\lambda} K\left(\frac{j}{n\lambda}\right) \left\{\ln X_{j:n} - \ln_{(j+1):n}\right\}\right)}$$

as an estimator of γ, where $\lambda = \lambda(n)$ is a bandwidth parameter, K is a kernel function satisfying the conditions:

(H1) $K(u) \geq 0$ for $0 < u < \infty$.

(H2) K(.) is non-increasing and right continuous on $(0,\infty)$.

(H3) $\int_0^{\infty} K(t)dt = 1$.

(H4) $\int_0^\infty t^{-1/2} K(t)dt < \infty$.

Within this setting, Csörgő et al. (1985) prove that

- $\hat{\gamma}_{Kernel} \to \gamma$ in probability as $\lambda(n) \to 0$ and $n\lambda(n) \to \infty$
- $\frac{\sqrt{n\lambda}}{\gamma} \left\{ \int_0^\infty K^2(u)du \right\}^{-1/2} \left(\hat{\gamma}_{Kernel} - \gamma - \beta_n \right) \to Normal(0,1)$

  under 2nd order conditions on F, and as $\lambda(n) \to 0$ and $n\lambda(n) \to \infty$,

  where β_n is related to the 2nd order behaviour of F.

Under the same 2nd order conditions on the underlying d.f. F, Csörgő et al. (1985) derive formulae for the optimal choice of bandwidth λ and the kernel function K.

The choice $K(u) = 1_{[0<u<1]}$ and $\lambda = k/n$ corresponds to the Hill estimator $\hat{\gamma}_H(k)$. The only real superiority of Hill estimator is that if we let $\lambda = k/n$ for a given asymptotic variance , then Hill uses the minimum number k of upper order statistics of the sample. Notice that, in general, the problem of choosing k has now be reformulated to the problem of choosing λ and K(.). Another point that we should mention is that in the kernel estimator, all the observations X_i are used, and not only the upper k order statistics as is the usual case, but the observations are taken into account with different weights depending on their rank order (i.e. how large they are). This idea may seem much more attractive than the ad-hoc choice of k upper order statistics, still, until recently, kernel estimators of extreme-value parameter have not been widely accepted and used.

4.6.3 QQ – Estimator

In the section concerning Hill's derivation we have quoted the 'QQ-plot' approach. At that point, we have mentioned that, approximately, Hill estimator is the slope of the line fitted to the upper tail of Pareto QQ plot. A more precise estimator, under this approach, has been suggested by Kratz and Resnick (1996). The main idea is as follows :

If F $\in MDA\left(H_{\gamma;\mu,\sigma}\right)$, $\gamma > 0$, then (1-F) is of Pareto type (regularly varying), i.e.

$1 - F(x) \sim x^{-1/\gamma} l(x)$, as $x \to \infty$ (*l* is a slowly varying function).

Consequently, conditional on $X_{(k+1):n}$ ($k \to \infty$), we have that ($X_{1:n}$, $X_{2:n}$, ..., $X_{k:n}$) have the distribution of the order statistics from a random sample of size k from a distribution concentrating on $\left(X_{(k+1):n}, \infty\right)$. Thus, conditional on $X_{(k+1):n}$, $\left(\frac{X_{i:n}}{X_{(k+1):n}},\ i = 1,\ldots,k\right)$ behave like the order statistics from a distribution concentrating on $\left(1,\infty\right)$ with tail $\frac{1-F(X_{(k+1):n}t)}{1-F(X_{(k+1):n})} \approx t^{-1/\gamma}$. So, it is reasonable to define the qq-estimator $\hat{\gamma}_{qq}$ based on the upper k order statistics to be the slope of the fitted line to the points $\left(-\ln\frac{k+1-i}{k+1}, \ln X_{i:n}\right)$, i=1, ...,k. That is,

$$\hat{\gamma}_{qq} = \frac{\sum_{i=1}^{k} \ln\frac{i}{k+1}\left\{\sum_{j=1}^{k} \ln X_{j:n} - k\ln X_{i:n}\right\}}{k\sum_{i=1}^{k}\left(\ln\frac{i}{k+1}\right)^2 - \left(\sum_{i=1}^{k}\ln\frac{i}{k+1}\right)^2}.$$

Kratz and Resnick (1996) prove the weak consistency and asymptotic normality of qq-estimator (under conditions similar to the ones imposed for the Hill estimator). It can be shown that the asymptotic variance of qq-estimator is twice the asymptotic variance of Hill estimator, while similar conclusions are drawn from simulations of small samples. However, Hill estimator exhibits considerable bias in certain circumstances and thus asymptotic variance is not a good criterion for superiority. Moreover, one of the advantages of qq-plotting over the Hill estimator is that the residuals (of the Pareto plot) contain information which potentially can be utilized to combat the bias in the estimates when the tail is not exact Pareto.

4.6.4 'k-records' Estimator

A statistical notion that is closely related to extreme-value analysis is that of records, or, more generally, k-records. The k-record times and k-record values, ($k \in \mathbb{N}$, $1 \le k \le n$), can be defined respectively as

$\tau^{(k)}(1) = k$, $\tau^{(k)}(i) = \min\left\{ j > \tau^{(k)}(i-1), R(j) \ge j-k+1\right\}$,

$X^{(k)}(i) = X_{k:\tau^k(i)}$,

where R(j) is the sequential rank of X_j in the sample (X_1,...,X_j), i.e. X_j=$X_{j-R(j)+1:j}$.
(for alternative definitions of k-records and a review on basic distribution theory and asymptotic results about records, see Nagaraja, 1988).
The k-record values are themselves revealing of the extremal behaviour of the d.f. F, so they can also be used to assess the extreme-value parameter $\gamma \in \Re$. Berred (1995) constructed the estimator :

$$\boxed{\hat{\gamma}_{rec} = \ln \frac{X^{(k)}(n) - X^{(k)}(n-k)}{X^{(k)}(n-k) - X^{(k)}(n-2k)}}.$$

Under the usual conditions on k(n) (though notice that now the meaning of k(n) is different from before), he proves weak and strong consistency of $\hat{\gamma}_{rec}$, while by imposing 2nd order conditions on F (similar to the previous cases) he also shows asymptotic normality of $\hat{\gamma}_{rec}$. More precisely he derives the result that :

$$\sqrt{k}\left(\hat{\gamma}_{rec} - \gamma\right) \to Normal\left(0, \frac{\left(e^{2\gamma}+1\right)\gamma^2}{\left(e^{\gamma}-1\right)^2}\right) \text{ in distribution}$$

as $n \to \infty$ and for more delicate conditions on k(n) and on F.

4.6.5 Other extreme-value index estimators

Another effort to estimate extreme-value index $\gamma \in \Re$ was made by Draisma and de Haan (1996). By defining the function $\Phi(\gamma) = 2\frac{\left(2^{\gamma-1}-1\right)/(\gamma-1)}{\left(2^{\gamma}-1\right)/\gamma} - 1$ and its empirical counterpart $\hat{\Phi}_n = \frac{k^{-1}\sum_{i=k}^{2k-1} X_{(i+1):n} - X_{(2k+1):n}}{X_{(k+1):n} - X_{(2k+1):n}}$, they propose the estimator

$$\boxed{\hat{\gamma}_n = \Phi^{\leftarrow}\left(\hat{\Phi}_n\right)},$$

for which they prove weak consistency and asymptotic normality. Still, their estimator turned out to be inferior (with respect to asymptotic variance as well as to sensitivity to the choice of k) to both Pickands and moment estimators.

An estimator related to the moment estimator $\hat{\gamma}_M$ is Peng's estimator

$$\boxed{\hat{\gamma}_L = \frac{M_2}{2M_1} + 1 - \frac{1}{2}\left(1 - \frac{(M_1)^2}{M_2}\right)^{-1}},$$

($M_i \equiv \frac{1}{k}\sum_{j=1}^{k}\left(\ln X_{j:n} - \ln X_{(k+1):n}\right)^i$, i=1, 2 same quantities as for $\hat{\gamma}_M$), which has been suggested by Deheuvels et al. (1997). This estimator has been designed to somewhat reduce the bias of the moment estimator.

Another related estimator suggested by the same authors is

$$\boxed{\hat{\gamma}_W = 1 - \frac{1}{2}\left(1 - \frac{(L_1)^2}{L_2}\right)^{-1}},$$

where $L_i \equiv \frac{1}{k}\sum_{i=1}^{k}\left(X_{i:n} - X_{(k+1):n}\right)^i$, i=1, 2.

As Deheuvels et al. (1997) mention, $\hat{\gamma}_L$ is consistent for any $\gamma \in \Re$ (provided, as usually, that k is moderately increasing), while $\hat{\gamma}_W$ is consistent only for $\gamma < 1/2$. Moreover, under appropriate smoothness conditions on F and a further bound on the rate of increase of $k(n)$, $\hat{\gamma}_L$ is asymptotically normal after normalization. Normality holds for $\hat{\gamma}_W$ only for $\gamma < 1/4$. On the other hand, $\hat{\gamma}_W$ is location and scale invariant, while $\hat{\gamma}_L$ is only scale invariant.

Drees (1998) introduces a general class of estimators of the extreme-value index that can be represented as a scale invariant functional applied to the empirical tail-quantile function.

As one can notice, apart from estimators applicable for any $\gamma \in \Re$, estimation techniques have been developed valid only for a specific range of values of γ. This is due to the fact that H_γ, for γ in a specific range may lead to families of d.f.'s of special interest. The most typical types are estimation methods for $\gamma > 0$, which correspond to d.f.'s with regularly varying tails (here the Hill estimator is included). Moreover, estimators for $\gamma \in (0,1/2)$ are of particular interst since H_γ, $\gamma \in (0,1/2)$ represents α-stable distributions (γ=1/α).

Estimators, for the index $\gamma > 0$ have also been proposed by Hall (1982), and Feuerverger and Hall (1999). More restricted estimation techniques for α-stable d.f.'s are described in Mittnik et al. (1998). Sometimes, the interest of authors is focused merely on the estimation of large quantiles, which in any case is what really matters in practical situations. Such estimators have been proposed by Davis and Resnick (1984) (for $\gamma > 0$) and Boos (1984) (for $\gamma = 0$).

We have mentioned several alternative estimators for the extreme-value index γ. All of these estimators share some common desirable properties, such as weak consistency and asymptotic normality (though these properties may hold under slightly different, unverifiable in any case, conditions on F and for different ranges of the parameter γ). On the other hand, simulation studies or applications on real data can end up in large differences among these estimators. In any case, there is no 'uniformly better' estimator (i.e. an estimator that is best under all circumstances). Of course, Hill, Pickands and moment estimators are the most popular estimators. This could be partly due to the fact that they are the oldest ones. The rest estimators have been introduced later. Actually, most of these have been introduced as alternatives to Hill, Pickands or moment estimator and some of them have been proven to be superior in some cases (but, again, not always). In the literature, there are some comparison studies of extreme-value index estimators (either theoretically or via Monte-Carlo methods), such as Deheuvels et al. (1997) and Rosen and Weissman (1996). Still, these studies are confined to a small number of estimators. Moreover, most of the authors that introduce a new estimator compare it with some of the standard estimators (Hill, Pickands, moment). In the following chapter, we include a rather extensive comparison study, via simulation, for most of the previously presented estimators.

4.7 'Peaks Over Thresholds' Estimation Methods

All the previously discussed semi-parametric estimation methods, where based on the notion of maximum domains of attractions of the generalized extreme-value d.f, i.e. they were based on the results of extreme value theory related to the magnitude of extreme observation. Still, further results in extreme-value theory describe the behaviour of large observations that exceed high thresholds, and these are the results which lend themselves to the so-called 'Peaks Over Threshold' (POT, in short) models. They address the question: "Given an observation is extreme, how extreme might it be?" The distribution which comes to the fore in this case is the generalized Pareto distribution (GPD). A thorough description of GPD and its exact role in the context of extreme value theory was provided in chapter 2.

The interpretation of GPD as the limiting distribution of the (scaled) excesses of a d.f. over a (high) threshold (F_u), was exploited in practice and led to some of the most widely-used methods in extreme-value analysis. Thus, the estimation of the extreme-value parameter γ or the large quantiles of the underlying d.f.'s can be alternatively estimated via the generalized Pareto distribution instead of the generalized extreme-value distribution.

The GPD can be fitted to data consisting of excesses of high thresholds by a variety of methods including the maximum likelihood method (ML) and the method of probability weighted moments (PWM). MLEs must be derived numerically because the minimal sufficient statistics for the GPD are the order statistics and there is no obvious simplification of the non-linear likelihood equation. Grimshaw (1993) provides an algorithm for estimating the MLEs for GPD. ML and PWM methods have been compared for data of GPD both theoretically and in simulation studies by Hosking and Wallis (1987) and Rootzén and Tajvidi (1997).

Hosking and Wallis (1987) have found that for generalized Pareto data with shape parameter in the range $0 \le \gamma \le 0.4$ and particularly for small sample sizes, the PWM method has advantages over the ML method, since PWM estimates show less dispersion around the true value (smaller mean squared error). However, as the sample size increases the difference becomes less pronounced. They also consider the simple method of moments, for which they conclude that it is preferable when $\gamma < 0$. On the other hand, Rootzén and Tajvidi (1997) show that for heavy tailed data with $\gamma \geq 0.5$ the PWM method gives seriously biased parameter estimates whereas ML estimates are consistent. Maximum likelihood has the further attraction to the statistician that models can be easily extended to encompass regression relationships between data and other explanatory variables.

For $\gamma > -0.5$ (which includes all heavy tailed applications) it can be shown (Smith, 1985) that maximum likelihood regularity conditions are fulfilled and that maximum likelihood estimates $(\hat{\gamma}_u, \hat{\sigma}_u)$ based on a sample of N_u excesses (over the threshold u) are asymptotically normally distributed (Hosking and Wallis, 1987).

Specifically we have

$$\sqrt{N_u}\begin{pmatrix}\hat{\gamma}_u \\ \hat{\sigma}_u\end{pmatrix} \rightarrow Normal\left[\begin{pmatrix}\gamma \\ \sigma\end{pmatrix}, \begin{pmatrix}(1+\gamma)^2 & \sigma(1+\gamma) \\ \sigma(1+\gamma) & 2\sigma^2(1+\gamma)\end{pmatrix}\right], \text{ in distribution as } u \rightarrow x_F .$$

This result enables us to calculate approximate standard errors for our maximum likelihood estimates.

A graphical method of estimation (Davison and Smith, 1990) is suggested by the following relation (property of GPD also mentioned in chapter 2). If $\gamma < 1$, then for $u < x_F$, it holds that

$$e(u) = E\left(X - u \middle| X > u\right) = \frac{\sigma + \gamma u}{1-\gamma}, \qquad \sigma + \gamma u > 0,$$

which suggests a 'mean residual life' plot in which the mean observed excess over u is plotted against u. If the generalized Pareto assumption is valid then the plot should follow a straight line with intercept $\sigma/(1-\gamma)$ and slope $\gamma/(1-\gamma)$; this suggests both graphical estimates of γ and σ and a goodness-of-fit test based on the linearity of the plot.

For points in the tail of the distribution ($x \geq u$) we note that

$$F(x) = P\{X \leq x\} = \left(1 - P\{X \leq u\}\right)F_u(x-u) + P\{X \leq u\},$$

where F_u is the conditional d.f. of X, given X>u.

We now know that we can estimate $F_u(x-u)$ by $G_{\gamma,u,\sigma}(x)$ for u large. We can also estimate $P\{X \le u\}$ from the data by $F_n(u)$, the empirical distribution function evaluated at u. Thus for $x \ge u$ we can use the tail estimate

$$\hat{F}(x) = (1 - F_n(u))G_{\gamma,u,\sigma}(x) + F_n(u)$$

to approximate the distribution function F(x). It can be shown that $\hat{F}(x)$ is also a generalized Pareto distribution, with the same shape parameter γ, but with scale parameter $\tilde{\sigma} = \sigma(1 - F_n(u))^{\gamma}$ and location parameter $\tilde{\mu} = u - \tilde{\sigma}\left((1 - F_n(u))^{-\gamma} - 1\right) / \gamma$.

Thus, the POT estimator of x_p is obtained by inverting the tail estimation formula $\hat{F}(x)$ given above and substituting unknown parameters of the GPD by estimates $\hat{\gamma}$ and $\hat{\sigma}$ to get

$$\hat{x}_p = \hat{F}^{\leftarrow}(p) = G^{-1}_{\hat{\gamma},u,\hat{\sigma}}\left(\frac{p - F_n(u)}{1 - F_n(u)}\right) = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left(\left(\frac{1-p}{1 - F_n(u)}\right)^{-\hat{\gamma}} - 1\right).$$

If we write N_u for the number of exceedances of the threshold u and n for the total number of realizations we have from the distribution F, our quantile estimator is

$$\hat{x}_p = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left(\left(\frac{n}{N_u}(1-p)\right)^{-\hat{\gamma}} - 1\right).$$

An important practical problem is the choice of the level u of the excesses. This is analogous to the problem of choosing k (number of upper order statistics) in the previous estimators. There are theoretical suggestions on how to do this, based on compromise between bias and variance – a higher level can be expected to give less bias, but instead gives fewer excesses, and hence a higher variance. However, these suggestions don’t quite solve the problem in practice. Practical aid can be provided by QQ plots, mean excess plots and on experiments with different levels u. If the model produces very different results for different choices of u, the results obviously should be viewed with more caution (Rootzén and Tajvidi, 1997).

Threshold methods as they are called have been developed by hydrologists over the last 30 years under the acronym POT. Davison and Smith (1990) describe models for exceedances and apply them to data-sets of river flows and wave heights. In addition they illustrate the ability of these models to combine several data-sets using regression methods, to deal with serially dependent and seasonal data, though they also emphasize the sensitivity of ML and PWM estimators to the few largest observations. Moreover, Rootzén and Tajvidi (1997) provide methods for handling of trends and incorporating exogenous available information (data) in the context of POT models.

Recently, POT models have attracted the interest of actuaries. Indeed, the interpretation of GPD as limiting d.f. of the scaled excesses over a high threshold, seems to fit perfectly the rationale of excess-of-loss reinsurance. So, lately, a large literature of POT methods and applications in actuarial issues has been developed. Rootzén and Tajvidi (1997) apply such models in data of wind storm losses, while McNeil (1997) and McNeil and Saladin (1997) make use of GPD and POT methods to estimate the tails of loss severity distributions that are necessary for pricing and positioning high-excess loss layers in reinsurance. An application in claims of fire insurance is presented in Embrechts et al. (1997), who also review POT methodology from a ‘point process’ point of view. Another application in Greek insurance data is given by Stamoulis (1999).