# 2

# HEAVY TAIL DISTRIBUTIONS

Motivated by the instances of extreme events and heavy tail distributions encountered in the first chapter, we present the most important theoretical results underpinning the estimation of the probabilities of these extreme and rare events. The basics of extreme value theory are presented as they pertain to estimation and risk management of extremes observed in financial applications. Our goal is to explain the connection between the generalized extreme value distributions and the generalized Pareto distributions, and illustrate the implementation of the theory into a set of practical tools for the detection and estimation of heavy tail distributions. In preparation for some of the applications considered later in the book, the chapter concludes with a discussion of measures of risk, both from a theoretical and a practical point of view.

## 2.1 A Primer on Extreme Value Theory

We present the parametric families of Pareto and extreme value distributions, very much in the spirit of the parametric families discussed in Chap. 1, and we show how the properties of the latter can be used to detect and identify the characteristics of the former.

### 2.1.1 Empirical Evidence of Extreme Events

We already argued that histograms and kernel density estimators could not give a good account of the tail properties of distributions, and we insisted that Q-Q plots offered the best graphical way to get a reasonable feeling for these properties. We emphasize one more time the non-normality of the distribution of daily financial

R. Carmona, *Statistical Analysis of Financial Data in R*, Springer Texts in Statistics,
DOI 10.1007/978-1-4614-8788-3_2, 

returns by considering their extreme values. Since we do not plan to give a precise mathematical definition of an extreme value, we shall simply say that a value is extreme if its distance to the mean location of the data (as given by the mean for example) is large when measured in standard deviation units, say greater that three or four standard deviations. For the purpose of illustration, we consider the daily log-returns on the S&P 500 index. Their values are encapsulated in the numeric vector `DSPLRet` included in the library `Rsafd`. Using the functions `mean` and `sd`, we compute the mean and the standard deviation of the daily log-returns.

```
> mean(DSPLRet)
[1] -0.0002729406
> sd(DSPLRet)
[1] 0.009727974
```

Looking at the sequential plot of the daily log-return (as reproduced in the right pane of Fig. 1.20) we notice a few very large negative values. Looking more closely at the largest of these down-moves we see that:

```
> min(DSPLRet)
[1] -0.2289972
> (min(DSPLRet)-mean(DSPLRet))/sd(DSPLRet)
[1]  -23.56813
```

which shows that this down move was over

$$23 \textbf{ standard deviations away from the mean}$$

daily move! So much for the normal distribution as a model for the daily moves of this index. The log-return on this single day of October 1987, as well as many others since then (though less dramatic in sizes) cannot be accounted for if the Gaussian distribution is used as a model for the daily log-returns. The tails of the normal distribution are too thin to produce such extreme values. However, other families of distributions could be used instead, and stable or Pareto distributions have been proposed with reasonable success. Pareto distributions are studied in detail in this chapter. For the time being, it suffices to say that, like Pareto distributions, stable distributions have polynomial tails, and moreover, they have useful scaling properties. However, their usefulness as statistical models for heavy tail distribution is limited by the fact that the rates of polynomial decay of their densities are restricted to an interval. Moreover, their scaling properties are of very little use since at least in the first part of the book, we are mostly interested in marginal distributions of financial returns, and hence we rarely use dynamical models involving time evolution of prices. Finally, the main shortcoming of the stable distributions is the lack of a closed form formula for the density and/or the cdf. The Cauchy distribution, is the only exception. Recall formula (1.13) for the definition of the Cauchy distribution which is sometime used as an alternative to the Gaussian distribution in the presence of extreme values. Indeed, like the Gaussian density, it is bell-shaped, but unlike the Gaussian density, its tails are so *thick* that the moments of the distribution such as the mathematical expectation (or mean) and the variance do not even exist.

The theory of probability distributions giving rise to unusually large numbers of extremes in each sample is called the theory of extreme-value distributions, or extreme-value theory. It is a well developed mathematical theory. The remainder of this chapter is devoted to an informal presentation of the most fundamental facts of this theory. For the purpose of illustration we demonstrate the practical implementations in the library `Rsafd`, providing versatile tools to fit heavy tail distributions to data, and generate random Monte Carlo samples from the fitted distributions.

### 2.1.2 **Pareto Distributions**

We first introduce a class of distributions which will play a fundamental role in our modeling of heavy tails. The present subsection could have been included in the previous chapter and provide one more example of a family of distributions. However, because of its pivotal role in the theory presented in this chapter, we chose to introduce it here.

#### *2.1.2.1* ***Ordinary Pareto Distributions***

The classical Pareto distribution is a distribution on the positive axis $[0, \infty)$ (i.e. the distribution of a positive random variable) with density given by the formula

$$f_\alpha(x) = \begin{cases} (1 + \frac{x}{\alpha})^{-(1+\alpha)} = \frac{1}{(1+x/\alpha)^{1+\alpha}} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

for some positive real number $\alpha > 0$. Like the exponential and lognormal distributions, this distribution has only one tail extending to $+\infty$. For this reason, it is often called a *one-sided* Pareto distribution. The above definition of the one-sided Pareto distribution can be found in many probability textbooks. For geosciences applications, especially in hydrology where heavy tail distributions were introduced first in order to estimate the frequencies of floods, the Pareto distributions are parameterized by $\alpha$. However for some strange reason, in financial applications, these distributions are parameterized by $\xi = 1/\alpha$ which is called the shape parameter of the distribution. Both parametrizations are implemented in the library `Rsafd`, but as we concentrate on the analysis of financial data, we shall use the $\xi$ – parameterization in this book. This choice is *passed* to the routines of the library `Rsafd` by setting the parameter `SHAPE.XI` to `TRUE`. For the sake of convenience, we restate the definition of the classical Pareto distribution using the shape parameter $\xi$.

$$f_\xi(x) = \begin{cases} (1 + \xi x)^{-(1+1/\xi)} = \frac{1}{(1+\xi x)^{1+1/\xi}} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

We shall discuss later the role of the shape parameter $\xi$, and when we do, we shall emphasize that even though $\xi$ will have to be a non-negative number in most of the applications we are interested in, from a mathematical point of view, we can

extend the definition of the Pareto distribution to include negative values of the shape parameter $\xi$. See below for details.

In any case, it is easy to compute the cdf of an ordinary Pareto distribution in closed form. Indeed straightforward integration gives:

$$F_\xi(x) = \int_{-\infty}^{x} f_\xi(x')dx' = \begin{cases} 1 - (1+\xi x)^{-1/\xi} = 1 - \frac{1}{(1+\xi x)^{1/\xi}} & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Changing location and scale, (remember our discussion of affine transformations), we can define and study ordinary (one-sided) Pareto distributions with location parameter $m \in \mathbb{R}$ and scale parameter $\lambda > 0$. Such a distribution is supported by the half line $[m, \infty)$ (i.e. it is the distribution of a random variable which is always greater than or equal to $m$). Its density will be denoted by $f_{m,\lambda,\xi}$. It is given by the formula:

$$f_{m,\lambda,\xi}(x) = \begin{cases} \frac{1}{\lambda}\left(1 + \frac{\xi}{\lambda}(x-m)\right)^{-(1+1/\xi)} & \text{if } x \geq m, \\ 0 & \text{otherwise.} \end{cases}$$

As before, we can compute the corresponding cdf. It is given by:

$$F_{m,\lambda,\xi}(x) = \begin{cases} 1 - (1 + \xi\frac{x-m}{\lambda})^{-1/\xi} = 1 - \frac{1}{(1+\xi\frac{x-m}{\lambda})^{1/\xi}} & \text{if } x > m, \\ 0 & \text{otherwise.} \end{cases}$$

#### 2.1.2.2 *More General Shape Parameters*

We now consider a first generalization of the parametric family of one-sided Pareto distributions which we shall call Generalized One-Sided Pareto distributions, GOSPD for short. It still relies on three parameters: a location parameter $m$, a scale parameter $\lambda$ and a shape parameter $\xi$. The cumulative distribution function of a GOSPD is given by

$$F_{m,\lambda,\xi}(x) = \begin{cases} 1 - \left(1 + \xi\frac{x-m}{\lambda}\right)^{-1/\xi} & \text{for } \xi \neq 0, \\ 1 - \exp\left\{-\frac{x-m}{\lambda}\right\} & \text{for } \xi = 0. \end{cases} \tag{2.3}$$

the above formulae defining the GOSPD on the domains:

$$\begin{array}{ll} m < x \leq m - \lambda/\xi & \text{for } \xi < 0, \\ m < x \leq \infty & \text{for } \xi \geq 0. \end{array}$$

In other words, we extended the family of one-sided Pareto distributions to include distributions with a negative shape parameter $\xi$. This is done for the sake of generality. It will not be used in the financial applications we consider in this book.

Notice that if $\xi > 0$, the generalized Pareto distribution with cdf $F_{m,\lambda,\xi}$ is nothing but the distribution of a random variable $m + \lambda X$ where $X$ has the ordinary Pareto distribution with parameter $\alpha$ provided we set $\xi = 1/\alpha$ as shape parameter.

Notice also that the case $\xi = 0$, $m = 0$ corresponds to the exponential distribution with scale parameter $\lambda$. In general, the case $\xi = 0$ corresponds to an exponential distribution with scale $\lambda$ *shifted* by the amount $m$, i.e. to the distribution of a random variable $X + m$ where $X \sim E(1/\lambda)$.

#### *2.1.2.3* ***Existence of Moments (or Lack Thereof)***

Another important fact concerning the size of the tail of a one sided Pareto distribution (generalized or not) is given by the existence (or lack thereof) of moments. Indeed, the above definition implies that, if $X \sim F_{m,\lambda,\xi}$ with $\xi \geq 0$, then

$$\mathbb{E}\{|X|^p\} < \infty \Leftrightarrow p < \frac{1}{\xi}. \tag{2.4}$$

Here are a few consequences for a non-negative random variable $X$ with a one-sided GOSPD.

- If $\xi = 0$, $X$ has moments of all orders, i.e. $\mathbb{E}\{X^p\} < \infty$ for all $p > 0$;
- The mean of $X$ exists (i.e. $\mathbb{E}\{X\} < \infty$) if and only if $\xi < 1$;
- The variance of $X$ exists if and only if $\xi < 0.5$.

Figure 2.1 shows the graphs of the densities of three one-sided Pareto distributions with default values $m = 0$ and $\lambda = 1$ for the location and scale parameters, and values $\xi = 0.2$, $\xi = 1.6$ and $\xi = 2.5$ for the shape parameter. The plots were produced with the following commands.

```
> X <- seq(from=-.2,to=15,length=5000)
> plot(X,dpareto(X,xi=.2),type="l",ylab="GPD densities",
                                        ylim=c(-.05,1.1))
> points(X,dpareto(X,xi=1.6),type="l",lty=3)
> points(X,dpareto(X,xi=2.5),type="l",lty=5)
> abline(h=0); abline(v=0)
```

**Remark.** The number of finite moments of a distribution is a good indication of the *thickness* of its tail. This number has been estimated for the marginal distribution of financial returns over different periods ranging from minutes, to days, weeks, months, ... and there is a heated debate concerning the values of these estimates in the so-called eocnophysics community. Indeed, it is claimed by some that this number of finite moments is universal across financial indices and asset classes. Others use self-similarity arguments to claim that this exponent should not change with time horizon, and that it should remain the same when computed with returns over 1 day, 1 week, 1 month, ... The rationale behind the universality of this exponent is beyond the scope of this book. However, we shall give examples (both in the text and in the problem sets) indicating that this universality conjecture does not stand some of the empirical analyzes made possible by the tools presented in this book.

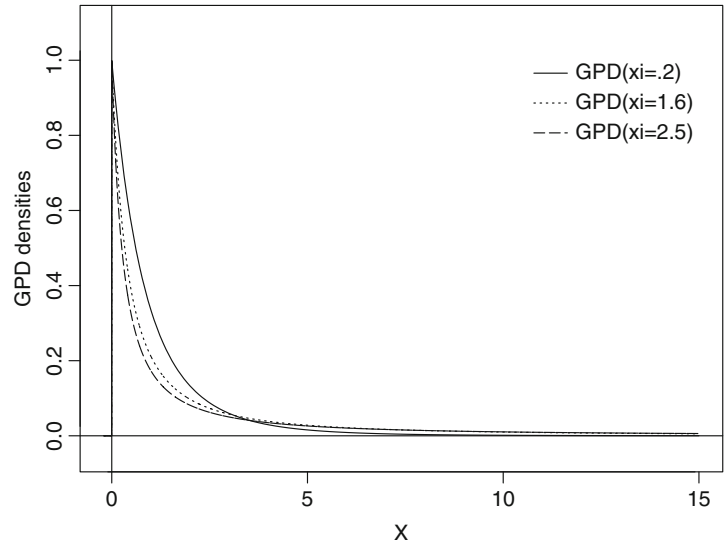


**Fig. 2.1.** One-sided Pareto densities with $m = 0$ and $\lambda = 1$ for the location and scale parameters, and values $\xi = 0.2$, $\xi = 1.6$ and $\xi = 2.5$ for the shape parameter

#### *2.1.2.4* ***Implementation***

As in the case of the classical distributions considered earlier, recall Table 1.2, on can compute values of the density, quantile, and cumulative distribution functions, as well as generating random samples with a set of functions adhering to the naming convention used in `R`. They are listed in Table 2.1

| Distribution | Random samples | Density | cdf | Quantiles |
|---|---|---|---|---|
| One-sided Pareto | `rpareto` | `dpareto` | `ppareto` | `qpareto` |

**Table 2.1.** `Rsafd` commands for the manipulation of one-sided Pareto distributions

#### *2.1.2.5* ***(Two-Sided) Generalized Pareto Distributions (GPD)***

We now define the class of (heavy tail) distributions which we fit to sample data exhibiting thick tails as detected by empirical Q-Q plots. At an intuitive level, our fitting procedures will search for heavy tails (typically densities with inverse polynomial decays) at $+\infty$, and $-\infty$ in the case of a left tail. Roughly speaking, these distributions should behave like

- The distribution of a **one-sided Pareto** random variable to the right of a specific threshold;

- The distribution of the negative of a **one-sided Pareto** random variable to the left of a specific threshold;
- **Nothing special in between**.

To be more specific, these distributions will be characterized by

- A location parameter $m_+$, a scale parameter $\lambda_+$ and a shape parameter $\xi_+$ specifying the one-sided Pareto distribution which applies to the right of the threshold $m_+$;
- A location parameter $m_-$, a scale parameter $\lambda_-$ and a shape parameter $\xi_-$ specifying the one-sided Pareto distribution which applies to the left of the threshold $m_-$ whenever the distribution has a left tail;
- Any distribution in the interval $[m_-, m_+]$.

Clearly, estimating such a distribution amounts to the estimation of the three parameters (possibly six when the distribution has tails extending to both $+\infty$ and $-\infty$) of the one sided Pareto distribution(s), and to the estimation of the density in between the thresholds. The latter will be done by a plain histogram.

The class of (possibly two-sided) Generalized Pareto Distribution (GPD for short) defined above is used in all the applications of extreme value theory considered in this book. The mathematical results we state and use for GPDs hold for a slightly more general class of distributions, namely those distributions with densities at $+\infty$ and/or $-\infty$ which, up to a slowly varying function (concept which we define later), behave like inverse powers.

### 2.1.3 Tidbits of Extreme Value Theory

There are several ways to investigate the statistical properties of extremes. The classical approach is based on the analysis of the statistics of the maxima over large blocks of data. It is most elegant mathematically, and we briefly review it below. However, because it requires large data samples, and involves much too often inefficient computations, the *block maxima approach* fell out of grace with practitioners who prefer relying on *threshold exceedance models* which lead to a more efficient use of limited data. We present the former first.

#### *2.1.3.1* ***The Fisher-Tippett Theorem***

As usual we start from a sample of values

$$x_1, x_2, \cdots\cdots, x_n, \cdots$$

which we envision as realizations of independent identically distributed random variables $X_1, X_2, \cdots\cdots, X_n, \cdots$ with a common distribution which we try to estimate.

*Remark 1.* Most of the results reviewed in this chapter remain valid without this independence assumption. Indeed, under various forms of dependence between the $X_j$'s,

similar conclusions can be reached. These extensions have great practical relevance, as real life data, and especially financial returns, are not strictly independent from one period to the next. However, we refrain from considering these generalizations by fear that their technical nature may obscure the ideas underpinning the theory.

As earlier, we use numerical statistics computed from the data in order to infer properties of the common distribution of these random variables. The limit theorems discussed in the first chapter are involved with the limiting behavior of the partial sums $S_n = X_1 + \cdots + X_n$ for large values of $n$. In particular, the Central Limit Theorem (CLT) states that

$$\lim_{n\to\infty} \mathbb{P}\left\{\frac{S_n - m_n}{\lambda_n} \leq x\right\} = \Phi(x), \qquad x \in \mathbb{R} \tag{2.5}$$

provided we define the normalizing (centering and scaling) constants $m_n$ and $\lambda_n > 0$ as

$$m_n = n\mu \qquad \text{and} \qquad \lambda_n = \sigma\sqrt{n}$$

where $\mu$ and $\sigma$ denote the mean and the standard deviation of the common distribution of the $X_j$'s. Extreme Value Theory (EVT for short) is concerned with the search of centering and scaling constants $m_n$ and $\lambda_n > 0$ for which limiting results of the form (2.5) hold for some limiting distribution functions $\Psi(x)$ when one replaces the partial sums $S_n$ by partial maxima

$$M_n = \max\{X_1, \cdots, X_n\}. \tag{2.6}$$

Obviously, switching from partial sums to maxima shifts the emphasis from aggregation to extremes.

*Remark 2.* The theory presented below is geared toward the analysis of upper tails of statistical distributions as it is formulated in terms of maxima of random samples. Obviously, similar results hold true for minima, and the same theory can be used for the analysis of lower tails of statistical distributions. For the sake of simplicity, we focus our discussion on results on maxima of sequences of random variables, even though we shall eventually turn the results of this theory into computing tools for the analysis of both upper and lower tails of statistical distributions.

The cornerstone of the block maxima approach is the following theoretical result known as the Gnedenko or Fisher-Tippett theorem.

**Theorem 1.** *If the cdf*

$$x \hookrightarrow \mathbb{P}\left\{\frac{M_n - m_n}{\lambda_n} \leq x\right\}$$

*converges as $n \to \infty$ toward a (non-degenerate) cdf for some normalizing sequences $\{m_n\}_n$ and $\{\lambda_n\}_n$ of centering and positive scaling constants, then the limiting distribution necessarily belongs to the family of Generalized Extreme Values (GEV for short) distributions defined below in formula (2.7).*

Before we give such a definition, we present a couple of enlightening examples, for which the result of this theorem can be checked with elementary calculus.

**The Case of the Exponential Distribution.** If the $X_j$ are independent random variables with the exponential distribution with rate $r > 0$, then $F_X(x) = 1 - e^{-rx}$ for $x \geq 0$, and the cdf of normalized $M_n$ is equal to

$$F_{(M_n - m_n)/\lambda_n}(x) = F_X(m_n + \lambda_n x)^n = (1 - \exp[-(m_n + \lambda_n x)])^n$$

for $x > -m_n/\lambda_n$, so that with the choices $m_n = (\log n)/r$ and $\lambda_n = 1/r$ for the normalizing constants we get

$$\lim_{n\to\infty} F_{(M_n - m_n)/\lambda_n}(x) = \lim_{n\to\infty} (1 - \frac{1}{n}\exp[-x])^n = 1 - \exp(-e^{-x})$$

for all $x \in \mathbb{R}$ since $-\log n = -m_n/\lambda_n$. This limiting distribution is known as the Gumbel distribution.

**The Case of the Ordinary Pareto Distribution.** Using the ordinary Pareto distribution with shape parameter $\xi > 0$ instead of the exponential distribution, we can still illustrate the result of the Fisher-Tippett-Gnedenko theorem with explicit computations. Indeed, if we choose the centering and the scaling constants $m_n$ and $\lambda_n$ as $m_n = n^\xi - 1$ and $\lambda_n = \xi n^\xi$, then:

$$F_{(M_n - m_n)/\lambda_n}(x) = F_X(m_n + \lambda_n x)^n = \left(1 - \frac{1}{n}(1 + \frac{x}{\alpha})^{-\alpha}\right)^n$$

for $\alpha = 1/\xi$ and $x > \alpha(-1 + n^{-1/\alpha})$, and consequently

$$\lim_{n\to\infty} F_{(M_n - m_n)/\lambda_n}(x) = \lim_{n\to\infty} \left(1 - \frac{1}{n}(1 + \frac{x}{\alpha})^{-\alpha}\right)^n = \exp\left[-(1 + \frac{x}{\alpha})^{-\alpha}\right]$$

for $x > -\alpha$. This distribution is known as the Fréchet distribution.

#### *2.1.3.2* ***Generalized Extreme Value Distributions (EVD)***

The families of extreme value distributions (EVD for short) which have been studied in the classical statistical literature comprise the Gumbel distribution (also known as EVI distribution), the Fréchet distribution (also known as EVII distribution), and the Weibull distribution (also known as EVIII distribution). These three distribution families can be combined into a single parametric family which is usually called the Generalized Extreme Value (GEV) distribution family. Its cumulative distribution function is given by the following formula:

$$G_{m,\lambda,\xi}(x) = \begin{cases} \exp\left[-\left(1 + \frac{\xi(x-m)}{\lambda}\right)^{-1/\xi}\right] & \text{for } \xi \neq 0, \\ \exp\left[-\left(e^{-\frac{(x-m)}{\lambda}}\right)\right] & \text{for } \xi = 0. \end{cases} \tag{2.7}$$

A GEV distribution is characterized by three parameters: a location parameter $m$, a scale parameter $\lambda > 0$, and a shape parameter $\xi$. In the above formula it is assumed that:

$$\begin{array}{ll} -\infty < x \leq m - \lambda/\xi & \text{for } \xi < 0, \\ -\infty < x < \infty & \text{for } \xi = 0, \\ m - \lambda/\xi \leq x < \infty & \text{for } \xi > 0. \end{array}$$

The Gumbel distribution corresponds to the case $\xi = 0$, the Fréchet distribution corresponds to the case $\xi > 0$, while the Weibull distribution corresponds to the case $\xi < 0$.

Figure 2.2 shows the graphs of the densities of three GEV distributions with default values $m = 0$ and $\lambda = 1$ for the location and scale parameters, and values $\xi = 0.2$, $\xi = 0.6$ and $\xi = 1.5$ for the shape parameter. Note that the left hand point of the distribution changes with the parameters. The plots were produced with the following commands.

```
> X <- seq(from=-2.2,to=8,length=5000)
> plot(X,dgev(X,xi=.2),type="l",ylab="GEV densities",
                                                ylim=c(-.05,.9))
> points(X,dgev(X,xi=.6),type="l",lty=3)
> points(X,dgev(X,xi=1.5),type="l",lty=5)
> abline(h=0); abline(v=0)
```

We use the fact that, like in the case of GPDs, the library `Rsafd` provides functions to generate random samples and compute densities, cdfs and quantiles of the GEV distributions. Table 2.2 gives these commands, and as we can see, they follow the

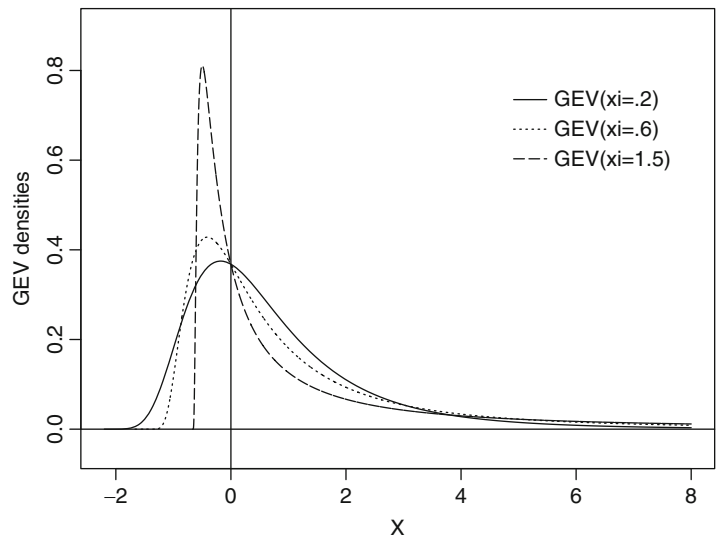


**Fig. 2.2.** Densities of GEV distributions with $m = 0$ and $\lambda = 1$ for the location and scale parameters, and values $\xi = 0.2$, $\xi = 0.6$ and $\xi = 1.5$ for the shape parameter

| Distribution | Random samples | Density | cdf | Quantiles |
|---|---|---|---|---|
| GEV distribution | `rgev` | `dgev` | `pgev` | `qgev` |

**Table 2.2.** Commands for the manipulation of the generalized extreme value distributions

standard `R` naming convention. We do not give plots of densities corresponding to negative values of $\xi$ for they are typically not used in the analysis of financial data.

Formula (2.7) is explicit and simple enough to be inverted explicitly. Doing so, we obtain the following formula for the quantile function of a GEV distribution.

$$Q_{m,\lambda,\xi}(p) = \begin{cases} m - \frac{\lambda}{\xi}\left[1 - (-\log p)^{-\xi}\right] & \text{for } \xi \neq 0, \\ m - \lambda \log(-\log p) & \text{for } \xi = 0. \end{cases} \tag{2.8}$$

This closed form formula makes the generation of random samples from GEV distributions quite easy and efficient. It also shows that estimates of the quantiles of a GEV distribution can be obtained from estimates of the parameters $m$, $\lambda$ and $\xi$ of the distribution by substitution of these estimates in (2.8). We address the estimation of the parameters of a GEV distribution later in this chapter.

*Remark 3.* Roughly speaking, when the Gnedenko, Fisher-Tippett theorem holds, it says that if $n$ is large enough

$$\mathbb{P}\{M_n \leq x\} \approx G_{m,\lambda,\xi}(x)$$

for some set $\{m, \lambda, \xi\}$ of parameters. But since the $X_j$'s are assumed to be independent, we have

$$\mathbb{P}\{M_n \leq x\} = F_X(x)^n$$

and consequently

$$\mathbb{P}\{M_n \leq \pi_p\} = F_X(\pi_p)^n = p^n$$

(if we recall the notation $\pi_p$ for the $p$-quantile of the common distribution of the $X_j$'s) which in turn implies that

$$\pi_p \approx m + \frac{\lambda}{\xi}\left[\left(n \log \frac{1}{p}\right)^{-\xi} - 1\right].$$

This approximation for the quantiles of the distribution of the $X_j$'s should be compared to the formula

$$\pi_p = \mu + \sigma \Phi^{-1}(p)$$

which holds in the Gaussian case. This remark should shed some light on the consequences of the Gnedenko, Fisher-Tippett theorem on tail sizes and properties of the quantiles of the common distribution of the individual random variables $X_j$. We shall revisit the significance of this remark when we discuss the estimation of Value at Risk in the presence of heavy tails.

The Gnedenko, Fisher-Tippett theorem is a very nice theoretical result, but in order to be useful in practical situations, we need to know for which distribution functions $F_X$ it does hold, and even better, we need a hash table pointing out which distributions $F_X$ lead to which GEV distributions. In mathematical terms, this last request is screaming for the identification of the so-called domain of attraction of a given GEV distribution. In other words, given a GEV distribution $G_{m,\lambda,\xi}$, can we characterize the distribution functions $F_X$ for which the distributions of maxima $M_n$ converge (after proper normalization), toward the GEV distribution in question as stated in Theorem 1. This *wishful thinking* is at the origin of several results of great practical usefulness. They go under the names of Gnedenko, Pickands, Balkema and de Haan. We state them in an informal way to avoid being distracted by the technical nature of some of the mathematical assumptions under which these results hold. The interested reader is directed toward the Notes and Complements at the end of the chapter for references of textbooks where these theories are presented in detail.

#### *2.1.3.3* ***Illustration***

The following commands illustrate the convergence of the distribution of block maxima of ordinary Pareto variates toward the Fréchet distribution, fact which we proved rigorously earlier.

```
> XX <- rpareto(1000000,xi=.4)
> dim(XX) <- c(1000,1000)
> MAX <- apply(XX,2,max)
> qqplot(MAX,rgev(1000,xi=.4))
> title("Q-Q Plot Evidence of Block Maxima Convergence,
                                                    xi=.4")
```

The first command creates a sample of size $10^6$ of independent random samples from the ordinary Pareto distribution with location $0$, scale $1$ and shape parameter $\xi = 0.4$. The second command splits this sample into $1{,}000$ blocks of lengths $1{,}000$ each by organizing them in a $1{,}000 \times 1{,}000$ data matrix. The next command computes the maximum of each of these blocks, creating in this way a sample of size $1{,}000$ of maxima $M_n$ with $n = 1{,}000$. The `qqplot` command produces a Q-Q plot of this sample of maxima against a random sample from the GEV distribution with the same shape parameter $\xi = 0.4$. This plot is reproduced in the left pane of Fig. 2.3. The fact that the points line up on a straight line is an indication that we are in the limiting regime of the theorem of Gnedenko, Pickands, Balkema and de Haan. This fact is a particular case of a more general result which we state as a theorem for later reference.

**Theorem 2.** *The distribution of the maxima $M_n$ converge after appropriate centering and scaling, toward a GEV distribution with shape parameter $\xi > 0$ if and only if the common cdf $F_X(x)$ of the $X_j$'s converges toward $1$ as $x \to \infty$ at the rate $x^{-1/\xi}$.*

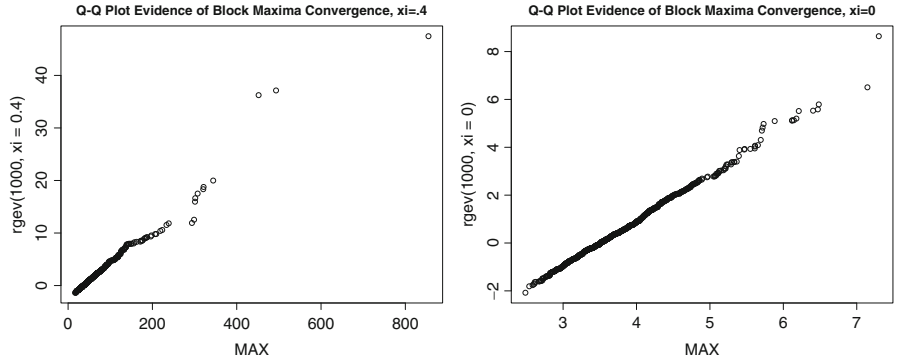


**Fig. 2.3.** Q-Q plots of a sample of 1,000 maxima over 1,000 disjoint blocks in a sample from a GPD, against a sample from the GEV distribution with the same shape parameter. *Left*: case of the Fréchet distribution with $\xi = 0.4$. *Right*: case of the Gumbel distribution (i.e. $\xi = 0$) from an exponential sample with rate $r = 2.0$

The precise mathematical statement is that the function $L(x) = x^{1/\xi}(1 - F_X(x))$ is slowly varying at $+\infty$ in the sense that

$$\lim_{x \to \infty} \frac{L(\lambda x)}{L(x)} = 1, \qquad \text{for all } \lambda > 0.$$

The case of the Gumbel distribution is unfortunately not as clearly delineated by a theoretical result such as Theorem 2 above. We proved in Sect. 2.1.3.1 that the Gumbel distribution was the limit of the distributions of block maxima of increasing sizes of independent exponential variates. As before, we can illustrate this theoretical fact with the help of random simulations.

```
> XX <- rexp(1000000,r=2)
> dim(XX) <- c(1000,1000)
> MAX <- apply(XX,2,max)
> qqplot(MAX,rgev(1000,xi=0.0))
> title("Q-Q Plot Evidence of Block Maxima Convergence,
                                                    xi=0")
```

The resulting plot is reproduced in the right pane of Fig. 2.3, and as before, the fact that the points line up on a straight line is an indication that we are in the limiting regime of the theorem of Gnedenko, Pickands, Balkema and de Haan. The exponential distribution is not the only distribution $F_X$ for which the distributions of the block maxima converge toward the Gumbel distribution. These distributions $F_X$ are not easily characterized. However, it can be proved that they all have finite moments of all orders in the sense that $\mathbb{E}\{X_j^p\} < \infty$ for all $p > 0$. So if the $X_j$'s have a common density $f_X(x)$, then this density goes to zero faster than any inverse polynomial. Exponentials do, but Gaussian and log-normal densities do as well. So in the case $\xi = 0$, the information content of the fact that the limit distribution of the

normalized block maxima is the Gumbel distribution is not as precise: we know that the tail decays faster than any inverse polynomial, but we cannot pin-point the exact rate of decay!

**Remark.** Notice that, because we are interested in extremes, and especially in rare and unexpected large values of financial returns or losses, we shall not consider the Weibull case $\xi < 0$ which forces the distribution to be limited, and prevents the tail from extending to infinity.

#### *2.1.3.4* ***Block Maxima Approach to Extreme Values Estimation***

We now formulate in an algorithmic fashion, the tail size estimation procedure based on the Gnedenko, Pickands, Balkema and de Haan theory which we reviewed earlier and illustrated by examples. This will provide us with a natural transition to the topics presented later on.

- In order to infer properties of the upper tail of the common distribution of the entries of a data sample $x_1, x_2, \cdots\cdots, x_m$, we partition the sample into blocks $B_1$, $B_2$, $\cdots$, $B_M$, and we compute the maxima $M_n = \max_{j \in B_n} x_j$ in each of these blocks.
- Assuming that each block size is large enough, we treat the set $\{M_n\}_n$ of maxima as a sample from a GEV distribution, and assuming that the number of blocks is large enough, we estimate the parameters of this hypothetical GEV distribution from the sample $\{M_n\}_n$
- We infer the size of the tail of the common distribution of the $x_j$'s (in particular the shape parameter $\xi$) from the values of the estimated parameters and the results of the Gnedenko, Pickands, Balkema and de Haan theory.

It is obvious from the second bullet point above that the inference procedure is justified if the block size is large since we rely on an asymptotic result holding in the limit of the block size going to $\infty$. Moreover, the estimation of the parameters of the limiting distribution also requires the blocks to be in large numbers. Having both large blocks, and a large number of maxima, requires a very large data set to start with. This sample size requirement is the major shortcoming of this block maxima method. Band-aids have been suggested, the most natural one being to use overlapping blocks. However, the gain in sample size is compensated by a loss in accuracy since the block maxima are not independent any longer, and as a consequence, the parameter estimation procedure looses efficiency. Quantifying the effects of dependencies due to block overlap as well as in the original data has been a concern of many researchers in the field, and the interested reader is referred to the books mentioned in the Notes and Complements at the end of the chapter.

For the time being, we note that the important second bullet point above stresses the need for procedures capable of estimating the parameters of a GEV distribution. This is the task we tackle next. Then, and only then, will we be able to implement the block maxima method and conclude on specific tail size alternatives.

## 2.2 GEV & GPD Parameter Estimation

The previous section has singled out two distribution families playing an important role in the analysis of extremes and heavy tails: the generalized extreme value and Pareto distributions. It also showed the need for estimating the parameters of these distributions. Maximum likelihood and method of moments are classical statistical procedures frequently used in estimating parameters. This section explains how these methods can be extended to fit these two important parametric distribution families.

### 2.2.1 **The Method of L-Moments**

Because many heavy tail distributions do not have enough finite moments (after all, the Cauchy distribution does not even have a first moment!) the classical method of moments cannot be used to estimate the parameters of GPD and GEV distributions. Keeping with the spirit of this time honored estimation procedure, researchers have devised work-arounds by *renormalizing* the traditional statistical moments in order to get analogs which could be used for data with extreme values. With this simplistic strategy in mind, we introduce the notion of theoretical L-moment.

#### *2.2.1.1* ***Theoretical Definitions***

L-moments are defined in terms of the so-called probability weighted moments. These generalized moments are defined for non-negative random variables $X$ with finite expectations and continuous cdf $F(x)$ in the following way. For each integer $r \geq 0$, the $r$-th probability weighted moment $\alpha_r$ is defined as the number

$$\alpha_r = \mathbb{E}\{XF(X)^r\} = \int_0^\infty x\, F(x)^r\, dF(x), \qquad r = 0, 1, 2, \cdots . \tag{2.9}$$

In other words, in computing the $r$-th probability weighted moment, we sum the possible values $x$ of the random variable $X$, but instead of weighting them by their probability of occurrence, we weight them by this probability times the cdf $F(x)$ raised to the power $r$. Recall that assuming that $X$ is has finite expectation means that

$$\mathbb{E}\{X\} = \int_0^\infty x dF(x) < \infty,$$

which guarantees that all the probability weighted moments $\alpha_r$ make sense as finite numbers since $0 \leq F(x)^r \leq 1$.

As usual we denote the corresponding quantile function by $F^{-1}(x)$, and a simple substitution in the integral appearing in (2.9) gives:

$$\alpha_r = \int_0^1 F^{-1}(y) y^r dy.$$

The L-moments are defined as specific linear combinations of the probability weighted moments with the intent to capture the descriptive features of the distribution in question, namely location, dispersion and other shape parameters. The first few L-moments are defined by the following equations

$$\lambda_1 = \alpha_0 = \int_0^1 F^{-1}(p)\,dp$$
$$\lambda_2 = 2\alpha_1 - \alpha_0 = \int_0^1 F^{-1}(p)(2p-1)\,dp$$
$$\lambda_3 = 6\alpha_2 - 6\alpha_1 + \alpha_0 = \int_0^1 F^{-1}(p)(6p^2-6p+1)\,dp$$
$$\lambda_4 = 20\alpha_3 - 30\alpha_2 + 12\alpha_1 - \alpha_0$$

The coefficients of these linear combinations are nothing but the coefficients of the "shifted Legendre polynomials"

$$P^*_{r-1}(y) = \sum_{k=0}^{r}(-1)^{r-k}\binom{r}{k}\binom{r+k}{k}y^k, \qquad r = 1, 2, \cdots$$

For the sake of definiteness we give the values of the first four Legendre polynomials $P^*_j(y)$:

$$P^*_0(y) = 1$$
$$P^*_1(y) = 2y - 1$$
$$P^*_2(y) = 6y^2 - 6y + 1$$
$$P^*_3(y) = 20y^3 - 30y^2 + 12y - 1$$

An alternative definition of L-moments can be given in terms of order statistics. Such form of the definition will be useful for empirical estimation from data samples. For any given integer $r \geq 1$ and sample $X_1, \cdots, X_r$ of i.i.d. random variables with the same distribution $F$, we use momentarily the notation

$$X_{(1:r)} \leq X_{(2:r)} \leq \cdots\cdots \leq X_{(r:r)}$$

for the order statistics which we usually denote by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r)}$. We use this notation to emphasize the dependence of these order statistics on the sample size. Given these preliminaries, the $r$-th L-moment can be alternatively defined as

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}\mathbb{E}\{X_{((r-k):r)}\}, \qquad r = 1, 2, \cdots \tag{2.10}$$

and using this definition we get the formulae

$$\lambda_1 = \mathbb{E}\{X\}$$
$$\lambda_2 = \frac{1}{2}\left(\mathbb{E}\{X_{(1:2)}\} - \mathbb{E}\{X_{(2:2)}\}\right)$$
$$\lambda_3 = \frac{1}{3}\left(\mathbb{E}\{X_{(1:3)}\} - 2\mathbb{E}\{X_{(2:3)}\} + \mathbb{E}\{X_{(3:3)}\}\right),$$

Descriptive statistics such as skewness and kurtosis play an important role in the analysis of statistical distributions. Since they are defined in terms of moments and their ratios, they have natural analogs in the present framework. An L-moment ratio is a dimensionless quantity defined as the ratio of an L-moment to the second L-moment. L-skewness, $\tau_3$, is the third L-moment ratio,

$$\tau_3 = \frac{\lambda_3}{\lambda_2},$$

and L-kurtosis, $\tau_4$, is the fourth L-moment ratio,

$$\tau_4 = \frac{\lambda_4}{\lambda_2}.$$

**Examples.**

- For the uniform distribution $U(0,1)$ we have

$$\lambda_1 = 1/2, \quad \lambda_2 = 1/6, \quad \tau_3 = 0, \quad \tau_4 = 0.$$

- In the case of the standard normal distribution $N(0,1)$ we have

$$\lambda_1 = 0, \quad \lambda_2 = 1/\sqrt{\pi}, \quad \tau_3 = 0, \quad \tau_4 \approx 0.123.$$

- In the case of the exponential distribution with unit rate we have

$$\lambda_1 = 1, \quad \lambda_2 = 1/2, \quad \tau_3 = 1/3, \quad \tau_4 = 1/6.$$

#### *2.2.1.2 First L-Moments Empirical Estimation*

Given the ordered statistics

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

of a sample $x_1, x_2, \ldots, x_n$ of size $n$, the estimate $l_r$ defined by

$$l_r = \frac{1}{\binom{n}{r}} \sum_{0 \leq i_1 < i_2 < \ldots < i_r \leq n} r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} x_{(i_{r-k})}$$

is an unbiased estimator of the theoretical $r$-th L-moment $\lambda_r$. Moreover, it has been shown that $l_r$ can be computed, from the order statistics as

$$l_r = (-1)^r \sum_{k=0}^{r-1} (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} a_k, \qquad r = 0, 1, 2, \ldots \tag{2.11}$$

where the numbers $a_k$ are the so-called probability weighted moments defined by

$$a_0 = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad a_k = \frac{1}{n}\sum_{j=k+1}^{n} \frac{(j-1)(j-2)\cdots(j-k)}{(n-1)(n-2)\ldots(n-k)} x_{(i)}, \quad k \geq 1. \tag{2.12}$$

Notice that, consistent with our earlier discussion, the coefficients appearing in expression (2.11) are the coefficients of the shifted Legendre polynomials introduced above.

The function `sample.LMOM` gives and implementation of formula (2.12). For the sake of illustration, we compute the L-moments of a Monte Carlo sample of size 1,000 from a GPD.

```
> X <- rpareto(1000)
> sample.LMOM(X)
Mean (l_1) L-mom 2 (l_2)    L-skewness    L-kurtosis
 1.0144528     0.5045187     0.3278689     0.1626310
```

For the sake of comparison we check with the theoretical L-moments of such a GPD. Indeed, since the GPD with location parameter $m = 0$, scale parameter $\lambda = 1$, and shape parameter $\xi = 0$ is nothing but the standard exponential distribution with rate one, we already gave its L-moments L-skewness and L-kurtosis. They are

$$\lambda_1 = 1, \qquad \lambda_2 = \frac{1}{2}, \qquad \tau_3 = \frac{1}{3}, \qquad \tau_4 = \frac{1}{6},$$

which shows that, at least in this case, the estimation procedure gets reasonable values for the parameters. Quite expectedly, the estimates $t_3$ of L-skewness and $t_4$ of L-kurtosis computed by the function `sample.LMOM` are obtained as the ratios $l_3/l_2$ and $l_4/l_2$, respectively.

**Important Remark.** Even though a sample mean can be computed from any sample irrespective of the distribution which governs the generation of the values appearing in the sample, it is used as an estimator, only when the theoretical distribution is at least of order one, namely when the mean actually exists. We recalled these facts in our discussion of the law of large numbers in Chap. 1. In particular, the empirical mean can always be computed for a sample from the Cauchy distribution, however, it cannot have the interpretation of an estimate of the mean in that case. A similar state of affairs holds in the case of L-moments. The empirical estimates introduced in this section can always be computed. However, as we said in their introduction, L-moments make sense only for distributions with a first moment. In particular, when we talk about L-moments of GPDs and GEV distributions, we shall always implicitly assume that $\xi < 1$ so the theoretical moment of order one does exist.

#### *2.2.1.3 **Small Sample Alternative***

Because of the very definition of L-moments, estimation involves the approximation of an integral whose integrand depends upon the entire cdf. It is intuitively

clear that the error produced by approximating the integral using a numerical quadrature method is much smaller than the error due to the approximation of the cdf from sample data when the sample size is small. For that reason, practitioners have searched for alternative estimates which could perform better with small samples.

The following procedure was proven to give good estimates for the L-moments of GPD in the case of small samples when the parameters $\gamma$ and $\delta$ are chosen appropriately. It is based on the notion of *plotting position.* For each integer $n$ (which will be chosen as the size of the sample under study) the plotting positions are defined as the numbers

$$p_i = \frac{i+\gamma}{n+\delta}$$

and the corresponding estimates of the $r$-th L-moments are given by

$$l_r = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{r-1}(-1)^{r-1-k}\binom{r-1}{k}\binom{r-1+k}{k}p_i^k x_i.$$

It has been shown that the plotting position estimators with $\gamma = 0.35$ and $\delta = 0$ produce good approximations of the L-moments for small GPD samples. This method is implemented in the library `Rsafd` by the function `plotting.positions` whose use is illustrated by the following display.

```
> X <- rpareto(50,xi = 0.4)
> PPLM <- plotting.positions(X)
> PPLM
ell_1     ell_2     tau_3     tau_4
2.0628630 1.7630784 0.7029661 0.5526929
> SLM <- sample.LMOM(X)
> SLM
ell_1     ell_2     tau_3     tau_4
2.3626477 1.7845944 0.7196471 0.5867451
```

#### *2.2.1.4* ***Distribution Estimation by the Method of L-Moments***

We now explain how estimates of the L-moments can be used to estimate the parameters of generalized extreme value and Pareto distributions.

#### *2.2.1.5* ***Estimating the Parameters of a GEV Distribution***

We now concentrate on the case of GEV distributions. Recall that, since the existence of L-moments requires that the common distribution of the observations has at least a first moment, we need to restrict ourselves to the case $\xi < 1$. Under this condition, the L-moments of a GEV distribution can be computed in closed form, leading to the following expressions:

$$\lambda_1 = m - \frac{\lambda(1 - \Gamma(1-\xi))}{\xi}, \tag{2.13}$$

$$\lambda_2 = -\frac{\lambda(1 - 2^{\xi})\Gamma(1-\xi)}{\xi}, \tag{2.14}$$

$$\lambda_3 = \frac{\lambda}{\xi}(1 - 3 \cdot 2^{\xi} + 2 \cdot 3^{\xi})\Gamma(1-\xi), \tag{2.15}$$

where $\Gamma(\alpha)$ is the Gamma function whose definition was recalled in (1.12). Taking the ratio of (2.15) to (2.14) we get:

$$\tau_3 = \frac{2(1 - 3^{\xi})}{(1 - 2^{\xi})} - 3.$$

Assuming that the first three L-moments $\lambda_1$, $\lambda_2$ and $\lambda_3$ were estimated as $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\lambda}_3$ from an empirical sample, we set $\hat{\tau}_3 = \hat{\lambda}_3/\hat{\lambda}_2$ and we plug the latter in the above equation in lieu of $\tau_3$. Since the equation so obtained involves only the unknown parameter $\xi$, we can use it to extract a value, say $\hat{\xi}$, for the shape parameter $\xi$. Obviously, this equation cannot be solved in a closed form, so we use a numerical method to do so. Once this is done, the computation of the remaining estimates is straightforward. The estimate of $\hat{\lambda}$ is easily derived from Eq. (2.14),

$$\hat{\lambda} = -\frac{\hat{\lambda}_2\hat{\xi}}{(1 - 2^{\hat{\xi}})\Gamma(1-\hat{\xi})}, \tag{2.16}$$

and after that, $\hat{m}$ is obtained from Eq. (2.13) by

$$\hat{m} = \hat{\lambda}_1 + \frac{\hat{\lambda}}{\hat{\xi}}\left(1 - \Gamma(1-\hat{\xi})\right). \tag{2.17}$$

**Remark.** Since we aim at computing the values of three parameters, we should only need three equations. Not surprisingly, the above methods requires only the knowledge of the first two L-moments $l_1$ and $l_2$ and the L-skewness $\tau_3$.

The above method of L-moment estimation of a GEV distribution is implemented in the function `gev.lmom`. Starting with a set of L-moments (as produced for example by the functions `sample.LMOM` or even the function `plotting.positions` discussed above) this function computes estimates of the three parameters of the GEV distribution suspected to have produced these L-moments. We demonstrate its use with the following simulation example where we first estimate the L-moments from a random sample from a GEV distribution which we choose.

```
> X <- rgev(500, lambda = 3.5, xi = 0.4)
> LMOMX <- sample.LMOM(X)
> LMOMX
ell_1         ell_2         tau_3         tau_4
4.2065658 3.9795450 0.4372888 0.3242249
```

```
> gev.lmom(LMOMX)
 $param.est
m          lambda     xi
0.1417766 3.4847654 0.3781168
```

#### 2.2.1.6 *Estimating the Parameters of a GPD*

In the case of GPDs, different methods are used depending upon whether or not the location parameter $m$ is known. The reason for considering these two alternatives will become clear in the next section. When using the POT method to estimate the size of a tail, the estimation procedure consists in fitting a GPD to the exceedances over an appropriately chosen threshold. By construction, the location parameter of a sample of exceedances is automatically zero. If $m$ is known, the GPD L-moment estimators are:

$$\hat{\xi} = 2 - \frac{l_1}{l_2}, \qquad \text{and} \qquad \hat{\lambda} = \left(\frac{l_1}{l_2} - 1\right) l_1 \,.$$

Notice that, since we assume that $m$ is known, we need to compute values for two parameters only, and hence, two equations are sufficient. In this case, we need only the knowledge of the first two L-moments $l_1$ and $l_2$ to estimate the entire GPD.

If $m$ is unknown, we need to compute three parameters. We expect to need three equations. However, instead of using the L-skewness as in the case of GEV distributions, the GPD L-moment estimation procedure which we implemented in `Rsafd` uses the first two L-moments $l_1$ and $l_2$ and the first order statistics $x_{(1)}$. The resulting estimates are given by the formulae:

$$\hat{\xi} = -\frac{2(n-1)l_2 - n(l_1 - x_{(1)})}{(n-1)l_2 - (l_1 - x_{(1)})}, \quad \hat{\lambda} = (1-\hat{\xi})(2-\hat{\xi})l_2, \text{ and } \hat{m} = x_{(1)} - \frac{\hat{\lambda}}{n-\hat{\xi}},$$

where $x_{(1)}$ is the smallest value of the sample.

The above method of L-moment estimation of a GPD is implemented in the function `gpd.lmom`. Starting with a set of L-moments and a value for the location parameter $m$ or a sample data set (from which the first order statistic will be computed) this function computes estimates of the three parameters of the GPD suspected to have produced these L-moments. As before, we demonstrate its use with a simulation example where we first estimate the L-moments from a random sample from a GPD which we choose. We give two examples, showing the results both when the location argument is provided and when the sample is provided instead.

```
> X<- rpareto(500,xi = 0.4)
> SLM <- sample.LMOM(X)
> gpd.lmom(SLM,location=0)
$param.est
```

```
       m      lambda       xi
0.0000000 0.9178838 0.4531872
 > gpd.lmom(SLM,sample=X)
 $param.est
         m       lambda       xi
0.002611717 0.912422137 0.455593845
```

### 2.2.2 Maximum Likelihood Estimation

We now present the most widely used method of parameter estimation. In the situations of interest, the parameter $\theta$ is multivariate since it comprises the location parameter $m$, the scale parameter $\lambda$ and the shape parameter $\xi$, so $\theta = (m, \lambda, \xi)$. Since explicit formulae for the density functions of GEV distributions and GPDs can be derived in a straightforward manner from the definition expressions we gave in (2.7) and (2.3), the strategy of the classical maximum likelihood estimation seems appropriate. The only slight difference with the classical cases handled by this method is the fact that the domain of definition of the density function $f_\theta$ changes with the parameter. This is a minor hinderance which can be overcome in practice.

#### *2.2.2.1 Likelihood and Log-Likelihood Functions*

We first consider the case of the GEV distributions. For the sake of notation, we give separate formulae for the cases $\xi = 0$ and $\xi \neq 0$. When $\xi = 0$, taking derivatives of both sides of (2.7) gives:

$$g_{m,\lambda,0}(x) = \frac{1}{\lambda} e^{-(x-m)/\lambda} \exp[-e^{-(x-m)/\lambda}] \tag{2.18}$$

which implies that the likelihood of a sample $x_1, \cdots, x_n$ is given by

$$L(m, \lambda | x_1, \cdots, x_n) = \frac{1}{\lambda^n} \exp[-\frac{1}{\lambda} \sum_{i=1}^{n} (x_1 - m)] \exp[-\sum_{i=1}^{n} e^{-(x_i - m)/\lambda}] \tag{2.19}$$

and the corresponding log-likelihood by:

$$\mathcal{L}(m, \lambda | x_1, \cdots, x_n) = -n \log \lambda + nm - \frac{1}{\lambda} \sum_{i=1}^{n} x_1 - \sum_{i=1}^{n} e^{-(x_i - m)/\lambda}. \tag{2.20}$$

The case $\xi \neq 0$ leads to similar computations. The density of the GEV distribution is given by:

$$g_{m,\lambda,\xi}(x) = \frac{1}{\lambda} \left(1 + \frac{\xi}{\lambda}(x - m)\right)^{-(1+1/\xi)} \exp\left[-\left(1 + \frac{\xi}{\lambda}(x - m)\right)^{-1/\xi}\right] \tag{2.21}$$

if $x \leq m - \lambda/\xi$ for $\xi < 0$ or $x \geq m - \lambda/\xi$ for $\xi > 0$, and 0 otherwise. This in turn implies that the likelihood of a sample $x_1, \cdots, x_n$ is given by

$$L(m, \lambda, \xi | x_1, \cdots, x_n) = \frac{1}{\lambda^n} \prod_{i=1}^{n} \left(1 + \frac{\xi}{\lambda}(x_i - m)\right)^{-(1+1/\xi)} \exp\left[-\sum_{i=1}^{n} \left(1 + \frac{\xi}{\lambda}(x_i - m)\right)^{-1/\xi}\right] \quad (2.22)$$

if $\max\{x_1, \cdots, x_n\} \leq m - \lambda/\xi$ for $\xi < 0$ or $\min\{x_1, \cdots, x_n\} \geq m - \lambda/\xi$ for $\xi > 0$, and 0 otherwise. Finally, the corresponding log-likelihood is given by:

$$\mathcal{L}(m, \lambda, \xi | x_1, \cdots, x_n) = -n \log \lambda - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left(1 + \frac{\xi}{\lambda}(x_i - m)\right) - \sum_{i=1}^{n} \left(1 + \frac{\xi}{\lambda}(x_i - m)\right)^{-1/\xi} \quad (2.23)$$

with the same domain restrictions as before. Consequently, maximum likelihood estimates of the parameters of a GEV distribution are obtained by solving the optimization problem

$$(\hat{m}, \hat{\lambda}, \hat{\xi}) = \arg \sup_{\lambda > 0.\ \lambda + \xi(x_i - m) \geq 0,\ i=1,\cdots,n} L(m, \lambda, \xi | x_1, \cdots, x_n) \quad (2.24)$$

The above constraints guarantee that the density is non-negative at the observations $x_1, \cdots, x_n$. Such an optimization problem could have presented difficulties years ago, but with the advent of modern computers and the development of efficient solvers, it can be solved in a very reliable manner on most every platforms. `S` and `R` come with solvers for nonlinear optimization based on quasi-Newton methods. The library `Rsafd` uses these solvers to produce maximum likelihood estimates of the parameters.

Next, we consider the case of the GPDs. As before, we give separate formulae for the cases $\xi = 0$ and $\xi \neq 0$. The case $\xi = 0$ is well known since it reduces to the classical analysis of exponential samples. Indeed, the density function is given by:

$$f_{m,\lambda,0}(x) = \frac{1}{\lambda} e^{-(x-m)/\lambda} \quad (2.25)$$

if $x \geq m$ and 0 otherwise. This implies that the likelihood of a sample $x_1, \cdots, x_n$ is given by

$$L(m, \lambda | x_1, \cdots, x_n) = \frac{1}{\lambda^n} \exp[-\frac{1}{\lambda} \sum_{i=1}^{n} (x_1 - m)] \quad (2.26)$$

if $\min\{x_1, \cdots, x_n\} \geq m$ and 0 otherwise. Hence, the corresponding log-likelihood is given by:

$$\mathcal{L}(m, \lambda | x_1, \cdots, x_n) = -n \log \lambda + \frac{nm}{\lambda} - \frac{1}{\lambda} \sum_{i=1}^{n} x_i \tag{2.27}$$

which leads to the classical maximum likelihood estimates of the location and scale of an exponential sample.

Computations are simpler in the case $\xi \neq 0$. Indeed, taking derivatives on both sides of (2.3) gives a density of the form:

$$f_{m,\lambda,\xi}(x) = \frac{1}{\lambda} \left(1 + \frac{\xi}{\lambda}(x - m)\right)^{-(1+1/\xi)} \tag{2.28}$$

if $x \leq m - \lambda/\xi$ for $\xi < 0$ or $x \geq m - \lambda/\xi$ for $\xi > 0$, and 0 otherwise. This in turn implies that the likelihood of a sample $x_1, \cdots, x_n$ is given by

$$L(m, \lambda, \xi | x_1, \cdots, x_n) = \frac{1}{\lambda^n} \prod_{i=1}^{n} \left(1 + \frac{\xi}{\lambda}(x_i - m)\right)^{-(1+1/\xi)} \tag{2.29}$$

if $\max\{x1, \cdots, x_n\} \leq m - \lambda/\xi$ for $\xi < 0$ or $\min\{x_1, \cdots, x_n\} \geq m - \lambda/\xi$ for $\xi > 0$, and 0 otherwise. Finally, the corresponding log-likelihood is given by:

$$\mathcal{L}(m, \lambda, \xi | x_1, \cdots, x_n) = -n \log \lambda - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{n} \log \left(1 + \frac{\xi}{\lambda}(x_i - m)\right) \tag{2.30}$$

with the same domain restrictions.

#### 2.2.2.2 *MLE of the Parameters of a GPD and GEV Distributions*

Maximum Likelihood Estimates (MLE for short) of the parameters of a GEV distribution and a GPD are provided by the functions `gev.ml` and `gpd.ml`. Since by definition of a maximum likelihood estimate, the result is obtained by solving an optimization problem, one needs to initialize the procedure with a first guess for the set of arguments (i.e. the three parameters of the distribution family). In the `gev.ml` and `gpd.ml` implementations, if no initial guess is provided, a vector of parameter estimates obtained by a different method is used by the function as starting point for the optimization routine attempting to maximize the likelihood. Indeed, if no such argument is specified, L-moment estimates are computed by the functions `gev.ml` and `gpd.ml` and used for initialization purposes. As in the case of L-moment estimation, if the location parameter `m` of a GPD is known, it may be specified, in which case, only the remaining two parameters will be estimated by maximum likelihood.

As before, we demonstrate the use of the functions of the package `Rsafd` with a simulation example where we choose the GEV distribution.

```
> X <- rgev(500, lambda = 3.5, xi = 0.4)
> gev.ml(X)
```

```
 $param.est
         m      lambda          xi
-0.1714924  3.4420736  0.4243908

$converged
[1] TRUE
```

Similarly, in the case of a GPD:

```
 > X <- rpareto(500, lambda = 3.5, xi = 0.4)
 > gpd.ml(X)$param.est
 $param.est
          m      lambda        xi
0.001238288 3.171523526 0.467466969
```

### 2.2.3 **An Example Chosen for Pedagogical Reasons**

It is possible to propose mathematical models for the time evolution of the PCS index. We described one of them in the Notes & Complements at the end of Chap. 1. These models are most often quite sophisticated, and they are difficult to fit and use in practice. Instead of aiming at a theory of the dynamics of the index, a less ambitious program is to consider the value of the index on any given day, and to perform a static analysis of its marginal distribution. This gives us a chance to illustrate how one uses the tools introduced above to fit a Pareto distribution to the data. The purpose of this exercise is to emphasize the limitations of a blind application of the general theory, and to motivate the modifications introduced and implemented in the following section on semi-parametric estimation.

The Q-Q plots produced in Chap. 1 clearly showed that the upper tail of the PCS index data was heavier than the tail of the exponential distribution. We use the function `gpd.lmom` to fit a GPD to the `PCS.index`, and we print the estimated location, scale and shape parameters with the following commands:

```
> PCS.lmom <- gpd.lmom(PCS.index)$param.est
> PCS.lmom
          m      lambda          xi
0.06824616 0.66521009 0.71314021
```

To visualize the properties of the fit we choose to plot the histogram of the original data set `PCS.index` together with the density of the estimated GPD.

```
> hist(PCS.index,breaks=25,density=20,freq=F)
> X <- seq(from=-1,to=160,length=1000)
> points(X,dpareto(X,m=PCS.lmom[1],lambda=PCS.lmom[2],
                       xi=PCS.lmom[3]),type="l")
```

The plot is given in Fig. 2.4. The fit does not look very good, especially in the left part of the plot where the histogram shows significant positive values.

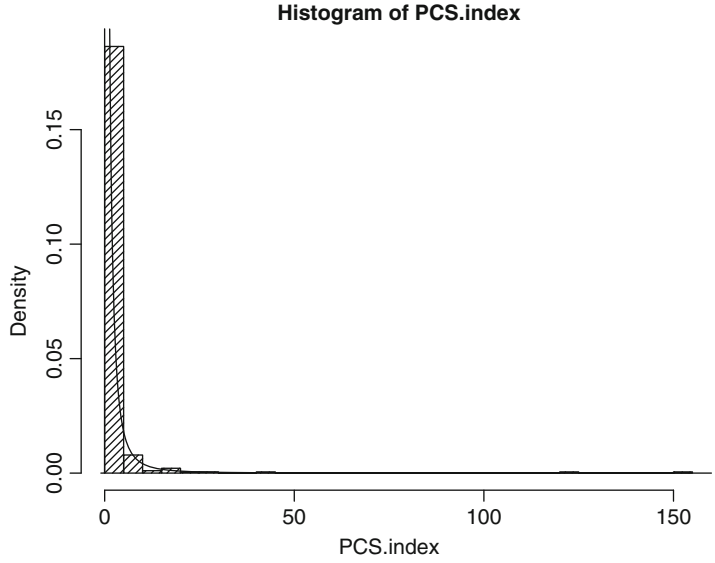


**Fig. 2.4.** Histogram of the PCS index, together with the density of the Pareto distribution estimated by the method of L-moments

As we explained in the first chapter, histograms and density plots do not give a clear picture of what is happening *in the tail.* So in order to check the goodness of the fit in the tail, we generate a large random sample from the distribution fitted to the data, and we produce a Q-Q plot of the Monte Carlo sample against the original data set `PCS.index`.

```
> PCS.rlmom <- rpareto(n=10000,m=PCS.lmom[1],
                       lambda=PCS.lmom[2],xi=PCS.lmom[3])
> qqplot(PCS.index,PCS.rlmom)
```

The result is reproduced in the left pane of Fig. 2.5.

As with Fig. 2.4, the result is disappointing. However, the plot in Fig. 2.5 points to a possible reason for the poor fit. Up until the large values, the quantiles of the simulated sample seem to align reasonably well with the quantiles of `PCS.index`. However the last quantile – quantile point being out of line seems to indicate that the thickness of the tail was not captured properly by the estimated distribution. It happens often that moment estimates are not as good as maximum likelihood estimates, so knowing that, we compute the GPD estimate produced by the function `gpd.ml`.

```
> PCS.ml <- gpd.ml(PCS.index)
> PCS.ml <-  PCS.ml$param.est
> PCS.ml
        m    lambda        xi
0.0700000 0.7095752 0.6359470
```

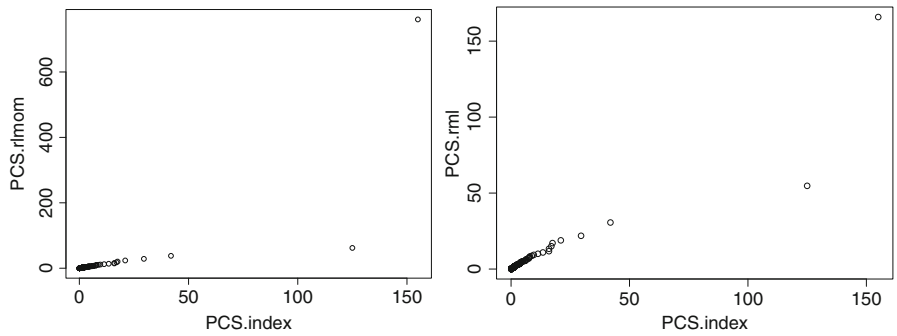


**Fig. 2.5.** Q-Q-plot of the sample of the one-sided Pareto distribution generated from the parameters estimated with the method of L-moments (*left*) and by maximum likelihood (*right*) against the original PCS index data

As before, we can try the same random generation experiment as before, using the maximum likelihood estimates of the location, scale and shape parameter instead.

```
> PCS.rml <- rpareto(n=10000,m=PCS.ml[1],lambda=PCS.ml[2],
                                                  xi=PCS.ml[3])
> qqplot(PCS.index,PCS.rml)
```

The result shown in the right pane of Fig. 2.5 are much better, strikingly good in fact. But a warning is in order as these results are very much dependent upon the actual random sample generated, and as such, they vary from one Monte Carlo experiment to another.

The following final remark uses the example of the PCS index given above to explain some of the reasons why one should not be surprised by the poor performance of these statistical estimation procedures.

**Final Remark.** Fitting a parametric distribution family as specific as the Pareto family cannot accommodate at the same time the features of the bulk of the data (i.e. the small values of the index in the example treated above), and of the tail (i.e. the extremely large values of the index). It is quite conceivable that the tail of the distribution has a polynomial decay while the left part of the distribution behaves in a non-polynomial way. The estimation procedure tries to find one single set of parameters to fit all the different parts of the distribution, and the resulting compromise often penalizes the tail because by definition, the latter is represented by a small number of data values. This conundrum is at the root of the semi-parametric approach presented in the next section.

### 2.2.4 **Implementation of the Block-Maxima Method**

We closed the previous section with a discussion of the block-maxima method, and we explained why its implementation required the estimation of the parameters of a

GEV distribution. The maximum likelihood method and the method of L-moments can now be brought to bear to solve a problem which we could not resolve then. We use the function `block.max` of the package `Rsafd` to illustrate the performance of the method on simulated data sets. Knowing the true shape parameter, and being able to afford as large a data set as needed make it possible to illustrate the shortcomings of the block-maxima method.

We first generate a sample of size $n = 5{,}000$ from the Pareto distribution with shape parameter $\xi = 0.3$. In the context of daily financial data, such a sample size would correspond approximately to 20 years worth of daily data.

Besides the data vector, the function main parameters of `block.max` are the variable `overlap` which is 0 by default and which should be an integer between 0 and 50, and the common length of all the block passed to the function as parameter `block.size` which needs to be an integer greater than or equal to 100. We study the influence of these parameters separately.

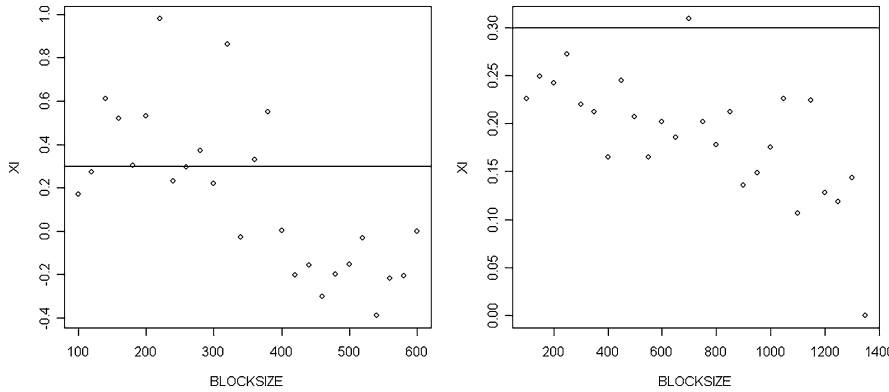


**Fig. 2.6.** Block-maxima shape parameter estimate as a function of the block size, for a sample of size 5,000 (*left*) and 50,000 (*right*). In both cases the true parameter was $\xi = 0.3$

The left pane of Fig. 2.6 shows the estimates $\hat{\xi}$ given by the block-maxima method when non-overlapping blocks are used. We vary the common length of the blocks from 100 to 600 by increments of 20, and for each fixed block size, we compute and plot the estimate of the shape parameter. The resulting points are scattered, indicating that the method fails in most cases: either the block size is not large enough, or when it is large enough, we do not have enough blocks to get a good estimate of the GEV shape parameter. The right pane gives the plot of the shape parameter estimates for the same block sizes when the data sample is 50,000. The results are obviously much better (notice the differences of the ticks on the vertical axes). However, if the data were arising from daily measurements, one would have to collect 200 years worth of data to have such a sample size. Needless to say, this does not happen often in financial applications.

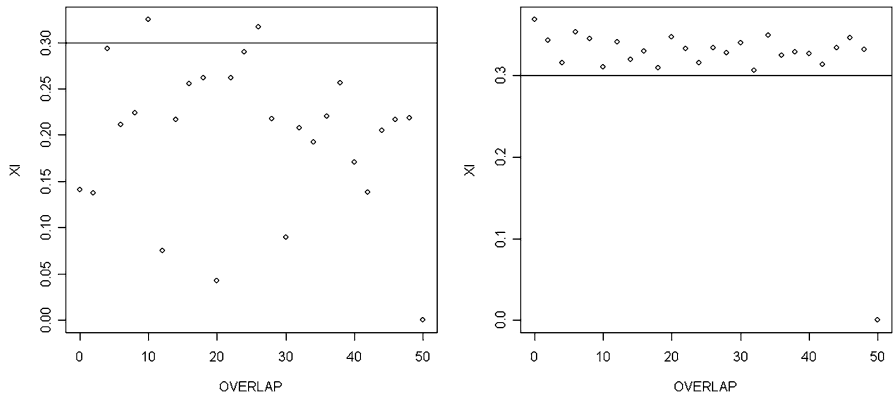


**Fig. 2.7.** Block-maxima shape parameter estimate as a function of the block overlap, for a sample of size 5,000 (*left*) and 50,000 (*right*). In both cases we plotted a horizontal line at the true value $\xi = 0.3$ of the parameter

The left pane of Fig. 2.7 shows the estimates $\hat{\xi}$ given by the block-maxima method when overlapping blocks are used. We vary the overlap of the blocks from 0 to 50 by increments of 2, and for each overlap, we compute and plot the estimate of the shape parameter $\xi$. The results are not very good, for essentially the same reasons as before. They improve dramatically when we increase the sample size to 50,000 as shown in the right pane.

## 2.3 Semi Parametric Estimation

This section is the culmination of the density estimation procedures introduced in this chapter. It combines the benefits of the non-parametric estimation when data are plentiful, and of the parametric methods to estimate generalized Pareto distributions in the tails when the latter are heavier than normal.

### 2.3.1 **Threshold Exceedances**

As before, we consider a sample $x_1, \cdots, x_n$ from the distribution of a random variable $X$ with cdf $F$ which we try to *estimate.* In most insurance and financial applications, $F$ is the loss distribution of a portfolio of contracts.

For any given level $\ell$, we define the excess distribution over the threshold $\ell$ as the conditional distribution of $X - \ell$ given $X > \ell$. The corresponding cdf is given by

$$F_\ell(x) = \mathbb{P}\{X - \ell | \leq xX > \ell\} = \frac{F(x+\ell) - F(\ell)}{1 - F(\ell)}, \qquad x \geq 0.$$

The mean of $F_\ell$ is called the mean excess over the level $\ell$, and viewed as a function of the level $\ell$, it is called the mean excess function.

$$\ell \hookrightarrow e(\ell) = \mathbb{E}\{X - \ell | X > \ell\}.$$

In the next section, we study risk measures computed from loss distributions. When $X$ represents a loss, the mean excess function gives the expected loss above a given level $\ell$.

**Examples.** When $F$ is an exponential distribution, the memoryless property implies that the excess distribution $F_\ell$ does not depend upon the level $\ell$ since $F_\ell(x) \equiv F(x)$. The excess distribution can also be computed explicitly in the case of GPD's. Indeed, for any $\ell$ we have:

$$\begin{aligned} F_\ell(x) &= \frac{F_{m,\lambda,\xi}(x+\ell) - F_{m,\lambda,\xi}(\ell)}{1 - F_{m,\lambda,\xi}(\ell)} \\ &= \frac{(1+\xi(x+\ell-m)/\lambda)^{-1/\xi} - (1+\xi(\ell-m)/\lambda)^{-1/\xi}}{(1+\xi(x+\ell-m)/\lambda)^{-1/\xi}} \\ &= 1 - \left[\frac{1+\xi(x+\ell-m)/\lambda}{1+\xi(\ell-m)/\lambda}\right]^{-1/\xi} \\ &= F_{m',\lambda',\xi'}(x) \end{aligned}$$

with $m' = 0$, $\lambda' = \lambda + \xi(\ell - m)$ and $\xi' = \xi$. So for a GPD, the excess distribution is another GPD located at 0 and with the same shape parameter $\xi$. This stability property is a remarkable property of the GPD's. Notice also that the mean excess function can only be defined when $\xi < 1$. It can be shown that in this case, it is linear in $\ell$ since

$$e(\ell) = \frac{\xi}{1-\xi}\ell + \text{cst} \tag{2.31}$$

as can be seen by a direct integration from the explicit form of $F_\ell(x)$ given above.

**Empirical Estimation.** Given a sample $x_1, \cdots, x_n$ and a level $\ell$, we denote by $n_\ell$ the number of $x_j$'s which are greater than $\ell$, i.e. the number of exceedances above the level $\ell$, and we denote by $x_1^{(e,\ell)}, \cdots, x_{n_\ell}^{(e,\ell)}$ the actual overshoots over the level $\ell$ obtained by subtracting $\ell$ from the $n_\ell$ values $x_j$'s which are greater than $\ell$. In this way, we can think of $x_1^{(e,\ell)}, \cdots, x_{n_\ell}^{(e,\ell)}$ as a sample from the excess distribution above $\ell$ and the excess function $e(\ell)$ can be estimated by the empirical mean

$$\widehat{e_n}(\ell) = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} x_j^{(e,\ell)} \tag{2.32}$$

Formula (2.31) shows that, when the sample $x_1^{(e,\ell)}, \cdots, x_{n_\ell}^{(e,\ell)}$ comes from a GPD, then the empirical estimate of the mean excess function given by formula (2.32) should be approximately **linear in the level** $\ell$.

We now state in a rather informal way, the main theoretical result of this section. It is known as the Balkema-de Hann-Pickands theorem. It is in the same vein as the

main result of the block maxima approach presented in the previous section. However, the practical estimation procedure which it leads to makes a more parsimonious use of the data, hence the reason of its success with practitioners.

**Theorem 3.** *The distribution of the block maxima $M_n$ converge toward a GEV with shape parameter $\xi$ if and only if the excess distribution $F_\ell(x)$ over a level $\ell$ converges uniformly in $x$ as $\ell$ increases, toward a GPD with shape parameter $\xi$ and a scale parameter possibly varying with $\ell$.*

The above result is at the root of the Peaks Over Threshold (POT for short) method described below. In particular, it has the following consequence. If the excess distribution $F_\ell(x)$ is essentially a GPD with shape parameter $\xi$, then the mean excess function over the levels higher than $\ell$ should be approximately linear. This justifies the use of the *mean excess plot* as a diagnostic for the POT approach. This plot is obtained by graphing the couples

$$(x_j, \widehat{e_n}(x_j))_{j=1,\cdots,n} \tag{2.33}$$

of the empirical estimate of the mean excess function computed at the sample values. Except for the expected fact that the right most points may be randomly varying because of the smaller number of exceedances used to compute the mean excess estimate, this plot should show a linear trend in case the Balkema-de Hann-Pickands theorem holds. We shall use this graphic diagnostic extensively in what follows.

#### *2.3.1.1* ***Peaks Over Threshold Modelling***

We now explain how the theoretical facts reviewed above can be used to estimate the tail of a distribution function which behaves like a GPD beyond a certain threshold. So, if we remember the disappointing results obtained in Sect. 2.2.3 when we tried to fit a GPD to the whole PCS data, the main difference is that instead of forcing a GPD on the entire range of the random samples, we only fit a GPD to the large values in the sample. This seemingly innocent difference will turn out to have drastic effects on the usefulness of the estimates.

As usual we describe the statistical procedure starting from a sample $x_1, \cdots, x_n$ of realizations of random variables $X_1, \cdots, X_n$ which we assume to be independent and with the same cdf $F$. Our main assumption will be that the Balkema-de Haan-Pickands result stated above as Theorem 3 applies to this distribution. In other words, this common distribution gives rise to a distribution of block maxima converging toward a GEV with shape parameter $\xi$. The theory presented in the previous section says that this shape parameter $\xi$ determines the size of the upper tail of the distribution, and controls the size and the frequency of the extreme values occurrences.

- The first step is a graphical check that the method is appropriate for the data at hand. Based on the rationale identified in the previous subsection, we check that we are dealing with a generalized Pareto distribution by checking that the mean excess plot is mostly linear (except may be for the few right most points).

- Now, according to Balkema-de Haan-Pickands theorem, for each threshold $\ell$ high enough, the sample $x_1^{(e,\ell)}, \cdots, x_{n_\ell}^{(e,\ell)}$ of exceedances over the level $\ell$ form a sample from a distribution which is uniformly close to a GPD with shape parameter $\xi$ and a scale parameter $\lambda = \lambda(\ell)$ which may depend upon $\ell$.
- Using this sample of exceedances, we estimate the shape and scale parameters $\xi$ and $\lambda$ by the method of L-moments, or by maximum likelihood. Let us denote by $\hat{\xi}$ the estimate of the shape parameter and by $\hat{\lambda}$ the estimate of the scale parameter. Note that the location estimate is irrelevant since we consider only exceedances, so the location parameter is necessarily 0.
- The final estimate of the unknown cdf above the level $\ell$ is then given by the formula

$$\hat{F}(x) = 1 - \frac{n_\ell}{n}\left(1 + \hat{\xi}\frac{x-\ell}{\hat{\lambda}}\right)^{1/\hat{\xi}}, \qquad x \geq \ell \tag{2.34}$$

The rationale for this estimate is the following. If $x \geq \ell$ we have

$$\begin{aligned} 1 - F(x) &= \mathbb{P}\{X > x | X \geq \ell\}\mathbb{P}\{X \geq \ell\} \\ &= \mathbb{P}\{X - \ell > x - \ell | X \geq \ell\}(1 - F(\ell)) \\ &= (1 - F_\ell(x - \ell))(1 - F(\ell)) \end{aligned} \tag{2.35}$$

from which the choice of formula (2.34) is now clear. The factor $1 - F(\ell)$ appearing in the right hand side of (2.35) is estimated empirically by the ratio $n_\ell/n$ giving the empirical frequency of the exceedances. This is usually a reasonable estimate since by definition of the tail of a distribution, most of the data values in the sample are below the level $\ell$. The estimate of the first factor of (2.35) is taken from the fact that the sample of exceedances above $\ell$ is a sample from a GPD whose shape and scale parameters have been estimated.

This estimation strategy is extended in the next subsection to handle the estimation of entire distributions.

### 2.3.2 **Semi Parametric Estimation**

After reviewing the classical parametric and non-parametric methods of density estimation, we introduce our method of choice to estimate heavy tail distributions.

By definition of the tails of a distribution, most of the sample values do bundle up in the *center* or *bulk* of the distribution. On this part of the domain, the density and the cumulative distribution functions can efficiently be estimated by non-parametric methods. Appealing again to the definition of the tails of a distribution, one knows that observations in the tails, even if they end up being extreme, may not be plentiful, and as a consequence, parametric estimation methods will make a better use of the scarce data. This is exactly the philosophy promoted by the POT approach: identify a threshold to the left of which the distribution can be estimated non-parametrically, and beyond which it is estimated parametrically.

**Remark: Going beyond the Data**. Another advantage of the parametric estimation of the tails is the possibility to *go beyond the data.* Indeed, non-parametric methods are limited by the scope of the data. Except for minor leakage produced by the smoothing of kernel-like methods, (especially when the bandwidth is too large) a non-parametric estimate of a distribution will not assign probability to values which are not part of the sample (i.e. *have not been observed in the past*). So in terms of extreme events, nothing more extreme than what has already been observed will carry any probability: so non-parametric methods cannot foresee events more extreme than those that have already been observed. Parametric methods can. Indeed, having used the data at hand to estimate the shape parameter $\xi$, the density estimate will extend beyond the most extreme observed data values, and extreme events will be given a positive probability (depending on the estimate of $\xi$) even if they never occurred in the past.

**Identifying the Tails**. Before getting into the gory details of the estimation procedures, the first question to address is:

*where does the center of the distribution end, and where do the tails start?*

As in most cases, common sense will be required to make sure that poor choices do not bias the estimates of the shape parameters in a significant way. The POT implementation of `fit.gpd` can be used without having to make this delicate choice. If values of the thresholds are not provided, the program uses values which guarantee that the tail contains $15\,\%$ of the points when the data set is small, and $150$ observations when the original data set is large. But we should be clear on the fact that there is no panacea, and that any automatic threshold choice will fail from time to time. The solution we recommend is to use the plots provided by the function `shape.plot` to choose the thresholds.

As we shall see throughout the remainder of this chapter, the results of many analyzes depend upon the choices of these thresholds. So we encourage the reader to get a sense of the sensitivities of his or her results with respect to the choices of the thresholds. To this effect we propose an enlightening simulation example in Problem 2.8 below. It was designed for pedagogical reasons to illustrate the possible biases in the estimates of the tail shape parameters with poor choices of the thresholds. We show that the POT method can fail in two ways: either by not including enough observations in the tail (this is typically the case when the absolute value of the threshold is too large), or by including too many observations from the center of the distribution in the tails when the absolute value of the threshold is not large enough. This simulation example also shows that the graphical diagnostics offered by the function `shape.plot` are our best weapon against the dangers of poor threshold choices.

#### *2.3.2.1* ***The POT Strategy***

We first recall the main steps in this strategy to estimate the tail(s) of a distribution. We concentrate on the upper tail for the sake of definiteness. Let $X$ be a random

variable with distribution $F_X$, let $x_1, x_2, \ldots, x_n$, be a random sample of observations of $X$, and let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be its order statistics. We assume that we already gathered evidence (usually from descriptive statistics and plots such as Q-Q plots) that the tail of the distribution is of a *generalized Pareto* type. Not only does that implies that $F_X$ is in the domain of attraction of a GEV distribution in the sense that the distributions of properly normalized block maxima converge toward a GEV distribution, but the POT theory does also apply. In other words, we can use the fact that, provided the level $\ell$ is appropriately chosen, the conditional distribution of excesses over $\ell$ can be closely approximated by a GPD:

$$F_\ell(x) = \mathbb{P}\{X \leq x + \ell | X > \ell\} = \frac{F_X(\ell + x) - F_X(\ell)}{1 - F_X(\ell)} \sim F_{m=0,\lambda(\ell),\xi}(x).$$

As explained in formula (2.34), given a threshold level $\ell$, $F_X(x)$ is estimated by a non-parametric empirical cdf for $x \leq \ell$, and by a GPD for $x > \ell$. To be specific, we choose the estimate

$$\hat{F}(x) = \begin{cases} \frac{i-0.5}{n} & \text{if } x_{(i)} \leq x < x_{(i+1)} \text{ and } x \leq \ell, \\ 1 - \frac{n_\ell}{n}\left(1 - \frac{\hat{\xi}(x-\ell)}{\hat{a}}\right)^{1/\hat{k}} & \text{if } x > \ell, \end{cases}$$

where $n_\ell$ is the number of points greater than $\ell$ in the sample. This estimate is implemented in the function `fit.gpd` of the package `Rsafd`. Strictly speaking, the above non-parametric part is implemented in the way described above when the optional parameter `linear` is set to `FALSE`. If `linear = TRUE`, then $\hat{F}(x)$ is linearly interpolated for $x \leq \ell$.

Moreover, as we can see from a quick look at the explanations in the help file, this function can handle distributions with two tails. In that case, instead of one single level $\ell$, we need to identify two thresholds which we call `upper` and `lower`. The non-parametric estimation of the cdf is now restricted to the interval limited by the thresholds `lower` and `upper`. Furthermore, the exceedances above the threshold `upper` are used as described above to estimate the shape parameter of the upper tail, while similarly, the excursions below the threshold `lower` are treated in the same way to estimate the shape parameter of the lower tail. Obviously the two shape parameter estimates can be different, this is a result of the flexibility of the method of estimation.

### 2.3.3 The Example of the PCS Index Revisited

We now revisit the estimation of the distribution of `PCS.index` already considered in Sect. 2.2.3 where we attempted to fit a one-sided ordinary Pareto distribution. Here, we try to fit a GPD with the tools of the library `Rsafd`. As noticed at the start of Sect. 2.3.2, the first order of business is to choose a cut-off value to separate the tail from the bulk of the distribution. This choice should be driven by the following two seemingly contradictory requirements. The cut-off point should be large enough

so that the behavior of the tail is homogeneous beyond this threshold. But at the same time, it should not be too large, as we need enough data points beyond the threshold to guarantee a reasonable estimation of $\xi$ by the POT method. For the sake of the discussion, we make a specific choice without justification, leaving the discussion of a reasonable procedure to choose the threshold to our explanations about the function `shape.plot` later in this subsection.

```
> PCS.est <- fit.gpd(PCS.index, tail="upper",upper=4)
```

This command creates an object `PCS.est` of class `gpd` which contains all we need to know about the estimation results. As a side effect, it also generates a plot. We reproduce the latter in Fig. 2.8. We shall also give examples of ways to extract information from the objects thus created. We used the parameter `tail="upper"` because the distribution does not have a lower/left tail (remember that all the values of the index are positive). According to our earlier discussion of the mean excess plots, the fact that the points appearing in the left part of the plot in Fig. 2.8 are essentially in a straight line is an indication that a generalized Pareto distribution may be appropriate.

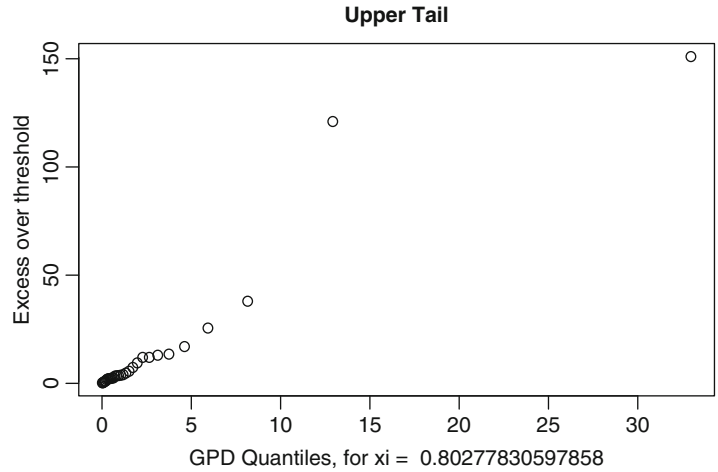


**Fig. 2.8.** Mean excess plot from the use of the function `fit.gpd` on the PCS index data

Plotting an object of class `gpd` with the command `plot(PCS.est)` would produce four plots: a plot of the excesses, a plot of the tail of the underlying distribution, and also a scatterplot and a Q-Q plot of the residuals. Since we are mostly interested in the second of these plots, we use instead the command `tailplot` to visualize the quality of the fit. For the sake of illustration we run the commands:

```
> tailplot(PCS.est)
```

and reproduce the result in the bottom pane in Fig. 2.9. Notice that the vertical axis is for the *survival function* $1 - F(x)$, instead of the cdf $F(x)$. The use of the

option `optlog` forces `R` to use the natural scale instead of the logarithmic scale which is used by default. This is done for the first plot reproduced on the top of Fig. 2.9. Unfortunately, the curve sticks very early to the horizontal axis and it is extremely difficult to properly quantify the quality of the fit. In other words, this plot is not very instructive. It was given for illustration purposes only. Plotting both the values of the index, and the values of the survival function on a logarithmic scale makes it easier to see how well (or possibly how poorly) the fitted distribution gives an account of the data. The second command (using the default value of the parameter `optlog`) gives the plot of the survival function in logarithmic scales. Both plots show that the fit is very good. Our next inquiry concerns the value of the shape parameter $\xi$. Remember that this number is what controls the *power decay* of the density in the tail of the distribution at $\infty$. The choice of a threshold indicating the beginning of the tail, forces an estimate of $\xi$. The value of this estimate is printed on the plot produced by the function `fit.gpd` and it can be read off Fig. 2.8. Since the location parameter is passed to the function as the (upper) threshold determining the beginning of the tail, only two parameters are fitted. The estimated values for the parameters are included in the object `PCS.est` and can be extracted in the following way:

```
> PCS.est@upper.par.ests
   lambda        xi
4.5014927 0.8027783,
```

the command `$upper.par.ests[2]` giving the single shape parameter `xi`.

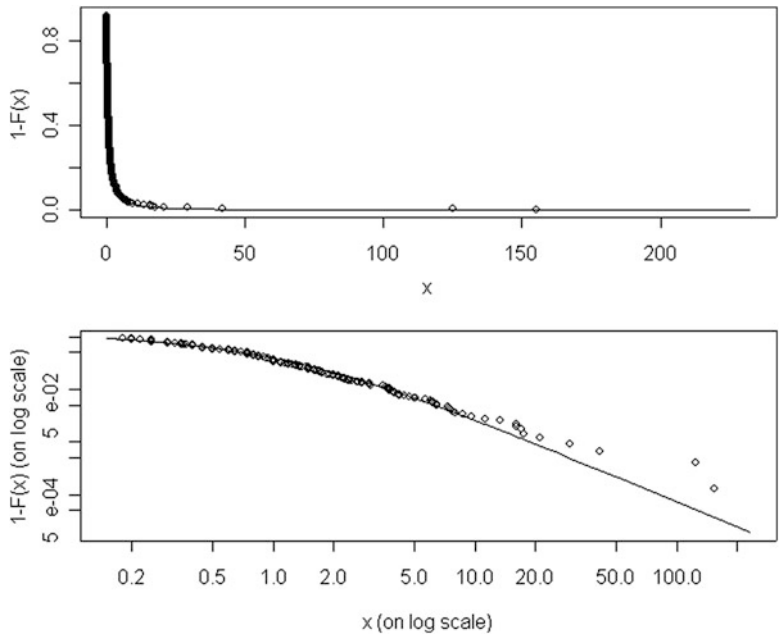


**Fig. 2.9.** Plot of the tail of the GPD fitted to the PCS data together with the empirical tail given by the actual data points, in the natural scale (*top*) and in logarithmic scale (*bottom*)

Changing the value of the threshold `upper` in the call of the function `fit.gpd` changes the value of the estimate of $\xi$, so we should be concerned with the stability of the result: we would not want to rely on a procedure that is too sensitive to small changes in the choice of the threshold. Indeed, since there is no obvious way to choose this threshold, the result of the estimation of the shape parameter should remain robust to reasonable errors/variations in the choice of this threshold. The best way to check that this is indeed the case is graphical. It relies on the use of the function `shape.plot` which gives a plot of the estimates of the shape parameter $\xi$ as they change with the values of the threshold used to produce these estimates. The command:

```
> shape.plot(PCS.index)
```

produces a plot of all the different estimates of $\xi$ which can be obtained by varying the threshold parameter `upper`. This plot is reproduced in Fig. 2.10. The leftmost part of the plot should be ignored because, if the threshold is too small, too much of the bulk of the data (which should be included in the center of the distribution) contributes to the estimate of the tail, biasing the result. The rightmost part of the plot should be ignored as well because, if the threshold is too large, not enough points contribute to the estimate. A horizontal axis was added to the upper part of the plot to give the percentage of points included in the estimate. This information is extremely

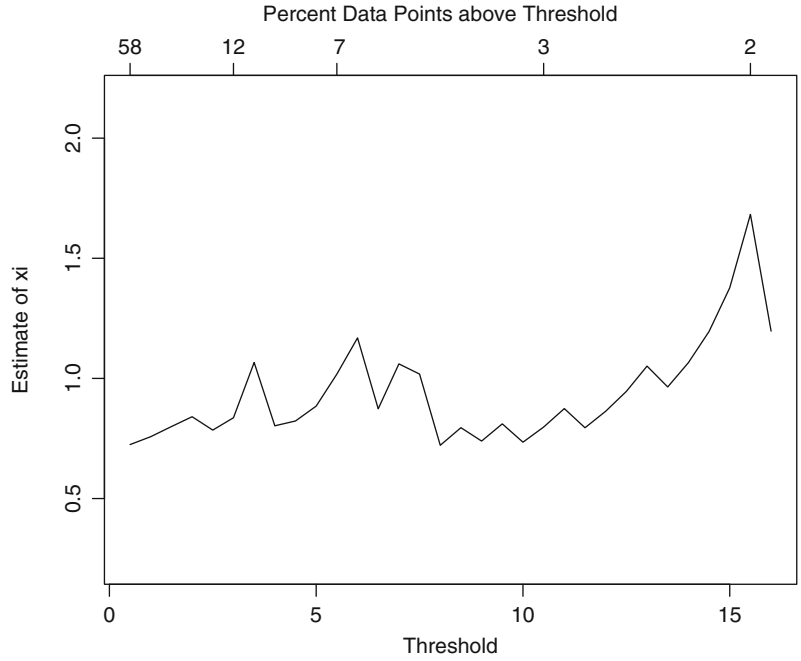


**Fig. 2.10.** PCS data shape parameter $\xi$ (*vertical axis*) as function of the upper threshold (*lower horizontal axis*) and the corresponding percentage of point in the subsequently defined tail (*upper horizontal axis*)

useful when it comes to deciding whether one should take seriously some of the estimates of $\xi$ which appear on the left and right ends of the plot. The central part of the graph should be essentially horizontal (though not always a straight line) when the empirical distribution of the data can be reasonably well explained by a GPD. This is indeed the case in the present situation, and a value of $\xi = 0.8$ seems to be a reasonable estimate for the intercept of a horizontal line fitting the central part of the graph. Also from this plot we see that the particular choice `upper=4` we made for the location threshold gives a tail containing approximately $10\,\%$ of the sample points, which gives a sample of size 38 (since the size of the vector `PCS.index` is 381) for the estimation of the scale and shape parameters $\lambda$ and $\xi$ which is reasonable.

Our last test of the efficiency of our extreme value toolbox is crucial for risk analysis and stress testing of stochastic systems suspected to carry extreme rare events. It addresses the following important question: can we generate random samples from a generalized Pareto distribution fitted to a data set? The function `qgpd` was included in the library `Rsafd` for the sole purpose of answering this question. If `X` is a vector of numerical values, and `gpd.object` is a gpd.object, then `qgpd(gpd.object,X)` gives the vector of the values computed at the entries of `X`, of the quantile function (i.e. the inverse of the cdf) of the GPD whose characteristics are given by `gpd.object`. If we recall our discussion in Chap. 1 of the way Monte Carlo samples from a given distribution can be generated if one can evaluate the quantile function, we see that, replacing the numerical vector `X` by a sample from the uniform distribution will give a sample from the desired distribution. We now show how this is done in the case of the PCS index. The command

```
> PCSsim <- qgpd(PCS.est,runif(length(PCS.index)))
```

produces a random sample of the same size as the original PCS data from the GPD fitted to the data. The plots produced by the following commands are reproduced in Fig. 2.11.

```
> par(mfrow=c(1,2))
> plot(PCS[,1],PCS.index)
> plot(PCS[,1],PCSsim)
> par(mfrow=c(1,1))
```

When the `R` function `plot` is called with a couple of numerical vectors with the same numbers of rows say $n$, as arguments, it produces a plot of $n$ points whose coordinates are the entries found in the rows of the two vectors. Putting next to each other the sequential plots of the original data, and of this simulation, shows that our simulated sample seems to have the same statistical features as the original data. This claim is not the result of a rigorous test, but at this stage, we shall consider ourselves as satisfied! See nevertheless Problem 2.4 for an attempt at quantifying the goodness of fit.

### 2.3.4 The Example of the Weekly S&P Returns

The following analysis is very similar to the previous one, the main difference being the presence of two tails instead of one. We include it in the text to show the details

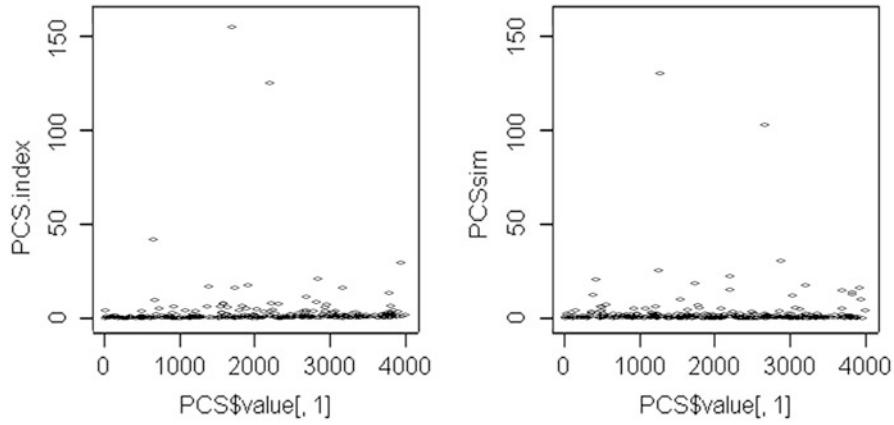


**Fig. 2.11.** PCS original data (*left*) and simulated sample (*right*)

of all the steps necessary to perform a complete analysis in this case, i.e. when the distribution is unbounded both from above *and* below. We choose the thresholds designating the end points of the tails from the output of the function `shape.plot`. From the results of the command:

```
> shape.plot(WSPLRet, tail="two")
```

reproduced in Fig. 2.12 we see that $0.02$ and $-0.02$ are reasonable choices for the upper and lower thresholds to be fed to the function `fit.gpd`. So the fundamental object of the fitting procedure is obtained using the command:

```
> WSPLRet.est <- fit.gpd(WSPLRet,lower=-0.02,upper=0.02)
```

Notice also that the shape plots in Fig. 2.12 confirm the differences in the sizes of the left and right tails: the frequency and the size of the negative weekly log-returns are not the same as the positive ones. The threshold parameters `lower` and `upper` do not have to be given "opposite" values, i.e. they do not need to have the same absolute values. This is likely to be the case for symmetric distributions, but it does not have to be the case in general. Finally, notice that we did not have to set the parameter `one.tail` by including `one.tail=FALSE` in the command because this is done by default. The above command produced the two plots given in Fig. 2.13.

Both sets of points appear to be essentially in a straight line, so a generalized Pareto distribution is a reasonable guess. Notice that the two estimates of the shape parameter $\xi$ are not the same. The estimates obtained from the particular choices of the threshold parameters `lower` and `upper` are $\xi_{left} = 0.24$ and $\xi_{right} = -0.01$. If the distribution is not symmetric, there is no special reason for the two values of $\xi$ to be the same, in other words, there is no particular reason why in general the polynomial decays of the right and left tails should be identical! As before, we can check visually the quality of the fit by superimposing the empirical distribution of the points in the tails onto the theoretical graphs of the tails of the fitted distributions. This is done with the command:

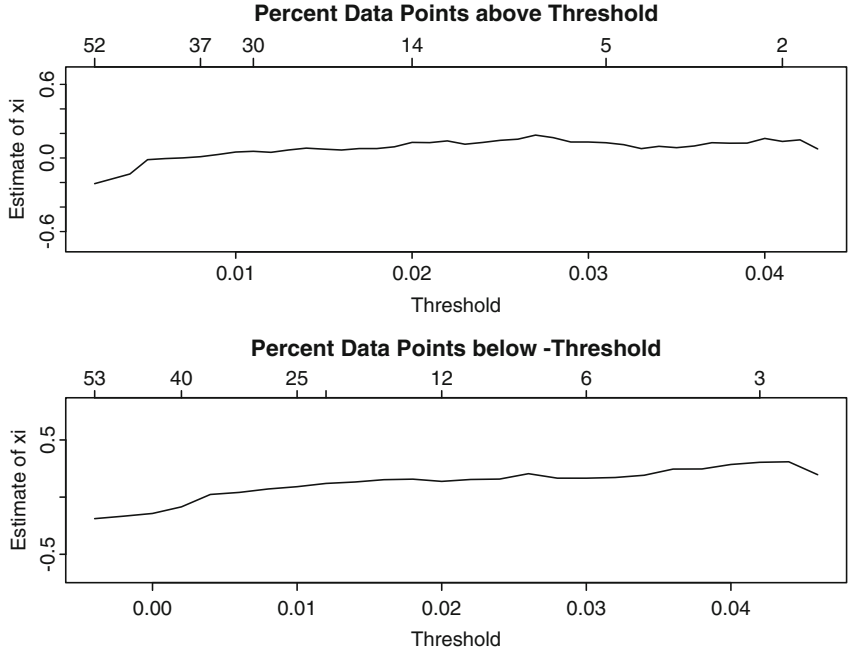


**Fig. 2.12.** Values of the shape parameter $\xi$ for the right tail (*top*) and left tail (*bottom*) of the distribution of the weekly log-returns of the S&P 500 index, as functions of the values of the thresholds marking the ends of the tails

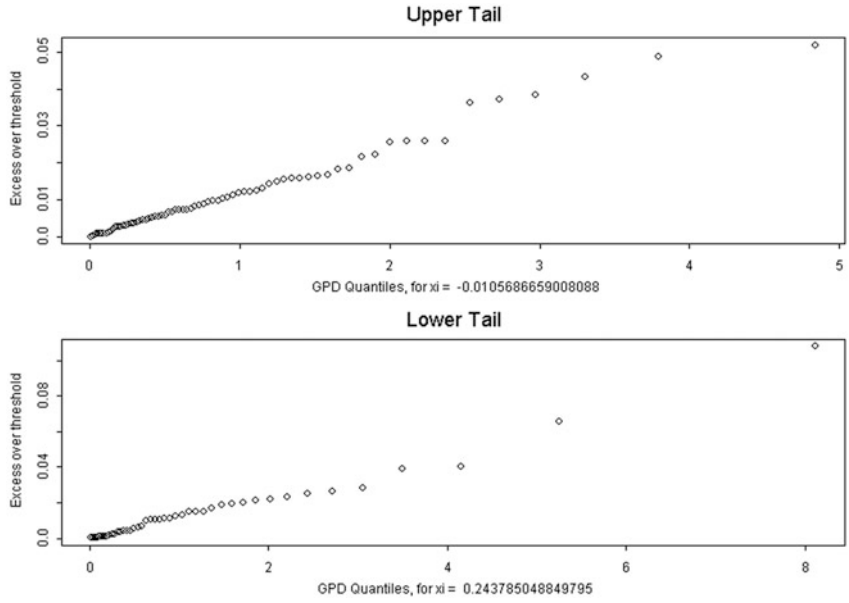


**Fig. 2.13.** Mean excess plots for the right/upper tail (*top*) and left/lower tail (*bottom*) resulting from the fit of a GPD distribution to the weekly S&P log-return data

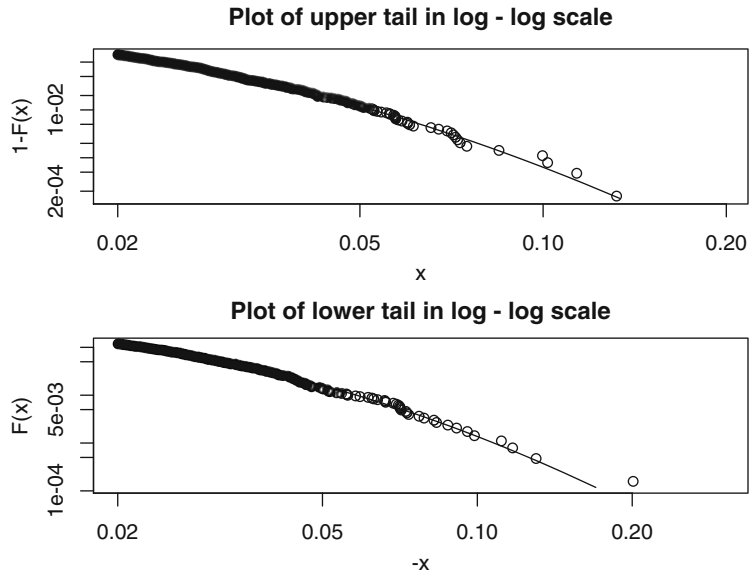


**Fig. 2.14.** Plot of the tails of the GPD fitted to the `WSPLRet` data together with the empirical tails given by the actual data points, for the upper tail (*top*) and the lower tail (*bottom*)

```
> tailplot(WSPLRet.est,tail="two")
```

which produces the plots given in Fig. 2.14, showing the results (in logarithmic scale) for both tails. Using the quantile function `qgpd(WSPLRet.est, . )` as before, we can generate a sample of size `N` from the fitted distribution with the command:

```
> WSPLRetsim <- qgpd(WSPLRet.est,runif(N))
```

## Appendix: Risk Measures: Why and What For?

The goal of this appendix is to give a more mathematical account of the notion of measure of risk as it emerged in the development of mathematical models for applications in the financial and insurance industries.

Historically, and especially in the financial and insurance industries, risk has been equated to the size of the fluctuations of random outcomes as quantified by the standard deviations of these outcomes. Markowitz' mean-variance portfolio theory is the epitome of such a risk-reward modelling. In line with our introduction of the value at risk, the modern approach to risk measure is based on efforts to quantify capital requirements of financial institutions, and risk measures are now used as yardsticks

for insurance underwriting, to allocate capital, to identify prudent investment strategies and acceptable future net worths. It is now a commonly accepted view that one should think of a risk measure as a way to estimate the

*minimum extra capital which makes the future position acceptable*.

**Axiomatic Set-Up**

The basic objects of the theory are random quantities intended to describe all the states of nature at a future date (for example the future values of a portfolio). The possible outcomes of these random variables represent the scenarios which could occur depending upon market changes and other random events. The set of *acceptable* positions and portfolios is decided by the regulator (states of the world requiring government resources – *guarantor of last resort*), an exchange or clearing firm, the investment manager in charge of the portfolio, the board of Directors, etc. For the purpose of illustration of the importance of heavy tail distributions in the quantification of risk, we use a very simple model in order to capture some of the most important stylized facts needed to make our point. We only consider one period static models and we denote by $\Omega$ the set of all possible outcomes/scenarios. A *risk* $X$ is a function on $\Omega$ giving the possible values $X(\omega)$ of a position at the end of the period. The set $\mathcal{A}$ of *acceptable risks* is a subset of the set of risks satisfying a set of *axioms* which will be articulated later, and a *risk measure* $\rho$ is a function associating a real number $\rho(X)$ to each risk $X$. The interpretation of $\rho(X)$ is captured in the following bullet points:

- If $\rho(X) > 0$, $\rho(X)$ is the minimum extra cash one has to add to the position (and invest prudently in the instrument) in order to make the position acceptable;
- If $\rho(X) < 0$, as much as $-\rho(X)$ can be withdrawn from the position without making it unacceptable.

With this interpretation in mind, the following theoretical properties become natural:

- **Shift Invariance** for all real number $m$ and $X$

$$\rho(X + m) = \rho(X) - m$$

  Intuitively, this axiom means that adding cash to a position reduces the risk by the same amount;
- **Monotonicity** for all $X_1$ and $X_2$

$$\text{if} \quad X_1 \leq X_2 \quad \text{then} \quad \rho(X_2) \leq \rho(X_1)$$

  Intuitively, this condition means that the higher the value of the asset or portfolio, the smaller the risk;
- **Convexity** for all $X_1$ and $X_2$ and real numbers $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ such that $\lambda_1 + \lambda_2 = 1$,

$$\rho(\lambda_1 X_1 + \lambda_2 X_2) \leq \lambda_1 \rho(X_1) + \lambda_2 \rho(X_2)$$

This last axiom has a clear geometric interpretation. Financially speaking, this axiom has a very important consequence: it implies that a measure of risk satisfying this axiom should **encourage diversification** as the risk of an aggregate portfolio (in the left hand side of the above inequality) is lower than the aggregation of the risks of the individual components (the above right hand side).

In modern textbooks on quantitative risk management, a risk measure satisfying these four axioms is called a *convex risk measure*.

### First Example

After choosing a benchmark instrument whose returns over the given horizon we denote $R$ and a set $\mathcal{P}$ of probabilities (agent beliefs) on the set $\Omega$ of outcomes, for each (random variable) risk $X$ we set:

$$\rho_{\mathcal{P}}(X) \;=\; \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}\{-X/R\}$$

The interpretation of $\rho_{\mathcal{P}}(X)$ is the following: for a given risk $X$, $\rho_{\mathcal{P}}(X)$ represents the worst expected discounted loss computed from an a priori set of beliefs. So-defined, $\rho_{\mathcal{P}}(X)$ is a coherent measure of risk.

### Second Example: Value at Risk (VaR)

We jump in directly to the formal definition of a mathematical notion of value at risk, referring to the discussion of Sect. 1.1.3 in Chap. 1 for a discussion of the practical applications leading to the abstract definition in terms of quantile of a distribution. Given a probability level $p \in (0,1)$, the value at risk $VaR_p$ (at level $p$ and for the given horizon) of the final net worth $X$ is the negative of the $100p$ percentile of $X$

$$VaR_p(X) = -\inf\{x;\ \mathbb{P}\{X \le x\} > p\}$$

Clearly this measure of risk is given by the amount of capital needed to make the position $X$ acceptable with probability $1-p$ if acceptability is understood as being positive. Value at Risk is widely used as internal risk control (in accordance with Basel I), however in practice, it is **not clear** which probability to use in order to compute the percentile quantifying the risk: should one use a quantile estimated from historical data, or should one use a probability model calibrated to be risk neutral? This is only one of the very many practical problems associated with the use of VaR as a measure of financial risk.

In order to emphasize the dramatic effect that the choice of a particular distribution can have, we recall the comparison of the percentiles of two distributions presented given in Chap. 1. Choosing the probability level $p = 2.5\,\%$ for the sake of definiteness, if a portfolio manager assumes that the P&L distribution is Gaussian, then she will report that the value at risk is $1.96$ (never mind the units, just bear with me), but if she assumes that the P&L distribution is Cauchy, the reported

value at risk will jump to 12.71. Quite a difference!!! This shocking example illustrates the crucial importance of the choice of a model for the P&L distribution. As we just proved, this choice is not innocent, and as a consequence, open to abuse.

### VaR Computation from Empirical Data

For the sake of illustration, we give a simple example based on data already analyzed in this chapter. We shall discuss less stylized and less contrived examples in Chap. 3 next. Even in this simple situation, several avenues are possible:

- One can use empirical VaR given by the empirical estimate of the percentile;
- One can also assume that the portfolio returns are reasonably explained by a Gaussian model in which case we
  - Estimate the mean and the variance of the sample returns;
  - Compute the quantile of the corresponding Gaussian cdf;
- Finally, one can also use the tools developed earlier in the chapter to fit heavy tail distributions, in which case we
  - Fit a GPD to the returns;
  - Compute the quantile of the estimated distribution.

To illustrate the differences between these three procedures on a specific example, we choose the weekly S&P 500 log return data already studied in this chapter. The numerical results reported in Table 2.3 were obtained by running the commands:

```
> -quantile(WSPLRet,0.01)
> -qnorm(0.01,mean=mean(WSPLRet),sd=sd(WSPLRet))
> -qgpdt(WSPLRet.est,0.01)
```

| | **Empirical** quantile | **Gaussian** model | **GPD** model |
|---|---|---|---|
| $VaR_{0.01}$ | 0.05582396 | 0.0471736 | 0.0582578 |

**Table 2.3.** One week 1 % Values at Risk from the S&P 500 index data

Clearly VaR computed under the Gaussian hypothesis is the smallest of the three, offering the most optimistic vision of the risk over a period of one week. The most conservative vision is offered by the fit of a GPD to the weekly returns, while the empirical VaR is reasonable because of the large size of the data set and the presence of a few crashes. The reader is encouraged to rerun this analysis with data prior to October 1987 to understand the role of the semi-parametric fitting procedure in the estimation of the tail of the distribution.

The above example is still rather academic as most practical situations involve multiple underlying instruments (baskets including stocks, bonds, and derivatives)

or even aggregations at the fund or company level of the risk profiles of many desks or business units. The estimation is more difficult in this case. Indeed, as we shall see in the next chapter, estimating separately the risks of the individual desks or units does not help much in deciding how to aggregate these risks while integrating their interdependencies. This is a very *touchy business* and except for the Gaussian case for which one can perform analytic computations, not many tools are available and we will rely on the theory of copulas developed in Chap. 3 and on Monte Carlo computations. In any case, we want to issue the following warning: Using a Gaussian computation when heavy tails are present GROSSLY UNDERESTIMATES the value at risk!

### Troubling Example

We illustrate the main shortcomings of VaR with an example which, despite its rather artificial nature, captures well the features of VaR which we want to emphasize. Let us assume that the short interest rate is zero, that the spread on ALL corporate bonds is $2\,\%$, and that corporate bonds default with probability $1\,\%$, independently of each other. **First scenario** We assume that $1{,}000{,}000$ is borrowed at the base rate and invested in the bond of a single company. In this case:

$$VaR_{0.05} = -20{,}000$$

in other words, there is no risk. **Second scenario** Now let us assume that searching for risk diversification, the same type of investment is set up so that $1{,}000{,}000$ are borrowed at the base rate and invested in equal parts in the bonds of $100$ different companies. Then

$$\mathbb{P}\{\text{at least two companies default}\} > 0.18$$

So $\mathbb{P}\{X < 0\} > 0.05$ and consequently $VaR_{0.05}(X) > 0$ and the portfolio now appears to be risky. In conclusion we see that

- VaR did not detect **over-concentration** of risk
- VaR did not encourage (in fact, it sometimes discourages) **diversification**

Summarizing the shortcomings of VaR we see that

- At the intuitive level VaR only captures the minimal size of a "one in a hundred" event, and highlights merely the best of the rare extreme events to be feared;
- At the mathematical level, VaR is not sub-additive, so VaR is not a convex measure of risk as it does not encourage diversification.

### Conditional Value at Risk (CVaR) and Expected Shortfall (ES)

Despite its popularity, VaR's not encouraging diversification pushed academics and some practitioners to design and adopt risk measures free of this shortcoming. The

natural candidate for taking over VaR is the Expected Short Fall which, while keeping the spirit of VaR in considering only rare losses, takes into account the actual sizes of these losses, something that VaR does not do. This measure of risk is also called *TailVaR* or *Tail Conditional Expectation*

For a given probability level $p$, the shortfall distribution is the cdf $\Theta_p$ defined by:

$$\Theta_q(x) = \mathbb{P}\{X \leq x | X > VaR_p\}. \tag{2.36}$$

This distribution is just the conditional loss distribution given that the loss exceeds the Value at Risk at that level. The mean or expected value of this distribution is called the expected shortfall, and is denoted by $ES_p$. Mathematically, $ES_p$ is given by:

$$ES_p = \mathbb{E}\{X | X > VaR_p\} = \int x \, d\Theta_p(x) = \frac{1}{q} \int_{x > VaR_p} x \, dF(x). \tag{2.37}$$

It gives the expected loss size given that the loss is more extreme than VaR at the same level. Defined this way, it fixes most of the problems of VaR

- At the intuitive level, the sizes of the losses are taken into account,
- At the *theoretical* level it can be proven that it is *essentially* a coherent measure of risk.

A good part of risk analysis concentrates on the estimation of the value at risk $VaR_p$ and the expected shortfall $ES_p$ of various portfolio exposures. The main difficulty comes from the fact that the theoretical cdf $F$ is unknown, and its estimation is extremely delicate since it involves the control of rare events. Indeed, by the definition of a tail event, very few observations are available for that purpose. More worrisome is the fact that the computation of the expected shortfall involves integrals which, in most cases, need to be evaluated numerically.

## PROBLEMS

Ⓣ **Problem 2.1** *Explain (in two short sentences) the conflicting conditions which you try to satisfy when choosing the threshold in fitting a GPD to the tail of a distribution using the POT (Peak over Threshold) method.*

Ⓣ **Problem 2.2**

1. *For this first question we assume that $X$ is a random variable with standard Pareto distribution with shape parameter $\xi$ (location parameter $m = 0$, scale parameter $\lambda = 1$).*
   - 1.1. *Give a formula for the c.d.f. of $X$. Explain.*
   - 1.2. *Derive a formula for the quantile function of $X$.*
   - 1.3. *How would you generate Monte Carlo samples from the distribution of $X$ if you only had a random generator for the uniform distribution on $[0, 1]$ at your disposal?*

2. *Give a formula for the density $f_Y(y)$ of a random variable $Y$ which is equal to an exponential random variable with mean 2 with probability $1/3$ and to the negative of a classical Pareto random variable with shape parameter $\xi = 1/2$ (location $m = 0$ and scale $\lambda = 1$) with probability $2/3$. Explain.*
3. *How would you generate Monte Carlo samples from the distribution of $Y$?*

Ⓣ **Problem 2.3** *In this problem, we study the loss distribution of a portfolio over a fixed period whose length does not play any role in the analysis. Loss is understood as the negative part of the return defined as $L = \max(-R, 0)$. We assume that a fixed level $\alpha \in (0, 1)$ is given, and we denote by $VaR_\alpha$ the Value at Risk (VaR) at the level $\alpha$ of the portfolio over the period in question. In the present context, this VaR is the $100(1-\alpha)$-percentile of the loss distribution. This is consistent with the definition used in the text. The purpose of the problem is to derive a formula for the expected loss given that the loss is assumed to be larger than the value at risk.*

1. *For this question, we assume that the loss distribution is exponential with rate $r$.*
   1.1. *Give a formula for the c.d.f. of $L$. Explain.*
   1.2. *Derive a formula for $VaR_\alpha$.*
   1.3. *Give a formula for the expected loss given that the loss is larger than $VaR_\alpha$. Recall that, if a random variable $X$ has density $f$, its expected value given the fact that $X$ is greater than or equal to a level $x_0$ is given by the formula*

$$\frac{1}{\mathbb{P}\{X \geq x_0\}} \int_{x_0}^{\infty} x f(x) dx, \qquad \text{or equivalently} \qquad \frac{\int_{x_0}^{\infty} x f(x) dx}{\int_{x_0}^{\infty} f(x) dx}.$$

2. *For this question, we assume that the loss distribution is the standard Pareto distribution with shape parameter $\xi$, location parameter $m = 0$ and scale parameter $\lambda = 1$.*
   2.1. *Give a formula for the c.d.f. of $L$. Explain.*
   2.2. *Derive a formula for $VaR_\alpha$.*
   2.3. *Give a formula for the expected loss given that the loss is larger than $VaR_\alpha$.*
3. *The expected short fall (also known as the conditional VaR) at the level $\alpha$ is the expected loss conditioned by the fact that the loss is greater than or equal to $VaR_\alpha$. The goal of this question is to quantify the differences obtained when using it as a measure of risk in the two loss models considered in questions 1 and 2.*
   3.1. *For each $\alpha \in (0, 1)$, derive an equation that the rate parameter $r$ and the shape parameter $\xi$ must satisfy in order for the values of $VaR_\alpha$ computed in questions 1.2 and 2.2 to be the same.*
   3.2. *Assuming that the parameters $r$ and $\xi$ satisfy the relationship derived in question 3.1 above, compare the corresponding values of the expected short fall in the models of questions 1 and 2 and comment on the differences.*

Ⓔ **Problem 2.4** *This problem attempts to quantify the goodness of fit resulting from our GPD analysis of samples with heavy tails.*

1. *Use the method described in the text to fit a GPD to the PCS index, and generate a Monte Carlo random sample from the fitted distribution five times the size of the original data sample.*
2. *Produce a Q-Q plot to compare graphically the two samples and comment.*

3. *Use a two-sample Kolmogorov-Smirnov goodness-of-fit test to quantify the qualitative results of the previous question.* **NB***: Such a test is performed in* `R` *with the command* `ks.test`*. Check the help files for details on its use and the returned values.*
4. *Same questions as above for the weekly log-returns on the S&P data.*

Ⓔ **Problem 2.5** *This problem uses the data set* `PSPOT` *included in the library* `Rsafd`*. The entries of this vector represent the daily Palo Verde (firm on peak) spot prices of electricity between January 4, 1999 and August 19, 2002. Use exploratory data analysis tools to argue that the tails of the distribution are heavy, fit a GPD to the data, and provide estimates of the shape parameters.*
***NB:*** *We usually refrain from talking about the distribution of a financial time series, reserving fitting a distribution to the returns instead of the entries of the original series. You are asked to do just that in this problem. Even though an analysis of the returns in the spirit of what is done in the text would make perfectly good sense, a look at a time series plot of* `PSPOT` *shows a form of stationarity of the data (to be explained later in the book) justifying the analysis asked of you in this problem.*

Ⓔ **Problem 2.6** *This problem requires the data set* `DSP`*. The entries of this numeric vector represent the daily closing values of the S&P 500 index between the beginning of January 1960 and September 18, 2001.*

1. *Compute the vector of log-returns and call it* `DSPLRet`*.*
2. *We now use the data set* `MSP`*. The entries of this numeric vector represent minute by minute quotes of the* `S&P 500` *on September 10, 1998. Compute the corresponding log-return vector and call it* `MSPLRet`*.*
3. *Produce a Q-Q plot of the empirical distributions of the two log-return vectors, and comment. In particular, say if what you see is consistent with the claim that the properties of the daily series are shared by the minute by minute series. Such an invariance property is called* self-similarity*. It is often encountered when dealing with fractal objects.*
4. *Compute the empirical means and variances of the* `DSPLRet` *and* `MSPLRet` *data. Assuming that these data sets are Gaussian, would you say that the two distributions are the same in view of these two statistics?*
5. *Fit GPDs to the* `DSPLRet` *and* `MSPLRet` *data, and compare the distributions one more time by comparing the shape parameters.*

Ⓔ **Problem 2.7** *This problem deals with the analysis of the daily S&P 500 index closing values.*

1. *Create a vector* `DSPRET` *containing the daily raw returns. Recall that the raw return on a given day is the difference between the value on that day and the day before divided by the value on the previous day. Compute the mean and the variance of this daily raw return vector.*
2. *Fit a GPD to the daily raw returns, give detailed plots of the fit in the two tails, and discuss your results.*
3. *Generate a sample of size* $10{,}000$ *from the GPD fitted above. Call this sample* `SDSPRET`*, produce a Q-Q plot of* `DSPRET` *against* `SDSPRET`*, and comment.*
4. *Compute the VaR (expressed in units of the current price) for a horizon of 1 day, at the level* $\alpha = 0.005$ *in each of the following cases:*
   4.1 *Assuming that the daily raw return is normally distributed;*
   4.2 *Using the object of class* `gpd` *which you created in question 2 to fit a GPD distribution to the data;*

*4.3 Using the Monte Carlo sample* `SDSPRET` *you generated.*
*Explain the differences and the similarities between the three estimates of the VaR so obtained.*

5. *Redo the questions above after replacing the vector* `DSP` *with the vector* `SDSP` *containing only the first* 6,000 *entries of* `DSP`. *Compare the results, and especially the VaR's. Explain the differences.*

Ⓔ Ⓣ **Problem 2.8** *The goal of this problem is to highlight some of the properties of the estimates obtained with the command* `fit.gpd` *when fitting a GPD to a data sample* $x_1, \cdots, x_n$. *We assume that the distribution of the data has two tails (one extending to* $-\infty$ *and the other one to* $+\infty$*), and we are interested in understanding the effect of the choice of the thresholds* `lower` *and* `upper`.
*Remember that a distribution with an upper tail is a GPD if its density* $f(x)$ *is well approximated in the tail by a function of the form*

$$f_{\xi_+,m_+,\lambda_+}(x) = \frac{1}{\lambda_+\xi_+}\left(1 + \frac{x-m}{\lambda_+}\right)^{-(1+\frac{1}{\xi_+})} \tag{2.38}$$

*at least when* $x > m_+$ *for some large enough threshold* $m_+$, *where* $\lambda_+$ *is interpreted as a scale parameter, and where* $\xi_+ > 0$ *is called the shape parameter governing the size of the upper tail. If the distribution has a lower tail, one requires a similar behavior for* $x < m_-$ *for possibly different parameters* $m_-$, $\lambda_-$ *and* $\xi_-$.

*For the purpose of the problem, we assume that the true density of the sample* $x_1, \cdots, x_n$ *is given in Fig. 2.15. It is exactly equal to the function* $f_{\xi_+,m_+,\lambda_+}(x)$ *for* $x > 2$ *with* $m_+ = 2$ *and some value* $\xi_+ > 0$ *(to be estimated), and equal to the corresponding function* $f_{\xi_-,m_-,\lambda_-}(x)$ *for* $x < -2$ *with* $m_- = -2$ *and some* $\xi_- > 0$ *(to be estimated as well).*

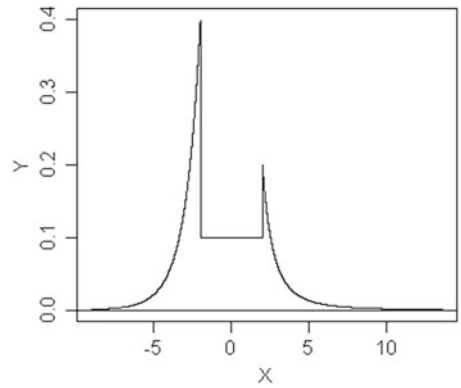


**Fig. 2.15.** Density of the GPD from which the sample $x_1, \cdots, x_n$ is generated

1. *What should you expect from the estimate* $\hat{\xi}_+$ *given by the function* `fit.gpd` *if you use a threshold* `upper`
   *1.1. Exactly equal to* 2.
   *1.2. Greater than* 5.
   *1.3. Between* 0 *and* 1.

2. *What should you expect from the estimate* $\hat{\xi}_-$ *given by the function* `fit.gpd` *if you use a threshold* `lower`
   2.1. *Exactly equal to* $-2$.
   2.2. *Smaller than* $-8$.
   2.3. *Between* $0$ *and* $-1$
   *and in each case, say how the estimate* $\hat{\pi}_{0.01}$ *of the 1 percentile compares to the true value* $\pi_{0.01}$.

## NOTES & COMPLEMENTS

What distinguishes our presentation of exploratory data analysis from the treatments of similar material found in most introductory statistics books, is our special emphasis on heavy tail distributions. Mandelbrot was presumably the first academic to stress the importance of the lack of normality of the financial returns. See [64, 65], and also his book [66]. He proposed the Pareto distribution as an alternative to the normal distribution. The theory of extreme value distributions is an integral part of classical probability calculus, and there are many books on the subject. We refer the interested reader to [29] because of its special emphasis on insurance applications. In particular, the discussion given in the Notes & Complements section of Chap. 1 of a possible mathematical model for the PCS index dynamics fits well in the spirit of [29].

The Fisher-Tippett theory reviewed in this chapter was enhanced and brought to the level of a complete mathematical theory in the fundamental works of Gnedenko. Many textbooks give a complete account of this theory. We refer the reader to the books of Leadbetter, Lindgren and Rootzen [63], Resnick [79] and Embrechts, Klüppelberg and Mikosch [29]. This last reference emphasizes the notion of maximum domain of attraction to delineate which distributions give rise to block maxima convergence, after proper normalization, toward a specific GEV distribution. This more modern point of view is also chosen in the more recent account of McNeil, Frey and Embrechts [72]. There are other reasons why a reader interested in the applications of the block maxima method should consult this text. Indeed, he or she will find there a detailed discussion of the effects of dependencies upon the estimates of the shape of the tail. These dependencies occur in two different and non-exclusive ways: as temporal correlation already contained in the data, or as artifacts of the overlap of blocks. Both these issues are addressed and further references to the relevant literature are given.

The block maxima approach to the estimation of extremes has its origin in hydrology where extreme value theory was used to study and predict flood occurrences. There the shape parameter $\xi$ is replaced by its negative $k = -\xi$. The package `Rsafd` gives the user the option to choose which parametrization of the GEV distributions and GPD's he would rather work with by setting a global variable `SHAPE.XI` to `TRUE` or `FALSE`. We did not mention the $k$-parametrization in the text because of our overwhelming interest in financial applications. Early examples of the use of the block maxima approach in the analysis of financial data were introduced by Longin in [61]. See the book by Embrechts, Frey and McNeil [28] for more examples of in the same spirit.

Details on the maximum likelihood fitting of GEV distributions can be found in Hosking [46] and Hosking, Wallis and Wood [48]. Asymptotic normality was proved by Smith in [89] in the case $\xi > -0.5$.

The method of L-moments seems to have its origin in hydrology. It was introduced by Hosking, Wallis and Wood in [48], and further developed in [47] and [46]. The probability weighted moments, introduced by Greenwood and collaborators in [33] were the precursors of the L-moments. This method of estimation of the parameters of GPD's and GEV distributions does not seem to have permeated the insurance and finance literature, and we purposely chose to include it in our analysis in order to add diversity to our estimation toolbox. Moreover, even if we decide to rely exclusively on maximum likelihood estimators, initializing the maximization search algorithm with the values of the empirical L-moments has proven to be an efficient method of increasing the chances of convergence, speeding up this convergence, and even converging toward a more reasonable local maximum. The derivation of formulae (2.13)–(2.15) giving the L-moments of a GEV distribution in terms of its natural parameters can be found in Hosking [46].

Except possibly for the maximum likelihood estimation of GPD's and GEV distributions, the material presented in this chapter is not systematically covered in the literature devoted to insurance and financial applications. This is especially true with the use of L-moments. See nevertheless the recent work of Seco et al. [70] which may indicate a renewal of interest for these methods for financial applications. We were made aware of the importance of L-moments via numerous enlightening discussions with Julia Morrison.

The fundamental result of this chapter is due to Pickands [75] and Balkema and de Haan. The estimation procedures presented in this chapter rely on the assumption that the data points $x_1, \cdots, x_n$ are realizations of independent and identically distributed random variables. The independence assumption is rarely satisfied in real life applications, and especially with series of financial returns which are of interest to us. However, in many instances, this assumption is not as restrictive as it may seem. Indeed, for many stationary time series, the exceedances over increasing levels can be shown to have a limiting Poisson distribution. So it seems that the independence assumption is restored in the limit of exceedances over high levels. However, most financial return data exhibit clustering properties captured by ARCH and GARCH models, and incompatible with the independence assumption. The reader concerned by these issues is referred to the book of Mc Neil, Frey and Embrechts [72], where further references to the literature can be found.

A time honored method to estimate the size of *power tails* is to compute the Hill's estimator of the exponent $\alpha$ (essentially the inverse of the shape parameter $\xi$). We purposely chose to ignore this method of estimation, because of horror stories about the misleading conclusions one can reach with this estimation method. The interested reader is referred to the textbook [29] for a discussion of the Hill estimator, and for a series of examples showing clearly its limitations.

Financial institutions started worrying about risk exposures long before regulators got into the act. The most significant initiative was RiskMetrics span off by J.P. Morgan in 1994. Even though the original methodology was mostly concerned with market risk, and limited to Gaussian models, the importance of Value at Risk (VaR) calculations was clearly presented in a set of technical documents made available on the web at the URL `www.riskmetrics.com`. A more academic discussion of the properties of $VaR$ can be found in the book by C. Gourieroux and J. Jasiak [42], and the less technical book by Jorion [53]. VaR is one among many possible ways to quantify a risky exposure to possible adverse moves. The seminal paper of Artzner, Delbaen, Eber and Heath [4] was the first instance of an attempt to formalize mathematically the notion of financial risk measure. Their original set of axioms included positive homogeneity stating that for all $\lambda \geq 0$ and $X$, one should have $\rho(\lambda X) = \lambda\rho(X)$, and a sub-additivity condition slightly weaker than convexity. Risk measures satisfying their

four axioms were called coherent risk measures. Because positive homogeneity is not obviously a natural requirement for a measure of risk, it was gradually abandoned in favor of the smaller set of axioms given in the text, and which was systematically advocated by Föllmer and Schied. For a set of axioms to capture properly the desirable properties of a rigorous risk quantification, some form of convexity or sub-additivity should be included in order for the risk of a diversified portfolio to be less than the sum of the individual risks. Unfortunately, VaR does not encourage diversification in this sense. However, Conditional VaR and Expected Shortfall (at least when the distribution is continuous) do.

A clear exposé of risk measures and their mathematical theory can be found in Foellmer and Schied's book [36]. The risk measures discussed in the text are static in the sense that they are based on models of the sources of risk over a fixed period limited by a fixed horizon, and that no provision is made to update the quantification of the risk as time goes by. In this sense they can be viewed as a first generation of risk measures. Very active research is now dealing with a new generation of risk measures which can capture the time evolution of risk. Such multi-period models have to deal with very technical consistency issues, and easy implementations of the first theoretical results which appeared in this area are still a long way.

The `R` methods used in this chapter to estimate heavy tail distributions, simulate random samples from these distributions, and compute risk measures are taken from the `R` package `Rsafd` based in part on the library `EVANESCE` originally developed in `S` by J. Morrison and the author.

http://www.springer.com/978-1-4614-8787-6

Statistical Analysis of Financial Data in R

Carmona, R.

2014, XVII, 588 p. 187 illus., 37 illus. in color.,
Hardcover
ISBN: 978-1-4614-8787-6