Exploring Heavy Tails
Pareto and Generalized Pareto Distributions

December 1, 2016

This vignette is designed to give a short overview about Pareto Distributions and Generalized Pareto Distributions (GPD). We will work with the `SPC.we` data of our **quantmod** vignette. Therefore we have to reproduce the `SPC.we` data in exactly the same way as described the **quantmod** vignette.

In financial data analysis stock indices as the S&P 500 index are typically analyzed by using the *returns* of the index. We use the log-returns

```
R> WSPLRet <- diff(log(SPC.we))
```

We start to analyze these by plotting a histogram

```
R> hist(WSPLRet)
```

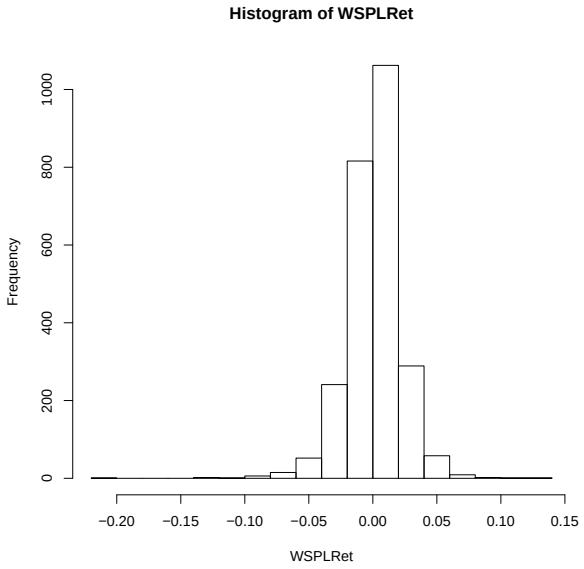

Figure 1: Histogram of the log-returns of the S&P 500 from 1960-01-04 to 2009-01-01.

This histogram shows a unimodal distribution of values with the peak around 0, which nourishes the hypothesis that the log-returns are normally distributed. A very intuitive method to test this is the Q-Q plot.

The slope of the (linear regression) line and its intercept determine the parameters of the corresponding Gaussian distribution. If the points are close to this line the empirical distribution of the sample can

```
R> qqnorm(WSPLRet)
R> qqline(WSPLRet)
```

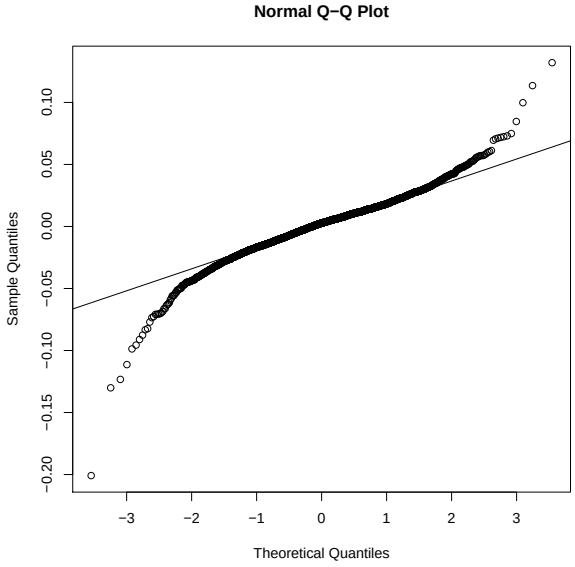

Figure 2: Q-Q plot of `WSPLRet` values.

very well be approximated by a normal distribution. Figure 2 shows that log-returns of the weekly S&P 500 index have heavy tails on both sides and are therefore not modeled well by a normal distribution. The tails of the normal distribution are too thin to produce enough extreme events to match those in the sample.

However, other families of distributions, like Pareto distributions can be used. One way to identify classes of distributions which produce wild events is to show that the density of the considered distribution decays polynomially and then to estimate the degree of such a polynomial decay (Note that for the normal distribution, decay is exponential). Such distributions are called *generalized Pareto distributions* (GPD).

In the following we give a short explanation of Pareto Distributions and GPDs, before we study the problem of estimating the tails of or S&P 500 returns.

1 Pareto distribution

The Pareto distribution (e.g., `https://en.wikipedia.org/wiki/Pareto_distribution`) is commonly used for quantities that are distributed with very long right tails. It is named after the Italian economist Vilfredo Pareto, who originally used this distribution to describe the allocation of wealth among individuals since it seemed to show rather well the way that a larger portion of the wealth of any society is owned by a smaller percentage of the people in that society.

A random variable X has a Pareto distribution with *scale* parameter $K > 0$ and *shape* parameter $\alpha > 0$ iff its cumulative distribution function is given by

$$F(x) = \begin{cases} 1 - (K/x)^{\alpha}, & x \geq K \\ 0, & x < K. \end{cases}$$

(If a family of probability distributions with parameter s and other parameters θ is such that the cumulative distribution functions satisfy $F_{s,\theta}(x) = F_{1,\theta}(x/s)$, then s is a scale parameter. In the above, note that for $x \geq K$, $F_{K,\alpha}(x) = 1 - (x/K)^{-\alpha} = F_{1,\alpha}(x/K)$.)

Hence, K is the minimum possible value of X. The density of X is then given by

$$f(x) = \begin{cases} \alpha K^{\alpha}/x^{\alpha+1}, & x \geq K \\ 0, & x < K. \end{cases}$$

For a shape parameter $\alpha > 1$ the expected value is given by

$$\mathbb{E}(X) = \frac{\alpha K}{\alpha - 1},$$

otherwise ($\alpha \leq 1$) the expected value is infinite.

How the probability distribution of the Pareto distribution changes when one varies the shape parameter is illustrated in the following example where we make use of function `dpareto()` included in package **VGAM**:

```
R> require("VGAM")
R> x <- seq(0.1, 10, length = 1000)
R> plot(x, dpareto(x, scale = 1, shape=1),
+      type = "l", xlab = "x", ylab = "dpareto(x)",
+      main = "Pareto Probability Density")
R> lines(x, dpareto(x, scale = 1, shape=.5), col = "red")
R> lines(x, dpareto(x, scale = 1, shape= .2), col = "blue")
```

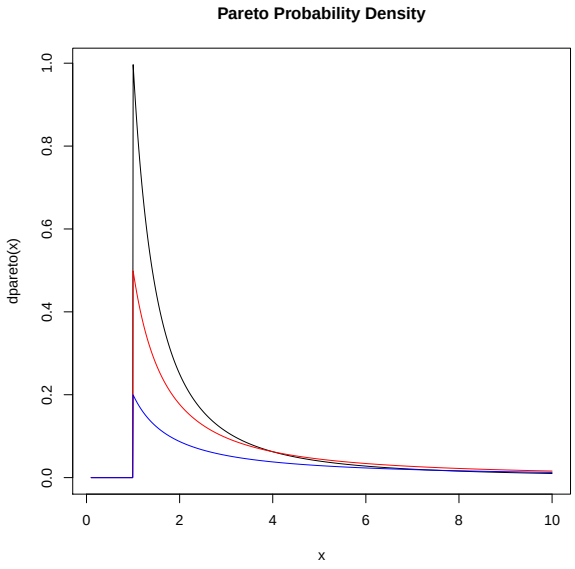

Figure 3: Pareto probability density for shape parameters equal to 1, 0.5, and 0.2.

2 Generalized Pareto Distribution

In comparison to the Pareto Distributions, the Generalized Pareto Distribution (GPD, e.g., `https://en.wikipedia.org/wiki/Generalized_Pareto_distribution` has three three parameters; one *location parameter* μ and two parameters for *scale* and *shape*, σ and ξ. The cumulative distribution function of the GPD is given by:

$$\mathbb{P}(X \leq x) = \begin{cases} 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right), & \xi = 0, \end{cases}$$

for $x \geq \mu$ when $\xi \geq 0$, and $\mu \leq x \leq \mu - \sigma/\xi$ when $\xi < 0$, where μ and ξ are arbitrary real numbers and $\sigma > 0$.

(Note that the distribution function must take values in $[0, 1]$. For $\xi > 0$, this needs $1+\xi(x-\mu)/\sigma \geq 1$, which is equivalent to $x \geq \mu$. For $\xi < 0$, this needs $0 \leq 1 + \xi(x - \mu)/\sigma \leq 1$, which is equivalent to $\mu \leq x \leq \mu - \sigma/\xi$.)

For a $\xi < 1$, the mean of a GPD is given by

$$\mathbb{E}(X) = \mu + \frac{\sigma}{1 - \xi}.$$

The GPD is generalized in the sense that it contains a number of special cases: When $\xi > 0$ and $\mu = 0$, the distribution function is that of an ordinary Pareto Distribution with $\alpha = 1/\xi$ and $K = \sigma/\xi$.

If we are interested in generating generalized Pareto random variables we can apply the following formula:

$$X = \mu + \frac{\sigma(U^{-\xi} - 1)}{\xi} \sim GPD(\mu, \sigma, \xi)$$

for a uniformly distributed variable $U \sim \text{unif}(0, 1)$.

Back to the S&P 500: Like the exponential distribution, the Generalized Pareto distribution is often used to model the tails of another distribution. Now we will use the GPD in order to understand the tails of the log-returns of the S&P 500 index as described in the **quantmod** vignette.

The main difficulty in identifying memberships in this class is that the traditional density estimation procedures (like histograms or kernel density estimators) cannot estimate the tails precisely enough even though they serve as a good estimator for the center of the distribution. Thus the estimation of the size of the tails has to be done in a parametric way, while the estimation of the center of the distribution can be done via histograms or kernel density estimators. To sum up,

- Standard methods (e.g., kernel density estimators) for the center of the distribution,
- parametric techniques to estimate the polynomial decay of the density in the tails.

What we brushed under the rug is the determination when and where the tails of the distribution start. This is a delicate problem and there is no universal way to determine the value where the tail starts. This cut off point should be large enough so that the behavior of the tail is homogeneous beyond the threshold but it should not be too large, as we need enough data points in the tail. On intuitive way to determine this cut off point is to use the Q-Q plot. This suggests that values around -0.04 and 0.04 could do the trick:

```
R> qqnorm(WSPLRet)
R> qqline(WSPLRet)
```

```
R> abline(a = 0.04, b=0, col= 2)
R> abline(a = -0.04, b=0, col= 2)
```

See Figure 4.

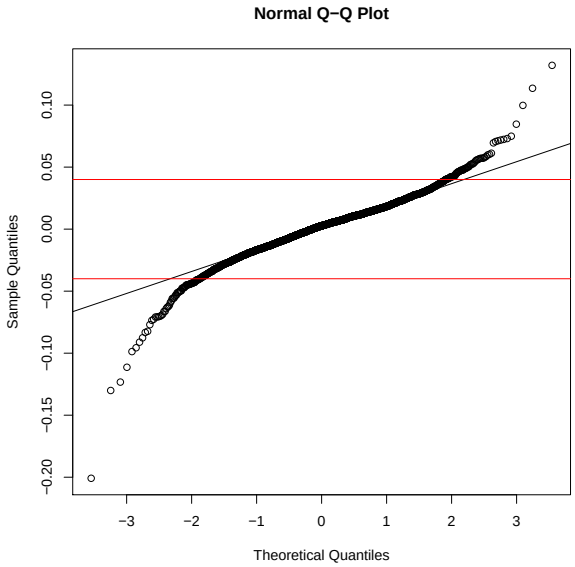

Figure 4: Q-Q plot of `WSPLRet` values with two possible threshold values.

Older versions of package **Rsafd** provide GPD functions (`dgpd()`, `pgpd()`, `qgpd()`, `rgpd()`) with parameters named named `m`, `lambda` and `xi` (corresponding to μ, σ and ξ), with defaults of 0, 1, and 0, respectively (so that using all default values for the parameters gives the exponential distribution), and a function `gpd.tail()` combining parametric estimation of upper and possibly also lower GPD tails with flexible non-parametric estimation of the rest. (Newer versions, accompanying "Statistical Analysis of Financial Data with R" use `fit.gpd()` for fitting the models, and functions `pgpd()` etc. for taking the probability function etc. for such fitted models.)

Given observations (data) $x_1, \ldots, x_n$, `gpd.tail()` fits a GPD to the upper tail by taking a (given) upper threshold θ_u, and then fitting a GPD with location parameter $\mu = 0$ to the positive upper exceedances ("excesses over the threshold") $x_i - \theta_u$. With $\hat{p}_u$ the estimate of the CDF of the data at θ_u (from the non-parametric part), for the upper tail one then uses the GPD with $\mu = 0$ and the fitted $\hat{\sigma}_u$ and $\hat{\xi}_u$ and a weight of $1 - \hat{p}_u$ (so that overall the composite CDF tends to 1 as $x \to \infty$). I.e., for $x \geq \theta_u$,

$$F(x) = \hat{p}_u + (1 - \hat{p}_u) F_{\mathrm{GPD}(0, \hat{\sigma}_u, \hat{\xi}_u)}(x - \theta_u) = \hat{p}_u + (1 - \hat{p}_u) F_{\mathrm{GPD}(\theta_u, \hat{\sigma}_u, \hat{\xi}_u)}(x).$$

Thus, the location parameter for the fitted GPD is always zero, and only the scale (`lambda`) and shape (`xi`) parameters are shown.

Similarly, if a lower tail is fitted as well, then one takes a threshold θ_l and fits a GPD with location parameter $\mu = 0$ to the positive lower exceedances $\theta_l - x_i$. With $\hat{p}_l$ the estimate of the CDF at θ_l, for the lower exceedences one then uses the GPD with $\mu = 0$ and the fitted $\hat{\sigma}_l$ and $\hat{\xi}_l$ and a weight $\hat{p}_l$. I.e., for $x \leq \theta_l$,

$$F(x) = \hat{p}_l F_{\mathrm{GPD}(0, \hat{\sigma}_l, \hat{\xi}_l)}(\theta_l - x)$$

(the weight ensures that $F(\theta_l) = \hat{p}_l$).

Using

```
R> require("Rsafd")
```

```
R> WSPLRet.est <- gpd.tail(as.vector(WSPLRet), lower = -.04, upper =.04)
```

we obtain Figure 5 which shows that the points appearing in a rather straight line indicating that a generalized Pareto distribution may be appropriate. (What is shown is a QQ-plot with the quantiles of the fitted GPD on the x axis and the empirical quantiles (i.e., the sorted excesses over the threshold) on the y axis.)

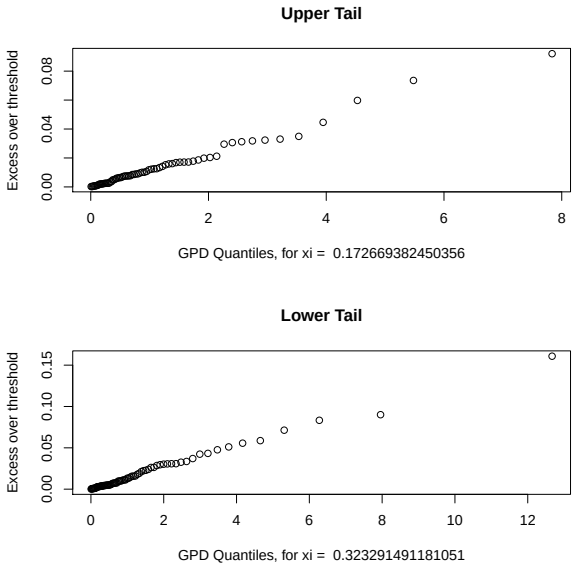

Figure 5: Quantile plot of the right/upper tail (top) and left/lower tail (bottom) resulting from the fit of a GPD distribution to the weekly S&P log-return data.

The parameter ξ is the estimated shape parameter which can be computed for the upper and for the lower tail, respectively:

```
R> WSPLRet.est$upper.par.ests
```

```
    lambda         xi
0.01144629 0.17266938
```

```
R> WSPLRet.est$lower.par.ests
```

```
    lambda         xi
0.01317755 0.32329149
```

which can also be plotted:

```
R> op <- par(mfrow = c(2,1))
R> shape.plot(WSPLRet, tail = "upper")
R> shape.plot(WSPLRet, tail = "lower")
R> par(op)
```

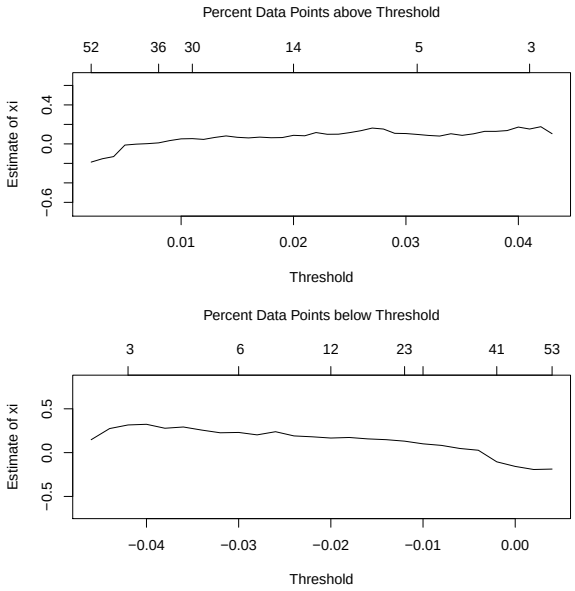

Figure 6: Shape parameter ξ for the right tail (top) and for the left tail (bottom) of the distribution of `WSPLRet`.

See Figure 6.

These plots are more or less consistent with the estimated values.

In a next step, we check the quality of our fit by superimposing the empirical distribution of the points in the tails onto the theoretical graphs of the tails of the fitted distributions:

```
R> op <- par(mfrow = c(2,1))
R> tailplot(WSPLRet.est, tail = "upper")
R> tailplot(WSPLRet.est, tail = "lower")
R> par(op)
```

See Figure 7.

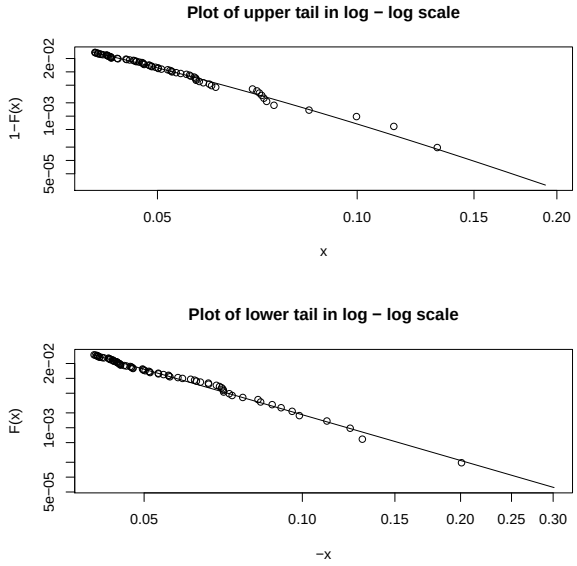

Figure 7: Plot of the tails of the fitted GPD together with the empirical values given by `WSPLRet`.

Notes for Interested Readers

Testing: If you are interested in testing whether a distribution follows a GDP, you can use the package **gPdtest**.

```
R> require("gPdtest")
```

The function `gPd.test()` can be used for testing the null hypothesis H_0 that a random sample has a GPD with unknown shape parameter ξ, which is a real number. `gPd.test()` requires a numeric data vector as an input parameter. Therefore we run the commands

```
R> GPD_WSPLRet <- as.numeric(WSPLRet)
R> gpd.test(GPD_WSPLRet[GPD_WSPLRet > 0.04])

$boot.test

        Bootstrap test for the generalized Pareto distribution

data:  GPD_WSPLRet[GPD_WSPLRet > 0.04]
p-value = 0.8929


$p.values
                                                        p.value R-statistic
H_0^-: x has a gPd with negative shape parameter  0.0000000   0.8858715
H_0^+: x has a gPd with positive shape parameter  0.8928929   0.9936368
```

to test whether the right tail (above the chosen threshold) of the distribution follows a GPD. Since a GPD is only defined for positve values we have to take the absolute value of the lower tail if we want to test the null hypothesis.

Risk Measures: Package **evir** provides a function, `riskmeasures()`, which can be used for rapid calculations of point estimates of prescribed quantiles and expected shortfalls. As an input parameter this function needs the output of the function `gpd()` from the same package. As an example we will illustrate these functions on our data `WSPLRet`.

```
R> require("evir")
R> RMgpd <- gpd(-WSPLRet, 0)

R> riskmeasures(RMgpd, 0.99)

        p   quantile     sfall
[1,] 0.99 0.06173685 0.0776616
```

The first function fits a GPD to negative return and the second one gives estimates of 0.99 quantiles of the `WSPLRet` distribution as well as the associated expected shortfall estimates.