Chapter 4

Extreme Value Theory

1 Motivation and basics

The risk management is naturally focused on modelling of the tail events ("low probability, large impact"). In the introductory lecture, we have already showed that the returns of the S&P 500 stock index are better modeled by Student's t-distribution with approximately 3 degrees of freedom than by a normal distribution. Hence it makes sense to further explore the class of statistical distribution which exhibits fat tails.

However, in an attempt to estimate the whole distribution of losses, we focus on the central part of the distribution and neglect the tails. Hence, it would be beneficial to have a model which is mainly focused on the tails of this distribution, taking into account the heavy-tails property.

One of the possible approaches is to use the Extreme Value Theory (EVT). This theory was originally motivated by a very practical question: How high should be a barrier to be built to prevent flooding with a probability e.g. of 99.99% ? Hence, actually we can be grateful to the flat Benelux landscape for the incentivising of this important contribution to the risk management. Indeed, the current regulatory framework requiring banks to hold sufficient amount of capital as a very close analogy to such anti flood defenses.

For a proper risk management, it would be very helpful to know the statistical distribution e.g. of the largest 10-day drop in the stock index in one year. However, data on such tail events are by definition rare. Nevertheless, some asymptotic properties of the tails can be derived (by an analogy to the Central Limit Theorem).

As we will describe in details later, the main result of the EVT is that the distribution of the tails is limited to only three types:

1. no tails (distribution function is cut off),

2. exponential function,

3. power function.

Given that the EVT represents a very broad topic, this lecture note is limited only to a brief overview focusing on the application of theoretical knowledge in the risk management.

2 Pareto distribution

2.1 Basic properties

One of the important representatives of fat-tailed distributions is the Pareto distribution. This distribution is named after Italian economist Vilfredo Pareto (1848 – 1923). Its distribution function is as follows:

$$P(X \leq x) = \begin{cases} 1 - cx^{-\gamma} & \text{if } x \geq x_m, \\ 0 & \text{if } x < x_m, \end{cases} \tag{4.1}$$

where $c = x_m^{\gamma}$.

The Pareto distribution can be used e.g. for the modelling of the following fenomena:

- size of cities,
- size of meteorites,
- the amount of oil reserves in oil fields,
- the biggest one-day rainfall in one year,
- number of hectares of forest affected by a fire,
- daily stock returns.

Recall that if X is a random variable with a Pareto distribution, then its expected value is finite only if $\gamma > 1$:

$$E(X) = \frac{\gamma}{\gamma - 1}.$$

In addition, its variance is finite only if $\gamma > 2$:

$$\text{Var}(X) = \frac{\gamma}{(\gamma - 1)^2(\gamma - 2)}. \tag{4.2}$$

Hence, if $\gamma \leq 2$, then X is has theoretically an infinite variance. However, it might not be clear what does this mean in real data. Indeed, when working with real data, we can always calculate a sample variance which is finite.

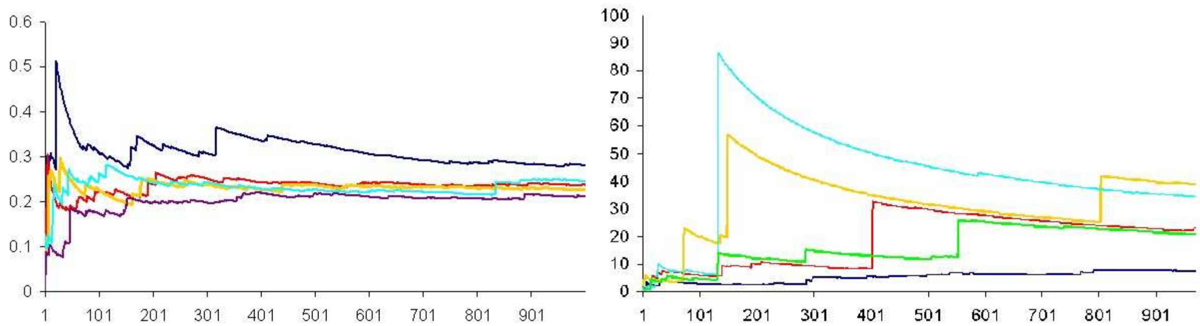

Figure 4.1: Plot of cumulative volatility σ_k for $\gamma = 6$ (theoretical volatility $\sigma = 0.24$, left-hand side) and $\gamma = 1.2$ (theoretical volatility $\sigma = \infty$, right-hand side).

To shed more light into this issue, let us generate data from a Pareto distribution and plot the cumulative sample volatility.

- First, generate 1000 data from the uniform distribution on [0,1].
- Second, use the inverse transformation

$$X = F^{-1}(U) = \left(\frac{1-U}{c}\right)^{-\frac{1}{\gamma}}$$

  to obtain a Pareto distribution.
- Third, compute and plot the cumulative volatility σ_k from $X_1, \ldots, X_k$ for $k = 1, \ldots, n$:

$$\sigma_k = \frac{1}{k}\sum_{i=1}^{k}(X_i - \bar{E}(X))^2.$$

As indicated on Figure 2.1, in case that the theoretical volatility is finite (i.e. $\gamma > 2$), the sequence of cumulative volatilities converges to the theoretical value given by (3.2). On the other hand, this sequence diverges for $k \to \infty$ in case that $\gamma \leq 2$ since in this case, the theoretical volatility does not exist. In addition, the cumulative volatility exhibits fractal behavior: Once we omit the largest observation, the character of the plot does not change.

2.2 Generalized Pareto Distribution

The Pareto distribution introduced above is one of the distributions with fat tails. In addition, it is a "standardized distribution" in the sense that its mean and variance depend only on the parameter γ.

Its generalization is called Generalized Pareto Distribution. This distribution plays an important role in the Extreme Value Theory.

Definition 1. *Generalized Pareto Distribution (GPD) is a distribution with the following distribution function:*

$$G_{\xi,\beta} = \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0, \\ 1 - e^{-x/\beta} & \text{if } \xi = 1. \end{cases}$$

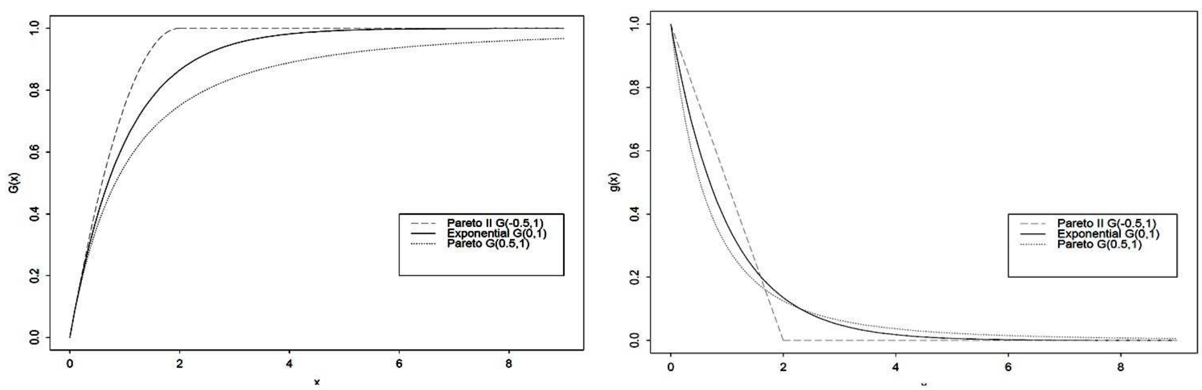

Figure 4.2: Distribution function (left) and probability density function (right) of the GPD.[2]

The parameter ξ (**shape parameter**) has an important impact on the character of the distribution:

- For $\xi > 0$ we have a Pareto distribution with $\gamma = 1/\xi$ (**tail index**),
- For $\xi = 0$ we have an exponential distribution,
- For $\xi < 0$ we have a distribution on a bounded interval $[0, -\beta/\xi]$ (so-called Pareto type II distribution).

3 Peaks over threshold

GPD can be used for the modelling of tails of distributions, i.e. for data exceeding certain threshold (peaks over threshold). To be more specific, let us denote by X the random variable representing the losses (taken by the positive sign). In addition, let us choose a threshold denoted by u.

Then the random variable $X - u$ represents the value of exceedances over threshold u, provided that this threshold has been exceeded. The distribution function of this random variable is called **excess distribution function** (denoted by F_u) and it is calculated as follows:

[2] Source: McNeil, A.J., Frey, R., Embrechts, P.: Quantitative Risk Management: Concepts, Techniques and Tools, Princeton University Press, 2005.

$$F_u(x) = P(X - u \le x \,|\, X > u) = \frac{P(X - u \le x \wedge X > u)}{P(X > u)} = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

One the most important results of the EVT says that this excess distribution function can be approximated by the GDP:

Theorem 1 (Balkema and de Haan, 1974, Pickands, 1975). *For a large class of distributions a function $\beta(u)$ can be found such that*

$$\lim_{u \to \bar{x}} \sup_{0 \le x < \bar{x} - u} |F_u(x) - G_{\xi, \beta(u)}| = 0,$$

where

- $\bar{x}$ *= rightmost point of the distribution,*
- u *= threshold,*
- F_u *= excess distribution function.*

The class of distributions for which this theorem is valid includes the majority of standard distributions (e.g. normal, lognormal, χ^2, t, F, beta, gamma, exponential, uniform etc.).

The interpretation of this theorem is as follows: The conditional distribution of sizes of exceedances above the threshold u can be "asymptotically" modeled by GPD.

Now, we will discuss several practical questions which are important when applying this method in the risk management:

- How to determine whether such a model corresponds with the real data?
- How to set the threshold (u)?
- How to estimate the parameters ξ and β in the GPD?
- How can these results be applied in practice (e.g. in VaR calculation)?

3.1 Quality of the model and setting the threshold

Both the model validation as well as setting the threshold is based on the so-called mean excess function.

Definition 2. Mean excess function *represents the conditional mean of the exceedance size over threshold (given that an exceedance occurred):*

$$e(u) = E(X - u \,|\, X > u).$$

The importance of this function lies in the fact that it is a linear function (in the variable u) in case of the GPD:

$$e^{GPD_{\xi,\beta}}(u) = \frac{\beta + \xi u}{1 - \xi}. \tag{4.3}$$

Hence, one can calculate the empirical mean excess function for different values of the threshold u and verify whether it is linear in u. Based on the definition, the empirical mean excess function is the average of all observations exceeding u decreased by u:

$$e^{EMP}(u) = \frac{\sum_{i=1}^{n} \max(X_i - u, 0)}{\sum_{i=1}^{n} I_{\{X_i > u\}}}.$$

In practice, this function is not linear for lower values of u but becomes approximately linear for higher values. Hence, the value of the threshold can be determined empirically as follows: draw chart $(X_i, e^{EMP}(X_i))$ and set $u := X_i$ for such i, from which the function is approximately linear. Let us note that this plot is called **sample mean excess plot**.

Note that if the GPD has fat tails with finite mean (i.e. $\xi \in (0,1)$), (3.3) implies that $e(u)$ is increasing in u.

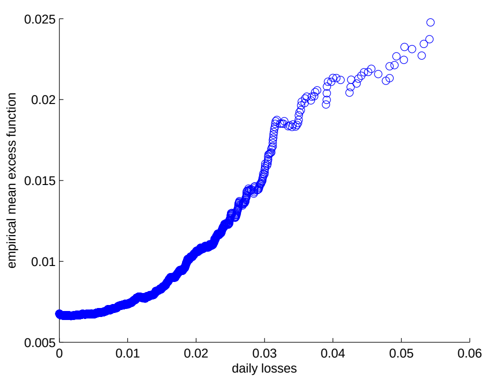

Figure 4.3: Sample mean excess plot for daily losses of S&P 500, 1950-2011 (last 0.25% data are omitted due to the instability of the estimations) This chart indicates that the threshold u might be set to 3.2 % (= 99 % quartile).

3.2 Estimation of the parameters by the maximum likelihood

The parameters ξ and β of the GPD (together with their confidence intervals) can be estimated using the maximum likelihood method. This method can be easily implemented in Matlab using the function `gpfit`:

```
losses=load('sp500.txt'); %Data for S&P 500, 1950 - 2011, losses only

q = quantile(losses,.99); % 99% quantile
y = losses(losses>q) - q; % exceedances

[parameters,confidence_interval]=gpfit(y);
xi=parameters(1)     % shape parameter
alpha=1/xi           % tail index
beta=parameters(2)   % scale parameter
```

For the given time series of S&P 500 returns (for 1950 – 2011), the estimated values of the parameters are $\hat{\xi} = 0.22$ (i.e. $\hat{\gamma} = 4.60$) and $\hat{\beta} = 0.015$, implying that the distribution has indeed a fat tail.

3.3 VaR calculation using GPD

Once we know the values of the parameters of the GPD distribution, we can use them to calculate the value at risk. We start from the fact that the GPD is a good approximation of the excess distribution function:

$$G_{\xi,\beta}(x) \approx F_u(x) = \frac{F(x+u) - F(u)}{1 - F(u)}.$$

In addition, $F(u)$ can be numerically approximated by

$$F(u) = \frac{N - N_u}{N},$$

where N denotes the total number of data (including profits) and N_u denotes the number of exceedances over the threshold u.

Hence

$$F(x+u) = (1 - F(u))\,G_{\xi,\beta}(x) + F(u) = \frac{N_u}{N}\,G_{\xi,\beta}(x) + \frac{N - N_u}{N}.$$

Let us denote the probability level of (1-day) VaR by α, i.e. we want to calculate $VaR^1(1-\alpha)$. Given that the random variable X denotes 1-day losses taken with the positive sign, we have

$$P(X \le VaR^1(1-\alpha)) = 1 - \alpha \qquad \Rightarrow \qquad F(VaR^1(1-\alpha)) = 1 - \alpha.$$

In addition, let x_p be such that $x_p + u = VaR^1(1-\alpha)$. Hence, one has

$$1 - \alpha = F(x_p + u) = \frac{N_u}{N}\,G_{\xi,\beta}(x_p) + \frac{N - N_u}{N},$$

implying that

$$x_p = G^{-1}_{\xi,\beta}\left(1-\alpha\frac{N}{N_u}\right) = \frac{\beta}{\xi}\left[\left(\alpha\frac{N}{N_u}\right)^{-\xi} - 1\right].$$

Finally

$$VaR^1(1-\alpha) = u + x_p = u + \frac{\beta}{\xi}\left[\left(\alpha\frac{N}{N_u}\right)^{-\xi} - 1\right].$$

3.4 Alternative method for estimating the parameters and calculating VaR

In this section, we will present an alternative method for estimating the parameters and calculating value at risk. The main advantage of this method is that it can be easily implemented in common spreadsheets application like Excel.

First, we will derive so called Hill estimator which is an estimator for the parameter ξ. Afterwards, the tail estimator will be introduced which allows for calculating VaR without knowing the value of the parameter β.

Hill estimator

The Hill estimator is derived as the maximum likelihood estimate of the power coefficient in the Pareto distribution. We use the fact that the distribution function $F(x)$ can be approximated by the Pareto distribution function given that x exceeds the threshold u.

Recall that from (3.1) the distribution function of the Pareto distribution for $x > x_m$ is

$$F(x) = 1 - cx^{-\gamma}$$

and the density function is

$$f(x) = \gamma c x^{-\gamma-1}.$$

The conditional density for x exceeding a sufficiently high value s is given by

$$f(x \mid x > s) = \frac{f(x)}{P(x > s)}.$$

Note that the unconditional density function has to be scaled by $1/P(x > s)$ in order to have

$$\int_u^\infty f(x \mid x > s)\, dx = 1.$$

Hence we have

$$f(x \mid x > s) = \frac{f(x)}{1 - F(s)} = \frac{\gamma c x^{-\gamma-1}}{cs^{-\gamma}} = \gamma\left(\frac{x}{s}\right)^{-\gamma} x^{-1}.$$

The maximum likelihood estimator is given by

$$\frac{\partial f}{\partial \gamma} = \frac{1}{\gamma} - \ln\frac{x}{s} = 0 \quad \Rightarrow \quad \frac{1}{\hat{\gamma}} = \ln x - \ln s.$$

Finally, x is replaced by the average of all losses exceeding s, in order to have the estimation more robust. Denoting the sorted values of $X_1, X_2, \ldots$ (in descending order) by $X_{(1)}, X_{(2)}, \ldots$ and denoting by k the index of the lowest observation exceeding s, one has

$$\frac{1}{\hat{\gamma}_k} = \frac{1}{k}\sum_{i=1}^{k} \ln X_{(i)} - \ln s.$$

In practice, we can calculate the Hill estimator for gradually decreasing s. Generally, the Hill estimator is rather unstable for several first highest values of s, following by a short stable region and by a large region of decreasing values afterwards. The estimated value of γ is then taken as the average value from the stable region.

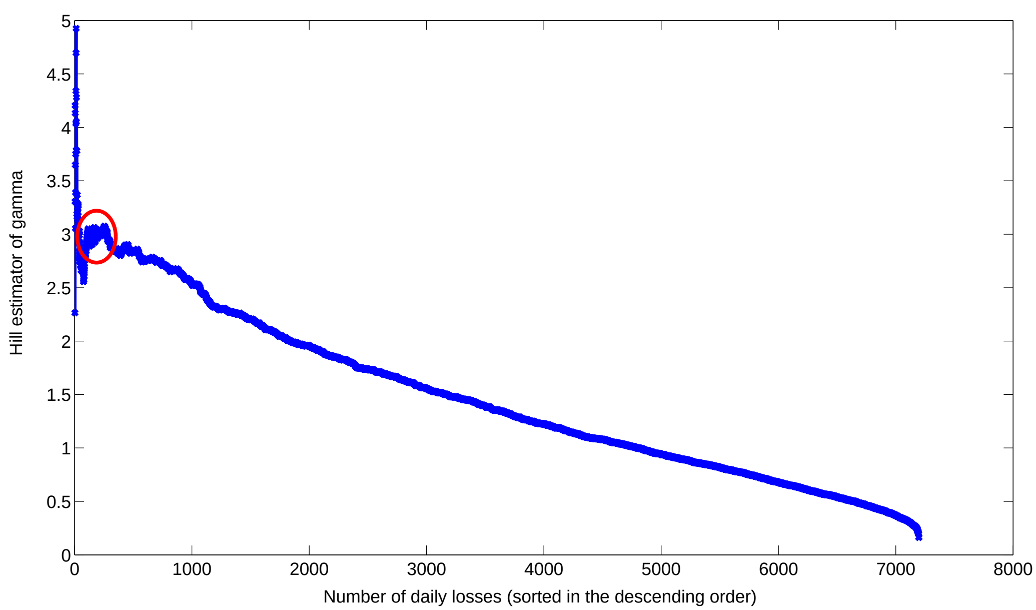

Figure 4.4: The estimated value of the parameter γ using the Hill estimator (S&P 500, 1950-2011)

Tail estimator

Once we have estimated the tail index γ using the Hill estimator, we can directly use it to estimate a high quantile (i.e. value at risk). The scale parameter (β) is not needed for this estimation.

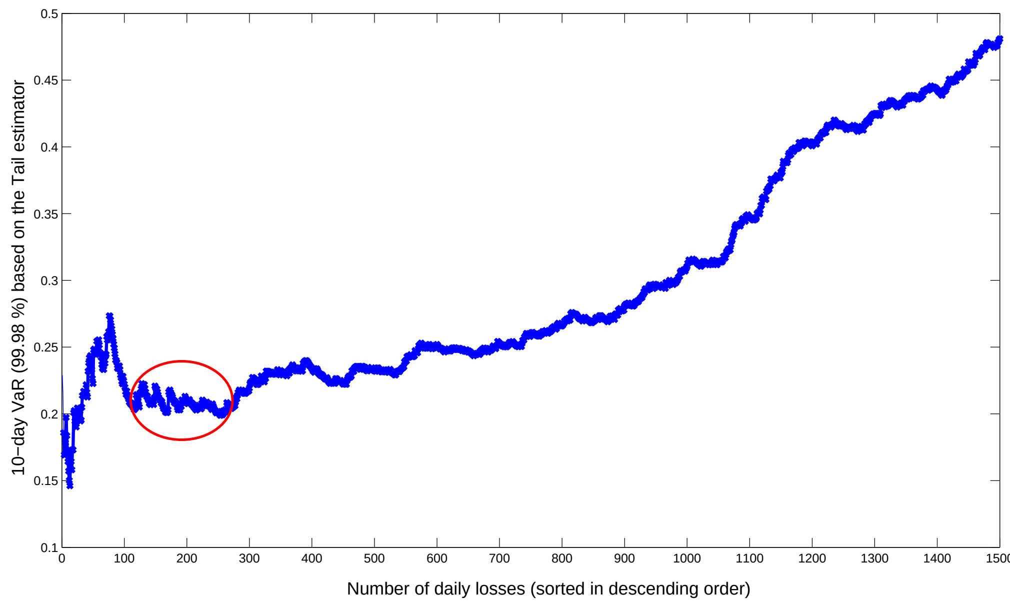

Figure 4.5: 10-day VaR based on the Tail estimator for daily returns S&P500 for 1950–2011 ($n = 15517$, $\alpha = 0.0002\%$, $m = 5$ (we take 5 observations))

The tail estimation can be derived as follows: For the probability level α, let us denote x_p as follows:

$$\alpha = P(x > x_p) = 1 - F(x_p) = c\, x_p^{-\gamma}.$$

In addition, let us denote an empirical quantile corresponding to the m-th largest loss by x_q. Hence, the probability level corresponding to this quantile is $\frac{m}{n}$, where n is the total number of observations including profits (assuming that $\frac{m}{n} > \alpha$). Hence, we have

$$\frac{m}{n} = P(x > x_q) = 1 - F(x_q) = c\, x_q^{-\gamma}.$$

Combining these two equations yields

$$x_p = x_q \left(\frac{m}{np}\right)^{\frac{1}{\gamma}}$$

The power coefficient in the previous equation $(1/\gamma)$ can be replaced by the Hill estimator corresponding to $s = x_q$. Again, the tail estimator is the average of the values in the stable region of the plot.

```
N=15519; % length of original data (both profits and losses)

for i=1:length(losses)
```

```
    MEF(i)=sum(losses(1:i))/i-losses(i);
    Hill(i)=1/(sum(log(losses(1:i)))/i-log(losses(i)));
    Tail(i)=losses(i)*(i/(N*p))^(1/Hill(i));
    VaR10EVT(i)=10^(1/Hill(i))*Tail(i);
end
```

4 Convolution rule

One of the theoretical results of the extreme value theory with an important application to the risk management is the convolution rule. This is related to the calculation of the distribution for the sum of random variables. The result is based on the following theorem:

Theorem 2 (Feller convolution theorem)**.** *If X and Y are both Pareto distributed (i.i.d.), then for a sufficiently large t one has*

$$P(X+Y>t)\approx 2ct^{-\gamma}=P(X>t)+P(Y>t).$$

This theorem establishes the additivity rule for Pareto-distributed random variables.

Its interpretation can be related to the well-known fact that a random variable with heavy tails has the following property: The sum of random variables is completely determined by the value of a single outlier.

This theoretical results can be used to calculate VaR with a holding period $d>1$ days (denoted by VaR^d) from VaR with a holding period 1 day (denoted by VaR^1). In the lecture notes on VaR, we have already mentioned that the scaling factor which is usually applied is based on the squared-root rule (i.e. $\mathrm{VaR}^d=\sqrt{d}\,\mathrm{VaR}^1$). The Feller convolution thoerem, however, implies a different scaling factor for VaRs based on extreme value theory:

Theorem 3 (Scaling of the value at risk based on EVT)**.** *Assuming iid Pareto-distributed losses, we have*

$$\mathrm{VaR}^d=d^{\frac{1}{\gamma}}\mathrm{VaR}^1,$$

where $\gamma>2$ is the tail index.

Proof:

Denoting the 1-day loss at time t by r_t^1, one has

$$\alpha=P(r_t^1>\mathrm{VaR}^1(1-\alpha))=c\left(\mathrm{VaR}^1(1-\alpha)\right)^{-\gamma}.$$

For d-day loss we have

$$\alpha = P(r_t^d > \text{VaR}^d(1-\alpha)) = P\left(\sum_{i=t}^{t+d-1} r_i^1 > \text{VaR}^d(1-\alpha)\right) \approx$$

$$\approx \sum_{i=t}^{t+d-1} P(r_i^1 > \text{VaR}^d(1-\alpha)) = \sum_{i=t}^{t+d-1} c\left(\text{VaR}^d(1-\alpha)\right)^{-\gamma} = dc\left(\text{VaR}^d(1-\alpha)\right)^{-\gamma},$$

where the approximation is based on the Feller convolution theorem. Since both the probabilities in the previous equations are equal to α, we can combine these two equations to obtain

$$c\left(\text{VaR}^1(1-\alpha)\right)^{-\gamma} = dc\left(\text{VaR}^d(1-\alpha)\right)^{-\gamma}.$$

Hence

$$\text{VaR}^d(1-\alpha) = d^{\frac{1}{\gamma}} \text{VaR}^1(1-\alpha). \quad \square$$

Note that the scaling factor implied by the previous theorem ($d^{\frac{1}{\gamma}}$) is lower than the usually used scaling factor ($d^{\frac{1}{2}}$). However, it should be also taken into account that the 1-day VaR based on the EVT tends to be higher than VaR calculated by other methods which are not based on the EVT. This is due to the fact that EVT addresses the heavy-tails property of the data.

5 Classical EVT

The theory on peaks over threshold which has been discussed so far represents an important part of EVT, mainly from the viewpoint of its application in the risk management. There is however the second important part of this theory which is called block maxima. In the common literature on EVT, this part is explained first and the theory on peaks over threshold is mentioned afterwards. Hence, it might be helpful to briefly introduce the block maxima method here together with its relation to the peaks over threshold method.

First, let us denote the block maximum: **Block maximum** is the maximum from n observations:

$$M_n := \max(X_1, \dots, X_n).$$

The important question is what is the distribution function of the block maxima, given the the distribution function of the random variables X_i is F (i.e. we assume that these random variables are iid).

One has:

$$P(M_n \le x) = P(X_1 \le x, ..., X_n \le x) = F^n(x).$$

The problem is that this distribution depends on the distribution of the underlying random variables. However, this underlying distribution is not known in practice. Hence, it would be helpful to know whether there is some kind of asymptotical distribution for a large value of n.

An analogous result is the central limit theorem which specifies the asymptotical distribution for sums of iid random variables. In particular, we know that the central limit theorem implies that the (appropriately normalized) sums $S_n := X_1 + ... + X_n$ have asymptotically the standard normal distribution:

$$\lim_{n\to\infty} P\left(a_n S_n + b_n \le x\right) = \Phi(x), \quad \text{where } a_n = \frac{1}{\sqrt{var(X_1)}} \text{ and } b_n = -\frac{nE(X_1)}{\sqrt{var(X_1)}}.$$

For block maxima, the asymptoticaly distribution exists as well and it is so called general extreme value distribution:

Theorem 4 (Fischer -Ű Trippet, 1928)**.** *If appropriately normalized block maxima (i.e. $a_n M_n + b_n$) converge in distribution to a non-trivial limit distribution, then this distribution is the* **generalized extreme value** *(GEV) distribution in the following form:*

$$H_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-1/\xi}\right) & \text{if } \xi \neq 0, \\ \exp\left(-e^{-x}\right) & \text{if } \xi = 0. \end{cases}$$

The parameter ξ determines three different types of GEV distributions:

(i) Fréchet ($\xi > 0$) [for any underlying distribution with fat tails],

(ii) Gumbel ($\xi = 0$) [for e.g. these underlying distributions: normal, lognormal, exponential, gamma, F, χ^2],

(iii) Weibull ($\xi < 0$) [e.g. for uniform underlying distribution].

Note that the distribution function of a GEV distribution is continuous in the parameter ξ.

An important result is that if the distribution of normalized block maxima of an underlying distribution F converges to the GEV distribution for block maxima with parameter ξ, then the distribution of exceedances over threshold converges to the GPD with the same parameter ξ.

Example 1. Example of standardization of maxima for the exponential distribution:

Distribution function: $F(x) = 1 - e^{-\lambda x}$

For $a_n = 1/\lambda$ and $b_n = \ln n/\lambda$, one has

$$F^n(a_n x + b_n) = \left(1 - \frac{1}{n}e^{-x}\right)^n,$$

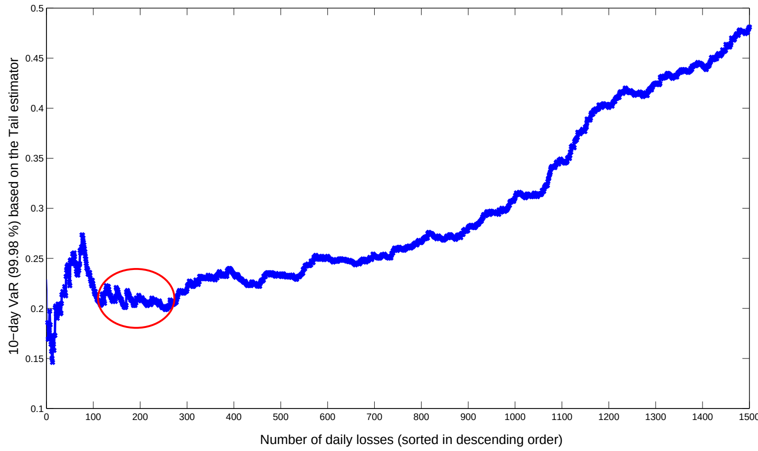

Figure 4.6: Distribution functions and densities of GEV distributions (dashed line – Weibull ($\xi = -0.5$), solid line – Gumbel ($\xi = 0$), dotted line – Fréchet ($\xi = 0.5$)).

hence

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-e^{-x}\right)$$

and the limit distribution is Gumbel distribution (i.e. $\xi = 0$).

Example 2. Example of standardization of maxima for the Pareto distribution:

Distribution function: $F(x) = 1 - cx^{-\gamma}$, where $\gamma > 0$.

For

$$a_n = (cn)^{\frac{1}{\gamma}} \qquad \text{and} \qquad b_n = 0$$

one has

$$F^n(a_n x + b_n) = \left[1 - \frac{1}{n} x^{-\gamma}\right]^n,$$

hence

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-x^{-\gamma}\right)$$

and the limit distribution is Fréchet distribution (i.e. $\xi = 1/\gamma > 0$).

6 Appendix

In this section, we will address two questions which might be important when applying the theoretical results to practice. In particular, so far we have assumed that the underlying data are iid. However, one should at least discuss the following two issues:

1. autocorrelation,

2. heteroskedasticity.

Autocorrelation

Assume that the random variables X_i are identically distributed but dependent. Intuition suggests that if data are dependent, the higher values tend to group together in clusters. Hence the effective size of the individual block might be considered to be smaller (e.g. θn instead of n, where $\theta \in (0,1)$). Therefore, one has

$$P(M_n \leq x) = [P(M_n^* \leq x)]^\theta = F^{\theta n}(x),$$

where $P(M_n* \leq x)$ denotes the value analogous to $P(M_n \leq x)$ for iid variables. Parameter θ is called **extremal index**.

Example: For ARCH(1) model with parameter $\alpha = 0.5$, the value of the extremal index is $\theta = 0.835$ (estimated value based on simulations). The main important result in this context is that the value of the extremal index does not influence on the value of parameter ξ in the Fréchet distribution.

Heteroskedasticity

If data are heteroskedastic where the volatility follows a GARCH model, we can apply the following methodology:

First estimate an (AR)-GARCH model, then apply a GEV distribution on the filtered (= standardized) residuals.[3]

7 Further reading

The following diploma thesis is a good reference for a basic overview of the EVT and its application:

Kovácsová (2008): Použitie teórie extrémnych hodnot vo finančníctve.

[3]For more details see e.g.: McNeil Ű Frey (2000): Estimation of tail-related risk measures for heteroskedastic financial time series: an extreme value approach, Journal of empirical finance.