Extreme Value Theory

Nicole B. Lipsky
Seminar in Computational Finance for CSE FS2016

July 20, 2016

1 Introduction

In this report, we summarize some results from Extreme Value Theory (EVT), following closely the findings in [2]. EVT focuses on extreme events, i.e. the events where the risk takes values from the tail of the distribution. In other words, we will examine extreme deviations from the mean.

In order to assess risk, it needs to be modeled. This is done using probability theory, where risks are random variables, following some unknown distribution. Historical data can be used to estimate the distribution.

Within EVT there are two methods to measure extreme event risk: Peaks-Over-Threshold (see Section 2) and Block Maxima Models (see Section 3). Note that when addressing extreme losses instead of gains, it is convention that a loss is a positive number and a profit is a negative number.

2 Peaks-Over-Threshold

2.1 Preliminaries

2.1.1 Excess Distribution

Peaks-Over-Threshold (POT) observes the excess distribution, i.e. the distribution of realizations with the cumulative distribution function (cdf) F over a high threshold u. The cdf of the excess distribution is:

$$F_u(y) := \mathbb{P}\{X - u \leq y | X > u\} = \frac{F(y+u) - F(u)}{1 - F(u)}, \tag{1}$$

for $0 \leq y < x_0 - u$, where $x_0 < \infty$ is the right endpoint of F. Intuitively, this is the probability that a loss exceeds the threshold u by at most an amount y, given that it exceeds the threshold.

2.1.2 Generalized Pareto Distribution

The distribution function which we will use to model the excess distribution is the Generalized Pareto Distribution (GPD), as defined as follows:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\beta), & \xi = 0, \end{cases}$$

where $\beta > 0$, $x \geq 0$ for $\xi \geq 0$, and $0 \leq x \leq -\beta/\xi$ for $\xi < 0$. Here, the parameter β is the *scaling* parameter and ξ is the *shape* parameter, so that F is heavy-tailed when $\xi > 0$. The special feature of the GPD is that it subsumes other distributions:

- If $\xi > 0$: $G_{\xi,\beta}$ is a reparametrized version of the ordinary Pareto distribution (heavy-tailed).
- If $\xi = 0$: $G_{\xi,\beta}$ is the exponential distribution.
- If $\xi < 0$: $G_{\xi,\beta}$ is a Pareto type II distribution.

2.1.3 GPD Limit Theorem

An important feature of the GPD is the following limit theorem, as described in [2], and which is a key result in EVT.

Theorem 1. *For a large class of underlying distributions there exists a function $\beta(u)$ such that*

$$\lim_{u \to x_0} \sup_{0 \leq y < x_0 - u} |F_u(y) - G_{\xi,\beta(u)}(y)| = 0$$

for $0 \leq y < x_0 - u$, where $x_0 < \infty$ is the right endpoint of F.

In words, GPD approximates the tails of these distributions above sufficiently high thresholds:

$$F_u(y) \approx G_{\xi,\beta(u)}(y). \tag{2}$$

Furthermore, we assume that for a certain u and for some ξ and β, our model for a risk X_i is exactly GPD. In the sense of this theorem, the GPD is the 'natural model' for the unknown excess distribution, which is essential for the POT technique.

2.2 Approach

The first step is to choose u, which is a balancing act between having to choose a sufficiently high threshold to follow the asymptotic theorem and a sufficiently low threshold to have sufficient material for estimation of the parameters. The second step is to fit historical observations over the threshold onto the GPD. When using the maximum likelihood estimation (MLE) the parameter values $\hat{\xi}$ and $\hat{\beta}$ are chosen in such a way that they maximize the joint probability density of the observations. Using the limit theorem from Equation 2, we now obtain $G_{\hat{\xi},\hat{\beta}(u)}(y) \approx F_u(y)$.

We set $s = u + y$. When combining Equation 1 and 2, we arrive at following result:

$$G_{\xi,\beta(u)}(x-u) \approx \frac{F(x)-F(u)}{1-F(u)}$$
$$\Rightarrow F(x) \approx (1-F(u))G_{\hat{\xi},\hat{\beta}(u)}(x-u) + F(u). \tag{3}$$

The last step is to estimate $F(u)$ in order to construct the tail estimator. [2] argues that the best way to estimate this parameter is the method of historical simulation, i.e. use the observations $X_1, \cdots, X_n$ to estimate $F(u) \approx \frac{n-N_u}{n}$, where n is the total number of observations and N_u is the number of observations above the threshold: $N_u = \#\{X_i > u\}$. [2] argues that the historical simulation method can, however, not be used to estimate the whole tail of $F(x)$ because in the tail of the distribution the data become too sparse. The threshold u is chosen is such a way that there are enough observations to enable a reasonable historical simulation estimate of $F(u)$, but beyond that point, historical simulation is a poor method.

The tail estimator can now be constructed by combining the estimate of $F(u)$ and MLE estimates of the GPD parameters:

$$\widehat{F}(x) = 1 - \frac{N_u}{n}(1 + \hat{\xi}\frac{x-u}{\hat{\beta}})^{-1/\hat{\xi}}, \tag{4}$$

given that $x > u$.

2.3 Application on Real Data

[2] provides an illustration of the POT method by taking an insurance example. In particular, their data consists of 2167 industrial fire insurance claims from Denmark covering a decade, from 1980 to 1990. The threshold u is set at 10 million Krone, which reduces the $n = 2167$ losses to $N_u = 109$ observations above the threshold. A visualization of the insurance claims and the losses above 10 million Krone can be seen in Figure 1.

While the author does not explain the rationale behind the decision to set u at that value, it becomes clear when observing the historical simulation estimate of $F(u)$ for this data, which turns out to be $\frac{n-N_u}{n} = \frac{2167-109}{2167} \approx 0.95$. Hence, 95% of the data are below the threshold, while the remaining 5% of the data form the excess distribution.

With $u = 10$ million, MLE estimates $\hat{\xi}$ and $\hat{\beta}$ to be 0.5 and 7.0, respectively. Figure 2 depicts the estimated GPD model for the excess distribution and one can easily see that the GPD model fits the excess losses nicely. Figure 3 shows the tail probabilities $1 - F(x)$ on a log scale. It is evident that the GPD tail estimation works well for extrapolation into the area where the data becomes sparse.

2.4 Value-at-Risk and Expected Shortfall

The goal of the POT method is to ultimately calculate the Value-at-Risk (VaR) and the Expected Shortfal (ES), two measures which attempt to describe the tail of a loss distribution.

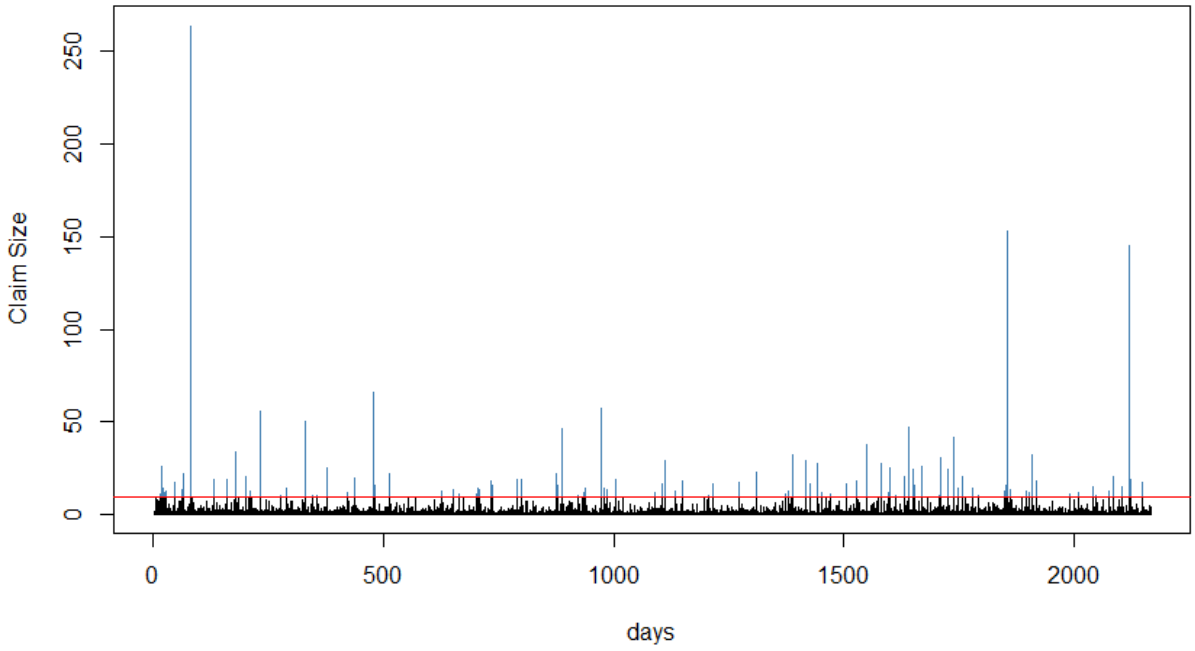

Figure 1: 2167 industrial fire insurance claims from Denmark from 1980 to 1990 with the red line representing the threshold u at 10 million Krone.

VaR is the $(100q)^{\text{th}}$ percentile of the distribution F:

$$\text{VaR}_q = F^{-1}(q),$$

where F^{-1} is the inverse of F and q a high quantile of the distribution of losses (typically 95^{th} or 99^{th} percentile). In words, VaR provides an upper bound for a loss for a certain confidence level. The shortcoming of the VaR is that no information on the severity of the loss is given when the VaR-limit is exceeded. Furthermore, VaR is not 'coherent', in the sense that diversification is penalized: For two independent random variables X and Y, $\text{VaR}_q(\frac{1}{2}X + \frac{1}{2}Y) \geq \frac{1}{2}\text{VaR}_q(X) + \frac{1}{2}\text{VaR}_q(Y)$.

The connection between POT and VaR is that for a given probability $q > F(u)$, the VaR estimate is calculated by inverting the POT tail estimator (Equation 4):

$$\widehat{VAR_q} = u + \frac{\hat{\beta}}{\hat{\xi}}((\frac{n}{N_u}(1-q))^{-\hat{\xi}} - 1)$$

The ES is the expected loss of a portfolio value given that a loss is occurring at or above the VaR-quantile. Conversely to VaR, ES is a coherent risk measure. ES is related to VaR by the fact that ES_q is just the mean of the excess distribution over the threshold VaR_q, as shown in the following equation:

$$ES_q = VAR_q + \mathbb{E}[X - VAR_q | X > VAR_q]$$

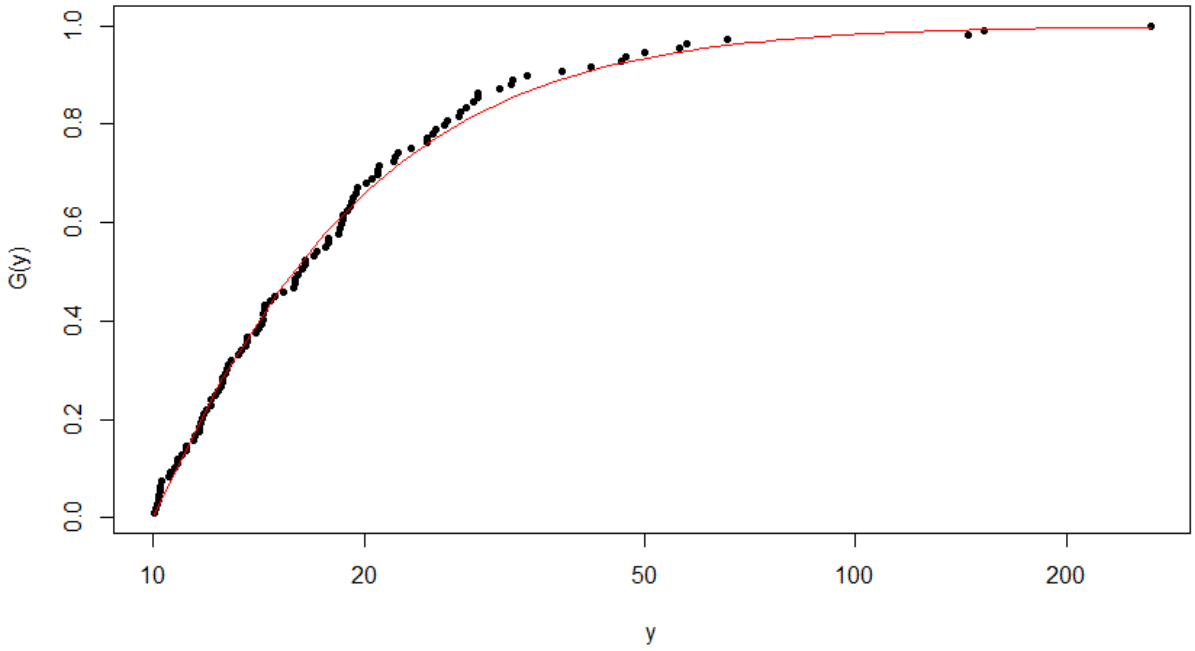

Figure 2: Estimated GPD model for the excess distribution above the threshold.

Rewriting this formula yields:

$$
\begin{aligned}
ES_q &= VAR_q + \frac{\beta + \xi(VaR_q - u)}{1-\xi} \\
\frac{ES_q}{VAR_q} &= 1 + \frac{\beta + \xi(VaR_q - u)}{(1-\xi)VaR_q} \\
&= \frac{1}{1-\xi} - \frac{\xi VaR_q}{(1-\xi)VaR_q} + \frac{\beta + \xi VaR_q - \xi u}{(1-\xi)VaR_q} \\
&= \frac{1}{1-\xi} + \underbrace{\frac{\beta - \xi u}{(1-\xi)VaR_q}}_{\to 0 \text{ for } x_0 \to \infty}
\end{aligned}
$$

Note that for an infinite right endpoint x_0, the second term of the right hand side of the previous equation goes to zero. In that case, the ratio of ES to VaR is then largely determined by the factor $\frac{1}{1-\xi}$, which shows the importance of the shape parameter ξ.

The VaR and ES of the demonstration data are estimated to be 27 and 58, respectively.

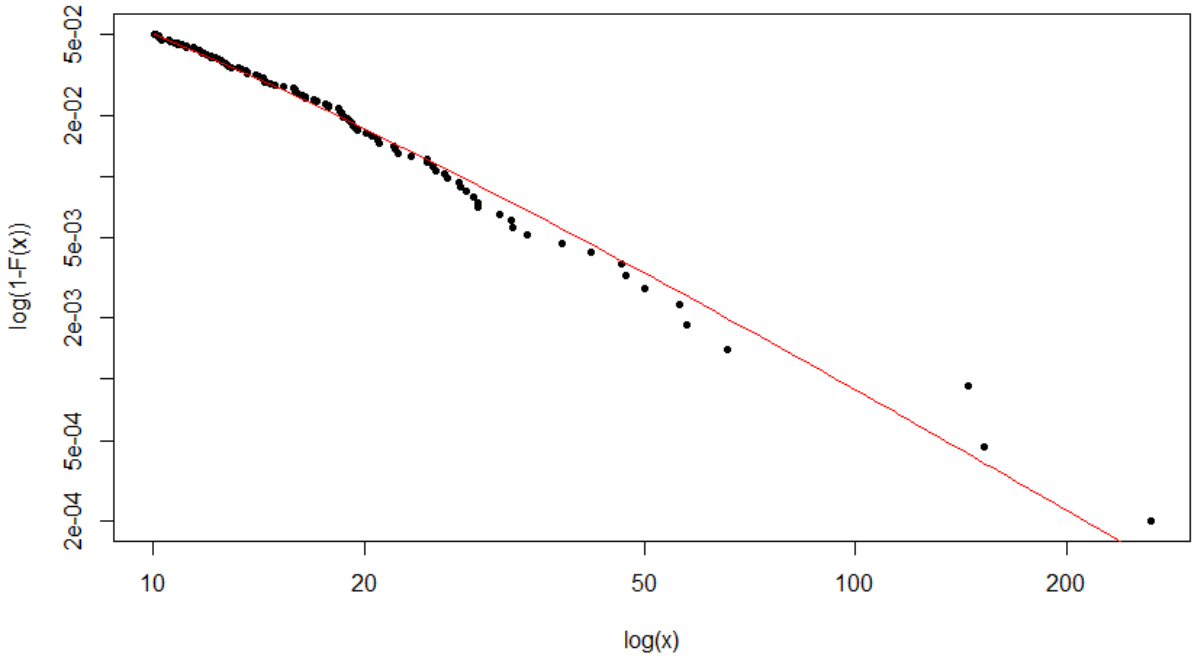

Figure 3: Tail probabilities $1 - F(x)$: The graph shows the probability of an extreme value of a given size, e.g. the threshold of 10 corresponds to a tail probability of 0.05 (top left corner).

3 Block Maxima Models

3.1 Preliminaries

3.1.1 Block Maximum

The second kind of model for extreme values is the Block Maxima Model (BMM). It is a model for the largest observations collected from blocks of identically distributed observations. BMM plays a large role for the analysis of stress losses.

We define M_n as the maximum values of a block of identically distributed losses: $M_n := \max(X_1, \ldots, X_n)$. The normalized block maximum is therefore: $(M_n - b_n)/a_n$, with a sequence of numbers $a_n > 0$ and b_n.

3.1.2 Generalized Extreme Value Distribution

Analogous to the POT method above, we make use of a certain distribution function to model extreme values, namely the generalized extreme value distribution (GEV). The distribution

function of the GEV is defined as:

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

where $1 + \xi x > 0$. As for the GPD, ξ is the shape parameter and GEV also subsumes other distributions:

- If $\xi > 0$: H_ξ is a Fréchet distribution.
- If $\xi = 0$: H_ξ is a Gumbel distribution.
- If $\xi < 0$: H_ξ is a Weibull distribution.

3.1.3 Fisher-Tippett Theorem

The reason why we use the GEV for the BMM is that GEV is the natural limit distribution for normalized maxima, which is described in the Fisher-Tippett Theorem ([2]):

Theorem 2. *If F is in the maximum domain of attraction of a non-degenerate H, then this limit must be an extreme value distribution of the form $H(x) = H_\xi((x-\mu)/\sigma)$, for some ξ, μ, and $\sigma > 0$.*

This theorem basically says that the distribution of the normalized maxima of the block observations converge to some limiting distribution H as the block size increases. Essentially, GEV is the only possible limiting distribution for normalized block maxima.

3.2 GEV Estimation

We take again the insurance claims data from above for demonstration purposes. The first step is to divide the sample data into blocks of equal length. As the data contains approximately one insurance claim every day, we chose to set the block length to 30, i.e. one block corresponds to samples from one month. The appropriate choice of a block length needs to balance the fact that the smaller the block sizes, the more blocks we have, and therefore more data points to estimate the GEV, while large enough block sizes are needed to obtain the limiting distribution H. As the calendar naturally suggests periods, like months or years, the choice of 30 as the block length seems appropriate. Figure 4 illustrates the same histogram of insurance claims as above, with vertical lines indicating the blocks.

The next step is to determine the maximal value of each block. Then, we apply MLE to estimate the parameter ξ of the GEV distribution function as well as μ and σ, the two normalization parameters: We obtain 0.5, 12.6, and 8.8 for ξ, μ, and σ, respectively. Inserting these values into the formula yields the estimated GEV model. In Figure 5 the estimated GEV model for the block maxima is again shown as a smooth curve, matching the empirical distribution of the block maxima values, which suggests that BMM is also an appropriate method for extreme risk assessment.

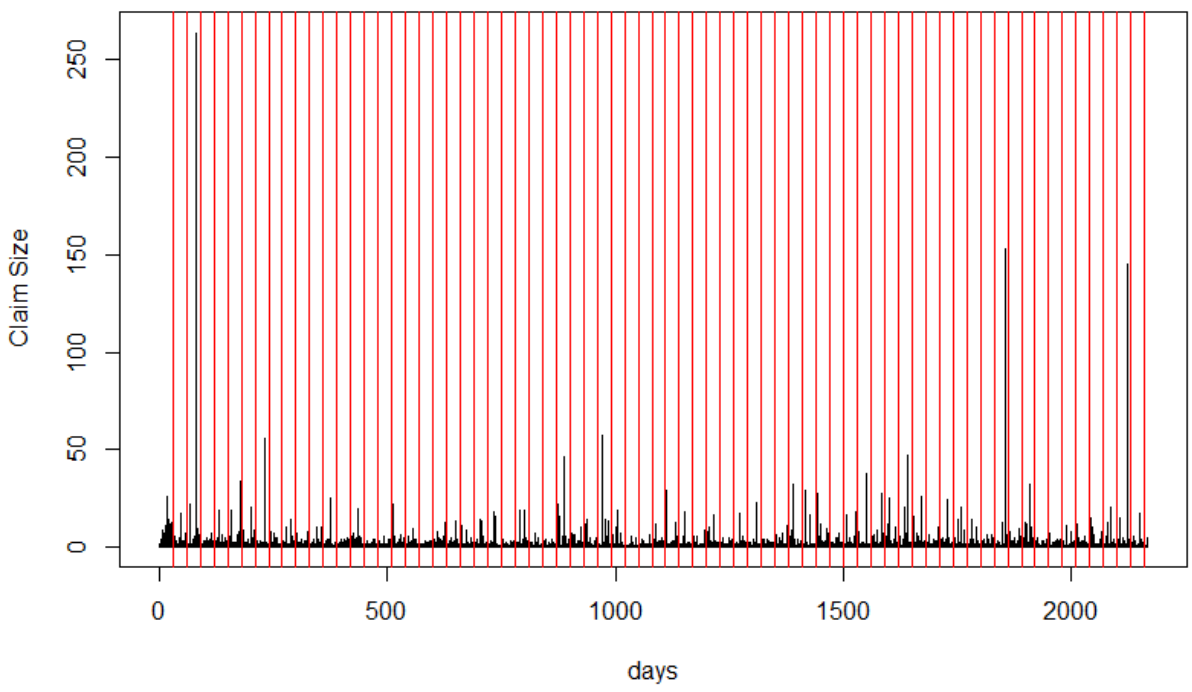

Figure 4: 2167 industrial fire insurance claims, divided in 72 blocks of length 30 each.

3.3 Stress Loss

A quantile of the distribution $H_{\xi,\mu,\sigma}$ is called a stress loss. This means, e.g. when fitting $H_{\xi,\mu,\sigma}$ to to our monthly maxima of daily negative returns, $H^{-1}_{\xi,\mu,\sigma}(0.95)$ gives the largest loss to expect every 20 months (which, of course, does not mean to expect the loss *in* 20 months; it might as well occur tomorrow).

4 Comparison

4.1 Differences

In this section we compare the two methods, POT and BMM. POT is a more modern method, and according to [2], "most useful for practical applications, due to their more efficient use of the (often limited) data on extreme values". Note that the BMM requires a lot of data, as blocks must be defined and then the data is reduced to block maxima only. The goal of the POT is to calculate the *Value-at-Risk* and *Expected Shortfall* from financial returns, while the BMM is useful for the analysis of *stress losses*. One key difference is that the POT method picks up all relevant high observations, while the BMM might miss some of these high observations and instead keep some lower ones ([1]).

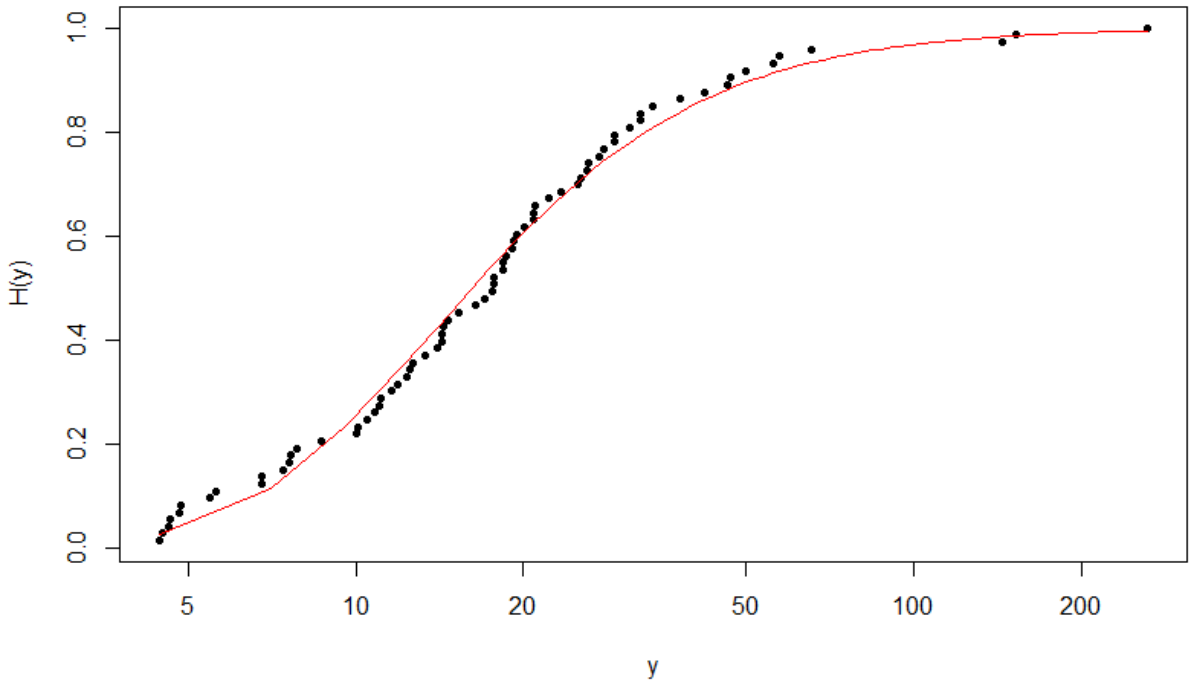

Figure 5: Estimated GEV model for the block maxima distribution.

However, BMM sometimes has advantages over the POT, which are described in detail in [1]: First, sometimes block maxima are the only available pieces of information, e.g. historical yearly records. Secondly, blocks might be naturally given, e.g. a seasonal periodicity in case of yearly maxima or a dependence within blocks but not in between blocks plays a role. In this case of seasonality, BMM may also be easier to be applied since no choice regarding block lengths must be made as the block periods appear naturally, while choosing a high threshold in the POT method can sometimes be difficult and have less predictive power.

4.2 Investigation in the plots almost aligning

When combining both the estimated GPD model and the estimated GEV model in one plot, as can be seen in Figure 6, one can observe that their curves are similar, but yet not perfectly aligned. While one might be tempted to assume that they should resemble each other more, we would like to present counter-arguments against this assumption and elaborate where the similarities and the differences stem from.

Out of all observations S of the data the following subsets are selected for POT and BMM respectively:

- $S_{\texttt{POT}} = \{s \in S | s \geq u\}$

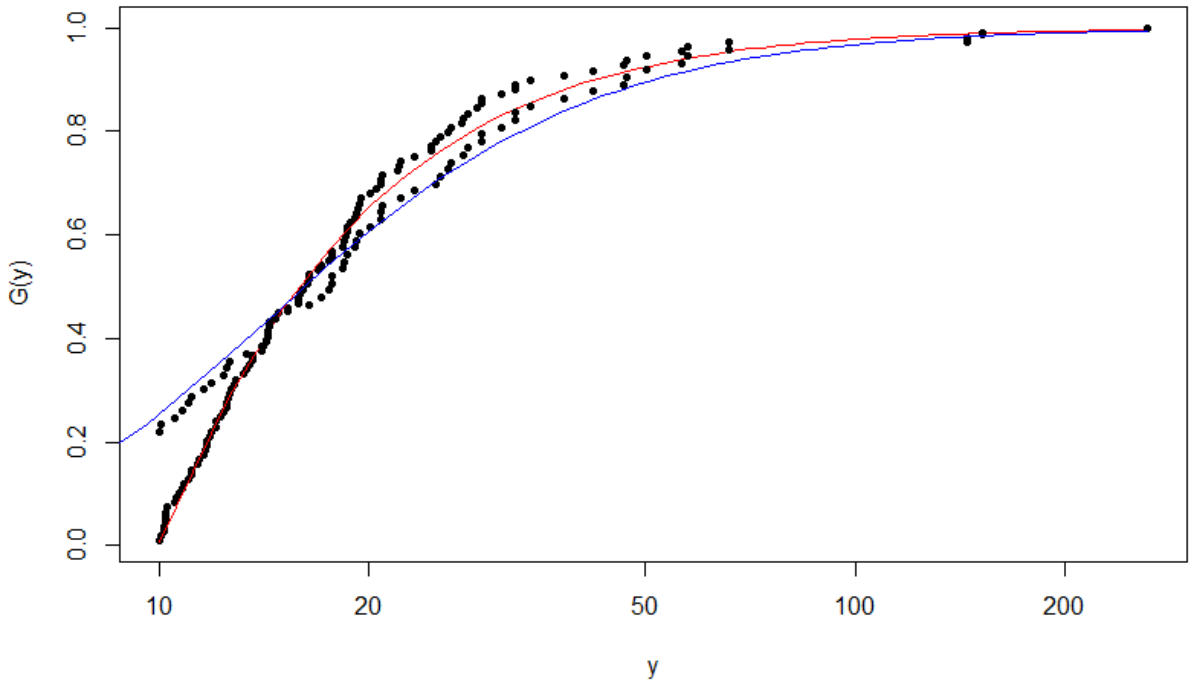

Figure 6: Red line: estimated GPD model; blue line: estimated GEV model.

- $S_{\texttt{BMM}} = \{s \in S | \exists\ Block\ B : s = \max_{b \in B}(b)\}$

Observation 1. The dots in the plot represent the following two mappings:

- $p : S_{\texttt{POT}} \to [0,1]$,
  $s \mapsto \hat{\mathbb{P}}[S < s | s \geq u]$
- $b : S_{\texttt{BMM}} \to [0,1]$,
  $s \mapsto \hat{\mathbb{P}}\left[S < s | \exists\ Block\ B : s = \max_{b \in B}(b)\right]$

Observation 2. Dots which align vertically therefore represent the same sample $s \in S_{\texttt{POT}} \cap S_{\texttt{BMM}}$. Dots which align horizontally represent samples at equal cumulative probability. Therefore, if two dots align in both axis this is equivalent to the observation that the sample $s \in S_{\texttt{POT}} \cap S_{\texttt{BMM}}$ they represent satisfies $\hat{\mathbb{P}}[S < s | s \geq u] = \hat{\mathbb{P}}\left[S < s | \exists\ Block\ B : s = \max_{b \in B}(b)\right]$.

Observation 3. If $S_{\texttt{POT}} \subseteq S_{\texttt{BMM}}$, then all block maxima are above the threshold u.
If $S_{\texttt{BMM}} \subseteq S_{\texttt{POT}}$, then each value of the excess distribution over u is also a block maximum. Set equality implies that only block maxima are above the threshold and every block maxima is above the threshold.

Observation 4. To assume that all points in the plot align is to assume that $\hat{\mathbb{P}}[S < s | s \geq u] = \hat{\mathbb{P}}[S < s | \exists \textit{ Block } B : s = \max_{b \in B}(b)]$ for all s (See Observation 2). This in turn implies by the GPD Limit Theorem and the Fisher-Tippett Theorem not least that there exists the appropriate parameters (which must coincide with the parameters referenced in either of these theorems) such that $G_{\xi,\beta}(x) = H_{\xi}(x)$. Neither of these quite extraordinary claims appears to be proven or intuitive.

Observation 5. The largest point of the POT model and the largest point of the BMM model are at exactly the same spot in the top right corner and have a probability of 1.0 because it is the estimate for the final point of both cdfs. Scaling either function with a constant factor or offset would destroy this structure and must therefore be ruled out, since the above observation is indeed provably needed.

Observation 6. The 'gradients' from one sample point to another are similar across the two curves because by construction of the cdf estimation each sample point adds $\frac{1}{N}$ to the estimated cdf, where N is the number of samples. Therefore, if N is approximately the same for both methods and $S_{\texttt{POT}} \approx S_{\texttt{BMM}}$ on some interval, then the 'gradient' is similar on that interval.
If and only if the largest such interval happens to be the entirety of the domain (i.e. encloses all samples of S), then the perceived curves will fit nicely since both endpoints are fixed at 0 and 1, respectively.

Observation 7. Clearly, for $u = 0$ and a block-width of 1, we would trivially have $S = S_{\texttt{POT}} = S_{\texttt{BMM}}$, which would show aligned points. However, at that point both GPD Limit Theorem and Fisher-Tippett Theorem would be out of the window. We have no indication that there should exist a non-trivial pair of threshold and block-width such that $S_{\texttt{POT}} \approx S_{\texttt{BMM}}$.

In conclusion, while both methods aim in assessing extreme value risk, they still do quantify different risks. While the POT model assesses the risk of a loss given that a threshold is exceeded, the BMM model assesses the risk of a high loss within a time interval. Ultimately, they will end up with many of the same sample points but not exactly the same, wherefore the estimated cdf will look differently. Finally, the GPD distribution $G_{\xi,\beta}(x)$ and the GEV distribution $H_{\xi}(x)$ are two different functions and will therefore in most cases yield two different curves.

References

[1] Ana Ferreira and Laurens de Haan. On the block maxima method in extreme value theory. *Annals of Statistics*, December 2014.

[2] Alexander McNeil. Extreme value theory for risk managers. *Departement Mathematik, ETH Zürich*, May 1999.