

3

Threshold methods

3.1 Background and theoretical motivation

Threshold methods use a more natural way of determining whether an observation is extreme - *all* observations greater than some high value (*threshold*) are considered. This allows more efficient use of data and avoids the problems that can arise as a result of blocking (see Section 2.6), but brings its own problems. We must first go back and consider the asymptotic theory appropriate for this new situation.

Suppose once more that X_1, X_2, \dots, X_n is a sequence of IID random variables having marginal distribution F , and – once again – let

$$M_n = \max \{X_1, \dots, X_n\}.$$

We know from Chapter 2 that

$$\Pr \{M_n \leq x\} \approx G(x),$$

where

$$G(x) = \exp \left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]_+^{-1/\xi} \right\}$$

is the Generalised Extreme Value distribution with location, scale and shape μ , σ and ξ respectively.

Theorem (*Distribution of threshold excess*)


For a large enough threshold u , the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)_+^{-1/\xi}, \quad (3.1)$$

defined on $y > 0$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \quad (3.2)$$

Proof

 ...proof continued...

The family of distributions defined by Equation (3.1) is known as the *Generalised Pareto family*; the distribution itself is often referred to as the Generalised Pareto Distribution, or GPD for short.

Comments

- If block maxima have approximate distribution G , then threshold excesses have a corresponding distribution given by the Generalised Pareto family.
- The parameters of the GPD are uniquely determined by those of the GEV:
 - The parameter ξ in Equation (3.1) is equal to that of the corresponding GEV;
 - The GPD scale parameter is a function of the GEV location and shape parameters.
- Estimates of the GEV parameters are sensitive to the size of block chosen to identify extremes; estimates of the GPD parameters are ‘stable’.
- The duality between the GEV and GPD means that the shape parameter ξ is dominant in determining the qualitative behaviour of the GPD:
 - If $\xi < 0$ the distribution of excesses has an upper bound;
 - if $\xi > 0$ the distribution has no upper limit;
 - the case $\xi = 0$ is also unbounded, and is taken as the limit $\xi \rightarrow 0$, giving

$$H(y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0;$$

i.e. an exponential distribution with rate $1/\tilde{\sigma}$.

- Until this point, we have used the notation $\tilde{\sigma}$ to denote the scale parameter of the GPD, so as to distinguish it from the corresponding parameter of the GEV. For notational convenience we now drop this distinction, using σ to denote the scale parameter within either family.

Example 3.1

Suppose X_1, X_2, \dots, X_n is a sequence of independent $\exp(1)$ random variables. Show that the limiting distribution of threshold excesses belongs to the generalised Pareto family.



Example 3.2

Suppose X_1, X_2, \dots, X_n is a sequence of independent $U(0, 1)$ random variables. Show that the limiting distribution of threshold excesses belongs to the generalised Pareto family.



3.2 Application: rainfall extremes in New York

The file `newyork.txt`, available to download from the course webpage, gives daily rainfall accumulations (in mm) for New York City for the years 1914–1961 (inclusive). Much of the northeastern United States is relatively low-lying and so prone to flooding; this is made much worse on the odd occasion that a hurricane travels this far north (for example, “Superstorm Sandy” in 2012). Thus, analysing extreme rainfall data has a real practical motivation here, in terms of river and sea flood defence systems. In this Section, we will illustrate a complete threshold-based analysis of the rainfall extremes observed at New York.

The data have been scanned into R from the course webpage and stored in the vector `rain`:

```
> rain = scan('newyork.txt')
```

3.2.1 Threshold choice

The *threshold stability property* of the GPD means that if the GPD is a valid model for excesses over some threshold u_0 , then it is valid for excesses over all thresholds $u > u_0$. Denoting by σ_{u_0} the GPD scale parameter for excesses over threshold u_0 , the expected value of our threshold excesses, conditional on being greater than the threshold, is

$$E[X - u | X > u] = \frac{\sigma_{u_0} + \xi u}{1 - \xi}. \quad (3.3)$$

Thus, for all $u > u_0$, $E[X - u | X > u]$, is a linear function of u . Furthermore, $E[X - u | X > u]$ is simply the mean of the excesses of the threshold u , for which the sample mean of the threshold excesses of u provides an estimate. This leads to the *mean residual life plot*, a graphical procedure for identifying a suitably high threshold for modelling extremes via the GPD. In this plot, for a range of candidate values for u we identify the corresponding mean threshold excess; we then plot this mean threshold excess against u , and look for the value u_0 above which we can see linearity in the plot.

We can easily do this from first principles in R. First of all, we set up a vector of possible thresholds, starting at zero and going up to the maximum value in our dataset:

```
> u = seq(0, max(rain), 0.1)
```

The vector `x` is then set up to take the corresponding values for the mean excess over each value in `u`:

```
> x = vector('numeric', length(u))
```

The following code computes the mean excess for each value in u and stores it in x :

```
> for(i in 1:length(x)){
+   threshold.exceedances = rain[rain>u[i]]
+   x[i] = mean(threshold.exceedances-u[i])
+ }
```

The MRL plot is produced using the following code, giving the plot in Figure 3.1:

```
> plot(x~u, type='l', main='MRL plot', ylab='mean excess')
```

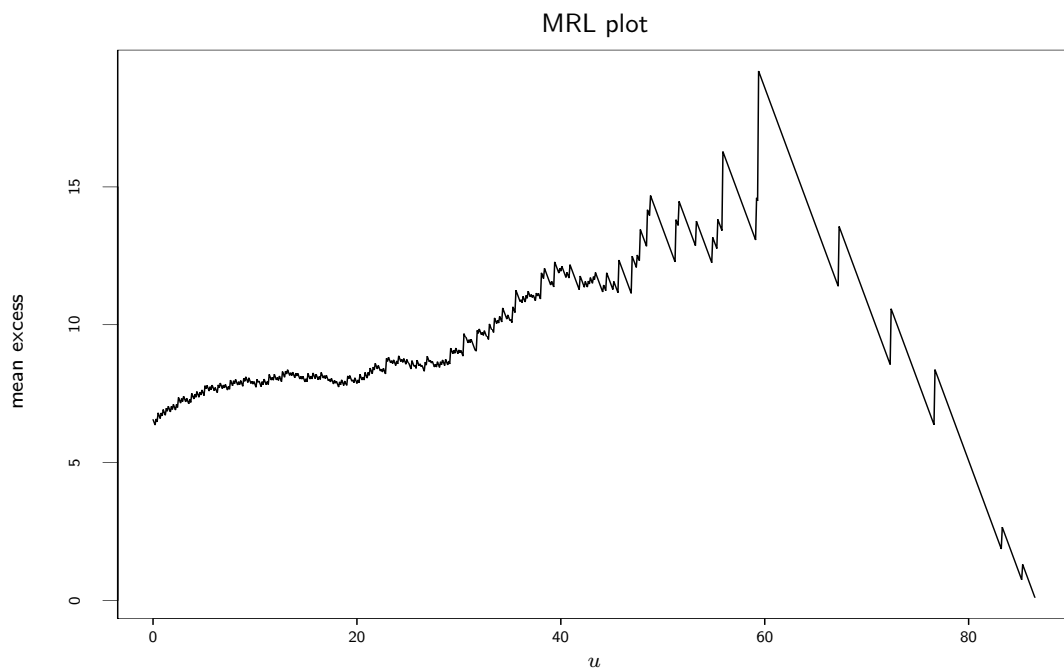


Figure 3.1: Mean residual life plot for the New York rainfall data.

Though interpretation of these plots can be subjective, linearity in Figure 3.1 might be suggested above $u_0 = 30\text{mm}$ (information in the far right-hand-side of these plots is unreliable; here, variability is high due to the limited amount of data above such high thresholds). Of course, there is a function in the `ismev` package that produces exactly the same plot, but including 95% confidence bounds for the mean excess:

```
> library(ismev)
> mrl.plot(rain)
```


3.2.2 Data pre-processing

In problems sheet 1, we considered how to use R to extract the set of block maxima to be modelled by the generalised extreme value distribution. Writing R code to do this can be a time-consuming process, and the code needs to be written specifically for the data being analysed. In a threshold-based analysis the data pre-processing is far more straightforward. Using $u_0 = 30$ as our threshold for identifying extremes (see Figure 3.1), we can easily obtain our set of threshold exceedances for modelling with the generalised Pareto distribution:

```
> above.threshold = rain[rain>30]
> threshold.exceedances = above.threshold-30
```

We can look at our set of threshold exceedances by typing:

```
> threshold.exceedances
 [1]  1.8  2.5  1.8 14.5  0.5 13.2  5.6  8.1  2.0  1.8  3.0  9.1  0.5
1.8  2.3
[16]  3.0  0.5  2.5 18.5  5.3 10.6  0.5  4.3  2.8  0.5 15.7  1.8  3.5
3.5  1.8
[31]  4.8  5.3  7.8 46.7  2.3  4.0  3.8  6.6  0.5 15.7 56.6  5.6 17.8
17.5  4.3
[46] 18.5  0.7 13.4 29.4  5.1 23.3  3.5  0.5  0.2 10.9 12.7 53.3 24.9
29.2  1.8
[61]  7.3  2.5  4.0 37.3  1.2  0.2  6.1  6.8  8.4  1.0  3.3 17.0  2.0
3.0  8.1
[76]  0.5 42.4  4.3  7.1  3.0 10.9  9.9 17.0  6.3  0.5  0.5 25.9  1.8
21.3 55.3
[91] 11.9  0.5  3.0  5.6 25.9 14.2  8.1  4.3  1.8  2.0  1.8  5.6 15.2
0.5  9.4
[106] 0.2 14.5  1.8  3.8 21.6  5.3 29.4  3.5  5.3  0.5  6.8 17.8 12.9
7.6 25.4
[121] 5.3 12.4  3.0  3.0 10.1  4.8  8.1  9.4  4.0  5.6  4.3  3.5  1.0
6.6  6.3
[136] 8.4  8.1 17.0  1.0  0.5  1.2  5.6 18.8 11.9  1.7  1.2 21.3  3.5
7.6  9.4
[151] 9.4 15.7
```

Thus, we have identified 152 observations as being extreme. Compare this to an analysis of annual maxima in which we would have only 48 observations to work with – one from each year.

3.2.3 Fitting the GPD

The most commonly-used approach to fit the GPD to the set of threshold excesses is that of maximum likelihood. The GPD log-likelihood function can be derived in the usual way; this is left as an exercise (actually, you are asked to do this in problems sheet 2!), but can be shown to be:

$$\ell(\sigma, \xi; \mathbf{y}) = -152 \log \sigma - (1 + 1/\xi) \sum_{i=1}^{152} \log \left(1 + \frac{\xi y_i}{\sigma} \right)_+, \quad (3.4)$$

where $\mathbf{y} = (y_1, \dots, y_{152})$ are the set of exceedances above threshold $u_0 = 30$. For the case $\xi = 0$, interpreted as $\xi \rightarrow 0$, we have the log-likelihood for an exponential distribution with rate $1/\sigma$ (again, see problems sheet 2). In R, we could now proceed as we did in Section 2.2.2 when fitting the GEV; that is, we could write a function which computes the (negative) log-likelihood for the GPD (using the 152 threshold excesses), and then use the `nlm` routine to minimise this with respect to $\theta = (\sigma, \xi)$. Actually, this is exactly what you will have to do in problems sheet 2! Here, we will obtain maximum likelihood estimates of the GPD parameters for our rainfall extremes by using the `gpd.fit` function in `ismev`:

```
> gpd.fit(rain, 30)
$threshold
[1] 30

$nexc
[1] 152

$conv
[1] 0

$nullh
[1] 485.0937

$mle
[1] 7.4422639 0.1843027

$rate
[1] 0.008669861

$se
[1] 0.9587773 0.1011714
```

Thus, our fitted model is given by

$$\hat{\sigma} = 7.440(0.958) \quad \hat{\xi} = 0.184(0.101)$$

Note also that the output gives the number of exceedances (`nexc` = 152), the threshold exceedance rate (`rate` = 0.0087; just the number of exceedances divided by the total length of the series, that is, 152/17532), and the value of the minimised negative log-likelihood (`nllh`; thus, the maximised log-likelihood is -485.0937).

3.2.4 Model adequacy

We can use probability plots and quantile plots to check the suitability of the fitted GPD to the set of extracted threshold exceedances; as demonstrated in Chapter 2 (Section 2.2.3), such plots are easily done from first principles in R; however, the function `gpd.diag` in `ismev` does for the GPD exactly what `gev.diag` did for the GEV in Chapter 2 (Section 2.2.5):

```
> A = gpd.fit(rain, 30, show = FALSE)
> gpd.diag(A)
```

The resulting plots are shown overleaf in Figure 3.2; all diagnostics seem to indicate a reasonable fit of the GPD to our rainfall extremes.

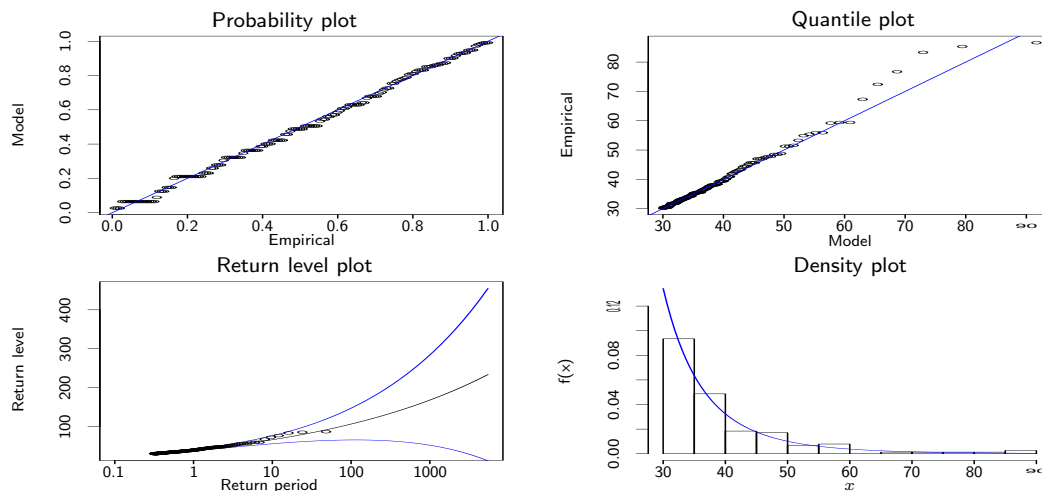


Figure 3.2: Diagnostic plots indicating the goodness-of-fit of the GPD to the New York rainfall extremes.

3.2.5 Return level estimation

Figure 3.2 seems to indicate that the GPD is suitable for our set of threshold exceedances y_1, \dots, y_{152} ; that is,

$$\Pr(X > u + y | X > u) \approx \left[1 + \frac{\hat{\xi}y}{\hat{\sigma}} \right]_+^{-1/\hat{\xi}}, \quad (3.5)$$

for $\xi \neq 0$. Working with the left-hand-side of (3.5), we see that

$$\Pr(X > u + y | X > u) = \frac{\Pr(X > u + y)}{\Pr(X > u)},$$

giving

$$\Pr(X > u + y) = \Pr(X > u) \Pr(X > u + y | X > u).$$

After substitution of (3.5), we get

$$\Pr(X > u + y) \approx \hat{\lambda}_u \left[1 + \frac{\hat{\xi}y}{\hat{\sigma}} \right]_+^{-1/\hat{\xi}}, \quad (3.6)$$

where $\lambda_u = \Pr(X > u)$ and is estimated as the empirical threshold exceedance rate $\hat{\lambda}_u$. Now each y_i , $i = 1, \dots, 152$, are the raw rainfall observations (exceeding the threshold) minus the threshold ($x_i - u$), as the GPD models the magnitude of excess over u ; substitution of $y_i = x_i - u$ into (3.6) gives

$$\Pr(X > x) \approx \hat{\lambda}_u \left[1 + \hat{\xi} \left(\frac{x - u}{\hat{\sigma}} \right) \right]_+^{-1/\hat{\xi}}. \quad (3.7)$$

Thus, an estimate of the level z_t that is exceeded on average once every t observations is obtained as the solution of

$$\hat{\lambda}_u \left[1 + \hat{\xi} \left(\frac{\hat{z}_t - u}{\hat{\sigma}} \right) \right]_+^{-1/\hat{\xi}} = \frac{1}{t},$$

giving

$$\hat{z}_t = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[(t\hat{\lambda}_u)^{\hat{\xi}} - 1 \right] \quad (3.8)$$

for $\xi \neq 0$, and

$$\hat{z}_t = u + \hat{\sigma} \log(t\hat{\lambda}_u)$$

when $\xi = 0$ (see problems sheet 2). By construction, z_t is the t -observation return level; however, it is often more convenient to give return levels on an annual scale, so that the r -year return level is the level expected to be exceeded once every r years. If there

are n_y observations per year, this corresponds to the t -observation return level with $t = r \times n_y$. Hence, an estimate of the r -year return level z_r is defined by

$$\hat{z}_r = u + \frac{\hat{\sigma}}{\hat{\xi}} \left[(rn_y \hat{\lambda}_u)^{\hat{\xi}} - 1 \right], \quad (3.9)$$

unless $\xi = 0$, in which case

$$\hat{z}_r = u + \hat{\sigma} \log(rn_y \hat{\lambda}_u).$$

Thus, for the New York City rainfall extremes, we have

$$\hat{z}_{50} = 30 + \frac{7.44}{0.184} \left[(50 \times 365.25 \times 0.00867)^{0.184} - 1 \right] = 92.24 \text{mm}$$

as an estimate of the 50-year return level, where $n_y = 365.25$ to account for leap years (we have daily observations).

The delta method can, once again, be used to obtain estimated standard errors for such return levels. We should, however, also include uncertainty in our estimate of λ_u in the calculation (since z_r is a function of λ_u). Since the number of threshold exceedances follows a binomial distribution $\text{Bin}(N, \lambda_u)$, where N is the total number of observations in the series, we know (from MAS2302) that

$$\text{Var}(\hat{\lambda}_u) = \hat{\lambda}_u(1 - \hat{\lambda}_u)/N = 0.0007^2$$

in our rainfall example. Then, by the delta method,

$$\text{Var}(\hat{z}_r) \approx \nabla z_r^T V \nabla z_r.$$

Here, V is now the variance-covariance matrix of the triple $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi})^T$; in our rainfall example, this is

$$V = \begin{pmatrix} 0.0007^2 & & \\ 0 & 0.958^2 & \\ 0 & -0.0655 & 0.101^2 \end{pmatrix},$$

assuming that $\text{Var}(\hat{\lambda}_u, \hat{\sigma}) = \text{Var}(\hat{\lambda}_u, \hat{\xi}) = 0$. The value -0.0656 is the estimated covariance between $\hat{\sigma}$ and $\hat{\xi}$, and is found in `A$cov` (recall that we stored the fit of the GPD to our rainfall extremes in `A`):

```
> A$cov
      [,1]      [,2]
[1,] 0.91925394 -0.06550662
[2,] -0.06550662 0.01023566
```

We can now proceed as in Section 2.2.4 to obtain standard errors.

For example, obtain the standard error for the 50-year return level estimate of rainfall at New York.



3.2.6 Profile likelihood

Recall that confidence intervals for return levels constructed using standard errors are usually not appropriate; rather, we should construct intervals using *profile likelihood*. We have already considered the idea of using the profile likelihood to obtain more realistic confidence intervals for return levels in Section 2.4; we now implement the `ismev` command `gpd.prof` to construct the profile log-likelihood for our estimate of the 50 year return level at New York.

Recall that we previously stored our results of fitting the GPD to the set of rainfall extremes observed at New York in `A`, that is:

```
> A = gpd.fit(rain, 30)
$threshold
[1] 30

$nexc
[1] 152

$conv
[1] 0

$nllh
[1] 485.0937

$mle
[1] 7.4422639 0.1843027

$rate
[1] 0.008669861

$se
[1] 0.9587773 0.1011714
```

Now typing:

```
> gpd.prof(A, 50, xlow=70, xup=150, npy=365.25)
If routine fails, try changing plotting interval
```

produces the plot shown in Figure 3.2. Notice that you need to provide the `gpd.prof` function with a range of values at which to fix the return level, before maximising the log-likelihood function with respect to the remaining parameters. Here, we have chosen

a range from 70→150, although this sometimes requires experimentation. Notice also that you need to tell the function the (average) number of observations per year (365.25 here, to account for leap years).

Just to check, you can also superimpose your estimated return level on the plot, just to make sure all is well. Recall from Section 3.2.5 that this was

$$\hat{z}_{50} = 92.24 \text{ mm},$$

and so the code

```
> abline(v=92.24)
```

inserts the vertical line that can be seen on Figure 3.2. As you can see, this is at the mode of the profile log-likelihood – as it should be! The default confidence interval is the 95% confidence interval, and so the `gpd.prof` function automatically places a horizontal line at a distance of $\frac{1}{2}\chi_1^2(0.05) = 1.921$ from the maximised value of the profile log-likelihood (see Section 2.4). This gives us a 95% confidence interval, based on profiling the log-likelihood, of about (74.1mm, 143mm). Planners and civil engineers often design to the upper end-point of such confidence intervals, just to make sure they don’t “under-protect”. When feeding back such information to non-statisticians, we often say something like “once every fifty years, we might expect daily rainfall accumulations in New York City to reach up to about 143mm”.

Compare the 95% confidence interval for the 50-year return level obtained from profiling the log-likelihood to that you would obtain by using the standard error. Comment.



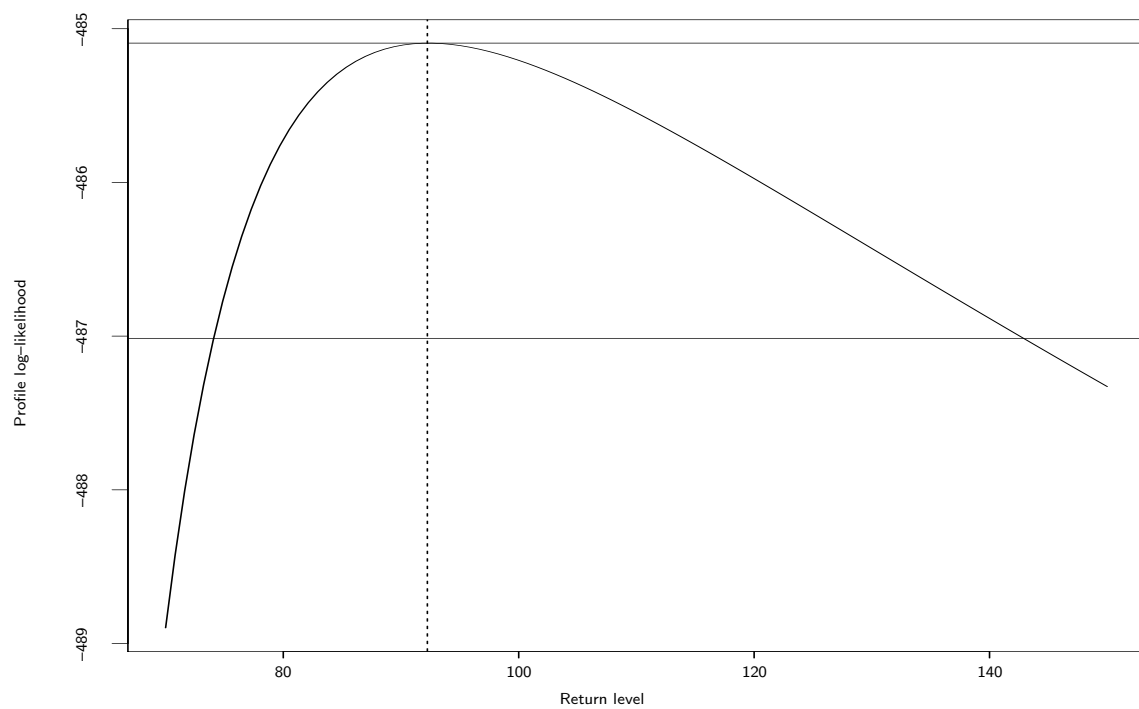


Figure 3.3: Profile log-likelihood curves for the 50 year return level daily rainfall accumulation at New York.