Chapter 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...

...but brings its own problems (see later).

We must first go back and consider the asymptotic theory appropriate for this new situation.

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

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

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

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

defined on $y > 0$, where

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

Outline proof (1/5)

Denote an arbitrary term in the X_i sequence by X. Then it follows that a description of the behaviour of extreme events is given by

$$\begin{aligned}
\Pr\{X > u + y | X > u\} &= \frac{\Pr(X > u + y, X > u)}{\Pr(X > u)} \\
&= \frac{\Pr(X > u + y)}{\Pr(X > u)} \\
&= \frac{1 - F(u + y)}{1 - F(u)}. \qquad (\dagger)
\end{aligned}$$

Outline proof (2/5)

From Section 2.1, we know that

$$\Pr\{M_n \leq x\} = F^n(x) \approx G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\},$$

for some parameters μ, σ and ξ. Hence

$$n\log F(x) \approx -\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} \qquad (\ddagger)$$

Outline proof (3/5)

For large values of x, a Taylor series expansion implies that

$$\log F(x) \approx -\{1 - F(x)\}.$$

Substitution into ($\ddagger$) gives

$$\begin{aligned} n(-\{1 - F(u)\}) &\approx -\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi} \\ 1 - F(u) &\approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi} \end{aligned}$$

for large u. Similarly, for $y > 0$,

$$1 - F(u+y) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u+y-\mu}{\sigma}\right)\right]_+^{-1/\xi}.$$

Outline proof (4/5)

Substitution into (†) then gives

$$\begin{aligned}
\Pr\{X > u+y | X > u\} &= \frac{1-F(u+y)}{1-F(u)} \\
&\approx \frac{\frac{1}{n}\left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]_+^{-1/\xi}}{\frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]_+^{-1/\xi}} \\
&= \left[\frac{1+\xi\left(\frac{u-\mu}{\sigma}\right)+\xi\frac{y}{\sigma}}{1+\xi\left(\frac{u-\mu}{\sigma}\right)}\right]_+^{-1/\xi} \\
&= \left[\frac{1+\xi\left(\frac{u-\mu}{\sigma}\right)}{1+\xi\left(\frac{u-\mu}{\sigma}\right)}+\frac{\xi\frac{y}{\sigma}}{1+\xi\left(\frac{u-\mu}{\sigma}\right)}\right]_+^{-1/\xi} \\
&= \left[1+\frac{\xi y}{\sigma+\xi(u-\mu)}\right]_+^{-1/\xi}
\end{aligned}$$

Outline proof (5/5)

This gives

$$\Pr\{X > u + y | X > u\} = \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]_+^{-1/\xi},$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$, as required (compare with Equation 3.1 on page 45).

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.

3.1 Background and theoretical motivation

- 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

3.1 Background and theoretical motivation

- 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 $\xi \to 0$, giving

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

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 $X_1, X_2, \ldots, X_n$ is a sequence of independent $exp(1)$ random variables.

Show that the limiting distribution of threshold excesses belongs to the generalised Pareto family.

Example 3.1: Solution

If $X_i \sim exp(1)$, then $F(x) = 1 - e^{-x}$ for $x > 0$.

By direct calculation,

$$\begin{aligned} \frac{1 - F(u+y)}{1 - F(u)} &= \frac{1 - \left(1 - e^{-(u+y)}\right)}{1 - (1 - e^{-u})} \\ &= \frac{e^{-(u+y)}}{e^{-u}} = e^{-y}, \qquad y > 0. \end{aligned}$$

Thus, the limit distribution is an exponential distribution; i.e. a GPD with $\xi = 0$. Further, we know that when $\xi = 0$, we have

$$1 - H(y) = \exp\left(-\frac{y}{\sigma}\right);$$

hence, we also have $\sigma = 1$. That is, the limit distribution of threshold exceedances is $GPD(1, 0)$.

Example 3.2

Suppose $X_1, X_2, \ldots, X_n$ is a sequence of independent $U(0,1)$ random variables.

Show that the limiting distribution of threshold excesses belongs to the generalised Pareto family.

Example 3.2: Solution

If $X_i \sim U(0,1)$, then $F(x) = x$ for $0 \leq x \leq 1$.

Hence

$$\frac{1 - F(u+y)}{1 - F(u)} = \frac{1 - u + y}{1 - u} = 1 - \frac{y}{1 - u}.$$

For the GPD with $\xi \neq 0$, we have

$$1 - H(y) = \left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi}.$$

Thus, we have $\xi = -1$ and $\sigma = 1 - u$, i.e. the limiting distribution for threshold exceedances is $GPD(1–u,–1)$.

3.2 Application: rainfall extremes in NYC

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.

3.2 Application: rainfall extremes in NYC

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

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

Thus, for all $u > u_0$, $E[X - u|X > u]$, is a linear function of u.

3.2.1 Threshold choice

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.

3.2.1 Threshold choice

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))
```

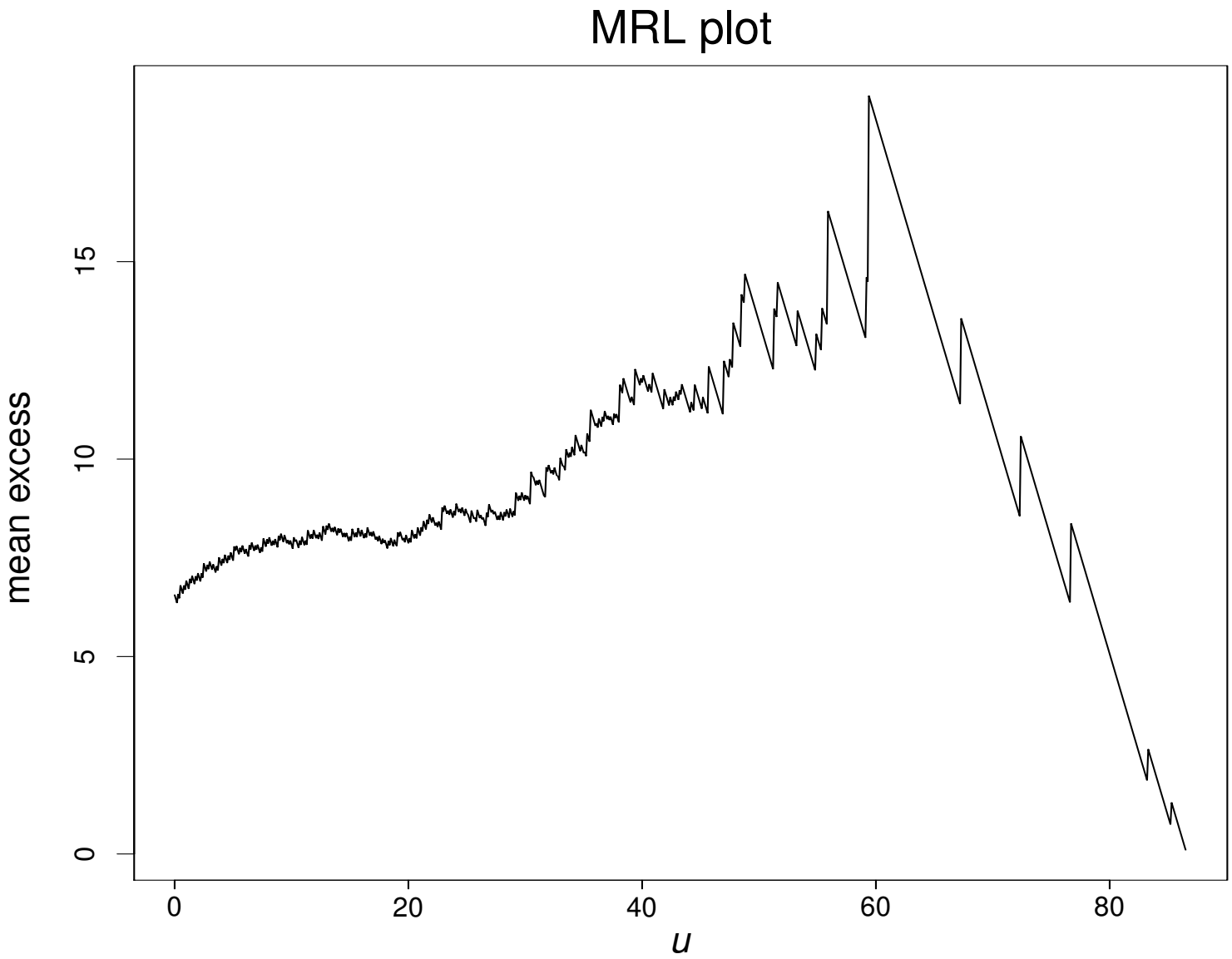
Then the following code computes the mean excess for each value in `u` and stores it in `x`:

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

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

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

3.2.1 Threshold choice



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 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
```

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, but can be shown to be:

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

where $\boldsymbol{y} = (y_1, \ldots, y_{152})^T$ are the set of exceedances above threshold $u_0 = 30$.

For the case $\xi = 0$, interpreted as $\xi \to 0$, we have the log–likelihood f or an exponential distribution with rate $1/\sigma$ (again, see problems sheet 2).

3.2.3 Fitting the GPD

R demo, including assessment of model adequacy

3.2.5 Return level estimation

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

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

for $\xi \neq 0$. Working with the LHS of (3.5), we see that

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

giving

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

3.2.5 Return level estimation

After substitution of (3.5), we get

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

where $\lambda_u = \Pr(X > u)$ and is estimated as the empirical threshold exceedance rate $\hat{\lambda}_u$.

3.2.5 Return level estimation

Now each y_i, $i = 1, \ldots, 152$, 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

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

3.2.5 Return level estimation

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

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

for $\xi \neq 0$, and

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

when $\xi = 0$ (see problems sheet 2).

3.2.5 Return level estimation

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**.

3.2.5 Return level estimation

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

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

unless $\xi = 0$, in which case

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

3.2.5 Return level estimation

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).

3.2.5 Return level estimation

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).

Now the number of threshold exceedances follows a binomial distribution Bin(N, λ_u), where N is the total number of observations in the series. We know (from MAS2302) that

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

in our rainfall example.

3.2.5 Return level estimation

Recall that, by the delta method,

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

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 $Var(\hat{\lambda}_u, \hat{\sigma}) = Var(\hat{\lambda}_u, \hat{\xi}) = 0$.

3.2.5 Return level estimation (1/3)

As in Chapter 2,

$$\nabla z_r^T = \left[\frac{\partial z_r}{\partial \lambda_u}, \frac{\partial z_r}{\partial \sigma}, \frac{\partial z_r}{\partial \xi}\right];$$

this can be shown to give

$$\begin{aligned}\nabla z_r^T = \Big[&\sigma(rn_y)^\xi \lambda_u^{\xi-1}, \xi^{-1}\left\{(rn_y\lambda_u)^\xi - 1\right\}, \\ &-\sigma\xi^{-2}\left\{(rn_y\lambda_u) - 1\right\} + \sigma\xi^{-1}(rn_y\lambda_u)^\xi \log(rn_y\lambda_u)\Big],\end{aligned}$$

which we evaluate at $(\hat{\lambda}_u, \hat{\sigma}, \hat{\xi})$.

3.2.5 Return level estimation (2/3)

This gives:

$$\nabla z_{50}^T = (2179.096, 8.366, 181.757),$$

and so

$$\begin{aligned} Var(\hat{z}_{50}) &= (2179.096, 8.366, 181.757) \\ &\quad \times \begin{pmatrix} 0.0007^2 & & \\ 0 & 0.958^2 & \\ 0 & -0.0655 & 0.101^2 \end{pmatrix} \\ &\quad \times \begin{pmatrix} 2179.096 \\ 8.366 \\ 181.757 \end{pmatrix} \\ &= 397.1853 \end{aligned}$$

3.2.5 Return level estimation (3/3)

And so

$$s.e.(\hat{z}_{50}) = \sqrt{397.1853} = 19.930.$$

3.2.6 Profile likelihood

As discussed in Chapter 2, confidence intervals for return levels are better constructed via the method of profile likelihood, owing to the asymmetry in the likelihood surface often observed.

A plot of the profile log–likelihood for the 50–year return level for daily rainfall accumulations at New York is shown in Figure 3.3 of the lecture notes.

This gives

$$(74.1 \text{ mm}, 143 \text{ mm}).$$

Interpretation: "Once every fifty years, we might expect daily rainfall accumulations in New York City to reach up to about 143mm".

3.2.6 Profile likelihood

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

3.2.6 Profile likelihood

From the profiled log–likelihood:

$$(74.1 \text{ mm}, 143 \text{ mm}).$$

Using the standard error, we get:

$$92.24 \pm 1.96 \times 19.93 \longrightarrow (53.2 \text{ mm}, 131.3 \text{ mm}).$$

Comment: the confidence interval based on the profiled log–likelihood is much more conservative, giving a substantially higher upper bound for the return level.