Example 2.4

The magnitudes of the vertical forces produced by seismic *degassing bursts* are known to be associated with destructive volcanoes.

Table 5 shows the maximum force produced by degassing bursts, every quarter in the years 2015–2021 (inclusive), for the Kilauea Volcano in Hawaii.

	Jan–Mar	Apr–Jun	Jul–Sep	Oct–Dec
2015	99983	100067	99905	100367
2016	99980	99970	100086	99988
2017	99912	100084	100432	99921
2018	100123	99913	100240	100448
2019	100162	100296	100015	100122
2020	99930	99997	100710	100136
2021	99946	99918	99962	100052

Example 2.4

These data were stored the in the vector `haw` in R, and the following output obtained:

```
gev.fit(haw)
```

```
$ conv
[1] 0

$nllh
[1] 178.2502

$mle
[1] 9.998028e+04 8.710824e+01 5.921945e-01

$se
[1] 21.6185212 21.3598264 0.3027163
```

Example 2.4

Use this information to estimate the degassing bursts we would expect to see (i) once every year, (ii) once every fifty years, and (iii) once every 100 years at this volcano (do not attempt to obtain standard errors for these estimates).

How does your estimate in (i) compare to its empirical counterpart?

Example 2.4: Solution (1/5)

From Equation (10) we know that

$$\hat{z}_r = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(-\log\left(1 - r^{-1}\right) \right)^{-\hat{\xi}} - 1 \right].$$

The quantity $\hat{z}_r$ is an estimate of the level that is exceeded, on average, once every r **observations** – if we have only one observation per year, this is exactly the the same as the r–<u>year</u> return level.

In this example, we have **four** observations per year!

Example 2.4: Solution (2/5)

Thus, an estimate of the one–year return level is given by

$$\begin{aligned}\hat{z}_4 &= 99980.28 + \frac{87.10824}{0.5921945}\left[\left(-\log(1-4^{-1})\right)^{-0.5921945} - 1\right] \\ &= 100140.8 \text{ kg.}\end{aligned}$$

Similarly, an estimate of the 50–year return level is given by

$$\begin{aligned}\hat{z}_{200} &= 99980.28 + \frac{87.10824}{0.5921945}\left[\left(-\log(1-200^{-1})\right)^{-0.5921945} - 1\right] \\ &= 103218.6 \text{ kg.}\end{aligned}$$

Example 2.4: Solution (3/5)

Also, an estimate of the 100–year return level is given by

$$\begin{aligned}\hat{z}_{400} &= 99980.28 + \frac{87.10824}{0.5921945}\left[\left(-\log(1-400^{-1})\right)^{-0.5921945} - 1\right] \\ &= 104940.6 \text{ kg.}\end{aligned}$$

Example 2.4: Solution (4/5)

We can obtain the empirical estimate of the one–year return level by looking at the **empirical distribution function** (see Section 2.2.3 on quantile–quantile plots).

For example, ordering the data gives:

$x_{(i)}$	$i/(n+1)$
$\vdots$	$\vdots$
99946	0.241
99962	0.276
$\vdots$	$\vdots$

An empirical estimate of the one–year return level is the value which is exceeded once every 4 observations – the value for which $i/(n+1) = 0.25$.

Example 2.4: Solution (5/5)

$i/(n+1) = 0.25$ is 25.71% of the way between 0.241 and 0.276; 25.71% of the way between 99946 and 99962 gives

$$\hat{z}_{4,\text{emp}} = 99950.11 \text{ kg}.$$

The model gives a slight over–estimate.

Example 2.5

The table below shows the annual maximum number of sunshine hours (per day) for Santiago de Compostela, Spain, for the years 1993–2017.

12.8	10.7	11.9	13.1	13.3	10.7	11.1	11.9	12.1
11.3	12.5	12.1	12.0	12.4	11.8	10.5	10.9	10.1
11.7	12.7	10.6	11.6	9.9	9.6	12.0		

(a) Use the ismev package in R to fit the GEV distribution to these data. Report estimates of the location, scale and shape parameters, with associated standard errors.

(b) Explain why the output in part (a) shows that the Weibull (Type *III* extreme value distribution) is most appropriate here, and give a practical justification for this.

Example 2.5: Solution

See R demo

The GEV: Words of warning

- Block size?
  - Convention is to work with blocks equal in length to the calendar year, but what if you have hourly data collected over just a few years?
  - Block length too small – limiting arguments will not hold (the GEV is a limiting result, which holds *approximately* for large n)
  - Block length is too large – not enough maxima to work with!
  - Possible sensitivity of GEV parameter estimates to block length

The GEV: Words of warning

- Extremely wasteful of data:
  - Discard all but the block maxima
  - Often results in throwing away tens of thousands of observations – some of which might be 'extreme' – just not as extreme as the block maxima!
  - Results in large standard errors and extremely wide confidence intervals for return levels

The GEV: Words of warning

- Standard asymptotic likelihood results not always applicable for the GEV:
  - When $\xi < 0$, the end–points of the GEV are functions of the parameter values: e.g. $\mu = \sigma/\xi$ is the upper end–point
  - This violates the usual regularity conditions of MLEs
  - Effect: When $-1 < \xi \leq -0.5$, MLEs for ξ can be obtained, but are **super–efficient** (i.e. have variance smaller than the Cramér–Rao lower bound)
  - Another consequence: Cannot obtain MLE for ξ when $\xi \leq -1$

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.

3.1 Background and theoretical motivation

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.

3.1 Background and theoretical motivation

Theorem *(Distribution of threshold excess)*

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

$$H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)_{+}^{-1/\xi}, \tag{13}$$

defined on $y > 0$, where

$$\tilde{\sigma} = \sigma + \xi(u - \mu). \tag{14}$$

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 (‡) gives

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

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.

The family of distributions defined by Eq. (13) 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 (13) 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 $1/\tilde{\sigma}$.

3.1 Background and theoretical motivation

- 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).