MAS8306

Topics in Statistics: Environmental Extremes

Dr. Lee Fawcett

Semester 2 2017/18

1

Background and motivation

1.1 Introduction

Finally, there is almost[1] a global consensus amongst scientists that our planet's climate is changing. Evidence for climatic change has been collected from a variety of sources, some of which can be used to reconstruct the earth's changing climates over tens of thousands of years. Reasonably complete global records of the earth's surface temperature since the early 1800's indicate a positive trend in the average annual temperature, and maximum annual temperature, most noticeable at the earth's poles.

Glaciers are considered amongst the most sensitive indicators of climate change. As the earth warms, glaciers retreat and ice sheets melt, which – over the last 30 years or so – has resulted in a gradual increase in sea and ocean levels. Apart from the consequences on ocean ecosystems, rising sea levels pose a direct threat to low–lying inhabited areas of land. Less direct, but certainly noticeable in the last fiteen years or so, is the effect of rising sea levels on the earth's weather systems. A larger amount of warmer water in the Atlantic Ocean, for example, has certainly resulted in stronger, and more frequent,

[1] Almost...

tropical storms and hurricanes; unless you've been living under a rock over the last few years, you would have noticed this in the media (e.g. Hurricane Katrina in 2005, Superstorm Sandy in 2012). Most recently, and as reported in the New York Times in January 2018, the 2017 hurricane season was "... hyperactive, deadly and extremely destructive, featuring seventeen names storms and ranking alongside 1936 as the fifth most active season since records began". Most notable, perhaps, was Hurricane Maria, which devastated the islands of Puerto Rico and Dominica. The same can be said over the Pacific Ocean, where low–lying areas of Bangladesh, India and Malaysia are more frequently than ever battered by tropical cyclones. Closer to home, we have seen an increase in flooding events in the U.K. over the last fifteen years, with 2012 being the second wettest year on record and December 2015 being the wettest month since records began (largely due to Storm Desmond).

The world's changing climate is also likely to give rise to an increase in the frequency of periods of intense heat and drought. Though some parts of the world are familiar with the effects of extreme heat, we are beginning to see an increase in severe hot spells in more temperate latitides – for example, the 2003 European heatwave which resulted in thousands of deaths. In drier parts of the world, this has resulted in an increase in devastating forest fires, and in developing counties an increase in famine.

1.2 Examples

1.2.1 Hurricane Katrina: Late summer 2005

Hurricane Katrina made landfall as a category 3 hurricane (on the Saffir–Simpson hurricane scale, which goes from 1 (weakest) to 5 (strongest)) just south–west of New Orleans, on 31st August 2005. The tropical storm grew in intensity over the Atlantic ocean from mid–August, almost reaching hurricane category 5 status over the warm waters of the Gulf of Mexico at one point. Apart from wreaking havoc in Cuba and other Caribbean islands, the storm surge – the heightened levels of the sea as a result of extremely strong winds caused by record air pressure lows – over–topped the New Orleans flood defence systems ("levees" – see Figure 1.1) causing widespread flooding throughout the city and the surrounding area. There was a mandatory evacuation of the city, although many of the city's poorest people were left without transport and were trapped. It is estimated that the storm killed over 1,800 people (some are still unaccounted for), but displaced over a quarter of a million. Though still counting the cost, it is estimated that the storm caused over 100 billion dollars worth of damage. The highest one minute sustained wind speed recorded was 175mph, and the maximum recorded sea–surge at New Orleans reached nearly 14 feet (not verified).

Figure 1.1: Clockwise, from top–left: Satellite image of Hurricane Katrina as it approaches New Orleans; sea–surge breaching the London Street levee in the Lower Ninth Ward; flooded street after the hurricane had passed; my own photograph taken in the St. Bernard district five years later.

1.2.2 Superstorm Sandy: October 2012

Hurricane Sandy, billed in the media as a "superstorm" or "perfect storm" because of its level of organisation, devasated parts of the Caribbean and the mid–Atlantic and northeastern states of the USA. Sandy was the tenth hurricane of a busy hurricane season in 2012 (although not quite as busy as the 2017 hurricane season), and was a category 2 at its peak intensity. Causing an estimated 65.6 billion dollars worth of damage, this is the second costliest hurricane in US history, after Katrina, and 253 people were killed. The highest one minute sustained wind speed recorded was 110mph. Sandy's strength and angle of approach combined to produce a record storm surge of water into New York City, with a maximum sea–surge of around 10 feet recorded just off Long Island. The path of Sandy was extremely unusual: after causing havoc in the Caribbean, Florida and the southeast United States, rather than racing out to sea like most coastal storms Sandy turned hard to the left and made landfall much further north than most hurricanes usually do. See Figure 1.2.

Figure 1.2: Clockwise, from top–left: Satellite image of Hurricane Sandy; downtown Havana, Cuba, three days after the storm; a flooded subway station one week after the storm; the next Sandy?

1.2.3 Extreme weather in the U.K.

The U.K. is no stranger to changing climate. The year 2012 was the second wettest year in the U.K. since records began, and six years since 1998 are in the top ten wettest years (1998, 1999, 2000, 2002, 2008 and 2012). December 2015 was the wettest single month on record. Flash–flooding in Glasgow in 2002 resulted in the evacuation of over 200 people from their homes, landslides, and the contamination of drinking water. Flooding in Boscastle (southwest England) in 2004 resulted in the destruction of over 100 homes and business properties; flooding across the entire UK in 2007, 2008, 2009 and 2012 – including severe flash–flooding in notoriously dry (!) Newcastle in June 2012 – resulted in a number of deaths, as well as millions of pounds worth of damage to prpoerties and infrastructure. Observational studies and climate models indicate that the occurrence, and magnitide, of such extreme rainfall events will continue to increase in the future.

On a much more destructive scale was the "Great North Sea Flood" of 1953. On the night of 31st January a major European windstorm combined with a high spring tide

which resulted in a storm surge along the coasts of Holland, Belgium and Eastern Scotland/England of up to 18 feet. In total, 2,551 people were killed. The Great Storm of 1987, in which record–breaking hurricane–strength wind speeds battered parts of southeast England and northern France, killed 22 people – on a night in which the BBC weather forecaster, Michael Fish, infamously said: "Don't worry, there won't be a hurricane"! The "Burns Day Storm" on 25th January 1990 proved more deadly than the Great Storm of 1987, mainly because it hit during the day. It caused chaos, killing some children in their schools when the roof collapsed and throwing cars off the road. 47 people died and around half a million homes were left without electricity. However, unlike 1987, the UK Met Office had predicted the storm's progress and had issued several warnings. Over the last decade or so the U.K. has seen a gradual shift in its wind climate, with gales, tropical storm–strength wind speeds and even tornadoes (e.g. Birmingham in 2005). See Figure 1.3.

Figure 1.3: Clockwise, from top–left: Effects of flooding in Cockermouth, Cumbria, 2008; Lightning strike of the Tyne Bridge during a flash–flood in July 2012; flooding in Zeeland, Holland, during the 1953 North Sea storm surge; the next big north sea storm surge?

1.2.4 Heatwaves and drought

The world's changing climate is also likely to give rise to an increase in the frequency of periods of intense heat and drought. Though some parts of the world are familiar with the effects of extreme heat, we are beginning to see an increase in severe hot spells in more temperate latitides – for example, the 2003 European heatwave which resulted in tens of thousands of deaths (including around 2000 deaths in the U.K., mostly through heat stroke or dehydration). In drier parts of the world, this has resulted in an increase in devastating forest fires, and in developing counties an increase in famine.

> Now – perhaps more than ever – the statistical analysis of the extremes of environmental data is in the spotlight. We need to get things right!

1.3 Why statistical modelling?

Statistical modelling of extreme weather has a very practical motivation: reliability — anything we build needs to have a good chance of surviving the weather/environment for the whole of its working life. This has obvious implications for civil engineers and planners. For example, in the context of some of the examples we have looked at so far, they need to know:

- How tall to build sea walls/flood defences – for example –
  - the levees surrounding New Orleans (Figure 1.1, top right);
  - flood defences in New York that were breached during Superstorm Sandy;
  - river and sea defences in the U.K.;
- How strong to make buildings – for example, in the U.K. the *British Standards Institute* specifies that new structures need to be strong enough to withstand the "Fifty year wind speed" ;
- How tall to build reservoir dams;
- How much fuel to stockpile;

This motivates the need to estimate what the:

- Highest tide;
- Heaviest rainfall;
- Strongest wind;
- most severe cold spell;

will be over some fixed period of future time. The only sensible way to do this is to use data on the variable of interest (wind, rain etc.) and fit an appropriate statistical model. The models themselves are motivated by asymptotic theory, and this will be our starting point in Chapter 2. Before this, let's start at a very basic level, and think about why we need statistical models at all.

1.4 Simple example: sea surges at Wassaw Island

Wassaw Island is a barrier island off the southeast coast of Georgia, in the USA. Any hurricanes about to make landfall on the mainland often hit such barrier islands first, substantially subduing the impending storm. Sea surge data is collected every hour; the data shown below are the extracted *annual maxima* for the years 1955–2004.

8.5	8.9	9.1	8.9	8.4	9.7	9.1	9.6	8.7	9.3
9.6	9.3	8.7	9.0	8.8	8.9	8.9	12.2	7.8	7.7
8.3	8.1	7.3	6.8	6.7	7.3	7.6	8.2	8.6	9.8
9.5	7.4	7.3	10.2	10.3	10.4	8.8	9.7	10.0	10.8
11.1	12.7	11.5	11.8	12.6	13.0	10.5	10.5	10.0	9.4

Table 1.1: Annual maximum sea–surges at Wassaw Island, 1955–2004 (inclusive). Units are feet.

1. Civil engineers are interested in various *exceedance probabilities*. For example, suppose they need to know the probability that sea–surge will exceed 8.75, 11.25 and 14 feet. Use the data to estimate these probabilities.

✎

2. How high should state authorities build flood defences at the nearby city of Savannah if they want to protect the city against the sea–surge they would expect to see:

(i) Once every 10 years;

(ii) Once every hundred years?

✎

Further comments

✎

2

Classical models for extremes

2.1 Background and theoretical motivation

Suppose that $X_1, X_2, \ldots, X_n$ is a sequence of independent and identically distributed (IID) random variables with common distribution function F. One way of characterising extremes is by considering the distribution of the maximum order statistic

$$M_n = \max\{X_1, X_2, \ldots, X_n\}. \tag{2.1}$$

Think back to the example in Section 1.4 concerning sea surges at Wassaw Island. Sea surge measurements were taken every hour; for each year (1955–2004) you were presented with the *annual maximum* sea surge. Thus, here $n = 365 \times 24 = 8760$ (for non–leap years, anyway), and we might use the notation:

$$M_{8760,i}, \qquad i = 1, \ldots, 50,$$

to denote, generally, the 50 annual maxima given in Table 1.1 (of course, for leap years we would write $M_{8784,i}$). The IID assumption might be reasonable here. For example, each annual maximum is likely to occur during the hurricane season (which is usually at its peak in September/October), and so it seems likely that maximum hourly observations from one year to the next will be far enough apart to be independent;

issues of non–stationarity, however, might arise in long–range datasets owing to the effects of climate change, for example, and we will come back to this in Chapter 4.

Assuming that our maxima are IID, how can we obtain the distribution of M_n?

✎

2.1.1 A brief history of Extreme Value Theory

One of the earliest books on the statistics of extreme values is **E.J. Gumbel** (1958, see Figure 2.1). Gumbel traces the origins back to 1709, when **N. Bernoulli** considers the problem of estimating he age of the longest survivor in a group of people.

Research into extreme values as a subject in it's own right began much later, between 1920 and 1940, when work by **E.L. Dodd**, **M. Fréchet**, **E.J. Gumbel**, **R. von Mises** and **L.H.C. Tippett** investigated the asymptotic distribution of the *largest order statistic*. This led to the main theoretical result: the *Extremal Types Theorem* (see Section 2.1.2), which was developed in stages by Fisher, Tippett and von Mises, and eventually proved in general by **B. Gnedenko** in 1943.

Until 1950, development was largely theoretical. In 1958, Gumbel started applying theory to problems in engineering. In the 1970s, **L. de Haan** and **J. Pickands** generalised the theoretical results, giving a better basis for statistical models.

Since the 1980s, methods for the application of Extreme Value Theory have become much more widespread. Current researchers who have played a significant role in developing applications and methodology include **Richard Smith** (University of North Carolina), **Anthony Davison** (Ecole Polytechnique Federale de Lausanne) **Jonathan Tawn** (Lancaster University) and **Stuart Coles** (University of Padua and SmartOdds).

There are still gaps between the theory and the models, and also between the models and common practice in applications – this is where research at Newcastle fits in, and this is the area of research we will focus on in the latter part of this course.

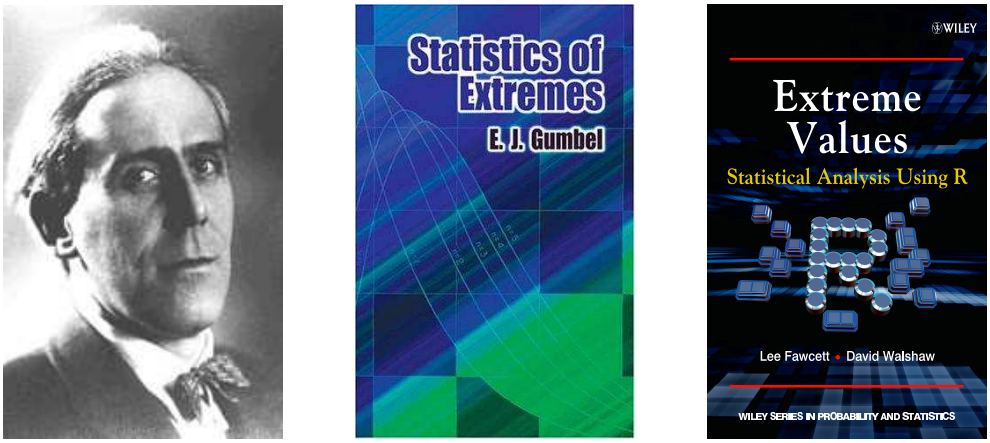

Figure 2.1: Emil Julius Gumbel (left), and his book "Statistics of Extremes" (centre); the latest book to hit the shops in August 2018 (right, pre–order on Amazon now!!).

2.1.2 The Extremal Types Theorem

The obvious question now, is, what possible distributions might be considered candidates for M_n as $n \to \infty$? It might also be convenient to find out if we can formulate this set of candidate distributions into a single class – say G – which is independent of F; if so, we can estimate the distribution of M_n using G, without any reference to F.

Clearly, the limiting distribution of M_n is *degenerate*, that is, the distribution converges to a single point on the real line with probability 1, this single point being the upper endpoint of F (which itself will be ∞ in some applications). This is analogous to the sample mean $\bar{X}$ converging to the population mean μ with certainty in the Central Limit Theorem; here, the degenerate limit is prevented by allowing a linear rescaling, so that

$$\frac{\bar{X} - b_n}{a_n} \xrightarrow{D} N(0,1)$$

where $b_n = \mu$ and $a_n = \sigma/\sqrt{n}$, where σ and n are the population standard deviation and sample size, respectively. Can we apply a similar linear rescaling to M_n to avoid convergence of the distribution to a single point? The answer, of course, is "yes", and is provided by the main result in classical extreme value theory – the *Extremal Types Theorem* – a result for the maximum M_n which is analogous to the Central Limit Theorem for the mean μ.

Theorem (The Extremal Types Theorem)

If there exist sequences of constants $a_n > 0$ and b_n such that, as $n \to \infty$,

$$\Pr\{(M_n - b_n)/a_n \leq x\} \to G(x) \tag{2.2}$$

for some non–degenerate distribution G, then G is of the same type as one of the following distributions:

$$I : G(x) = \exp\{-\exp(-x)\} \qquad -\infty < x < \infty; \tag{2.3}$$

$$II : G(x) = \begin{cases} 0 & x \leq 0 \\ \exp(-x^{-\alpha}) & x > 0, \alpha > 0; \end{cases} \tag{2.4}$$

$$III : G(x) = \begin{cases} \exp\{-(-x)^{\alpha}\} & x < 0, \alpha > 0 \\ 1 & x \geq 0. \end{cases} \tag{2.5}$$

The three types of distribution in the Extremal Types Theorem (I, II and III) have become known as the Gumbel, Fréchet and Weibull types (respectively), and are known collectively as the *extreme value distributions*. For both the Gumbel and Fréchet distributions the limiting distribution G is unbounded; that is, the upper–endpoint tends to ∞. The Weibull distribution has a finite upper bound. It should be noted that the Extremal Types Theorem does not *ensure* the existence of a non–degenerate limit for M_n; nor does it specify *which* of types I, II or III is applicable *if* a limit distribution *does* exist (i.e. in which *domain of attraction* the distribution of G lies). However, when such a distribution does exist, we find that, by analogy with the Central Limit Theorem, the limiting distribution of sample maxima follows one of the distributions given by the Extremal Types Theorem, *no matter what the parent distribution F.*

So we know that

$$\frac{M_n - b_n}{a_n} \xrightarrow{D} G,$$

where – if it exists – G is given by one of the extreme value distributions (either the Gumbel, Fréchet or Weibull distributions). But how do we know which one of these distributions to use?

2.1.3 The Generalised Extreme Value (GEV) distribution

In practice, working with, and having to choose between, three distributions is inconvenient. However, there exists a parameterisation which encompasses all three types of extreme value distribution. Von Mises (1954) and Jenkinson (1955) independently derived the *generalised extreme value distribution* (GEV), often denoted $\mathcal{G}(\mu, \sigma, \xi)$, whose cumulative distribution function is given by

$$G(x; \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{-1/\xi}\right\}, \tag{2.6}$$

where $a_+ = \max(0, a)$. The situation where $\xi = 0$ is not defined in (2.6), but is taken as the limit as $\xi \to 0$, given by

$$G(x; \mu, \sigma) = \exp\left\{-\exp\left(\frac{x-\mu}{\sigma}\right)\right\}. \tag{2.7}$$

The parameters μ ($-\infty < \mu < \infty$), σ (> 0) and ξ ($-\infty < \xi < \infty$) are location, scale and shape parameters respectively. The value of the shape parameter ξ differentiates between the three types of extreme value distribution in Theorem 2.1: $\xi = 0$, leading to (2.7), corresponds to the Gumbel distribution (type *I*); $\xi > 0$ and $\xi < 0$ correspond to the Fréchet (type *II*) and Weibull (type *III*) distributions respectively. In practice, when we estimate the shape parameter ξ, the standard error for ξ accounts for our uncertainty in choosing between the three models given by (2.3)—(2.5).

But what about the constants a_n and b_n? We know that

$$\frac{M_n - b_n}{a_n} \xrightarrow{D} \mathcal{G}(\mu, \sigma, \xi), \qquad \text{as } n \to \infty.$$

After some algebra, it turns out that

$$M_n \xrightarrow{D} \mathcal{G}(\mu^*, \sigma^*, \xi), \qquad \text{as } n \to \infty,$$

with a_n and b_n being absorbed into μ^* and σ^*. Since the GEV parameters need to be estimated anyway, in practice we just ignore the normalisation constants and fit the GEV directly to our set of maxima $M_{n,i}$. However, before we consider applications of the GEV to real data, let us first consider some theoretical examples which demonstrate that, with careful choices of a_n and b_n, one of the three extreme value distributions is always achieved when the parent distribution F is known.

Example 2.1

Suppose $X_1, X_2, \ldots, X_n$ is a sequence of independent $Exp(1)$ variables, that is

$$F(x) = 1 - e^{-x}, \qquad x > 0.$$

By letting $a_n = 1$ and $b_n = \log n$, show that the limit distribution of $(M_n - b_n)/a_n$ is of extreme value type, and identify the distribution.

Solution

✎

✎

Example 2.2

Suppose $X_1, X_2, \ldots, X_n$ is a sequence of independent *Fréchet*(1) variables, that is

$$F(x) = e^{-1/x}, \qquad x > 0.$$

By letting $a_n = n$ and $b_n = 0$, show that the limit distribution of $(M_n - b_n)/a_n$ is of extreme value type, and identify the distribution.

Solution

✎

2.1.4 Typical application

1. Data pre–processing

Unless you already have a filtered set of block/annual maxima, the first step in any application of the GEV is to obtain your $M_{n,i}$. This requires choosing a *block length* n (usually the calendar year), and then discarding all but the maximum value in each block. Be careful – if your block length is too short, the limiting arguments supporting the GEV might not be valid, and your maxima may be too close together to assume independence; if n is too large, you won't have enough data to work with!

2. Estimation of the GEV parameters

A typical application of the generalised extreme value distribution is to fit (2.6) to a series of block (often annual) maxima. Numerous ways of fitting this to a set of maxima have been considered, though numerical maximum likelihood estimation has become generally accepted as the most robust procedure. For example, differentiation of (2.6) gives the probability density function of the GEV; this can be found to be

$$g(x;\mu,\sigma,\xi) \quad = \quad \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-(1/\xi+1)} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}. \quad (2.8)$$

Use the GEV probability density function in (2.8) to form the likelihood function $L(\mu,\sigma,\xi;\boldsymbol{x})$. Also, obtain the GEV log–likelihood function $\ell(\mu,\sigma,\xi;\boldsymbol{x})$ and the corresponding (log) likelihood equations.

✎

How would you use the (log) likelihood equations in order to obtain maximum likelihood estimates of μ, σ and ξ? Why can't we obtain closed form solutions for $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$? How can we get around this?

✎

The usual techniques can be used to obtain standard errors for the estimated GEV parameters, i.e. via inversion of the information matrix: these can be used to construct confidence intervals and perform hypothesis tests.

3. Checking model adequacy

As with all statistical models, there are various goodness–of–fit properties that should be considered to check the overall adequacy of the fitted GEV. These include probability plots, quantile–quantile (Q-Q plots) and simply plotting a histogram of the data against the fitted density. Again, these will be reviewed shortly via a real–life data demonstration in R.

4. Return level estimation

As discussed in Section 1.3, interest usually lies not in estimates of the GEV parameters themselves, but in how we can use the fitted model to estimate other quantities, such as

- The height of a sea wall to protect against the once in a hundred year sea–surge;
- The "fifty year wind speed" to provide new structures enough protection against wind damage.

Such quantities, in extreme value terminology, are usually referred to as *return levels*. Provided we have faith in our fitted model being suitable beyond the range of our observed data, we can estimate the r–year return level z_r for any period by setting the GEV distribution function (Equation 2.6) equal to $1 - 1/r$ and solving for $x = \hat{z}_r$, provided we have annual maxima.

For example, suppose we fit the GEV to the set of annual maxima given in Table 1.1 and obtain estimates of the location, scale and shape as $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ (respectively). Suppose further that the authorities require an estimate of z_{100}, the sea surge we might expect to be exceeded once in a hundred years (this could be used to help estimate the corresponding height of a new sea wall). Then we can write down the following probability statement:

$$\Pr(\text{annual maximum} > z_{100}) \quad = \quad \frac{1}{100},$$

i.e.

$$1 - \Pr(\text{annual maximum} \leq z_{100}) \quad = \quad \frac{1}{100}. \tag{2.9}$$

Now the left–hand–side of (2.9), in terms of our fitted GEV, is

$$1 - G(\hat{z}_{100}; \hat{\mu}, \hat{\sigma}, \hat{\xi}),$$

giving

$$\begin{aligned} 1 - \exp\left\{ -\left[1 + \hat{\xi}\left(\frac{\hat{z}_{100} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}} \right\} &= 0.01 \qquad \text{i.e.} \\ \exp\left\{ -\left[1 + \hat{\xi}\left(\frac{\hat{z}_{100} - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}} \right\} &= 0.99. \end{aligned}$$

Solving for $\hat{z}_{100}$ gives an estimate of the 100–year return level as

$$\hat{z}_{100} = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[(-\log(0.99))^{-\hat{\xi}} - 1\right];$$

more generally, estimates of the r–year return level z_r are given by

$$\hat{z}_r \quad = \quad \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[\left(-\log\left(1 - r^{-1}\right)\right)^{-\hat{\xi}} - 1\right]. \tag{2.10}$$

Care needs to be taken when $\xi = 0$; in this case, as with forming the likelihood in the first place, we should work with the limiting form (as $\xi \to 0$) of the distribution function (i.e. Equation (2.7)).

Of course, a point estimate of the return level alone is not enough; any estimate requires a corresponding standard error. As the return level is a function of the GEV parameters,

which we have standard errors for, we can use the delta method to obtain standard errors for return level estimates. Further, confidence intervals formed in the usual way are often extremely misleading for return level estimates, owing to the (often severe) asymmetry encountered in the surface of the likelihood for these quantities; rather, the method of *profile likelihood* is preferred to construct confidence intervals here (see later).

2.2 Worked example: sea surge at Wassaw Island

Recall the example from Section 1.4. The southeastern coast of the USA is an area often hit by hurricanes. The historic city of Savannah, Georgia, has suffered direct hits from 22 Hurricanes since 1871, and can expect to 'brushed' or directly hit once every other year. Predictions show that the city is three years overdue it's next direct hit (keep an eye on the news this summer). The figures in Table 2.1 below show the annual maximum sea surges observed at Wassaw Island, a barrier island just off the coast of Savannah, between 1955 and 2004 (inclusive).

8.5	8.9	9.1	8.9	8.4	9.7	9.1	9.6	8.7	9.3
9.6	9.3	8.7	9.0	8.8	8.9	8.9	12.2	7.8	7.7
8.3	8.1	7.3	6.8	6.7	7.3	7.6	8.2	8.6	9.8
9.5	7.4	7.3	10.2	10.3	10.4	8.8	9.7	10.0	10.8
11.1	12.7	11.5	11.8	12.6	13.0	10.5	10.5	10.0	9.4

Table 2.1: Annual maximum sea–surges at Wassaw Island, 1955–2004 (inclusive). Units are feet.

Recall that, in Chapter 1, we tried to estimated quantities beyond the range of our observed data. For example, using simple empirical arguments,

$$\Pr(\text{Sea surge exceeds 14 feet}) = 0,$$

simply because, over the time-frame we have data for, we have not observed such an extreme event. Does this really mean this event is *impossible*? Probably not, it's just that we're trying to estimate the probability of an event more extreme than we've ever observed before. We will now proceed with a detailed extreme value analysis of these annual maxima.

2.2.1 Data pre–processing

In order to fit the GEV distribution, we need to pre–process our data so we have a filtered set of extremes – in particular, block, or annual, maxima. In this example

no pre–processing is necessary, as we are given a set of annual maxima. In the first computer practical session, we will consider how to pre–process a dataset to obtain the set of block maxima.

Figure 2.2 shows a time series plot and histogram of the 50 annual maxima. Notice there is no obvious trend in our dataset, although there could be an issue of dependence, with neighbouring years having similar values of maximum sea surge. These plots have been produced using R, and, in your own time, you should make sure you can reproduce them. The data are available to download/copy–and–paste from the MAS8304 webpage, in the file "`wassaw.txt`". To load them into R, save the file to your own workspace and then type:

```
> wassaw = scan('wassaw.txt')
```

The following code sets up a year counter (from 1955 to 2004):

```
> year = seq(1955, 2004, 1)
```

Then, to produce the plots shown in Figure 2.2, you can type:

```
> par(mfrow = c(1, 2))
> plot(wassaw ~ year, type = 'b')
> hist(wassaw)
```

Obviously, these are the basic commands; further arguments can be defined to choose different titles/axis labels etc.

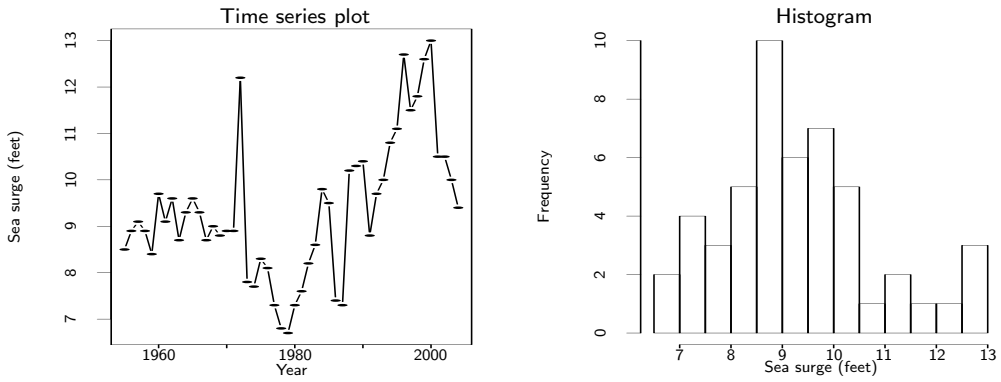

Figure 2.2: Time series plot (left) and histogram (right) of annual maximum sea–surges observed at Wassaw Island.

2.2.2 Fitting the GEV

As discussed in Section 2.1.4, the most commonly–used technique for fitting the GEV is maximum likelihood, which requires maximisation of the GEV log–likelihood function $\ell(\mu, \sigma, \xi; x)$. In particular, for our data we have

$$\begin{aligned} \ell(\mu, \sigma, \xi; \boldsymbol{x}) &= -50\log\sigma - (1 + 1/\xi) \sum_{i=1}^{50} \log \left[1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right]_+ \\ &\quad - \sum_{i=1}^{50} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right]_+^{-1/\xi}, \end{aligned} \tag{2.11}$$

As previously discussed, there is no analytical solution for obtaining the maximum likelihood estimates of the GEV parameters. Instead, we need to maximise the log–likelihood in (2.11) numerically; we do this using R.

Using R to estimate the GEV parameters

You should, at some point, try the following code for yourself. Having already stored our data in `wassaw`, the next step is to define the GEV parameter vector $\theta = (\mu, \sigma, \xi)$, and to provide R with "starting values" for the elements of θ. Since μ and σ are the GEV location and scale parameters, it is common to use the mean and standard deviation (respectively) as starting values for these parameters; the shape parameter ξ can be given any starting value, but a common choice is 0.1. Thus, in R:

```
> theta = c(mean(wassaw), sd(wassaw), 0.1)
```

Now, the fiddly bit is to define the GEV log–likelihood function. Read through the following code, and corresponding notes, carefully – you will be expected to do this sort of thing in Assignment 2!

Notes on the code overleaf

1. This line stores the following component in the scalar `m`:

$$\left[1 + \xi \left(\frac{x_i - \mu}{\sigma} \right) \right]_+ .$$

2. Recall that, if the component `m` is less than zero, we take it to be equal to zero, as $a_+ = \max(0, a)$. Thus, if for a particular combination of $\theta = (\mu, \sigma, \xi)$ we get `m<0.00001`, we return a huge value for the negative log–likelihood (10^6) ensuring

that this combination will almost surely never be the combination that minimises the negative log–likelihood (and so will almost surely never maximise the log–likelihood – see note 6).

3. This line performs a similar task for `sigma` as the code referred to in note 2 does for `m`. σ is the scale parameter and so can never be negative; thus, any attempt to minimise the negative log–likelihood using `sigma<0.00001` returns a huge negative log–likelihood (1e+06), again ensuring that this choice of σ can almost surely never minimise the negative log–likelihood (again, see note 6).

4. This line defines the GEV log–likelihood when $\xi = 0$, taken as the limit when $\xi \to 0$ (see Equation (2.7)).

5. This line defines the GEV log–likelihood when $\xi \neq 0$ (see Equation (2.6)).

6. The R routine `nlm` is a minimisation routine that minimises a supplied function (here `gev.loglik`) using a Newton–Raphson type algorithm. Since the aim is to find estimates of the GEV parameters that *maximise* the log–likelihood, this will be equivalent to finding estimates of the GEV parameters that *minimise* the *negative* log–likelihood, which is why we ask the function `gev.loglik` to return `-loglik`.

```
> gev.loglik = function(theta){
+   mu = theta[1]
+   sigma = theta[2]
+   xi = theta[3]
+   m = min((1+(xi*(dd-mu)/sigma)))                          #See note 1
+   if(m<0.00001)return(1000000)                             #See note 2
+   if(sigma<0.00001)return(1000000)                         #See note 3
+   if(xi==0){
+     loglik = -length(dd)*log(sigma)-sum((dd-mu)/sigma)-sum(exp(-((dd-
mu)/sigma)))}                                               #See note 4
+   else{
+     loglik = -length(dd)*log(sigma)-(1/xi+1)*sum(log(1+(xi*(dd-
mu)/sigma)))-sum((1+(xi*(dd-mu)/sigma))**(-1/xi))}          #See note 5
+   return(-loglik)                                          #See note 6
+ }
```

We implement the `nlm` routine in the following way, giving the output shown underneath:

```
> dd = wassaw
> nlm(gev.loglik, theta)
$minimum
[1] 89.52412

$estimate
[1]  8.7112781  1.3114844 -0.1084455

$gradient
[1] -1.465143e-06 -1.678775e-06  5.668710e-07

$code
[1] 1

$iterations
[1] 11
```

Thus, our estimates of the GEV parameters are

$$\hat{\mu} = 8.711 \qquad \hat{\sigma} = 1.311 \qquad \hat{\xi} = -0.108$$

The *minimised negative* log–likelihood is 89.524, giving a maximised log–likelihood at the above values for the GEV parameters of -89.524. `gradient` gives the gradient at the estimated minimum of the function `gev.loglik`, `iterations` gives the number of iterations performed (in the Newton–Raphson algorithm) before the solution to the minimisation problem was found, and `code` is an integer which indicates why the optimisation process terminated (`1` indicates the relative gradient is close to zero and so the current iterate is very close to the solution).

Using R to estimate standard errors of the GEV parameters

Using `nlm` to minimise the negative log–likelihood function is fine, but how can we obtain the corresponding standard errors for the corresponding maximum likelihood estimates? Invoking the `hessian` argument in the `nlm` routine returns the matrix of second–order partial derivatives (the *Hessian matrix*) of the function we are trying to minimise. In R, we type:

```
> A = nlm(gev.loglik, theta, hessian = TRUE)
```

This will give the same output as before, but now the Hessian matrix will also be returned. Note we have also stored the output in `A`. To look at the output, we type:

```
> A
$minimum
[1] 89.52412

$estimate
[1]  8.7112781  1.3114844 -0.1084455

$gradient
[1] -1.465143e-06 -1.678775e-06  5.668710e-07

$hessian
          [,1]      [,2]      [,3]
[1,] 27.375588 -4.837626  18.06141
[2,] -4.837626 56.151711  29.88968
[3,] 18.061406 29.889680 118.18045

$code
[1] 1

$iterations
[1] 11
```

Recalling theory from MAS2302, our observed information matrix will be

$$I_O = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial \mu^2} & & \\ -\frac{\partial^2 \ell}{\partial \sigma \partial \mu} & -\frac{\partial^2 \ell}{\partial \sigma^2} & \\ -\frac{\partial^2 \ell}{\partial \xi \partial \mu} & -\frac{\partial^2 \ell}{\partial \xi \partial \sigma} & -\frac{\partial^2 \ell}{\partial \xi^2} \end{pmatrix},$$

evaluated at $\mu = \hat{\mu}$, $\sigma = \hat{\sigma}$ and $\xi = \hat{\xi}$. Notice that this is just the negative of the hessian of the log–likelihood function; however, the hessian returned by R in the code above is for the *negative* log–likelihood (recall that `gev.loglik` returns `-loglik`), and so – in this example – the hessian is *exactly* the observed information matrix.

Recall also from MAS2302/MAS3320 that the variance–covariance matrix is obtained via inversion of the observed information matrix I_O. One way of doing this is to use the R command `solve`. This command solves, for X, a system of equations of the form

$$\mathbf{a}X = \mathbf{b},$$

where both **a** and **b** can be numeric or complex matrices, **a** corresponding to the coefficients of the linear system and **b** giving the right–hand–side of this system. In our example, **a** is the observed information matrix. If **b** is missing in the execution of `solve`, then the identity matrix is assumed and R will thus return the inverse of **a** which is exactly what we want here. In R, this gives:

```
> solve(A$hessian)
             [,1]         [,2]         [,3]
[1,]  0.043869852  0.008491577 -0.008852240
[2,]  0.008491577  0.022223112 -0.006918331
[3,] -0.008852240 -0.006918331  0.011564270
```

Thus, the square roots of the diagonal elements of the above matrix will give the expected standard errors for the GEV parameters (μ, σ and ξ respectively). We could obtain these standard errors in the following way:

```
> varcovar = solve(A$hessian)
> sqrt(diag(varcovar))
[1] 0.2094513 0.1490742 0.1075373
```

The first line of code stores the matrix shown above in the matrix `varcovar`; the second line of code finds the square roots of the diagonal elements of `varcovar`. The resulting expected standard errors are shown underneath. Thus, we now have the following inference for our annual maximum sea surges, in terms of the GEV distribution:

$$\hat{\mu} = 8.711 \ (0.209) \qquad \hat{\sigma} = 1.311 \ (0.149) \qquad \hat{\xi} = -0.108 \ (0.108)$$

From this, we can construct confidence intervals in the usual way. For example, standard symmetric 95% confidence intervals are found to be:

$$(8.301, 9.121) \qquad (1.019, 1.603) \qquad \text{and} \qquad (-0.320, 0.104)$$

for μ, σ and ξ (respectively).

2.2.3 Model adequacy

We usually fit a statistical model to data to draw conclusions about some aspect of the population from which the data were obtained. The more accurate the fitted model is, the more reliable these conclusions are likely to be. Since inferences are sensitive to the accuracy of the fitted model, it is important that we check that the model fits well. Ideally, we would like to check that our model describes well the variations in the wider population, but this is usually not possible unless there are other sources of data for

us to judge the model against. Thus, it is common to assess the goodness–of–fit of a particular model using the data that were used to estimate it in the first place.

A *probability plot* is a plot of the points

$$\left\{ \left(\hat{F}\left(x_{(i)}\right), \frac{i}{n+1} \right) : i = 1, \ldots, n \right\},$$

where $x_{(i)}, i = 1, \ldots, n$ is an ordered sample of independent observations and $\hat{F}$ is a candidate model for the true probability distribution F. The quantity $i/n + 1$ corresponds to the *empirical distribution function* evaluated at $x_{(i)}$. If $\hat{F}$ is a reasonable model for the true distribution, the points in the probability plot will lie close to the unit diagonal.

A *quantile–quantile plot* (q–q plot, or just quantile plot) is a plot of the points

$$\left\{ \left(\hat{F}^{-1}\left(\frac{i}{n+1}\right), x_{(i)} \right) : i = 1, \ldots, n \right\}.$$

The quantity $\hat{F}^{-1}(i/(n+1))$ gives a model–based estimate of the $i/(n+1)$ quantile provided by the candidate distribution $\hat{F}$ whilst $x_{(i)}$ itself provides an empirical estimate of this quantile. Again, a well–fitting model would provide points on this plot lying close to the unit diagonal.

Producing a probability plot in R

Recall that in the last section, the wassaw dataset was stored in `dataset`. In R, the code

```
> ordered = sort(dd)
```

then stores the ordered annual maximum sea surges observed at Wassaw Island, $x_{(i)}, i = 1, \ldots, 50$, in the vector `ordered`. For each point in our ordered sample, the following code then stores the empirical distribution function defined by $i/(n+1)$ in the vector `empirical`:

```
> empirical = (1: length(ordered))/(length(dd) + 1)
```

The function `GEV.DF` defines the distribution function for the GEV, as provided by equations (2.6) and (2.7):

```
> GEV.DF = function(input, mu, sigma, xi){
+     if(xi==0){
+         GEV = exp(-exp(-((input-mu)/sigma)))}
+     else{
+         GEV = exp(-(1+xi*((input-mu)/sigma))^(-1/xi))}
+     return(GEV)}
```

Then the following code stores a model–based estimate of the distribution function, evaluated at each point in the ordered sample `ordered`, in the vector `model` (recall that maximum likelihood estimates of the GEV parameters were stored in the object `A` previously):

```
> model = GEV.DF(ordered, A$est[1], A$est[2], A$est[3])
```

Plotting `model` against `empirical` produces the corresponding probability plot for the set of sea surge annual maxima; this can be done using the code:

```
> plot(model ~ empirical, main = 'Probability plot')
> abline(0,1)
```

where the code `abline(0,1)` superimposes the line with intercept 0 and gradient 1 (i.e. the unit diagonal). The resulting plot is shown in Figure 2.3 (left-hand-side). Notice that the model–based probability estimates match up very closely with what we see in our data, suggesting that the GEV is plausible.

Producing a quantile–quantile plot in R

In a similar fashion, we can produce a quantile plot using the following code in R:

```
> GEV.INV = function(input, mu, sigma, xi){
+     if(xi==0){
+         INV = mu-sigma*log(-log(1-input))}
+     else{
+         INV = mu+(sigma/xi)*(((-log(input))^(-xi))-1)}
+     return(INV)}
> model.quantile = GEV.INV(empirical, A$est[1], A$est[2], A$est[3])
> plot(model.quantile ~ ordered, main = 'Quantile plot')
> abline(0,1)
```

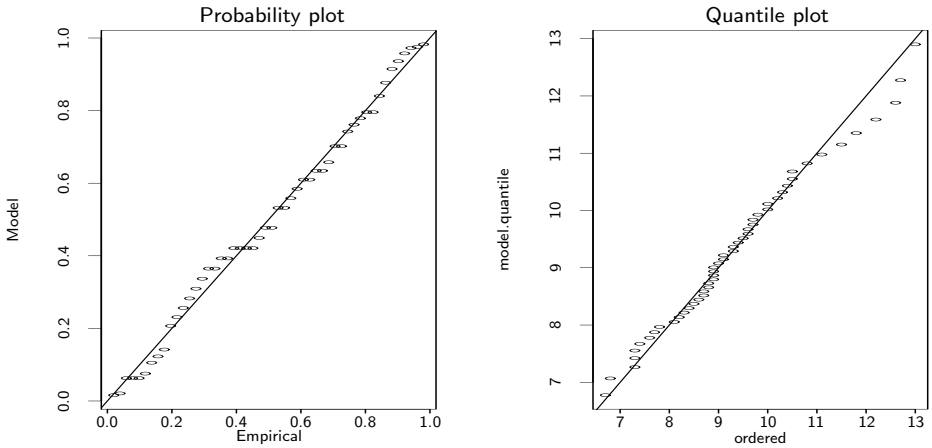

Figure 2.3: Probability plot (left) and quantile–quantile plot (right) for the GEV fitted to annual maximum sea–surges observed at Wassaw Island.

The function `GEV.INV` computes the inverse of the GEV distribution at each of the points in `data`; we evaluate this for each value in `empirical` to find $\hat{F}(i/(n+1))$ and store the results in the vector `model.quantile`. We then plot `model.quantile` against `ordered` to obtain the quantile plot, again superimposing the line of equality using `abline`. The resulting plot is shown in Figure 2.3 (right-hand-side). Again, this plot shows a reasonable fit of the GEV to the annual maximum sea surges at Wassaw.

2.2.4 Return level estimation

In question 2 of Section 1.4, you were asked to provide an estimate of the height of a new sea wall to protect the city of Savannah against the storm surge we would expect to see

(i) once in ten years;

(ii) once in a hundred years.

Using the data, and the data alone, we could not obtain a meaningful estimate of (ii) because we only have 50 years of data. However, we can now use our fitted GEV to extrapolate beyond the range of our data to estimate such return levels. Using Equation (2.10), we find that

$$\hat{z}_{10} = 8.711 - \frac{1.311}{0.108}\left[\left(-\log\left(1-10^{-1}\right)\right)^{0.108} - 1\right] = 11.33 \text{ feet;}$$

similarly,

$$\hat{z}_{100} = 8.711 - \frac{1.311}{0.108}\left[\left(-\log\left(1-100^{-1}\right)\right)^{0.108} - 1\right] = 13.46 \text{ feet.}$$

Thus, an estimate of the height of the sea–wall might be about 11.5 feet or 13.5 feet to protect against the once in ten year, or once in a hundred year, storm surges (respectively). In fact, due to the invariance property of maximum likelihood estimators, our estimates of z_{10} and z_{100} are also the maximum likelihood estimates of these quantities.

As with our inference for the GEV parameters, it is important to quote estimates of return levels along with their standard errors. Since z_r is a function of the GEV parameters μ, σ and ξ, we can use the *delta method* to find the approximate variance of $\hat{z}_r$. Specifically,

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

where V is the variance–covariance matrix of $(\hat{\mu}, \hat{\sigma}, \hat{\xi})^T$ and

$$\begin{aligned} \nabla z_r^T &= \left[\frac{\partial z_r}{\partial \mu}, \frac{\partial z_r}{\partial \sigma}, \frac{\partial z_r}{\partial \xi}\right] \\ &= \left[1, -\xi^{-1}(1 - y_r^{-\xi}), \sigma\xi^{-2}(1 - y_r^{-\xi}) - \sigma\xi^{-1} y_r^{-\xi} \log y_r\right], \end{aligned}$$

where $y_r = -\log(1 - r^{-1})$, evaluated at $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Using R to obtain standard errors for return levels

Recall that V is stored in the matrix `varcovar` in R. Also, recall that our estimates of the GEV parameters are stored in `A$est`. We can define ∇z_{10}, for example, in R as

```
> y10 = -log(1-(1/10))
> del = matrix(ncol=1, nrow=3)
> del[1, 1] = 1
> del[2, 1] = -((A$est[3])^(-1))*(1-(y10^(-A$est[3])))
> del[3, 1] = ((A$est[2])*((A$est[3])^(-2))*(1-((y10)^(-A$est[3]))))-
((A$est[2])*((A$est[3])^(-1))*((y10)^(-(A$est[3])))*log(y10))
> del.transpose = t(del)
```

Then the R command for matrix multiplication, `%*%`, can be used to obtain an estimate of the standard error for $\hat{z}_{10}$ in the following way:

```
> sqrt(del.transpose%*%varcovar%*%del)
          [,1]
[1,] 0.3614576
```

Estimated standard errors for other return levels can be obtained in a similar way. For example, for the standard error for the 100–year return level, we would replace `y10` in the above code with `y100`:

```
> y10 = -log(1-(1/10))
> y100 = -log(1-(1/100))
```

Table 2.2 shows a range of estimated return levels, with associated standard errors in parentheses. Of course, we could use these standard errors to construct confidence intervals for our return level estimates; however, as we shall shortly discuss, such confidence intervals can be misleading.

	Return level			
	z_{10}	z_{100}	z_{200}	z_{1000}
MLE (st. err.)	11.33 (0.361)	13.46 (0.938)	13.99 (1.182)	15.09 (1.821)

Table 2.2: Estimated return levels for sea surges at Wassaw Island, with standard errors (in parentheses).

2.2.5 Using the ismev package in R

The ismev package provides an add–on suite of functions for R users to support the examples and computations in *An Introduction to Statistical Modeling of Extreme Values* (Coles, 2001); the acronym of this book gives the package its name. Once installed, ismev gives the user access to the datasets used in Coles (2001), and the functions supplied by this package allow the user to reproduce the examples used in the book as well as perform rudimentary extreme value analyses on other datasets. All routines in the ismev package assume the data (where necessary) have been pre–processed so that, for example, the user has a set of block maxima ready to analyse. We shall give some thought to the practicalities of obtaining a set of block maxima and dealing with incomplete datasets in future computer practical sessions; for now, we assume any necessary data pre–processing has taken place.

Installing the ismev package requires a version of R no older than R 1.50. Provided you have R 1.50 (or something more recent) installed on your computer, you can load ismev using the command[1]:

```
> library(ismev)
```

[1] Provided you have installed the package on your machine in the first place, using the command `install.packages`

Doing so will give you access to functions which can be used to perform a basic analysis of annual maxima in R: for example, `gev.fit` and `gev.diag`. The command `gev.fit` provides a maximum likelihood fitting routine for the GEV that also allows generalised linear modelling of each parameter (we will return to this in Chapter 5). For example, with our set of annual maximum sea surges stored in `wassaw`, typing `gev.fit(wassaw)` gives the following output:

```
> gev.fit(wassaw)
$conv
[1] 0

$nllh
[1] 89.52412

$mle
[1]  8.7114765  1.3114046 -0.1083845

$se
[1] 0.2094549 0.1489799 0.1074819
```

If the convergence code, `conv`, takes the value zero, then successful convergence has been achieved, as is the case in our example above. `nllh` gives the value of the minimised negative log–likelihood, which has been found to be 89.524 – exactly the value we obtained when doing this 'from first principles' using the `nlm` routine in Section 2.2.2. The output also shows the maximum likelihood estimates, with their standard errors, giving (to three decimal places):

$$\hat{\mu} = 8.711 \ (0.209) \qquad \hat{\sigma} = 1.311 \ (0.149) \qquad \hat{\xi} = -0.108 \ (0.107)$$

Again, this is (almost!) exactly what we obtained when fitting the GEV 'from first principles' in Section 2.2.2; any differences can be attributed to the numerical procedure used to optimise the problem.

To check the goodness–of–fit of the GEV we can use the ismev function `gev.diag`. First of all, we need to store the output from `gev.fit`; for example:

```
> B = gev.fit(wassaw)
$conv
[1] 0

$nllh
[1] 89.52412

$mle
[1]  8.7114765  1.3114046 -0.1083845

$se
[1] 0.2094549 0.1489799 0.1074819
```

Then, typing

```
> gev.diag(B)
```

gives the plots shown in Figure 2.4. You should notice that the plots in the top row (probability plot and quantile plot) are exactly the same as those we constructed, from 'first principles', in Figure 2.3. However, `gev.diag` also gives a 'return level plot' (bottom left) and a plot of the fitted GEV density superimposed onto a histogram of the actual data. The return level plot is really just a quantile plot, but the x–axis is drawn on a logarithmic scale so as to compress the tail of the fitted GEV distribution and thus focus attention on long–period return levels, extrapolated far beyond the range of our observed data. Attached to these return level estimates are 95% confidence bounds using estimate $\pm 1.96 \times$ s.e., where the standard errors have been obtained via the delta method (as in Section 2.2.4).

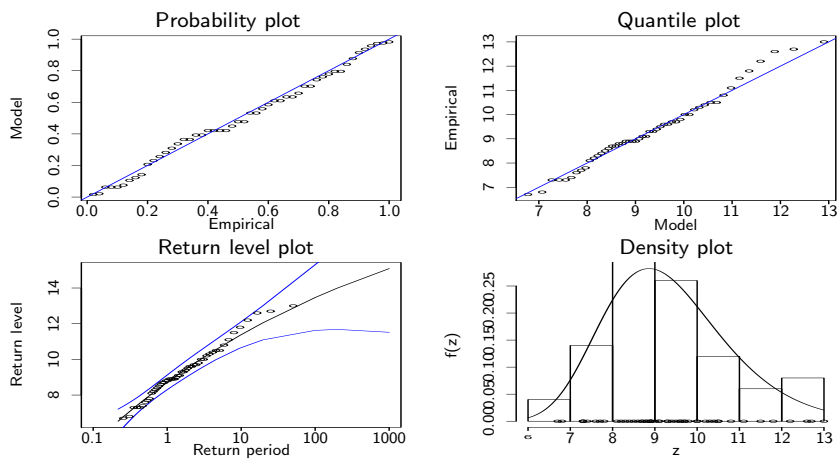

Figure 2.4: Diagnostic plots indicating the goodness–of–fit of the GEV to the Wassaw sea surge extremes.

2.3 Further examples

Example 2.3: Rainfall in the Lake District

The data shown in Table 2.3 are annual maximum rainfall accumulations, obtained from daily records, for a period of 21 years (1991–2011 inclusive) at Eskdale in the Lake District.

333	213	790	343	351	521	307	305	352	277	319
319	339	262	285	297	327	620	350	545	258	

Table 2.3: Annual maximum daily rainfall totals (mm) at Eskdale (1991–2011).

Shown below are the maximum likelihood estimates of the GEV parameters, obtained using R; also shown is the observed information matrix I_O.

$$\hat{\mu} = 304.242; \qquad \hat{\sigma} = 68.977; \qquad \hat{\xi} = 0.249;$$

$$I_O = \begin{pmatrix} 0.0062 & -0.0046 & 0.1962 \\ -0.0046 & 0.0091 & -0.2114 \\ 0.1962 & -0.2114 & 48.4122 \end{pmatrix}.$$

(a) Estimate the standard errors for $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$, and use these to construct 95% confidence intervals for the GEV parameters. Comment.

(b) Find the correlation between $\hat{\sigma}$ and $\hat{\xi}$.

(c) Estimate the 100 and 1000 year return levels for daily rainfall totals at Eskdale.

(d) Obtain standard errors for your estimates in (a), and use these to construct 95% confidence intervals in the usual way. Comment.

Solution to Example 2.3

✎

✎

✎

✎

Example 2.4: Sunshine hours in Santiago de Compostela

Table 2.4 shows the annual maximum number of sunshine hours (per day) for Santiago de Compostela, Spain, for the years 1983–2007.

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		

Table 2.4: Annual maximum sunshine hours (per day) observed at Santiago de Compostela, Spain, 1983–2007 (inclusive).

(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.

Solution to Example 2.4

✎

Example 2.5: Degassing bursts at the Kilauea Volcano, Hawaii

The magnitudes of the vertical forces produced by seismic *degassing bursts* are known to be associated with destructive volcanoes. Table 2.5 shows the maximum force produced by degassing bursts, every quarter in the years 2005–2011 (inclusive), for the Kilauea Volcano in Hawaii.

	Jan–Mar	Apr–Jun	Jul–Sep	Oct–Dec
2005	99983	100067	99905	100367
2006	99980	99970	100086	99988
2007	99912	100084	100432	99921
2008	100123	99913	100240	100448
2009	100162	100296	100015	100122
2010	99930	99997	100710	100136
2011	99946	99918	99962	100052

Table 2.5: Maximum vertical forces produced (kilograms) by Seismic degassing bursts at the Kilauea Volcano in Hawaii.

These data were stored the in the vector `bursts` 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
```

(a) 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 it's empirical counterpart?

(b) Use Figure 2.5 (overleaf) to assess the goodness–of–fit of the GEV to these data.

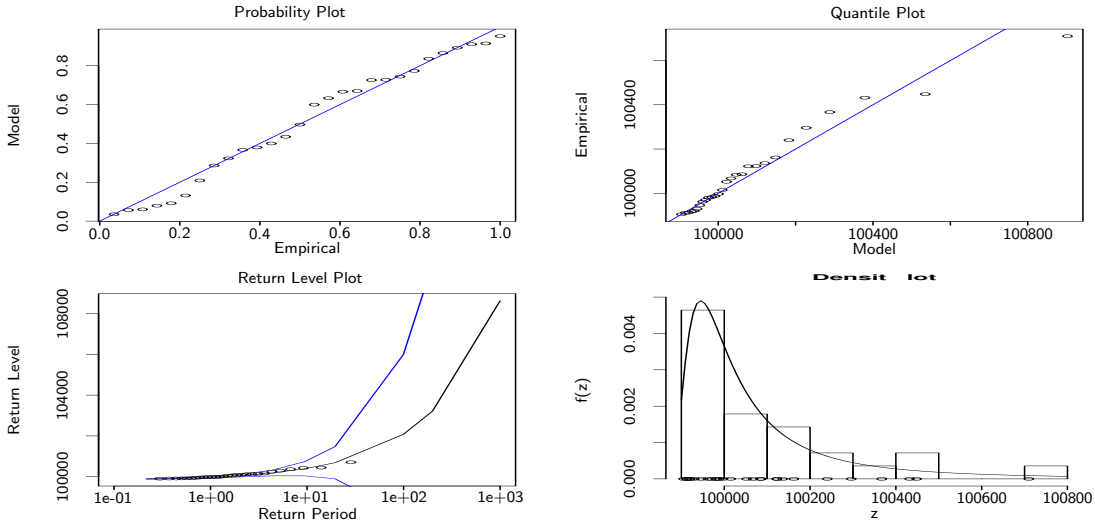

Figure 2.5: Diagnostic plots to assess the fit of the GEV to the degassing burst dataset in Table 2.5.

Solution to Example 2.5

2.4 Profile likelihood

As we saw in our solution to Example 2.3, standard errors for long range return levels, obtained via the delta method, can often be so large that confidence intervals become difficult to work with, or even meaningless. Actually, it turns out that constructing confidence intervals in the standard way (estimate $\pm$ 1.96 $\times$ s.e. for a 95% confidence interval, for example) is not advisable for return levels. This is because of the severe asymmetry often observed in the likelihood surface for the return level, suggesting that the assumption of normality may not be valid for z_r.

An alternative, and often more accurate, method for making inferences on a particular parameter can be found using the *profile log–likelihood*. Formally, the log–likelihood for a parameter vector $\boldsymbol{\theta}$ can be written as $\ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j})$, where $\boldsymbol{\theta}_{-j}$ corresponds to all components of $\boldsymbol{\theta}$ excluding $\boldsymbol{\theta}_j$. The *profile* log–likelihood for $\boldsymbol{\theta}_j$ is defined as

$$\ell_p(\boldsymbol{\theta}_j) \quad = \quad \max_{\boldsymbol{\theta}_{-j}} \ell(\boldsymbol{\theta}_j, \boldsymbol{\theta}_{-j}).$$

Thus, for each value of $\boldsymbol{\theta}_j$ the profile log–likelihood is the maximised log–likelihood with respect to all other components of $\boldsymbol{\theta}$.

For return levels using the GEV model, there are three main steps for the implementation of profile likelihood:

(1) Re–parameterise the GEV model so that z_r becomes one of the model parameters. For example, re–arrange equation (2.10) to write μ in terms of σ, ξ and z_r, i.e.:

$$\mu \quad = \quad z_r - \frac{\sigma}{\xi}\left[\left(-\log\left(1 - r^{-1}\right)\right)^{-\xi} - 1\right]; \tag{2.12}$$

then obtain an expression for the log–likelihood $\ell(\sigma, \xi, z_r)$ by substitution of (2.12) into

$$-m\log\sigma - (1 + 1/\xi)\sum_{i=1}^{m} \log\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+ - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{x_i - \mu}{\sigma}\right)\right]_+^{-1/\xi},$$

where m is the number of block maxima in the analysis.

(2) For some fixed lower value of $z_r = z_{r,\text{low}}$, maximise the GEV log–likelihood $\ell(\sigma, \xi, z_r = z_{r,\text{low}})$ with respect to the two remaining parameters (σ and ξ), to obtain $\ell_p(z_r)$ at $z_{r,\text{low}}$. Here, in terms of the more general notation above, $\boldsymbol{\theta}_j = z_r$ and $\boldsymbol{\theta}_{-j} = (\sigma, \xi)$.

(3) Repeat step (2) for a range of values z_r for $z_{r,\text{low}} \leq z_r \leq z_{r,\text{up}}$, and plot $\ell_p(z_r)$ against z_r to show the profile log–likelihood curve for z_r.

Figure 2.6 (left) shows a plot of the profile log likelihood for the 100 year return level $\ell_p(z_{100})$ for the Lake District rainfall data in Example 2.3, using $z_{100,\text{low}} = 550$ and $z_{100,\text{up}} = 2900$. The plot to the right–hand–side is for the 1000–year return level. Both plots reveal strong asymmetry in the (profile) log–likelihood for the return levels, and it should be clear from these plots that constructing return levels in the usual way (as we did in Example 2.3, part (c)) will be misleading. So how can we use these plots?

In this example, we are partitioning the GEV parameter vector $\boldsymbol{\theta} = (z_r, \sigma, \xi)$ into two components $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, where $\boldsymbol{\theta}^{(1)} = z_r$ and $\boldsymbol{\theta}^{(2)} = (\sigma, \xi)$, and the profile log–likelihood is now defined as

$$\ell_p(\boldsymbol{\theta}^{(1)}) \quad = \quad \max_{\boldsymbol{\theta}^{(2)}}(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}).$$

The following result leads to a procedure for making inferences on the maximum likelihood estimator of $\boldsymbol{\theta}^{(1)}$.

Result

Let $x_1, \ldots, x_n$ be independent realisations from a distribution within a parametric family $\mathcal{F}$, and let $\hat{\boldsymbol{\theta}}_0$ be the maximum likelihood estimator of the d–dimensional model parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, where $\boldsymbol{\theta}^{(1)}$ is a k–dimensional subset of $\boldsymbol{\theta}_0$. Then, under suitable regularity conditions, for large n

$$D_p(\boldsymbol{\theta}^{(1)}) \quad = \quad 2\left\{\ell(\hat{\boldsymbol{\theta}}_0) - \ell_p(\boldsymbol{\theta}^{(1)})\right\} \sim \chi^2_k.$$

Thus, for our single component $\boldsymbol{\theta}^{(1)} = z_r$, the set of values C_α for which $\{z_r : D_p(z_r) \leq c_\alpha\}$ provides a $(1-\alpha)$ confidence interval for z_r, where c_α is the $(1-\alpha)$ quantile of the χ^2_1 distribution.

We apply this result to the profile log–likelihood for z_{100} and z_{1000}; a cut–off point equal to $\frac{1}{2} \times \chi^2_1(0.05) = 1.921$ is shown in Figure 2.6 by the horizontal broken line. The points of intersection between this line and the profile log–likelihood for z_r define the set of values $C_{0.05}$ giving a more appropriate 95% confidence interval for z_r than obtained using the usual approach. These points of intersection are also shown on Figure 2.6 (dotted vertical lines), and give confidence interval for z_{100} and z_{1000} based on the profile likelihood approach of $(602.5, 2765)$mm and $(800, 103000)$mm respectively. These intervals are shown in the table below, along with the intervals we obtained the standard way in Example 2.3.

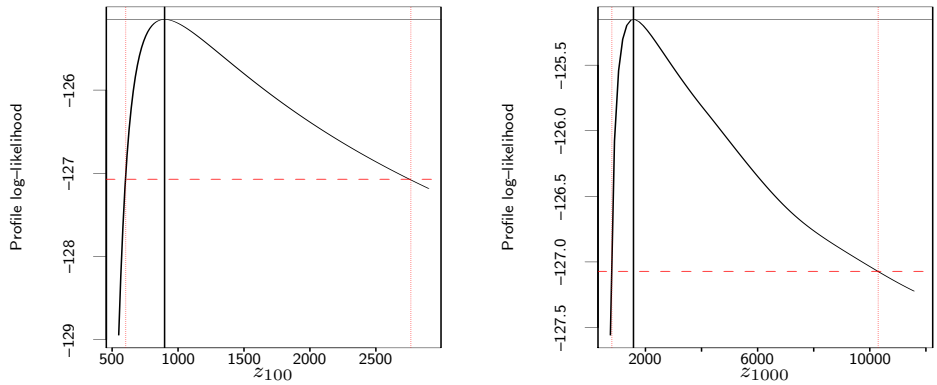

Figure 2.6: Profile log–likelihood curves for the 100 (left) and 1000 (right) year return levels for daily rainfall accumulations in Eskdale, the Lake District.

	$\hat{z}_{100}$	$\hat{z}_{1000}$
Standard 95% CI		
Profiled 95% CI		

The plots shown above in Figure 2.6 can be produced in R using the `gev.prof` function within the ismev package. For example, suppose the results of fitting the GEV to the Lake District rainfall data (obtained using `gev.fit`) are stored in `A`. Then the code `gev.prof(A, xlow=550, xup=2900, 100)` should produce something like the plot for the 100 year return level shown in Figure 2.6. We will see more of this function in a forthcoming computer practical session; use of this function is left as an exercise for you to experiment with!

2.5 Models for minima

Let $\bar{M}_n = \min\{X_1, \ldots, X_n\}$. If we can assume the X_i are independent and identically distributed, we can apply similar arguments to $\bar{M}_n$ as we applied to M_n in Section 2.1.2.

If there exist sequences of constants $a_n > 0$ and b_n such that, as $n \to \infty$,

$$\Pr\left\{(\bar{M}_n - b_n)/a_n \leq x\right\} \quad \to \quad \bar{G}(x)$$

for some non–degenerate distribution $\bar{G}$, then $\bar{G}$ is a member of the GEV family of distributions for minima:

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

where $a_+ = \max(0, a)$, $-\infty < \bar{\mu} < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. This result can be useful where we are interested in modelling extremely small, rather than extremely large, observations (e.g. annual minimum air temperatures). Alternatively, we could negate our set of block minima and then model the corresponding set of maxima, giving identical maximum likelihood estimates of the GEV parameters but for the sign correction $\hat{\bar{\mu}} = -\hat{\mu}$.

2.6 The GEV: Words of warning

✎