MAS8306: Problems sheet 2

Estimating Environmental Extremes

Answers to all starred questions should be placed in the homework submission box in the foyer of the Maths & Stats General Office (Herschel Building 3rd floor) by no later than **4pm, Thursday 15th March**.

1. The distribution function of the Generalised Pareto Distribution (GPD), as given in the lecture notes, is given by

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

where $a_+ = \max(a, 0)$, y corresponds to an exceedance of a high threshold and σ and ξ are the corresponding GPD scale and shape. As with the GEV, the case $\xi = 0$ is taken as the limit $\xi \to 0$, giving an exponential distribution with rate $1/\sigma$. Show that the GPD log–likelihood function is given by

$$\ell(\sigma, \xi; \boldsymbol{y}) \quad = \quad -k\log\sigma - (1 + 1/\xi)\sum_{i=1}^{k} \log(1 + \xi y_i/\sigma), \qquad (1)$$

when $\xi \neq 0$, and

$$\ell(\sigma; \boldsymbol{y}) \quad = \quad -k\log\sigma - \sigma^{-1}\sum_{i=1}^{k} y_i, \qquad (2)$$

for $\xi = 0$, where we have threshold exceedances y_i, $i = 1, \ldots, k$.

2*. The concentration of benzoa pyrene (x μg per 10^4m^3), a known carcinogenic air pollutant linked to exhaust fumes, was measured at the same time each day at a location in Mexico City for a period of 21 days; the data are shown below.

5	48	34	9	10	16	37
41	46	52	57	51	43	35
17	11	4	4	6	8	22

The authorities decided to adopt a threshold–based approach to model the extremes of this pollutant.

(a) Using a threshold of $u = 40\mu$g per 10^4m^3, calculate the threshold excesses $y = (x - u)|x > u$ that would be used in a typical threshold–based analysis of extremes.

(b) Assuming a Gumbel–type tail for the threshold excess distribution, show that the maximum likelihood estimator for σ is given by

$$\hat{\sigma} = k^{-1}\sum_{i=1}^{k} y_i,$$

and hence estimate the GPD scale parameter for your pollution concentration extremes.

(c) Calculate the observed information

$$I_O(\hat{\sigma}) = -\left.\frac{\partial^2 \ell}{\partial \sigma^2}\right|_{\hat{\sigma}},$$

where ℓ is the associated log–likelihood, and use this to find the standard error of $\hat{\sigma}$.

(d) Show that an estimate of the r–year return level is given by

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

where n_y is the number of observations per year and λ_u is the threshold exeedance rate. Hence, estimate the level of benzoa pyrene we can expect to be exceeded once per year at this location in Mexico City.

(e) Use the delta method to find the standard error for your estimate in part (d), accounting for *both* your uncertainty in σ *and* λ_u.

3*. Suppose $X_1, X_2, \ldots X_n$ is a sequence of independent *Fréchet*(1) random variables with

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

Show that the limiting distribution of threshold excesses belongs to the generalised Pareto family. *Hint: Consider* $Pr\{X > u + y | X > u\}$ *as* $u \to \infty$.

Practical work

Open R or Rstudio. You will also need to access data for some of these questions from the course website, so have this open:

`www.mas.ncl.ac.uk/∼nlf8/teaching/mas8306/home.html`

4. Venice, a city built on water, is no stranger to flooding. The seasonal *acqua alta*, a phenomenon which sees the city assaulted by high tides, often leave the historic streets and plazas under water. Go to the section of the course webpage relating to Problems sheet 2, and then click on the "Venice data" link. Save this file to your `H:` drive, or somewhere convenient to you, as venice.txt; then use the following code in R to read in the data:

```
> venice=read.table('venice.txt')
```

Look at the dataset in R. The first column is a year indicator, showing that data have been collected from 1961–2011 (inclusive). The data themselves, in columns 2–11, are the largest, second largest, third largest,..., tenth largest levels of the Venetian lagoon, in cm above the average level of the sea, in each of these years, taken from twice–daily measurements.

(a) Analyse the set of annual maxima by applying the ismev function gev.fit to the largest order statistics at Venice (i.e. the data in column 2). In your solutions, show the R code you have used, as well as the resulting output, which gives estimates of the GEV parameters and their standard errors, and the value of the negative log–likelihood. Store the results of this fit in `fitanmax`, as we will come back to this shortly.

(b) You were given some private reading from *Extreme Values: Statistical Analysis Using R* (Fawcett & Walshaw, 2013). Use the ismev command `rlarg.fit` to fit the r–largest order statistic model to: (i) the 2 largest sea levels observed each year at Venice; (ii) the 6 largest sea levels observed each year at Venice; (iii) all 10 largest sea levels observed each year at Venice. Store the results of these fits in, for example, `fit2`, `fit6` and `fit10`, respectively. In your solutions, do not include any R code, but include your estimates of the GEV parameters with their standard errors. For help using the `rlarg.fit` function, type

```
> ?rlarg.fit
```

(c) Why might it be preferable to use an r largest model over the standard annual maxima approach? Support your comments with reference to the results stored in `fitanmax`, `fit2`, `fit6` and `fit10`.

(d) If the asymptotic approximation to the GEV is valid for a particular order statistic r, then it is valid for all order statistics $r' < r$. This should mean we see *stability* in our estimates of the GEV parameters across all values of order statistics used. Comment, with reference to your estimated GEV parameters in `fitanmax`, `fit2`, `fit6` and `fit10`.

(e) Use the ismev commands `gev.diag` and `rlarg.diag` to produce diagnostics plots to assess the fit provided by `fitanmax`, `fit2`, `fit6` and `fit10`. In your solutions, show the corresponding probability plot, q–q plot, return level plot and fitted density plot for each of your fits, and comment.

(f) Produce a suitably–labelled time series plot of the set of annual maxima. What assumption underpinning the fitting of the GEV might be violated here?

5*. Hourly maximum wind gusts have been collected at High Bradfield, in the Peak District, over a period of 10 years from January 1st 2003 to December 31st 2012 (inclusive). These data can be saved from the course webpage and then loaded into R using the `scan` command.

(a) Load the wind speed data into R, storing them in the vector `bradfield`, and then perform simple exploratory analyses. In your solutions, include two graphical summaries of the wind speed data, as well as some simple numerical summaries (use the `summary` command). Comment.

A new visitor's centre is to be built near High Bradfield. The *British Standards Institute* specify design requirements for such structures such that they can withstand the wind speed we can expect to see, on average, once every 50 years.

(b) In part (a), you should have noticed that there are many missing values in the wind speed dataset. Before continuing this question, remove the missing values by typing

```
> bradfield2=bradfield[!is.na(bradfield)]
```

This will store data from `bradfield` which is *not* a missing value (`NA`) in the new vector `bradfield2`. There is nothing to show in your solutions for this part of the question – just remember that, for the remainder of this question, your data are now stored in `bradfield2`!

(c) As part of the team of planners involved in the building of the new visitors centre, you decide to perform an analysis of threshold exceedances in order to estimate the 50 year wind gust. Use the ismev command `mrl.plot` to produce a mean residual life plot, and use this as a tool to suggest a suitable threshold for identifying wind speeds as extreme. In your solutions, you should include (i) the mean residual life plot; (ii) your chosen threshold, and (iii) the threshold exceedance rate – that is, the proportion of observations which exceed your chosen threshold.

You should now use R to fit the Generalised Pareto Distribution (GPD) to the wind speed exceedances above your chosen threshold. You should do this from "first principles", using the `nlm` routine, and then confirm your results using the `gpd.fit` command in ismev.

(d) First of all, from "first principles":

(i) Extract the set of threshold exceedances. For example, with the wind speed data in `bradfield2` and a chosen threshold `u`, the code

```
> exceedances=bradfield2[bradfield2>u]-u
```

selects all observations from the wind speed data that exceed the threshold, subtracts the threshold and then stores these threshold exceedances in the vector `exceedances`.

(ii) Write your own function for the GPD log–likelihood you derived in question 1, and given by equations (1) and (2). Follow the GEV example on page 22 of the lecture notes, and include your function in your solution to this question. Make sure this is a function of `theta`, a parameter vector with two components relating to the GPD scale and shape parameters σ and ξ.

(iii) After initialising `theta` at some suitable starting values, use `nlm` to obtain estimates of the GPD scale and shape parameters. In your solutions, include the output from R which shows the parameter estimates, the negative log–likelihood and the corresponding Hessian matrix.

(iv) Use the command `solve` to obtain the variance–covariance matrix for the GPD parameters. In your solutions, include any relevant R code, and then use your answers to this question, and previous questions, to complete the following table:

	Scale parameter σ	Shape parameter ξ	Threshold exceedance rate λ_u
Estimate			
St. Err.			
95% CI			

(e) Verify your answers to part (d) by implementing the ismev command `gpd.fit`. Make sure you include your R code, and any relevant output, in your solutions.

(f) Use the ismev command `gpd.diag` to assess the fit of the GPD to your set of wind speed extremes. Include any relevant plots in your solutions, and comment.

(g) Using your answers in the table above, estimate the 50 year return level wind speed – the value required by the planners of the new visitor's centre. To account for leap years, you should use your number of observations per year $n_y = 24 \times 365.25$.

(h) Use the ismev command `gpd.prof`, with `xlow=92` and `xup=102`, to produce a plot of the profiled log–likelihood for the 50 year return level. Verify your answer to part (g) by using the command `abline(v=...)` to place a vertical line on this plot at your estimate of the 50 year return level.

(i) Use your plot in part (h) to obtain a 95% confidence interval for the estimate of the 50 year return level.

(j) Isolate the first month of data by typing

```
> jan2003=bradfield[1:(31*24)]
```

Now use the code

```
> par(mfrow=c(1,2))
```

to partition the plotting window into one row and two columns. Produce a time series plot of `jan2003`. Also produce a plot of each observation in `jan2003` against the next value in time, by typing

```
plot(jan2003[1:743],jan2003[2:744])
```

Include these plots in your solutions. What do you notice from both of these plots that might violate the usual assumptions when fitting the GPD?

(k) Now isolate the first three *years* of wind speed data and produce another time series plot. Include this plot in your solutions. Again, what do you notice from this plot that might suggest the usual assumptions are violated?