Extreme Value Analysis

Emma Eastoe

Department of Mathematics and Statistics
Lancaster University, UK

1. Introduction
2. Maxima and minima
   - The generalised extreme value (GEV) distribution
   - Regression-type models
3. Peaks over Threshold
   - The generalised Pareto (GP) distribution
   - Cluster identification
   - Extremal Index
4. Summary
   - Further topics

Introduction

- Statistical models aimed at capturing behaviour of very largest, or very smallest, observations in a data set.
- Many applications in the environmental sciences: air pollution, hydrology, temperatures, wind speed, precipitation, wave heights,...
- Goal: estimate return levels or the probability of an unusually large (small) event.

Return levels

- N-year return level, the value a process is expected to exceed, once every N years. Where N could be 50, 100, 500, ...
- If you are unlucky, this return level *could* be exceeded in two successive years even if N is large.

Probability of a large event

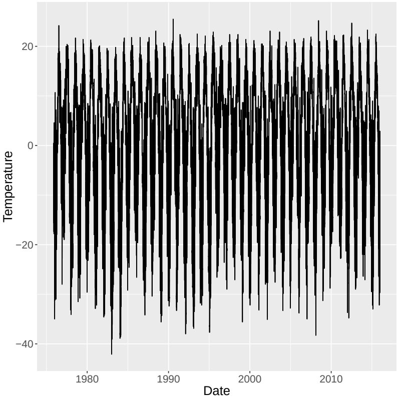
To estimate the probability that a process exceeds x:

- If x is 'moderate' do this empirically: calculate the proportion of observations (data points) above x;
- If x is high may have no observations above x so empirical method doesn't work: use a model and extrapolate.

Greenland temperatures

- Greenland is mostly ice sheet
- Why worry about extremely high temperatures?
- Rising temperatures $\rightarrow$
  more positive degree days $\rightarrow$
  greater melting $\rightarrow$
  increased global sea level

Daily max temperatures, Kangerlussuaq (1975–2015)



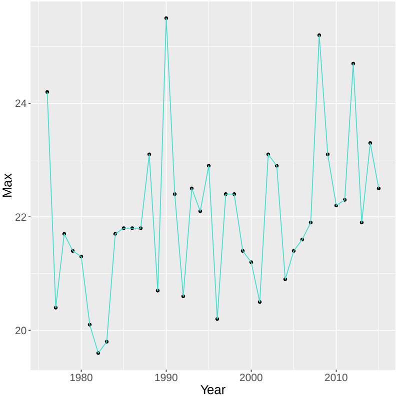
Annual maxima and minima

The simplest way to characterise an extreme event is to look at the annual maxima/minima. We may ask:

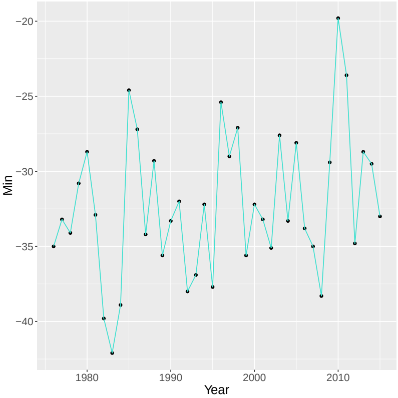
- What is the largest maxima (smallest minima) that we might expect to see
    - In 10-years?
    - In 100-years?
    - Ever?
- What is the chance that the annual maximum exceeds a certain high and previously unobserved value?
- Is the behaviour of the annual minima/maxima changing over time?

Kangerlussuaq

Annual maxima



Annual minima



Models for maxima and minima

- Generalised extreme value distribution
- Defined by it's cumulative distribution function. Let X represent an annual maximum (minimum) then

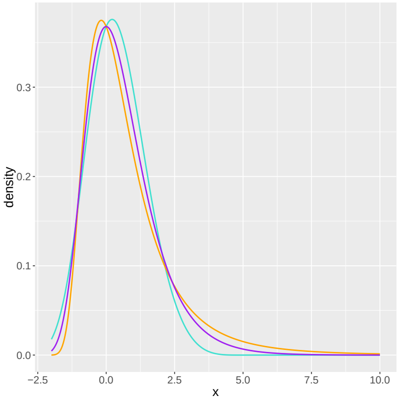
$$G(x) = \Pr[X \leq x] = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

- Three parameters (unknowns): location μ, scale $\sigma \in (0, \infty)$ and shape ξ
- Shape determines how fast the distribution decays i.e. how quickly the quantiles get large.

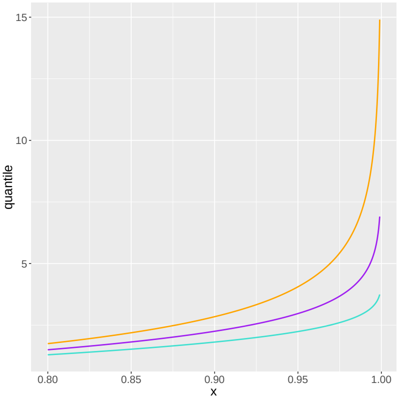
GEV$(0, 1, \xi)$

Shapes: $\xi = -0.2$ (turquoise), $\xi = 0$ (purple) and $\xi = 0.2$ (orange)

Densities



Upper quantiles



GEV as a statistical model

- Useful to model any data set which is contains maxima or minima
- Condition: maxima/minima have been taken over a 'large enough' number of underlying observations
  - Fine to use for annual max/min of daily observations
  - Not so good to use for daily max/min of hourly observations
- Estimation of parameters via any method of statistical inference
  - Maximum likelihood, Bayes, L-moments
  - All of these implemented in R package `extRemes`

Return levels

- For the GEV these are directly related to quantiles.
- Let x_N be the N-year return level, then to find x_N assuming a GEV model, solve

$$\Pr[X \leq x_N] = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} = 1 - \frac{1}{N}$$

- Gives

$$x_N = \frac{1}{\xi}\left\{[-\log(1-1/N)]^{-\xi} - 1\right\}$$

- Again implemented in `extRemes`

Implementation in R

- Quite a few packages including `extRemes`, `texmex`, `evd` and `ismev`
- For this course we will use `extRemes`
- Has a single function for fitting the various EVA models.

Kangerlussuaq again

- Load the data into R,

  ```
  > annMaxKanger <- read.csv("kangerMax.csv")
  ```

  Data frame with two columns: `Year` and `Max`
- Fit the model

  ```
  > max.fit.1 <- fevd(x=Max,data=annMaxKanger)
  ```

- To obtain parameter estimates, standard errors etc

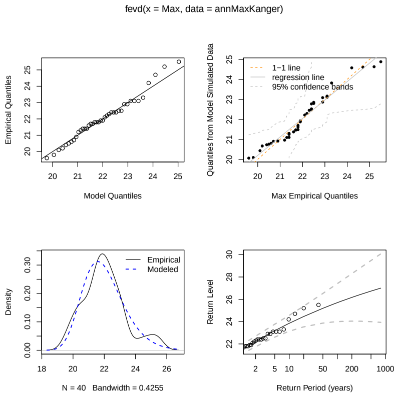
```
> summary(max.fit.1)
```

- Important parts of the output

```
 Estimated parameters:
  location       scale       shape
21.4478281  1.1863160 -0.1135755

 Standard Error Estimates:
 location      scale      shape
0.2095874 0.1471525 0.1088952

 AIC = 140.7024
```

Visual diagnostics of model fit:

```
> plot(max.fit.1)
```

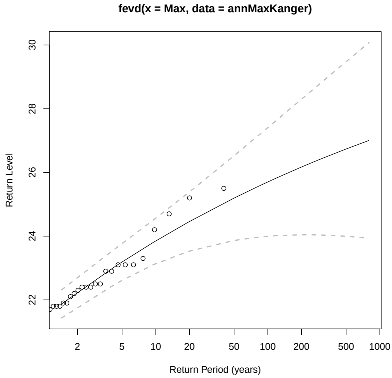


Estimate return levels using

```
> rl <- return.level(max.fit.1,return.period=seq(5,500,by=5))
```

and to plot

```
> plot(max.fit.1,type="rl")
```



Regression modelling

Model discussed above assumes data are stationary over time. It is more likely that either (or both):

- There is a trend in the maxima over time (climate change?)
- One or more physical variables could be used to help explain changes in the maxima

Model this using regression-type techniques.

- Write location and/or scale parameter as linear functions of covariate(s) e.g.

$$\begin{aligned} \mu(\text{time}) &= \mu_0 + \mu_1 \text{time} \\ \log \sigma(\text{time}) &= \sigma_0 + \sigma_1 \text{time} \end{aligned}$$

- 'log-link' for scale to ensure $\sigma(\text{time}) > 0$ for all time
- In general, keep shape ξ constant
- More unknown parameters to estimate!
- Compare models using likelihood ratio test or AIC to only include significant covariate(s)

Regression models in R

```
> annMaxKanger$Time <- annMaxKanger$Year-1974
> max.fit.2 <-
      fevd(x=Max,data=annMaxKanger,location.fun=~1+Time)
> summary(max.fit.2)
```

The fitted model is

$$
\begin{aligned}
\mu(\mathrm{year}) &= 20.4 + 0.049 \times (\mathrm{year} - 1974) \\
\sigma &= 0.99 \\
\xi &= 0.0056
\end{aligned}
$$

The standard error for the 'year' coefficient is 0.015.

Model selection

- If models are nested, use the likelihood ratio test
- Enter the simpler model first, in this case `max.fit.1`

Using the `extRemes` package:

```
> lr.test(max.fit.1,max.fit.2)
```

The output looks like this:

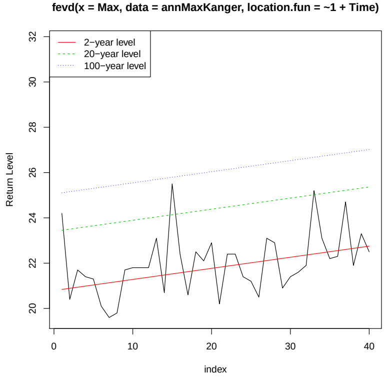
```
Likelihood-ratio Test


data:  MaxMax
Likelihood-ratio = 8.5274, chi-square critical value = 3.8415,
0.0500, Degrees of Freedom = 1.0000, p-value = 0.003498
alternative hypothesis: greater
```

$p = 0.0035 < 0.05$ therefore there is evidence of a time trend in the location parameter

Effective return levels

These are N-year return levels for each value of the covariate

```
> plot(max.fit.2,type="rl")
```



Peaks over Threshold (PoT)

- Alternative to modelling only maxima/minima
- Allows us to model all unusually large (or small) events
- Requires identification of these events
- But a more efficient use of the data than just taking maxima/minima

Overview of PoT approach

- Choose a high threshold: any observation above this is classified as an extreme event
- Model
  - Rate - how often do they occur?
  - Size - how large are they?

  of threshold exceedances.

Note on threshold identification: visual aids exist to help with this e.g. mean residual life plot.

Generalised Pareto distribution

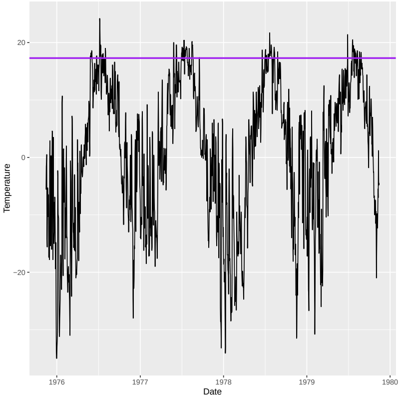
- The generalised Pareto (GP) distribution is used to model the size of threshold exceedances
- The full cumulative distribution function for an exceedance is given by

$$\Pr[X \leq x] = 1 - \phi \left[1 + \xi \left(\frac{x - u}{\psi} \right) \right]^{-1/\xi}$$

- Three unknown parameters: rate $\phi \in [0, 1]$, scale $\psi \in (0, \infty)$ and shape ξ
- As with GEV, parameter estimation by likelihood, Bayesian, L-moments,...
- Relies on threshold being very high

Kangerlussuaq: daily maxima temperatures

Choose the 90% quantile as a threshold (purple line)



Implementation of GP in R

Can use the same `fevd` function to fit the GP model as we used for the GEV model:

```
> kanger <- read.csv("kangerTemp.csv")
> thresh <- quantile(kanger$Temp,0.9) #define threshold
> gp.fit.1 <- fevd(x=Temperature,data=kanger,threshold=thresh,
> summary(gp.fit.1)
```

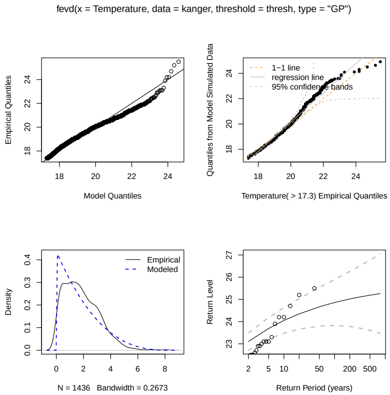
From the last command we see that $\hat{\psi} = 2.28$ and $\hat{\xi} = -0.27$. To find $\hat{\phi}$,

```
> gp.fit.1$rate
```

and $\hat{\phi} = 0.098$ (why is this not surprising?!)

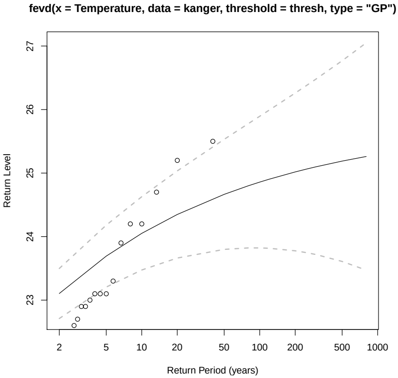
As for the GEV model, use visual diagnostics to check model fit,

```
> plot(gp.fit.1)
```



And return levels for the daily temperature (not the annual maxima now):

```
> plot(gp.fit.1,type="rl")
```



Modelling event maxima only

- The diagnostics and return level plot suggest that the model could do better at describing the highest temperatures
- Model could be too simplistic
- Perhaps behaviour changes over time
- Or extremes are not independent

Event identification

- So far have thought of each threshold exceedance as a separate event
- What if exceedances occur in groups (clusters)?
- Can model cluster *maxima* using the GP model
- Need a way to identify clusters: number of algorithms
- Use the *runs method*: Exceedances separated by
  - fewer than r consecutive non-exceedances belong to same cluster;
  - more than r consecutive non-exceedances belong to independent clusters.

Cluster Identification in R

Use the `decluster` function:

```
> kangerDecl <- decluster(kanger$Temperature,threshold=thresh
    ,method="runs",r=3)
```

To get a summary of the declustering

```
> print(kangerDecl)
```

- Shows 223 clusters.
- Should assess sensitivity to choice of different run lengths.

Extremal Index

Extremal Index

Often denoted θ is a measure of the strength of extremal dependence.

- Lies in the interval $[0, 1]$.
- Stronger dependence as θ gets closer to 0.
- Mean cluster size is reciprocal of extremal index.
- Independent series have $\theta = 1$ but $\theta = 1$ does not imply that the series is independent, merely that the threshold exceedances are independent.

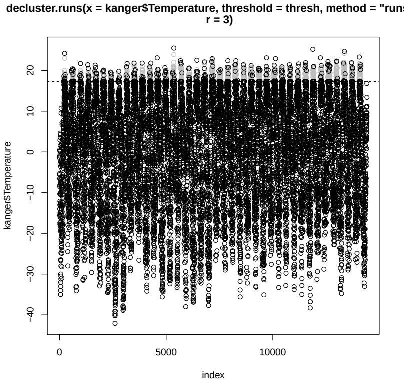
- The extremal index can be estimated using the runs or intervals method.
- `decluster` and `decluster.runs` in `extRemes` automatically output the intervals estimate even if the runs method is specified.
- Instead use `extremalindex` function:

```
> extremalindex(kanger$Temperature,threshold=thresh
      ,method="runs",run.length=3)
```

- Gives $\hat{\theta} = 0.158$, with a mean cluster size of $1/0.158 = 6.3$.

Can look at the clusters using

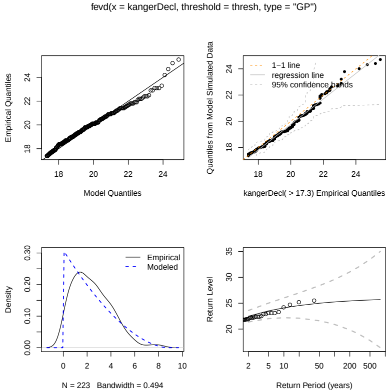
```
> plot(kangerDecl)
```



Refit GP to cluster maxima only, and check diagnostics.

```
> gp.fit.2 <- fevd(x=kangerDecl,threshold=thresh,type="GP")
> plot(gp.fit.2)
```

For cluster max only $\hat{\psi} = 0.24$ and $\hat{\xi} = 0.040$. Diagnostics much better!



Summary

- Two main EVA models: Generalised Extreme Value (GEV) and generalised Pareto (GP) distributions
- GEV appropriate for annual maxima or minima data
- GP used to model threshold exceedances in PoT approach
- Can extend both models to include regression terms
- For PoT approach might want to check for clustering and model only cluster maxima
- Return levels can be easily produced for both GEV and GP models; if regression terms are included then effective return levels will be produced

Further areas of interest

- Including regression terms in the GP model
- Statistical downscaling of extremes
- Spatial modelling of extreme events; Gaussian process based models not necessarily appropriate
- Multivariate modelling of extremes: joint extremal behaviour of e.g. two variables at a single location, the same variable at two locations,...