Extreme Value Analysis

Shikun Li

Contents

How do we predict rare events of extreme values, such as floods (high riverflow), tsunami (high sea level) and DDoS attack (anomalous traffic), when there are very few such observations? What do 100-year flood, 50-year wave and 200-day outbreak mean? Successful predictions of these rare events is important to prevent loss.

In this example, **extRemes** package (Gilleland and Katz 2016) will be used.

Simplest case: $X_1, X_2, X_3, ...X_n \overset{i.i.d.}{\sim} F$. Require accurate inference on the tail of F.

1 Block maxima

1.1 Definition

$X_1, X_2, X_3, ...X_n \overset{i.i.d.}{\sim} F$ and define:

$$M_n = max\{X_1, X_2, X_3, ...X_n\}$$

Then the distribution of M_n is

$$Pr\{M_n < z\} = (F(z))^n$$

1.2 Fisher–Tippett–Gnedenko theorem

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

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G(z)$$

then

$$G(z) \propto exp[-(1 + \xi z)^{-1/\xi}]$$

For some non-degenerate distribution, G belongs to one of the following:

Gumbel:

$$G(z) = exp\{-exp(-(\frac{z-b}{a}))\}, z \in \mathbb{R}$$

Weibull:

$$G(z) = \begin{cases} exp\{-exp(-(\frac{z-b}{a}))\} & z < b \\ 1 & z \geq b \end{cases}$$

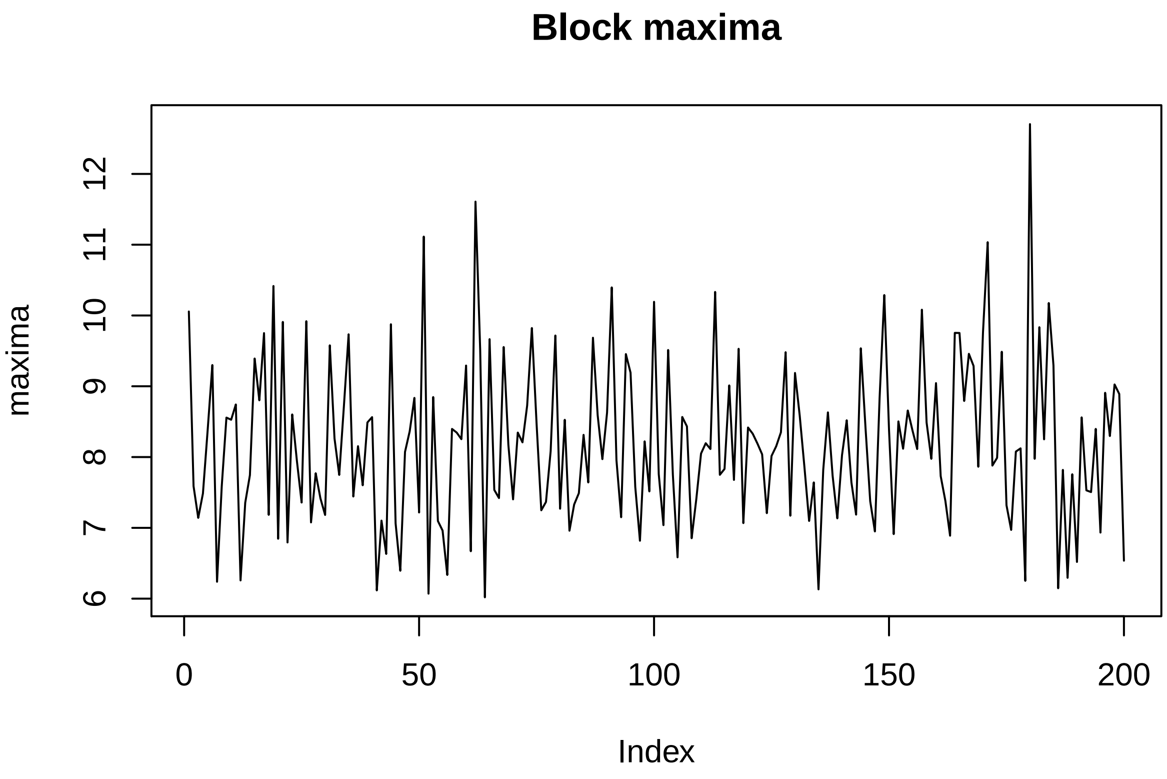
Frechet:

$$G(z) = \begin{cases} 0 & z \leq b \\ exp\{-(\frac{z-b}{a})^{-a}\} & z > b \end{cases}$$

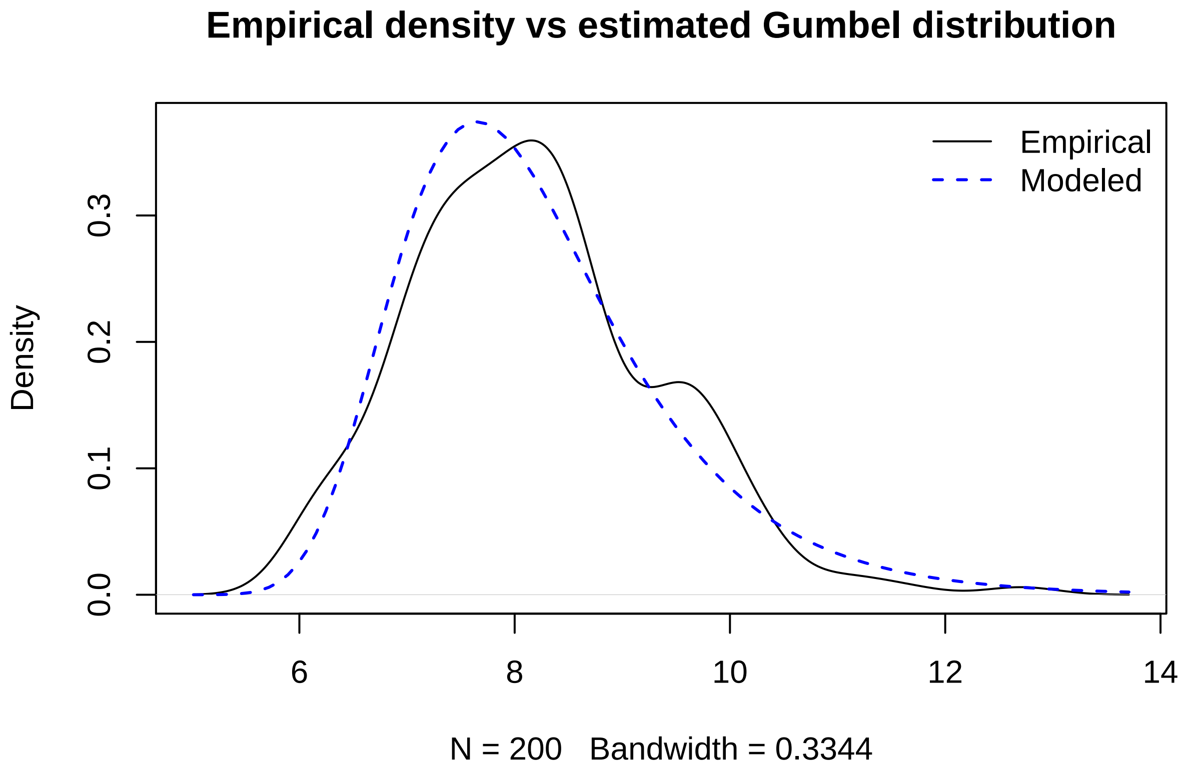
Illustrating the theorm by simulation:

```r
set.seed(123)
library(extRemes)
## block size
n <- 12
original_mean <- 5
original_sd <- 2

## Create a series of maxima
series_length <- 200
maxima <- c()
## Simulate blocks of data
data_series <- list()
for (i in 1:series_length) {
  data_series[[i]] <- rnorm(n = n, mean = original_mean, sd = original_sd)
}
maxima <- unlist(lapply(data_series, max))
plot(maxima, main = "Block maxima", type = "l")
```



```r
fit <- fevd(maxima, type = "Gumbel")
plot(fit, type = "density", main = "Empirical density vs estimated Gumbel distribution")
```



1.3 Quantiles and return levels

In terms of quantiles, take $0 < p < 1$ and define

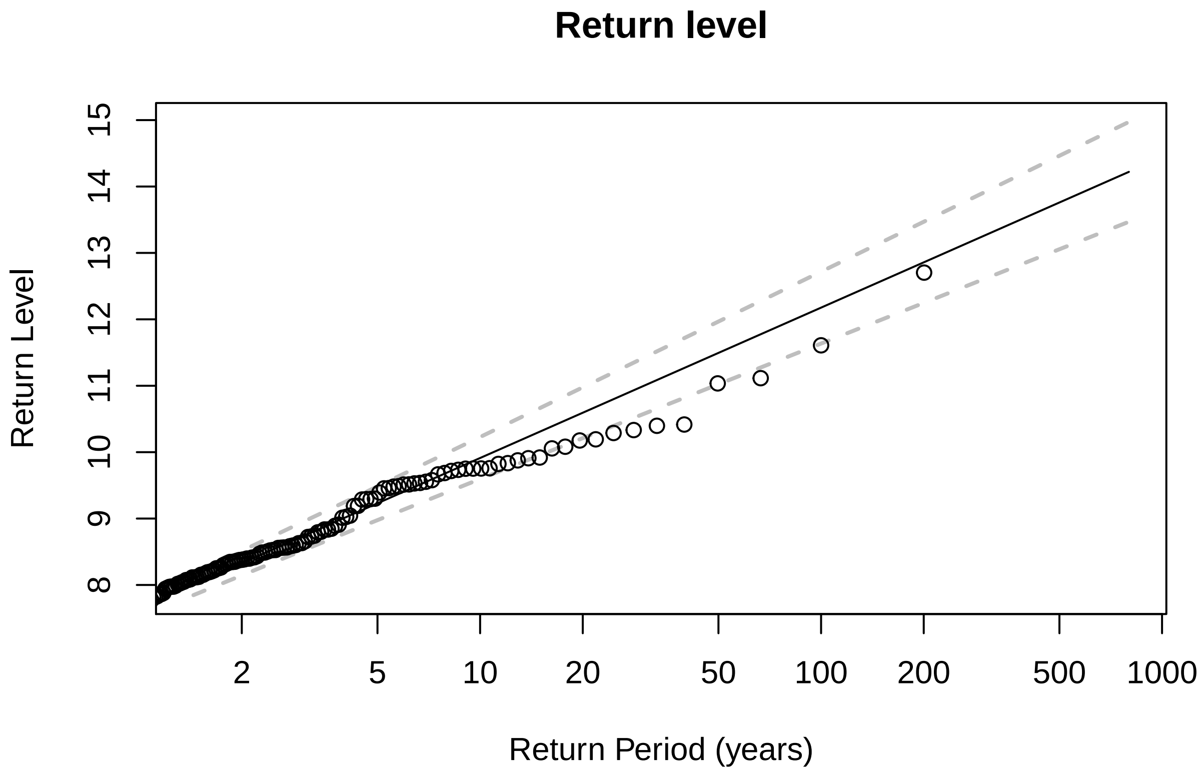
$$z_p = \mu - \frac{\sigma}{\xi}[1 - \{-log(1-p)\}^{-\xi}]$$

where $G(z_p) = 1 - p$.

In extreme value terminology, z_p is the return level associated with the return period $1/p$.

For annual maxima of rainfall, z_p is the amount of annual maximum rainfall with probability of occurrence $1 - p$ in any given year, and $1/p$ is the recurrence interval in years. (it can be considered as a Geometric distribution)

```r
plot(fit, type = "rl", main = "Return level")
```



(Coles and Davison 2008)

2 Peaks over thresholds

Since lots of data are thrown away, the block maxima method is inefficient. A more efficient method is called peaks over thresholds. The POT method uses observations above a given threshold, u. There are two ways of using POT data: one for the size of exceedances and a second for the number of events in a time period considered.

2.1 Generalized Pareto distribution

$X_1, X_2, X_3, ...X_n \overset{i.i.d.}{\sim} F$, and define a new variable for exceedances above u: $Y = X - u$ for $X > u$. We can write $Y_j = X_i - u$ such that i is the index of the jth exceedance, $j = 1, ..., n_u$. The distribution function of Y can be defined as

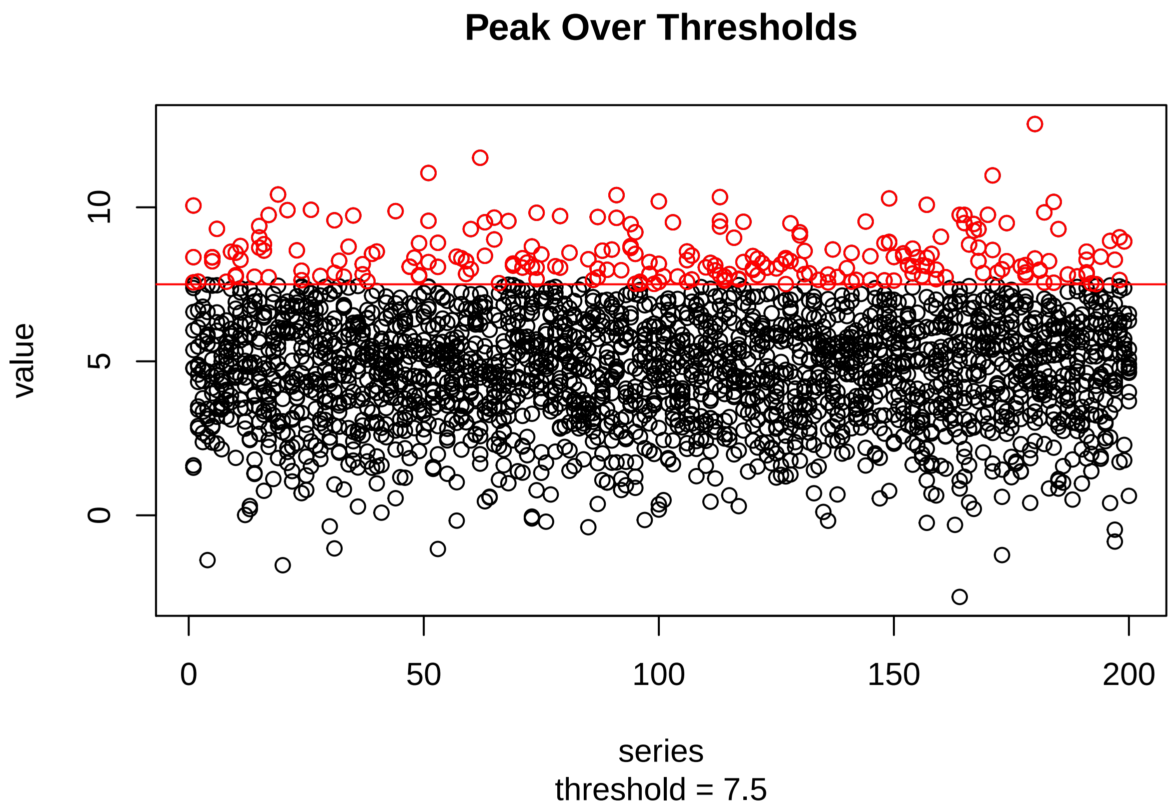
$$F(y) = Pr\{Y \leq y\} = Pr\{X - u \leq x | X > u\}, \ y \geq 0$$

This distribution can be approximated by a Generalized Pareto distribution (Pickands III 1975).

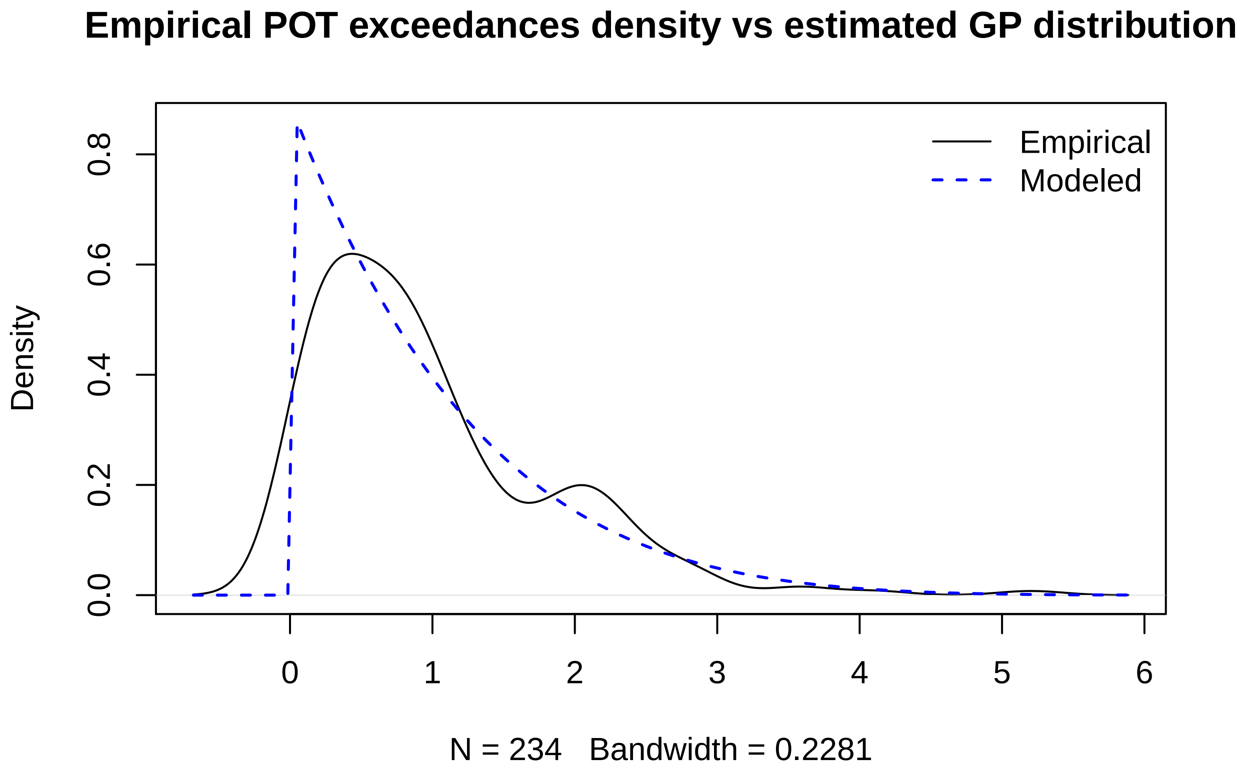
(Bommier 2014)

```
## Setting the threshold u
threshold <- 7.5
plot(x = rep(1:series_length, each = n), y = unlist(data_series), main = "Peak Over Thresholds",
     sub = paste("threshold =", threshold), xlab = "series", ylab = "value")
pot_points <- unlist(data_series) > threshold
```

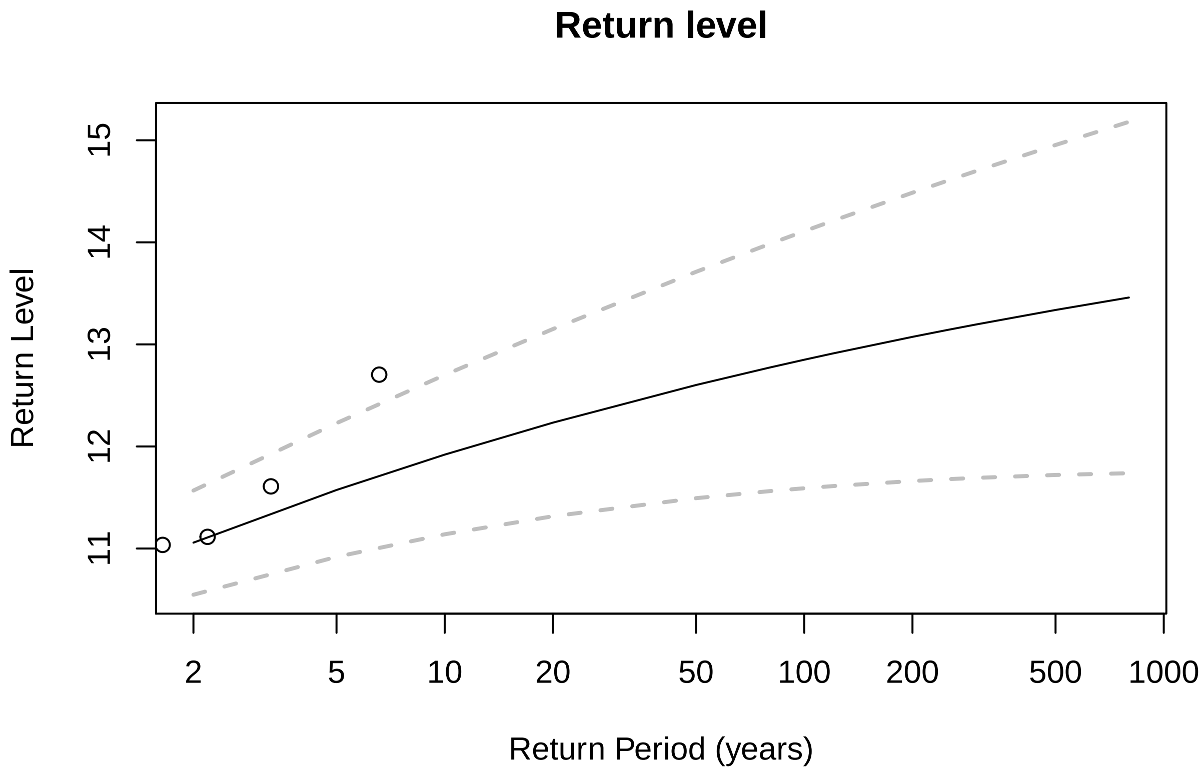
```
points(x = rep(1:series_length, each = n)[pot_points], y = unlist(data_series)[pot_points], col = "red"
abline(h = threshold, col = "red")
```



```
gp_fit <- fevd(unlist(data_series), threshold = threshold, type = "GP")
plot(gp_fit, type = "density", main = "Empirical POT exceedances density vs estimated GP distribution")
```



```r
plot(gp_fit, type = "rl", main = "Return level")
```



The selection of threshold is not straightforward: threshold too low – bias because of the model asymptotics being invalid; threshold too high – variance is large due to few data points.

2.2 Point process representation

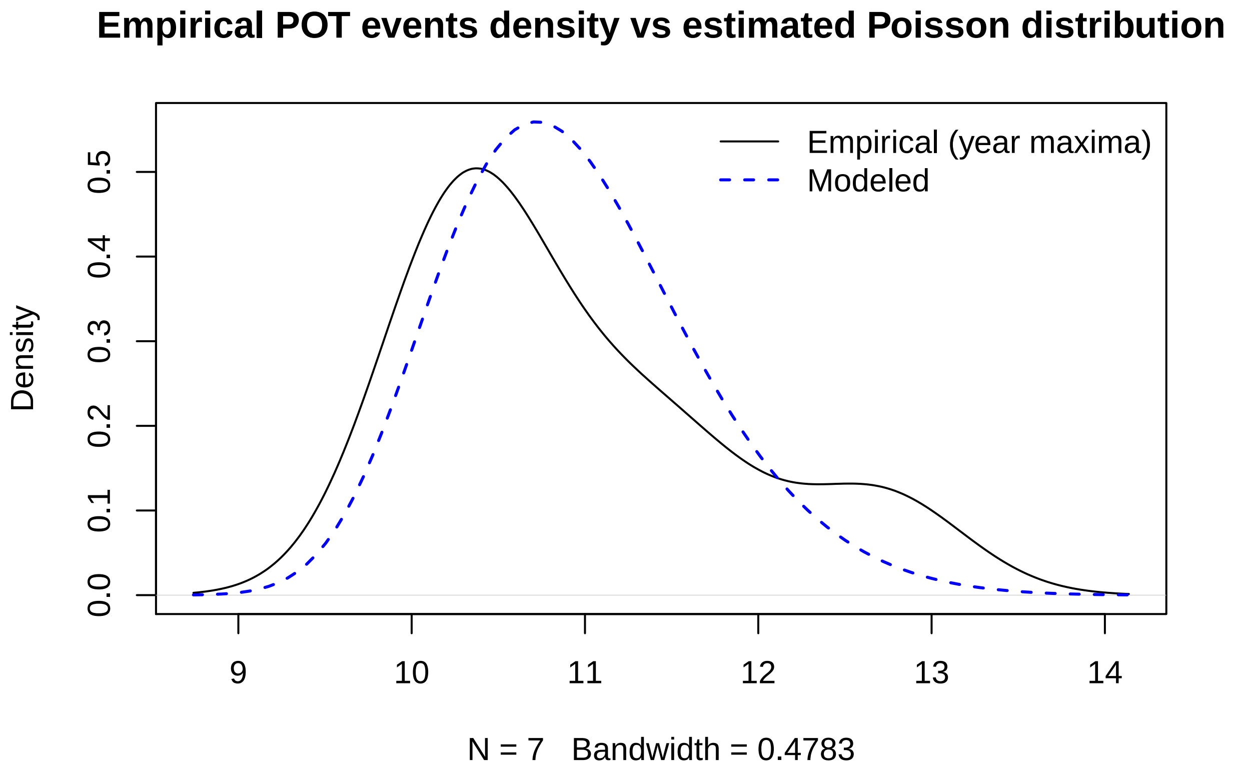
Suppose F unknown, $X_1, X_2, X_3, ...X_n \overset{i.i.d.}{\sim} F$. Form a 2-dimensional point process $\{(i, X_i); i = 1, ..., n\}$ and characterize the behaviour of this process in regions of the form $A = [t1, t2] \times [u, \infty)$.

Construct a sequence of point process on $\mathbb{R}^2$ by

$$P_n = \{(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} : i = 1, ..., n)\}$$

Then $P_n \to P$ as $n \to \infty$, where P is a Poisson process with intensity λ.

```
pp_fit <- fevd(unlist(data_series), threshold = threshold, type = "PP")
plot(pp_fit, type = "density", main = "Empirical POT events density vs estimated Poisson distribution")
```



References

Bommier, Esther. 2014. "Peaks-over-Threshold Modelling of Environmental Data."

Coles, Stuart, and Anthony Davison. 2008. "Statistical Modelling of Extreme Values." http://www.cces.ethz.ch/projects/hazri/EXTREMES/talks/colesDavisonDavosJan08.pdf.

Gilleland, Eric, and Richard W. Katz. 2016. "extRemes 2.0: An Extreme Value Analysis Package in R." *Journal of Statistical Software* 72 (8): 1–39. doi:10.18637/jss.v072.i08.

Pickands III, James. 1975. "Statistical Inference Using Extreme Order Statistics." *The Annals of Statistics*. JSTOR, 119–31.