Summary

- The three fits agree fairly well:
  - Maxima: $\widehat{\eta} = 20.4_{0.649}$, $\widehat{\tau} = 5.84_{0.483}$, $\widehat{\xi} = 0.08_{0.072}$;
  - Poisson process: $\widehat{\eta} = 19.8_{0.556}$, $\widehat{\tau} = 6.52_{0.379}$, $\widehat{\xi} = 0.07_{0.051}$;
  - POT: $\widehat{p}_u = 0.033$, $\widehat{\sigma}_u = 5.83_{0.394}$, $\widehat{\xi} = 0.07_{0.051}$.
- The location and scale parameters are estimated quite well, but the shape much less well.
- The shape parameter estimate is slightly positive, but not significantly so (some hydrologists claim that rainfall has $\xi \approx 0.1 \ldots$).
- The fit appears to be good.
- In applications one would need to check that the threshold fits are robust to the choice of u (above $u_{\min}$).
- It is tempting to fit the model with $\xi = 0$, which will give much smaller standard errors for the other parameters. But as we do not know that $\xi = 0$, this reduction in uncertainty may be unrealistic, and it may introduce bias in extrapolation.

3.4 Targets of Inference

Return levels and return periods

- In basic analyses, typically aim to estimate risk measures such as

$$\mathrm{P}(X > x) = 1 - F_X(x), \quad x_p = F_X^{-1}(1-p),$$

  where $X \sim F_X$ is a background observation and x and x_p are larger than any value yet observed.
- We often express risk in terms of blocks of m background observations, often daily measurements, with the blocks being years; then $m = 365.25$.
- We then call x_p a **T-year return level** with a **return period** of $1/p$ observations or T years (i.e., $N_p = Tm$ background observations),
  - e.g., the law states that nuclear installations should withstand the highest windspeed in $T = 10^7$ years(!), so if X is a daily maximum windspeed, then $N_p = 365.25 \times T$ and $p = 1/(365.25T)$.
- Hence a return level solves the equation

$$\mathrm{P}(X > x_p) = 1 - F_X(x_p) = p = 1/N_p \qquad (13)$$

  for some small p.

Return levels and return periods II

☐ If $x_p > u$, so p is less than the probability $\mathrm{P}(X > u) = p_u$ that a background observation exceeds threshold u, then solving $1 - F_X(x_p) = p$ in the POT model gives

$$x_p = \begin{cases} u + \dfrac{\sigma_u}{\xi}\left\{(p_u/p)^{\xi} - 1\right\}, & \xi \neq 0, \\ u + \sigma_u \log(p_u/p), & \xi = 0. \end{cases} \tag{14}$$

☐ The GEV applies to maxima of blocks of m background observations, so we approximate the upper tail of F by $G^{1/m}$, giving

$$1 - p = G^{1/m}(x_p), \tag{15}$$

which yields

$$x_p = \begin{cases} \eta + \dfrac{\tau}{\xi}\left[\{-m\log(1-p)\}^{-\xi} - 1\right], & \xi \neq 0, \\ \eta - \tau \log\{-m \log(1-p)\}, & \xi = 0. \end{cases} \tag{16}$$

☐ In both cases

- $-\log(1-p) \doteq p = 1/N_p$ for large N_p, giving simpler expressions,
- point estimates are obtained by replacing the unknown parameters by their estimates,
- uncertainty is best assessed using the profile log likelihood for x_p.

Note on computation of return levels

☐ For the POT model in which the GPD is fitted to exceedances of u, and provided $x_p > u$, we have

$$\begin{aligned} \mathrm{P}(X > x_p) &= \mathrm{P}(X > x_p \mid X > u)\mathrm{P}(X > u) \\ &= \mathrm{P}(X - u > x_p - u \mid X > u)\mathrm{P}(X > u) \\ &= \{1 + \xi(x_p - u)/\sigma_u\}_+^{-1/\xi} \times p_u, \end{aligned}$$

and we seek x_p such that

$$1 - p = \mathrm{P}(X \le x_p) = 1 - p_u\{1 + \xi(x_p - u)/\sigma_u\}_+^{-1/\xi},$$

which leads to the stated expression for x_p.

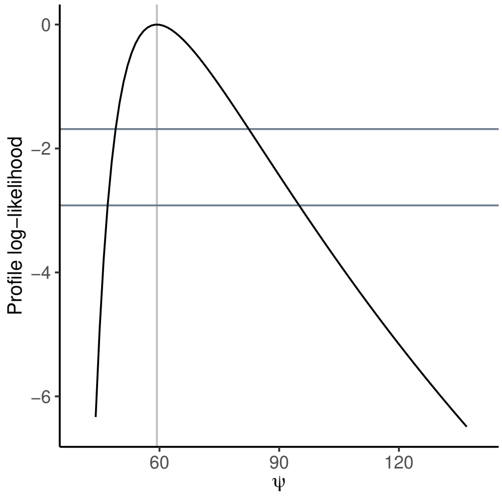
☐ If the GEV model is fitted to the maxima of blocks of m background observations then we have

$$1 - p = G^{1/m}(x_p) = \exp\left[-\{1 + \xi(x_p - \eta)/\tau\}_+^{-1/\xi}/m\right],$$

which gives the stated expression for x_p.

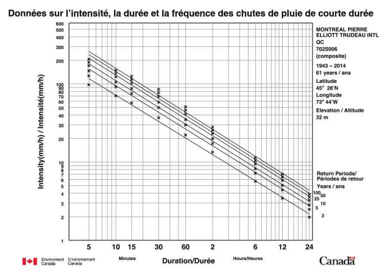
Profile log-likelihood

- Here $\psi = x_p$ is the 100-year return level for daily precipitation at Abisko based on the GEV fit.
- The strong asymmetry means that symmetric confidence intervals could be very misleading.



Return levels and return periods III

- In hydrology, an **intensity-duration-frequency (IDF)** curve describes the relationship between rainfall intensity, duration, and a given return period and is used for flood risk assessment and water management.
- For each duration D, the frequency and magnitude of extreme rainfall events are estimated.
- Relying on the GEV applied to the series of annual maxima, estimates of x_p, the T-year return level, are produced. For comparison purposes, we work with $I = x_p/D$.
- The Gumbel distribution is usually used for convenience but more general approaches have recently been proposed.



IDF curves for Montréal airport. Source: Environment and Climate Change Canada (ECCC)

Other measures of risk

- ☐ In environmental applications it may be important to estimate amounts of rain falling into an entire catchment area, or the length and impact of a heatwave, or . . .
- ☐ The Basel Accords regulate measures of risk to be used by financial institutions:
  - the **Value at Risk** VaR_p is another name for a quantile/return level x_p;
  - the **Expected Shortfall** is defined as the expected loss conditional on VaR_p being exceeded,

    $$\mathrm{E}(X - \text{VaR}_p \mid X > \text{VaR}_p),$$

    where in both cases X represents a potential loss.
  - More sophisticated measures such as **expectiles** are also used.

http://stat.epfl.ch slide 105

Comments

- ☐ The T-year return level is often called 'the level exceeded once on average every T years', and is easily misinterpreted:
  - 'on average' does not mean that disasters arise at regular T-year intervals!
  - selection is often discounted — if M independent time series are monitored, then we expect M/T T-year events each year;
  - the assumption of stationarity is rarely true, so large events may cluster together in periods of elevated risk.
- ☐ Preferable to refer to quantiles — but probably impossible to change a cultural icon!
- ☐ Return levels and return periods are parameters of distributions, but future events are as-yet unobserved random variables, and it may be useful to consider their distributions. The distribution of the largest value X_T to be observed over T blocks of future background observations is $G^T(y)$, and it may be better to use this for risk analysis, in a Bayesian approach (later, probably).

http://stat.epfl.ch slide 106

4 Complications slide 107

4.1 Introduction slide 108

Basic ideas

- ☐ Chapter 3 described 'vanilla' statistical analyses for rare events using the GEV, GPD and point process methods.
- ☐ The basic derivations of these models assume that

$$X_1, \ldots, X_m \overset{\text{iid}}{\sim} F, \quad m \to \infty.$$

- ☐ In applications these assumptions are generally false:
  - – m is finite;
  - – the background data may show trend, seasonality or other forms of **non-stationarity**, so $X_j \sim F_j$;
  - – time series are typically **dependent**, as cold weather, heatwaves, . . . occur over several days;
  - – some (maybe subtle) **selection** mechanism may apply, e.g., when an analysis is performed immediately after a rare event.
- ☐ This chapter will describe methods for detecting and dealing with these problems.

http://stat.epfl.ch slide 109

4.2 Nonstationarity slide 110

Vanilla analysis of maxima

- ☐ Our previous analyses supposed that
  - – block maxima satisfy $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim}$ GEV(η, τ, ξ),
  - – exceedances of a threshold u satisfy $X_1 - u, \ldots, X_n - u \overset{\text{iid}}{\sim}$ GPD(σ, ξ),

  but often we observe additional variation, either due to
  - – **systematic** changes in the background data (e.g., due to trend or seasonality), or to
  - – **haphazard** variation (e.g., due to weather conditions) that we have not accounted for.
- ☐ We'll pass most time looking at systematic changes.
- ☐ For an example of haphazard variation, consider annual maximum daily rainfall $M = \max(X_1, \ldots, X_{365})$, where X_j is total rainfall on day j. On many days $X_j = 0$, so

$$M = \max(X_1, \ldots, X_N),$$

  where $N \ll 365$ is the (random) number of rainy days. If N varies a lot from year to year, then M might be much smaller in some years than in others, so the GEV is a poor model (remember we derive it assuming that $X_j \sim F$, where F is continuous . . .).

http://stat.epfl.ch slide 111

A damp day in Venice

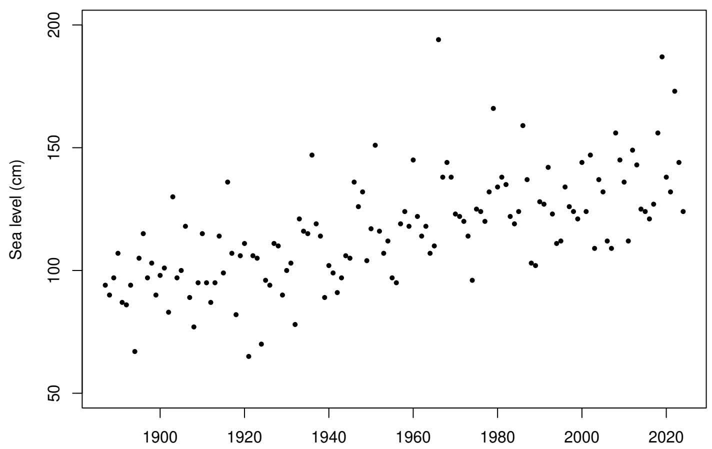
http://stat.epfl.ch slide 112

Punta della Dogana and Santa Maria della Salute

http://stat.epfl.ch slide 113

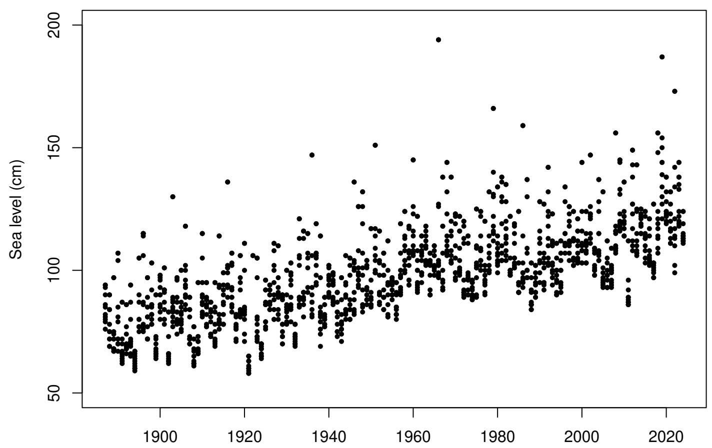
Annual maximum sea levels, 1887–2024

In October 2020, the MOdulo Sperimentale Elettromeccanico (MOSE) system was inaugurated: rows of mobile gates are raised when particularly high tides are predicted, in order to limit how much water from the Adriatic Sea can enter the Venetian lagoon. The record: 196 cm in 1996.



Ten largest annual sea levels, 1887–2024

In 1935, only the six largest values are available, and in 1922 only the largest value is available. The data sources for 1887–1981 and 1982 onwards are different.



Non-stationarity

- Obvious approach is to suppose that the GEV parameters can depend on external factors, i.e., $Y_t \sim \mathsf{GEV}(\eta_t, \tau_t, \xi_t)$, where the dependence might be specified as

$$\begin{aligned}
\eta_t(\beta) &= \beta_0 + \beta_1 t, \\
\eta_t(\beta) &= \beta_0 + \sum_{k=1}^{K} \{\beta_{2k-1}\cos(2\pi kt/365) + \beta_{2k}\sin(2\pi kt/365)\}, \\
\eta_t(\beta) &= \beta_0 + \beta_1 x(t), \\
\tau_t(\beta) &= \exp(\beta_0 + \beta_1 t), \\
\xi_t(\beta) &= \begin{cases} \beta_1, & t \le t_0, \\ \beta_2, & t > t_0, \end{cases}
\end{aligned}$$

  where $x(t)$ is some physical quantity that varies over time (e.g., ENSO, NAO, or global average temperature).
- In applications we typically find that
  - the location parameter η varies,
  - the scale parameter τ might or might not vary,
  - the shape parameter ξ is constant (it is difficult to estimate, and anyway often is regarded as an intrinsic aspect of the background process).

Parametric inference

- Example model specification: $y_t \stackrel{\text{ind}}{\sim} \mathsf{GEV}(\eta_t, \tau_t, \xi_t)$, where η_t, τ_t, ξ_t depend on parameters β.
- If $y_1, \ldots, y_n$ are assumed to be independent, then the log likelihood for β is

$$\ell(\beta) = \sum_{t=1}^{n} \log g\{y_t; \eta_t(\beta), \tau_t(\beta), \xi_t(\beta)\},$$

  where g is the GEV density.
- Maximization of $\ell(\beta)$ yields maximum likelihood estimates and the observed information matrix, from which we compute standard errors, confidence intervals, etc.
- We say that model $\mathcal{M}_0$ is nested within a model $\mathcal{M}_1$ if $\mathcal{M}_1$ reduces to $\mathcal{M}_0$ by fixing (say) d parameters. Then the corresponding maximised log likelihoods satisfy $\widehat{\ell}_1 \ge \widehat{\ell}_0$, and the likelihood ratio statistic (or equivalently difference in deviances) is

$$W = 2(\widehat{\ell}_1 - \widehat{\ell}_0).$$

- If $\mathcal{M}_0$ is adequate, then asymptotic likelihood theory implies that $W \dot\sim \chi^2_d$, so values of W larger than the $1-\alpha$ quantile of the χ^2_d distribution would lead to a rejection of $\mathcal{M}_0$ in favour of $\mathcal{M}_1$, at significance level α.

Model diagnostics

- If $Y_t \sim \text{GEV}(\eta_t, \tau_t, \xi_t)$ for $t = 1, \ldots, n$, then

$$Z_t = \frac{1}{\xi_t} \log\left(1 + \xi_t \frac{y_t - \eta_t}{\tau_t}\right) \quad \overset{\text{iid}}{\sim} \quad \text{standard Gumbel},$$

  i.e.,

$$\text{P}(Z_t \leq z) = \exp\{-\exp(-z)\}, \quad z \in \mathbb{R}, \quad t = 1, \ldots, n.$$

- If we replace the parameters by their estimates $\widehat{\eta}_t = \eta_t(\widehat{\beta})$, etc., these results should still hold (approximately) for the **Gumbel residuals**

$$\widehat{z}_t = \frac{1}{\widehat{\xi}_t} \log\left(1 + \widehat{\xi}_t \frac{y_t - \widehat{\eta}_t}{\widehat{\tau}_t}\right), \quad t = 1, \ldots, n.$$

- We use the $\widehat{z}_t$ in diagnostic plots, e.g.,
  - the **probability plot**, showing $\{j/(n+1), \exp\{-\exp(-\widehat{z}_{(j)})\};\ j = 1, \ldots, n\}$, or
  - the **quantile plot**, showing $\{(-\log[-\log\{j/(n+1)\}], \widehat{z}_{(j)})\ ;\ j = 1, \ldots, n\}$, or

  plots of the $\widehat{z}_j$ against appropriate variables, to see if any patterns remain after fitting the model.

Example: Venice sea levels, 1887–2019

- We ignore the data from 2020 onwards, when MOSE is operational.
- Analysis of maxima uses straight-line regression model,

$$\eta_t = \beta_0 + \beta_1 x_t, \quad t = 1, \ldots, n = 133,$$

  with $(x_1, \ldots, x_{133}) = (1887 - 1990, \ldots, 2019 - 1900)/100$ chosen so that
  - β_0 equals the location parameter in the year 1900,
  - β_1 denotes the change in maximum sea level over 100 years,
- We fit two nested models, both with constant scale and shape parameters, i.e.,

$$\mathcal{M}_0: \quad \eta_t = \beta_0, \quad \tau_t = \tau, \quad \xi_t = \xi,$$
$$\mathcal{M}_1: \quad \eta_t = \beta_0 + \beta_1 x_t, \quad \tau_t = \tau, \quad \xi_t = \xi.$$

- The code prints a 'deviance' $D = -2\widehat{\ell}$ for the fitted model, which allows model comparison using the likelihood ratio statistic:

$$w = 2(\widehat{\ell}_1 - \widehat{\ell}_0) = D_0 - D_1.$$

Example: Fitting models

```
y <- venice$y[venice$year<2020,1]
x <- (venice$year[venice$year<2020]-1900)/100
(fit0 <- evd::fgev(y))

Call: evd::fgev(x = y)
Deviance: 1193.487

Estimates
    loc    scale    shape
106.517   20.050   -0.139

Standard Errors
    loc    scale    shape
1.89487  1.29297  0.04412

Optimization Information
  Convergence: successful ...

(fit1 <- evd::fgev(y,nsloc=x))  # nsloc specifies the x variable for the non-stationary locat

Call: evd::fgev(x = y, nsloc = x)
Deviance: 1122.072

Estimates
     loc  loctrend     scale     shape
 89.8087   35.0291   15.0816   -0.1023

Standard Errors
     loc  loctrend     scale     shape
 2.34431   3.51218   0.96584   0.04071

Optimization Information
  Convergence: successful ...
```

Example: Venice sea levels, 1887–2019

Model-checking for fit to Venice maximum sea-level data. Left panel: Gumbel-scale residuals, $\widehat{z}_t$. Right: ordered $\widehat{z}_t$ plotted against Gumbel plotting positions, with pointwise (dot-dash) and overall (solid) 95% confidence bands obtained by simulating 10,000 Gumbel samples.