TP1 Statistical learning with extreme values: homework

Anne Sabourin, Antoine Doizé

October, 9, 2025

Objectives. You can just return a `.ipynb` file, with code cells for the code and Markdown cells for the "theoretical" questions.

2C. GPD fit and diagnostics

We recall that we denote X the random variable having same law as the supposedly i.i.d. random variables $(X_i)_{i=1...}$ denoting the random recorded daily rain intensities.

1. Fix u : $u = 30$ mm as the threshold. Fit a GPD on the exceedances $X - u$.
2. Display the GPD QQ-plot and comment on the fit.
3. Let's denote Z a random variable. Let us have $Z_1, Z_2 \dots$ i.i.d. replicas of Z
   (a) . Let's have T a duration. Let's have the associated quantile z_T such that $P(Z > z_T) = 1/T$. Denote τ_{z_T} the waiting time before first exceedance of z_T, i.e. $\tau_{z_T} := \min\{t \geq 1 : X_t > z_T\}$. What is the distribution of τ_{z_T} ?
   (b) Give the value of $\mathbb{E}[\tau_{z_T}]$. According to you why is z_T called the return level associated with the period T ?
   (c) Given a threeshold u, and a threshold level $z_T \geq u$, express $P(X > z_T)$ in terms of F_u defined as $F_u(y) := P(X - u \leq y \mid X > u), \quad y \geq 0$.
   (d) Let us have the quantity $\hat{\lambda} := \frac{N_u}{n}$ with $N_u := \#\{i = 1 \dots n | X_i > u\}$ counting the number of exceedance over u. Using two approximations, express $P(X > z_T)$ in terms of $\hat{\lambda}$ and of a GPD c.d.f. Comment quickly on what secures these approximations.
   (e) Compute (theoretically and numerically) the 100-years return level of daily rain intensity from the expression found in (d). Caution : Don't forget that our records $X_1 \dots X_n$ are daily records.

2D. Compute return level with block maxima method

1. Suggest another way to estimate the 100-years return level of rain intensity, but using the Block Maxima method. *Provide a structured and detailed answer. In particular you will comment on your block size choice : advantages and drawbacks. You will make the parallel between this hyperparameter and another choice we had to do in the GPD approach.*
2. Compare this estimation with the one obtained with the GPD approach.

Part 3 — Minima (Glass Fiber) GEV on lower tail (35–45 min)

Context. We now study *extreme minima* using experimental data on glass-fiber breaking strengths. The experiment consists of subjecting short glass fibers (each of length 1.5 cm) to mechanical vibrations whose power is gradually increased. The vibration intensity,controlled through both frequency and amplitude, is raised until each fiber sample breaks. The corresponding breaking power is then recorded as a measure of its strength.

To interpret these data, we use a simple conceptual model. We consider each 1.5 cm glass fiber sample as a small system composed of M microscopic glass sub-fibers. Each sub-fiber has its own intrinsic resistance (or breaking power), and the entire system fails as soon as the weakest sub-fiber breaks. In this framework, the observed breaking strength X_i of sample i corresponds to the minimum among the M sub-fiber strengths :

$$X_i = \min(X_{i,1}, X_{i,2}, \ldots, X_{i,M}),$$

where the $X_{i,j}$ are assumed to be i.i.d. random variables representing the strength of individual sub-fibers.

Hence, the recorded dataset can be viewed as a collection of such minima :

$$X_1, X_2, \ldots, X_n,$$

each corresponding to one experimental trial (one fiber system). This setup naturally leads us to study the statistical behavior of extreme minima. For convenience, we will also consider the transformation $Y = -X$, so that the minima of X become maxima of Y ; this allows us to reuse standard Generalized Extreme Value (GEV) tools developed for maxima.

3A. Data reading and exploration

1. Load the CSV *glassfiber.csv* (single column : `strength`). Let X denote the breaking strength. The dataset contains 63 observations corresponding to the breaking strengths of glass fibers under controlled experimental conditions, interpreted as the minimum resistance among the sub-fibers composing each system.
2. Plot a histogram of X and report $\min X$, median(X), and $\max X$.
3. Comment on the **left tail** : Give a simple physical intuition for the existence or not of a left bound.

3B. GEV fit for minima via sign flip

1. Define $Y = -X$ and fit a GEV to Y with `scipy.stats.genextreme.fit` (recall SciPy uses $c = -\xi$).
2. Report $(\hat{\mu}_Y, \hat{\sigma}_Y, \hat{\xi}_Y)$ and interpret the **sign of** $\hat{\xi}_Y$. Translate this to the *lower tail of* X : does it suggest a bounded lower endpoint (Weibull-type), light tail (Gumbel), or heavy tail (Fréchet) ? Could you have guessed this tail shape earlier ?

3C. Diagnostics for minima

1. **QQ-plot (GEV)** : Show empirical quantiles of Y vs theoretical quantiles of the fitted GEV. Focus your comments on the points corresponding to small X (i.e., large Y).
2. **"Return levels" for minima :** Can you give an expression of the quantile of Y with respect to a quantile level assuming the GEV modeling is relevant ? Using this expression and the fitted model on Y, compute low quantiles for X :

$$q_p^{(X)} = -\,Q_{1-p}^{(Y)} \quad \text{for } p \in \{0.01,\, 0.005,\, 0.001\},$$

where $Q_q^{(Y)}$ is the q-quantile under the fitted GEV for Y

Plot those quantiles (use log-scale axis for the x-axis of the quantile levels).

3D. Model selection

Context. In the previous section, we fitted a Generalized Extreme Value (GEV) distribution to the transformed variable $Y = -X$, obtaining parameter estimates

$$\hat{\mu}_Y,\ \hat{\sigma}_Y,\ \hat{\xi}_Y.$$

You found that the estimated shape parameter $\hat{\xi}_Y$ was close to zero. Since the case $\xi = 0$ corresponds to the *Gumbel distribution*, it is natural to ask whether the additional parameter ξ is truly needed.

Compare the restricted Gumbel model ($\xi = 0$) to the full GEV model (ξ free) using a Likelihood Ratio Test (LRT). Provide a structured answer in which you will explain each step and provide intermediate conclusions.

Reminder : Likelihood Ratio Test Let us consider two nested statistical models :

$$\mathcal{M}_0 \subset \mathcal{M}_1,$$

where the parameter vector in the full model can be written as

$$\theta = (\theta_0, \theta_1),$$

with

$$\theta_0 \in \Theta_0 \subset \mathbb{R}^{p_0}, \qquad \theta_1 \in \Theta_1 \subset \mathbb{R}^{p_1}.$$

The two models are then written as :

$$\mathcal{M}_1 = \{P_{\theta_0,\theta_1} : (\theta_0, \theta_1) \in \Theta_0 \times \Theta_1\},$$

$$\mathcal{M}_0 = \{P_{\theta_0,0} : \theta_0 \in \Theta_0\} \subset \mathcal{M}_1.$$

That is, the restricted model $\mathcal{M}_0$ fixes the sub-parameter $\theta_1 = 0$ (or some specific value), while the full model $\mathcal{M}_1$ allows it to vary freely.

Hypothesis to be tested :

$$H_0 : \theta_1 = 0 \quad \text{(restricted model } \mathcal{M}_0\text{)} \qquad \text{vs.} \qquad H_1 : \theta_1 \in \Theta_1 \quad \text{(full model } \mathcal{M}_1\text{)}.$$

Then under $\mathbb{H}_0$ and suitable regularity conditions (we will suppose those conditions are met) :

$$T := 2\ln\left(\frac{\sup_{\theta\in\Theta_1} \mathcal{L}(\theta) \mid y_n}{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta) \mid y_n}\right) \xrightarrow{n\to\infty,d} \chi^2_{p_1}$$

Thus, the likelihood ratio test relies on :

1. Compute the T statistics with the fitted parameters for each of the models
2. Check if it is over a given quantile of $\chi^2_{p_1}$
3. Reject H_0 with a given risk (the one associated with the quantile chosen earlier)