

TP2 Statistical learning with extreme values

Anne Sabourin, Antoine Doizé

October, 23, 2025

1 Building a simple model from scratch

We build and probe a simple bivariate model that exhibits *multivariate regular variation* (MRV) for the Peaks Over Threshold (POT) regime. We will reconnect with Proposition 3.3.2 (in particular item (5)) linking the domain of attraction of a simple max-stable vector and asymptotic radial-angular independence.

1.1 Model

Let R follow a *Pareto* distribution, such that $\mathbb{P}(R > r) = \frac{1}{2r}$ for $r \geq 2$. Let $\delta \in \{-1, +1\}$ with $\mathbb{P}(\delta = 1) = \mathbb{P}(\delta = -1) = 1/2$, independent of R . Define the angle (on the unit simplex)

$$\Theta = \left(\frac{1}{2} + \delta \frac{1}{2R}, \frac{1}{2} - \delta \frac{1}{2R} \right), \quad \Theta \in S_1 = \{(x_1, x_2) \in [0, 1]^2 : x_1 + x_2 = 1\}.$$

We can think of (R, Θ) as the *polar* representation of a 2D heavy-tailed vector $X = R\Theta$.

1.2 Instructions

1) **Proposition recap.** Recall Proposition 3.3.2 (item (5)):

Proposition 1. *Let $X = (X_1, \dots, X_d)$ be a random vector with distribution F and marginal distributions F_j , $j \leq d$. Let Y be a random vector with simple max-stable cdf $G^*(x) = \exp\{-\mu([0, x]^c)\}$ on $E = [0, \infty)^d \setminus \{0\}$, and angular measure $\Phi(B) = \mu\{tw : t \geq 1, w \in B\}$ for measurable $B \subset S_{d-1}$. The following conditions are equivalent:*

1. $\frac{1}{n} \bigvee_{i \leq n} X_i \xrightarrow{w} Y$.
2. $F_1^n(nx_1) \longrightarrow e^{-1/x_1}$, $x_1 > 0$, and letting $R = \|X\|$ and $W = \|X\|^{-1}X$, for $B \subset S_{d-1}$ and $r \geq 1$,

$$\mathbb{P}(W \in B, R > ur \mid R > u) \longrightarrow \frac{1}{r} \frac{\Phi(B)}{\Phi(S_{d-1})} \quad (u \rightarrow \infty).$$

Questions.

- (a) Show that Θ is on the simplex: $\Theta_1 + \Theta_2 = 1$ and $\Theta_i \in [0, 1]$ for $R \geq 1$.
- (b) What happens to the law of Θ conditionally on $R \rightarrow \infty$?
- (c) Show that the model respects one of the items of proposition 1.

2) **Sampling & exploration.**

- (a) Simulate i.i.d. samples of (R, Θ) .

- (b) Plot a histogram of R .
- (c) Choose thresholds corresponding to quantiles $q \in \{0, 0.5, 0.9, 0.99\}$ of R (i.e. u_q with $\mathbb{P}(R \leq u_q) = q$). For each threshold u_q , plot a histogram of Θ_1 using only observations with $R > u_q$. Describe what you see as q increases.
- 3) **Testing asymptotic independence of angle and radius.** The idea is to find a way to choose the "threshold" for which we reached the "asymptotic regime". We recall in section 2 the χ^2 independence test for categorical variables. For a growing sequence of thresholds u , keep only observations with $R > u$. Discretize Θ_1 into B bins and R (above u) into K bins, form the $K \times B$ contingency table, and apply the χ^2 test. For calculation of the test statistics and of its p-value, you can use `scipy.stats.chisquare` and checkout out the documentation at [chi-square documentation](#). Make a graph of the p -values versus the threshold quantile q .

2 Reminder χ^2 test of independance

2.1 What does the χ^2 test ask?

Are two *categorical* variables associated (not independent)? You need a contingency table of *counts*.

2.2 How it works (idea)

Assume independence (H_0). Compute expected counts

$$E_{ij} = \frac{(\text{row } i \text{ total})(\text{col } j \text{ total})}{\text{grand total}},$$

then sum discrepancies

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

with $\text{df} = (r - 1)(c - 1)$. A small p (e.g., $< .05$) suggests association.

2.3 Mini 2×2 example with full calculations

Sample size $n = 200$. Variables: Smoking status (*Smoker/Non-smoker*) and Disease (*Yes/No*).

Observed counts (O)			
	Disease: Yes	Disease: No	Row total
Smoker	40	60	100
Non-smoker	20	80	100
Column total	60	140	200

Expected counts under H_0 (E)			
	Disease: Yes	Disease: No	Row total
Smoker	30	70	100
Non-smoker	30	70	100
Column total	60	140	200

Step 1: compute E_{ij} precisely. Using $E_{ij} = \frac{(\text{row } i)(\text{col } j)}{n}$:

$$E_{\text{Smoker, Yes}} = \frac{100 \cdot 60}{200} = 30, \quad E_{\text{Smoker, No}} = \frac{100 \cdot 140}{200} = 70,$$

$$E_{\text{Non, Yes}} = \frac{100 \cdot 60}{200} = 30, \quad E_{\text{Non, No}} = \frac{100 \cdot 140}{200} = 70.$$

Step 2: compute each cell's contribution $\frac{(O-E)^2}{E}$.

$$\frac{(40 - 30)^2}{30} = \frac{100}{30} = 3.\bar{3}, \quad \frac{(60 - 70)^2}{70} = \frac{100}{70} = 1.4286,$$

$$\frac{(20 - 30)^2}{30} = \frac{100}{30} = 3.\bar{3}, \quad \frac{(80 - 70)^2}{70} = \frac{100}{70} = 1.4286.$$

Step 3: sum to get χ^2 and identify df.

$$\chi^2 = 3.3333 + 1.4286 + 3.3333 + 1.4286 \approx 9.5238, \quad \text{df} = (2 - 1)(2 - 1) = 1.$$

Step 4: decision rule (two equivalent approaches).

- *p-value approach:* $p = \Pr\{\chi^2_{(1)} \geq 9.5238\} \approx 0.002$. For a common $\alpha = 0.05$ (or even $\alpha = 0.01$), $p < \alpha \Rightarrow$ **reject H_0** .
- *Critical-value approach:* For $\text{df} = 1$, the 95% upper critical value is $\chi^2_{0.95,1} = 3.841$ (and $\chi^2_{0.99,1} = 6.635$). Since $9.524 > 3.841$ (and > 6.635), **reject H_0** .

Interpretation & effect size. There is evidence of association between smoking and disease in this sample. Report effect size with Cramér's V :

$$V = \sqrt{\frac{\chi^2}{n(k-1)}} = \sqrt{\frac{9.5238}{200 \cdot 1}} \approx 0.218 \quad (\text{small-to-moderate}).$$

Cells driving the result can be seen via standardized residuals $R_{ij} = (O_{ij} - E_{ij})/\sqrt{E_{ij}}$ (here, **Smoker, Yes** and **Non-smoker, Yes** have the largest $|R_{ij}|$).