Multivariate generalized Pareto distributions along extreme directions

A.Kiriliouk A.Mourahib J.Segers

UCLouvain, ISBA. FNRS.

Graphical models and Clustering, 15-17 May 2024, Montpellier.

Background and Motivation

Estimating small probabilities

- Observations of water level in Carnon beach between 1998-2023

Estimating small probabilities

- Observations of water level in Carnon beach between 1998-2023
- Question: how to estimate the probability of water level X exceeding 4 meters in a period of 1000 years

Estimating small probabilities

- Observations of water level in Carnon beach between 1998-2023
- Question: how to estimate the probability of water level X exceeding 4 meters in a period of 1000 years

- **Difficulty.**
  - But we only have less than 100 years data

$\rightarrow$ Estimate beyond data, i.e., an event that was never observed

$$\mathsf{P}(X > 4) = \frac{\#(X > 4)}{N} = 0$$

To infinity and beyond

Mission

To model rare events, beyond what we have observed so far.

To infinity and beyond

Mission

To model rare events, beyond what we have observed so far.

Guiding principle

To make as little assumptions as possible.

Table of Contents

Table of Contents

Central limit theorem for maxima

Framework

- We divide data by years
- We denote the observations of water level during a year t between 1998 and 2023 as $X_{1,t}, \ldots, X_{n,t}$ where these observations follow the distribution F

Central limit theorem for maxima

Framework

- We divide data by years
- We denote the observations of water level during a year t between 1998 and 2023 as $X_{1,t}, \ldots, X_{n,t}$ where these observations follow the distribution F
- Consider $M_{n,t} = \max(X_{1,t}, \ldots, X_{n,t})$ the maximum of water-level at year t

Central limit theorem for maxima

Framework

- We divide data by years
- We denote the observations of water level during a year t between 1998 and 2023 as $X_{1,t}, \ldots, X_{n,t}$ where these observations follow the distribution F
- Consider $M_{n,t} = \max(X_{1,t}, \ldots, X_{n,t})$ the maximum of water-level at year t

Assumption

There exists two sequences $a_n > 0,\ b_n$ and a non-degenerate distribution G such that

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty).$$

We say that F is in the domain of attraction of G.

Central limit theorem for maxima

Framework

- We divide data by years
- We denote the observations of water level during a year t between 1998 and 2023 as $X_{1,t}, \dots, X_{n,t}$ where these observations follow the distribution F
- Consider $M_{n,t} = \max(X_{1,t}, \dots, X_{n,t})$ the maximum of water-level at year t

Assumption

There exists two sequences $a_n > 0,\ b_n$ and a non-degenerate distribution G such that

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty).$$

We say that F is in the domain of attraction of G.

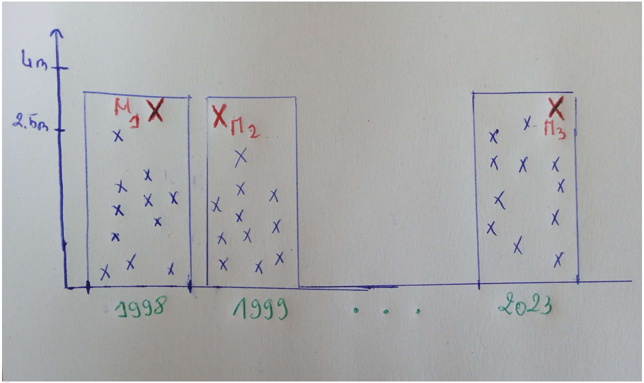
Intuition

We can approximate the distribution of the maximum-level water at year t if we have "enough" observations on that year

Caricature: Block maxima approach

Recall the assumption

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty).$$



In red, maximum of water level at each year between 1998 and 2023.

Extreme value distributions

Question: which question can we have about our assumption

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty)\,? \tag{1}$$

Extreme value distributions

Question: which question can we have about our assumption

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty)\,? \tag{1}$$

- Question : which distribution function G can arise from Equation (1)?

Extreme value distributions

Question: which question can we have about our assumption

$$\frac{M_{n,t} - b_n}{a_n} \xrightarrow{\mathrm{d}} G, \qquad (n \to \infty)\,? \tag{1}$$

- Question : which distribution function G can arise from Equation (1)?
  $\rightarrow$ Answer: G must be an extreme value distribution parametrized by some shape parameter $\gamma \in \mathbb{R}$, location parameter $\mu \in \mathbb{R}$, and scale parameter $\alpha > 0$

Table of Contents

Peaks over threshold

Framework

- This time, we do not divide observations by years
- We use only pick observations over a threshold b_N
- Recall $X_1, \ldots, X_N$ observations of the water level in Carnon beach
- Recall that we want to suppose as little assumptions as possible

Peaks over threshold

Framework

- This time, we do not divide observations by years
- We use only pick observations over a threshold b_N
- Recall $X_1, \dots, X_N$ observations of the water level in Carnon beach
- Recall that we want to suppose as little assumptions as possible

Assumption

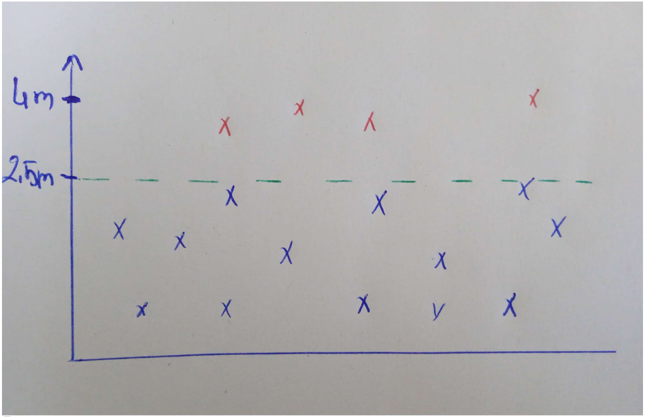
There exist two sequences $a_n > 0$ and b_n and a non-degenerate function H such that

$$\frac{X - b_N}{a_N} \mid X > b_N \xrightarrow{\mathrm{d}} H, \qquad (N \to \infty).$$

Caricature: Threshold exeedances approach

Recall the assumption

$$\frac{X - b_N}{a_N} \mid X > b_N \xrightarrow{\mathrm{d}} H, \qquad (N \to \infty).$$



In red, observations over the threshold which is fixed to 2.5 meters in this example

Generalized Pareto distributions

Question: which questions can we have about our assumption

$$\frac{X - b_N}{a_N} \mid X > b_N \xrightarrow{\mathrm{d}} H, \qquad (N \to \infty)\,? \tag{2}$$

Generalized Pareto distributions

Question: which questions can we have about our assumption

$$\frac{X - b_N}{a_N} \mid X > b_N \xrightarrow{\mathrm{d}} H, \qquad (N \to \infty)\,? \tag{2}$$

- Question: which distribution function H can arise from Equation (2) ?

Generalized Pareto distributions

Question: which questions can we have about our assumption

$$\frac{X - b_N}{a_N} \mid X > b_N \xrightarrow{\mathrm{d}} H, \qquad (N \to \infty)\,? \tag{2}$$

- Question: which distribution function H can arise from Equation (2) ?
  $\to$ Answer: H must be a generalized Pareto distribution parametrized by some shape parameter $\gamma \in \mathbb{R}$ and scale parameter $\alpha > 0$

Table of Contents

Table of Contents

Central limit theorem for multivariate maxima

Frmework

- During a year t between 1998 and 2023, we observe pairs $(X_{i,t}, Y_{i,t}),\ i = 1, \ldots, n$ of water level at two different locations: Carnon beach and Espiguette beach

Espiguette beach

Carnon beach

Central limit theorem for multivariate maxima

Frmework

- During year t, we observe pairs $(X_{i,t}, Y_{i,t}),\ i = 1, \ldots, n$ of water level at two different locations: Carnon beach and Espiguette beach

Central limit theorem for multivariate maxima

Frmework

- During year t, we observe pairs $(X_{i,t}, Y_{i,t}),\ i = 1, \dots, n$ of water level at two different locations: Carnon beach and Espiguette beach
- Maximum water level during year t in Carnon beach and Espiguette beach

$$M_{n,t}^{X} = \max(X_{1,t}, \dots, X_{n,t})$$
$$M_{n,t}^{Y} = \max(Y_{1,t}, \dots, Y_{n,t})$$

Central limit theorem for multivariate maxima

Frmework

- During year t, we observe pairs $(X_{i,t}, Y_{i,t}),\ i = 1, \ldots, n$ of water level at two different locations: Carnon beach and Espiguette beach
- Maximum water level during year t in Carnon beach and Espiguette beach

$$M_{n,t}^X = \max(X_{1,t}, \ldots, X_{n,t})$$
$$M_{n,t}^Y = \max(Y_{1,t}, \ldots, Y_{n,t})$$

Assumption

There exist sequences $a_n > 0$, b_n, $c_n > 0$ and d_n and a non-degenerate bivariate distribution G such that

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty$$

Central limit theorem for multivariate maxima

Frmework

- During year t, we observe pairs $(X_{i,t}, Y_{i,t}),\ i = 1, \ldots, n$ of water level at two different locations: Carnon beach and Espiguette beach
- Maximum water level during year t in Carnon beach and Espiguette beach

$$M_{n,t}^X = \max(X_{1,t}, \ldots, X_{n,t})$$
$$M_{n,t}^Y = \max(Y_{1,t}, \ldots, Y_{n,t})$$

Assumption

There exist sequences $a_n > 0$, b_n, $c_n > 0$ and d_n and a non-degenerate bivariate distribution G such that

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty$$

Let F denote the distribution of (X, Y), then F in the domain of attraction of G.

Multivariate extreme value distributions

Question: which questions can we have about our assumption

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty\,? \tag{3}$$

Multivariate extreme value distributions

Question: which questions can we have about our assumption

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\text{d}} G, \qquad n \to \infty\,? \tag{3}$$

- Question 1 : which distribution function G can arise from Equation (3)?

Multivariate extreme value distributions

Question: which questions can we have about our assumption

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n}\right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty\,? \tag{3}$$

- Question 1 : which distribution function G can arise from Equation (3)?
  $\to$ Answer: G must be a multivariate extreme value distribution

Multivariate extreme value distributions

Question: which questions can we have about our assumption

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty\,? \tag{3}$$

- Question 1 : which distribution function G can arise from Equation (3)?
  $\rightarrow$ Answer: G must be a multivariate extreme value distribution
- Question 2 : What are the parameters of G?

Multivariate extreme value distributions

Question: which questions can we have about our assumption

$$\left(\frac{M_{n,t}^X - b_n}{a_n}, \frac{M_{n,t}^Y - d_n}{c_n} \right) \xrightarrow{\mathrm{d}} G, \qquad n \to \infty\,? \tag{3}$$

- Question 1 : which distribution function G can arise from Equation (3)?
  $\rightarrow$ Answer: G must be a multivariate extreme value distribution
- Question 2 : What are the parameters of G?
- G is too big to be parameterized by a finite dimensional space.

Multivariate extreme value distributions: exponent measure

Challenge

How to decribe the extremal dependence structure between the maximum water level at Carnon beach and Espiguette beach?

Exponent measure

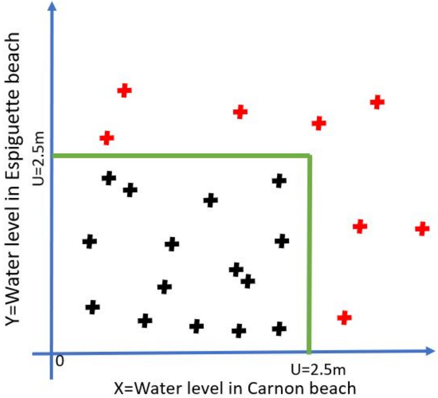
For each bivariate extreme value distribution G we can associate an exponent measure μ on $[0, \infty)^2 \setminus \{\mathbf{0}\}$.

Table of Contents

Multivariate peaks over threshold

Frmework

- Between 1998 and 2023, we observe pairs $(X_i, Y_i),\ i = 1, \dots, N$ of water level at two different locations: Carnon beach and Espiguette beach



Assumption

There exist sequences $a_N > 0$, b_N, $c_N > 0$ and d_N and a bivariate non-degenerate distribution H such that

$$\left(\frac{X - b_N}{a_N}, \frac{Y - d_N}{c_N} \mid X > b_N \text{ or } Y > d_N\right) \xrightarrow{\mathrm{d}} H, \qquad N \to \infty. \tag{4}$$

Assumption

There exist sequences $a_N > 0$, b_N, $c_N > 0$ and d_N and a bivariate non-degenerate distribution H such that

$$\left(\frac{X - b_N}{a_N}, \frac{Y - d_N}{c_N} \mid X > b_N \text{ or } Y > d_N\right) \xrightarrow{\mathrm{d}} H, \qquad N \to \infty. \tag{4}$$

- Question 1 : Which distribution function H can arise from (4)?

Assumption

There exist sequences $a_N > 0$, b_N, $c_N > 0$ and d_N and a bivariate non-degenerate distribution H such that

$$\left(\frac{X - b_N}{a_N}, \frac{Y - d_N}{c_N} \mid X > b_N \text{ or } Y > d_N\right) \xrightarrow{\mathrm{d}} H, \qquad N \to \infty. \tag{4}$$

- Question 1 : Which distribution function H can arise from (4)?
  $\rightarrow$ Answer: H must be a multivariate generalized Pareto distribution.

Assumption

There exist sequences $a_N > 0$, b_N, $c_N > 0$ and d_N and a bivariate non-degenerate distribution H such that

$$\left(\frac{X - b_N}{a_N}, \frac{Y - d_N}{c_N} \mid X > b_N \text{ or } Y > d_N \right) \xrightarrow{\mathrm{d}} H, \qquad N \to \infty. \tag{4}$$

- Question 1 : Which distribution function H can arise from (4)?
  $\rightarrow$ Answer: H must be a multivariate generalized Pareto distribution.

- Question 2 : Any relation between the multivariate extreme value distribution G and the multivariate generalized Pareto distribution H?

Assumption

There exist sequences $a_N > 0$, b_N, $c_N > 0$ and d_N and a bivariate non-degenerate distribution H such that

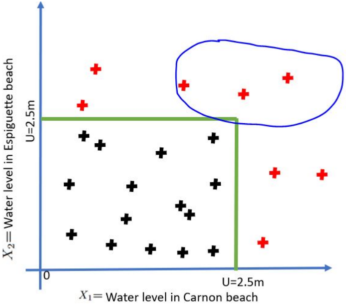
$$\left(\frac{X - b_N}{a_N}, \frac{Y - d_N}{c_N} \mid X > b_N \text{ or } Y > d_N \right) \xrightarrow{\mathrm{d}} H, \qquad N \to \infty. \tag{4}$$

- Question 1 : Which distribution function H can arise from (4)?
  $\rightarrow$ Answer: H must be a multivariate generalized Pareto distribution.

- Question 2 : Any relation between the multivariate extreme value distribution G and the multivariate generalized Pareto distribution H?
  $\rightarrow$ Answer: For any G, we can associate an H.

Multivariate generalized Pareto distributions: limitation on clusters

1. Some examples of multivariate generalized Pareto distributions studied in literature are: Hüsler–Reiss Pareto distributions[1], logistic Pareto distribution
2. All these models do not cover the case where high water levels occur in one station but not in other.



[1] See Engelke et al. (2015)

Key points

- Extremes are useful to estimate the probability of unusual events
- Two main approaches: Block maxima and Threshold exceedances
- Two main families: multivariate extreme value distributions and multivariate generalized Pareto distributions
- For each multivariate extreme value distribution, we can associate an exponent measure μ on $[0, \infty)^d \setminus \{\mathbf{0}_d\}$
- For each multivariate extreme value distribution, we can associate a multivariate generalized Pareto distribution
- Unfortunately, Threshold exceedances method does not cover the case where high water level occurs in one station but not in other

Flexible generalized Pareto distributions

Mission

Construct a model of multivariate Generalized Pereto distribution where high water levels occur in one station but not in others.

Mixture model

How to define a cluster Mathematically?

- Recall pairs $(X_i, Y_i),\ i = 1, \dots, N$ observations of water level at two different locations: Carnon beach and Espiguette beach
- Recall F the distribution of (X, Y)

How to define a cluster Mathematically?

- Recall pairs $(X_i, Y_i),\ i = 1, \ldots, N$ observations of water level at two different locations: Carnon beach and Espiguette beach
- Recall F the distribution of (X, Y)

Assumption

Suupose that F is in the domain of attraction of a bivariate extreme value distribution G; as we have seen before. ;)

How to define a cluster Mathematically?

- Recall pairs $(X_i, Y_i),\ i = 1, \ldots, N$ observations of water level at two different locations: Carnon beach and Espiguette beach
- Recall F the distribution of (X, Y)

Assumption

Suupose that F is in the domain of attraction of a bivariate extreme value distribution G; as we have seen before. ;)

- Question: how to interpret

  High water level in one station but not in the other

  in terms of the bivariate extreme value distribution G?

How to define a cluster Mathematically?

- Recall pairs $(X_i, Y_i),\ i = 1, \ldots, N$ observations of water level at two different locations: Carnon beach and Espiguette beach
- Recall F the distribution of (X, Y)

Assumption

Suupose that F is in the domain of attraction of a bivariate extreme value distribution G; as we have seen before. ;)

- Question: how to interpret

  High water level in one station but not in the other

  in terms of the bivariate extreme value distribution G?
  $\rightarrow$ Asnwer: using the exponent measure μ associated to G; as we have seen before

- Recall

$$\text{Water level } (X, Y) \overset{\text{Attracted by}}{\Rightarrow} G \overset{\text{associate}}{\Rightarrow} \mu$$

Table of Contents

Extreme directions

Definition (Goix et al. (2016))

A non-empty set $J \subset \{1, \ldots, d\}$ is an extreme direction of $\boldsymbol{X}$ if

$$\mu\left(\{\boldsymbol{x} \geqslant \mathbf{0} : x_j > 0 \text{ iff } j \in J\}\right) > 0.$$

- $J = \{1\}$. High water level only in Carnon
- $J = \{1, 2\}$. High water level in both Carnon and Espiguette
- $J = \{2\}$. High water level only in Espiguette

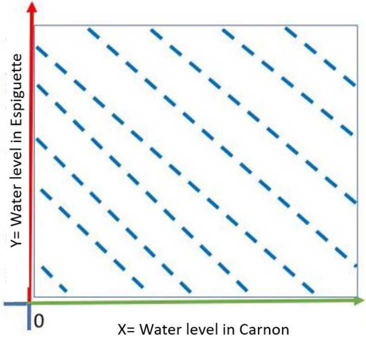


Table of Contents

Mixture model (Mourahib et al., 2023)

- Matrix $A = (a_{jk})_{j=1,\dots,d;k=1,\dots,r} \in [0,1]^{d\times r}$ s.t $\sum_{k=1}^{r} a_{jk} = 1$, $j \in \{1,\dots,d\}$
- Independent max-stable d-variate column-random vectors with unit Fréchet margins and single extreme direction $\{1,\dots,d\}$

$$\boldsymbol{Z}^{(1)} = \begin{pmatrix} Z_1^{(1)} \\ Z_2^{(1)} \\ \vdots \\ Z_d^{(1)} \end{pmatrix} \in \mathbb{R}^d, \qquad \boldsymbol{Z}^{(2)} = \begin{pmatrix} Z_1^{(2)} \\ Z_2^{(2)} \\ \vdots \\ Z_d^{(2)} \end{pmatrix} \in \mathbb{R}^d, \qquad \dots \qquad \boldsymbol{Z}^{(r)} = \begin{pmatrix} Z_1^{(r)} \\ Z_2^{(r)} \\ \vdots \\ Z_d^{(r)} \end{pmatrix} \in \mathbb{R}^d$$

- Model:

Complete dependence

$$\begin{array}{l} \text{Carnon} \longrightarrow \\ \text{Espiguette} \longrightarrow \\ \\ \text{Palavas} \longrightarrow \end{array} \begin{cases} M_1 = \max\left\{a_{11} Z_1^{(1)}, a_{12} Z_1^{(2)}, \dots, a_{1r} Z_1^{(r)}\right\} \\ M_2 = \max\left\{a_{21} Z_2^{(1)}, a_{22} Z_2^{(2)}, \dots, a_{2r} Z_2^{(r)}\right\} \\ \vdots \\ M_d = \max\left\{a_{d1} Z_d^{(1)}, a_{d2} Z_d^{(2)}, \dots, a_{dr} Z_d^{(r)}\right\} \end{cases}$$

Complete dependence

Question: can we identify the extreme directions of our model

The key is in the zero entries

Large simultaneously

$$\begin{array}{l} \text{Carnon} \longrightarrow \\ \text{Espiguette} \longrightarrow \end{array} \begin{cases} M_1 = \max\{\frac{1}{2} Z_1^{(1)}, \frac{1}{2} Z_1^{(2)}, 0 Z_1^{(3)}\} \\ M_2 = \max\{0 Z_2^{(1)}, \frac{1}{2} Z_2^{(2)}, \frac{1}{2} Z_2^{(3)}\} \end{cases}$$

Large simultaneously

- $J = \{1\}$. High water level only in Carnon
- $J = \{1, 2\}$. High water level in both Carnon and Espiguette
- $J = \{2\}$. High water level only in Espiguette

Extreme directions of the mixture model

- **Signatures**. For each column k in $\{1, \ldots, r\}$, let J_k be the set of those j in $\{1, \ldots, d\}$ such that $a_{jk} > 0$, that is $J_k = \{j \in \{1, \ldots, d\} : \ a_{jk} > 0\}$
- Recall that a non-empty set $J \subset \{1, \ldots, d\}$ is an extreme direction if

$$\mu\left(\{\boldsymbol{x} \geqslant \mathbf{0} : x_j > 0 \text{ iff } j \in J\}\right) > 0$$

Proposition

Extreme directions of the mixture model $\boldsymbol{M}$ *are exactly the signatures* $J_k, \ k = 1, \ldots, r.$

Example:

Large simultaneously

$$\begin{array}{r} \text{Carnon} \longrightarrow \\ \text{Espiguette} \longrightarrow \end{array} \begin{cases} M_1 = \max\{\frac{1}{2} Z_1^{(1)}, \frac{1}{2} Z_1^{(2)}, 0 Z_1^{(3)}\} \\ M_2 = \max\{0 Z_2^{(1)}, \frac{1}{2} Z_2^{(2)}, \frac{1}{2} Z_2^{(3)}\} \end{cases}$$

Large simultaneously

- Extreme directions of $\boldsymbol{M}$ are $J_1 = \{1\}, \ J_2 = \{1, 2\}, \ J_3 = \{2\}$

Density of the multivariate generalized Pareto associated with mixture model: difficulty

- Existence of an extreme direction different from $\{1, \ldots, d\}$
  $\Rightarrow$ the multivariate generalized Pareto random vector $\boldsymbol{Y}$ associated with the mixture model does not admit a density w.r.t. the Lebesgue measure λ_d
  $\Rightarrow$ need for a new measure υ in $[-\infty, \infty)^d$ that dominates $\boldsymbol{Y}$

Conclusion

Done

- Construct a threshold exceedances model that covers the case of lower dimensional extreme directions
- Simulate the model

To do

- Identify zeroes on the matrix A, i.e., identify extreme directions of the mixture model
- Estimate the other non-zero components of the matrix A
- Construct an extremal graphical model with lower dimensional extreme directions: Extension of the Hüsler–Reiss graphical model (Hentschel et al., 2022) or the logistic graphical model

Engelke, S., A. Malinowski, Z. Kabluchko, and M. Schlather (2015). Estimation of hüsler–reiss distributions and brown–resnick processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology 77*(1), 239–265.

Goix, N., A. Sabourin, and S. Clémençon (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pp. 75–83. PMLR.

Hentschel, M., S. Engelke, and J. Segers (2022). Statistical inference for h\" usler-reiss graphical models through matrix completions. *arXiv preprint arXiv:2210.14292*.

Mourahib, A., A. Kiriliouk, and J. Segers (2023). Multivariate generalized Pareto distributions along extreme directions. In preparation.