Bayesian and other approaches for extremes in one, many or infinitely many dimensions

Anne Sabourin

CNRS-LTCI, Télécom ParisTech

December 17^{th}, AppliBUGS, ENGREF, Paris

Outline

Univariate extremes

- Maxima, Excesses above threshold, Point process
- Inference

Multivariate extremes

- What are multivariate extremes
- Dirichlet mixture model

Spatial extremes

Towards large (finite) dimension

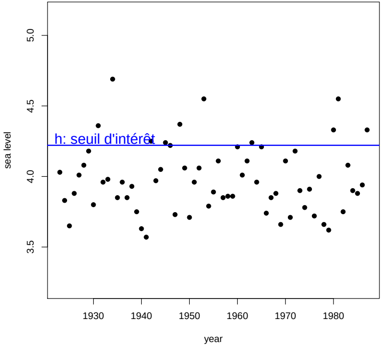
- Issues
- Anomaly detection
- Sparse support and estimation
- Estimation
- Results

Extremes and risks

FIGURE : Tempête Xynthia, La Faute-Sur-Mer, 1er Mars 2010.

Extremes and risks

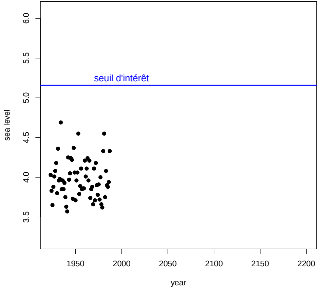
Quantity of interest : X (water level, temperature, insurance claims, ...)
$\longrightarrow$ time series $X_t, t \geq 0$.



- Given a high threshold h, find $p = \mathbb{P}(X \geq h)$
- Given p (e.g. $p = 10^{-4}$), find h such that $\mathbb{P}(X > h) \leq p$.
- Given a long duration T (e.g. 10^4), find $P(\max_{t \leq T} X_t \leq h)$.

Beyond the range of data

For $h \gg \max(X_{\text{obs}})$, or $T \gg T_{\text{obs}}$, or $p \ll 1/N_{\text{obs}}$ too small :



$$\text{Empirical estimator} \qquad \hat{P}(X > h) = \sum_{i=1}^{N_{\text{obs}}} \mathbb{I}_{X_i > h} = 0 \quad !!$$

Need an extrapolation model

Three complementary approaches to understand extremes

1. Block maxima

2. Excesses above a high threshold

3. Point process above a high threshold

The three approaches are equivalent in theory

Extreme value analysis

Theory : Under minimal assumptions, distributions of maxima/excesses converge to a certain class.

Modelling : Use those limits to model maxima/excesses above large thresholds.

$\mathbf{X}$: random object (variable / vector/ process) $\quad \mathbf{X}_i \overset{i.i.d.}{\sim} \mathbf{X}$.

$$\bigvee_{i=1}^{n} \mathbf{X}_i \quad \overset{d}{\approx} \quad \text{Max-stable} \qquad (n \text{ large})$$

$$\left[\, \mathbf{X} \;\middle|\; \|\mathbf{X}\| \geq r \,\right] \quad \overset{d}{\approx} \quad \text{Generalized Pareto} \qquad (r \text{ large})$$

$$\sum_{i=1}^{n} \delta_{\left(\frac{i}{n}, \frac{\mathbf{X}_i}{n}\right)} \quad \overset{d}{\approx} \quad \text{Poisson point process}$$

Dependence issues (see part 2)

- Sationary time series, not time-indepedent $\longrightarrow$ time declustering (separate clusters and keep the largest observation in each one)
- Non-stationary time series $\longrightarrow$ difficult to identify
- Spatial dependence / dependence between features (temperature, precipitation, wind , . . .) :
  - Max-stable models $\longrightarrow$ allows space extrapolation, with parametric assumptions on the dependence structure. Long range independence difficult to handle.
  - Multivariate extremes models (not necessarily spatial) $\longrightarrow$ learn the dependence structure of extremes

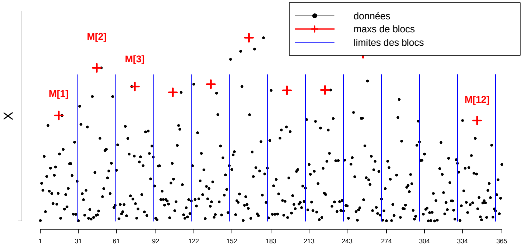
Block Maxima

- Maximum of a "block" of size n :

$$M_n = \max_{t=1,\ldots,n} X_t \overset{\text{notation}}{=} \bigvee_{1}^{n} X_t \,.$$

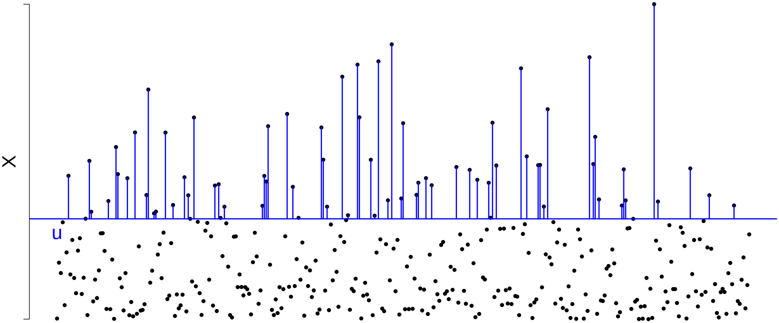
e.g. : monthly maximum of concentration for an air pollutant.

- Dividing the dataset into m blocks $\hookrightarrow$ m maxima $(M_n[1], \ldots, M_n[m])$; $M_n[i] = \bigvee_{t\in \text{ bloc } i} X_t$

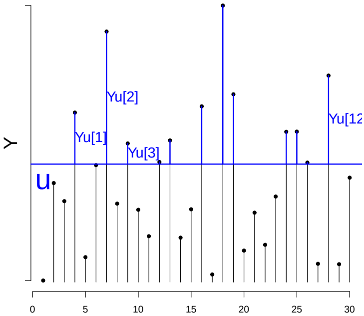


- $n * m$ data points (m blocks of size n) $\hookrightarrow$ only m maxima !

Peaks-Over-Threshold



Peaks-Over-Threshold

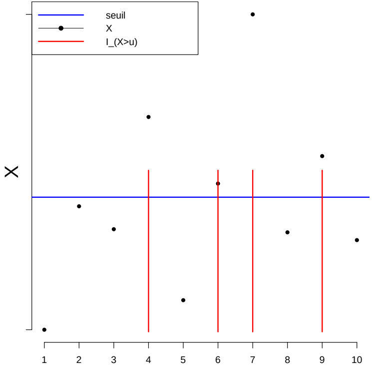


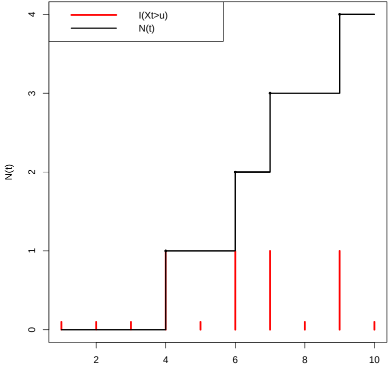
- *Excess* : $Y = X - u$, for $X > u$.
- *Conditional survival function*

$$\bar{F}_u(y) = P(X - u > y | X > u) = \frac{\bar{F}(u+y)}{\bar{F}(u)}$$

Point process (counting process) above threshold

$$N_u\big([t_1, t_2] \times [u, \infty)\big) = \sum_{t=t_1}^{t_2} \mathbb{I}_{\{X_t > u\}}$$

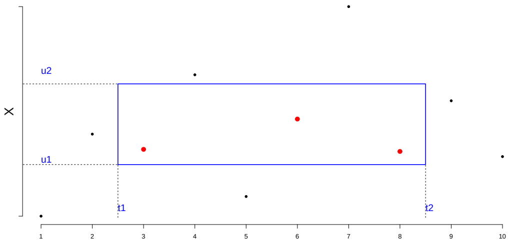




N_u counts the points above u

Bi-variate counting process

$$N([t_1, t_2] \times [u_1, u_2]) = \sum_{t=1}^{n} \mathbb{I}_{(t,X_t)}([t_1, t_2] \times [u_1, u_2])$$



$\ll$ Number of points in rectangle $[t_1, t_2] \times [u_1, u_2]$ $\gg$

- N : random measure, integer-valued, finite on compacts.
  $N = \{N(A),\ A \subset \mathbb{R}^2\}$.

Extreme values theorem

Theorem (Fisher et Tipett, 1928 ; Gnedenko 1943)

$(X_t)_{t\geq 0}$ *i.i.d random variables,* $M_n = \max_{t\leq n} X_t$*. If there exists sequences* $(a_n)_n > 0$, $(b_n)_n \in \mathbb{R}$*, and a non-degenerate r.v.* Y*, s.t.*

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} Y,$$

then, Y *is a « Generalized Extreme Value Distribution » (GEV), i.e.*

$$\forall x \in \mathbb{R}, \quad \mathbb{P}(Y \leq x) := G_{\mu,\sigma,\xi}(x) = e^{-[(1+\xi \frac{x-\mu}{\sigma})_+]^{-1/\xi}}$$

with $\xi \in \mathbb{R}$*,* $y_+ = \max(0, y)$*, and* $G_{\mu,\sigma,0}(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}$.

Maxima $\iff$ excesses $\iff$ point processes

$[x_*, x^*] = \mathrm{supp(G)}$. Let $\bar{H}(x) = -log(G(x))$.

Theorem

The following statements are equivalent :

- *(Maxima)* $F^n(a_n x + b_n) \underset{n\to\infty}{\longrightarrow} G(x) \quad (x_* < x < x^*)$
- *(Conditional law of excesses)* $\exists \sigma(t) > 0$, *s.t.*

$$\frac{\bar{F}(u + \sigma(u)x)}{\bar{F}(u)} \underset{u\to\infty}{\longrightarrow} \bar{H}(x) \qquad (x_* < x < x^*)$$

- *(Point process)*

$$\tilde{N}_n(\,\cdot\,) = \sum_{i=1}^{n} \delta_{\left(\frac{i}{n}, \frac{X_i - b_n}{a_n}\right)}(\,\cdot\,) \underset{n\to\infty}{\xrightarrow{d}} \tilde{N}$$

where $\tilde{N}$ *is a Poisson PP on* $(0,1) \times (x_*, x^*)$, *with intensity measure* $\tilde{\lambda}(t_1, t_2) \times (x, \infty) = (t_2 - t_1)\bar{H}(x)$

Inference methods, existing `R` packages

- Maximum likelihood, probability weighted moments
- `R` packages : `ismev`, `extRemes`, `evd`, `fExtremes`, `EVIM`, `Xtremes`, `HYFRAN`, `EXTREMES` , ...

  `http://cran.r-project.org/`
- Gilleland, Ribatet, Stephenson, 2013 : *A software review for extreme value analysis*
- Introductory book : Coles, 2001, An Introduction to Statistical Modeling of Extreme Values.

Assumption behind extreme values models

- For block maxima : for n large enough, $M_n \sim G_{\mu,\sigma,\xi}$.

- For Peaks-over-threshold : for u large enough, $[X - u | X > U] \sim GPD(u, \sigma, \xi)$

- Poisson process : for u, n large enough, $N = \sum_{i=1^n} \mathbb{I}_{\frac{i}{n}, \frac{X}{n}} \sim PP(\text{lebesgue} \otimes H_{u,\sigma,\xi})$

- goal : estimate μ, σ, ξ

- Bayesian inference : put a prior on μ, σ, ξ. Allows to take into account expert knowledge / historical information
  - Parent, Bernier, 2003, *Bayesian POT modeling for historical data*
  - Renard, 2011, *A Bayesian hierarchical approach to regional frequency analysis*
  - ...

...

example : POT model for univariate data

- GPD model above threshold :
  $\bar{F}_u(y|\xi,\sigma) := \mathbb{P}(X \geq u + y | X \geq u) \simeq \bar{H}_{\xi,\sigma}$
- data : excesses $(y_1, \ldots, y_{N_u})$ above $u \Rightarrow \left(\hat{\xi}, \hat{\sigma}\right)$?
- u moderate : enough data above. $\hat{\bar{F}}(u) = \frac{N_u}{n}$.
- $\bar{F}(u+y) \simeq \hat{\bar{F}}(u)\bar{F}_u(y)$

$$\mathcal{L}(\mathbf{y}, \xi, \sigma) \propto -\prod_{i=1}^{N_u} \frac{\mathrm{d}}{\mathrm{d}y} \bar{F}_u(y_i|\xi,\sigma)$$

MLE estimators :

$$\hat{\xi}, \hat{\sigma} \in \operatorname*{argmax}_{\sigma,\xi} \mathcal{L}(\mathbf{y}, \xi, \sigma)$$

Or Bayesian estimation $\rightarrow$ posterior sample.

Outline

Univariate extremes
- Maxima, Excesses above threshold, Point process
- Inference

Multivariate extremes
- What are multivariate extremes
- Dirichlet mixture model

Spatial extremes

Towards large (finite) dimension
- Issues
- Anomaly detection
- Sparse support and estimation
- Estimation
- Results

Why multivariate extremes ?

- Spatial or multivariate data : what is the dependence between features/ locations at extreme levels ?

- conditional probabilities of an excess : $\mathbf{X} = (X_1, X_2)$;

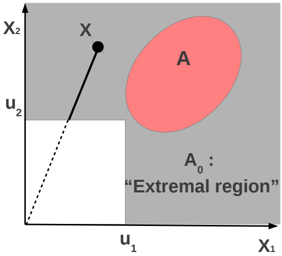
$$\mathbb{P}(X_2 > y | X_1 > x)? \quad (x \text{ large })$$

- probability of a joint excess :

$$\mathbb{P}(X_1 > x, X_2 > x)? \quad (x \text{ large })$$

Multivariate extremes

- Random vectors $\mathbf{Y} = (Y_1, \ldots, Y_{d,})$; $\quad Y_j \geq 0$
- Margins : $Y_j \sim F_j$, $1 \leq j \leq d$ (continuous).
- **Preliminary step : Standardization** $X_j = \frac{1}{1-F_j(Y_j))}$, $\mathbb{P}(X_j > v) = \frac{1}{v}$.
- Goal : $\mathbb{P}(\mathbf{X} \in A)$, A 'far from 0' ?



Fundamental assumption and consequences de Haan, Resnick, 70's, 80's

Intuitively : $\mathbb{P}(\mathbf{X} \in tA) \simeq \frac{1}{t}\mathbb{P}(\mathbf{X} \in A)$

Multivariate regular variation

$$0 \notin \bar{A} : \qquad t\,\mathbb{P}\left(\frac{\mathbf{X}}{t} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \qquad \mu : \text{ Exponent measure}$$

necessarily : $\mu(tA) = t^{-1}\mu(A)$ (Radial homogeneity)
$\rightarrow$ **angular measure** on the sphere : $\Phi(B) = \mu\{tB, t \geq 1\}$

General model for extremes

$$\mathbb{P}\left(\|\mathbf{X}\| \geq r \,; \quad \frac{\mathbf{X}}{\|\mathbf{X}\|} \in B \right) \simeq r^{-1}\,\Phi(B)$$

Φ is finite : $H := \frac{1}{\Phi(\mathbf{S}_d)}\Phi$ is a probability distribution : “angular distribution”.

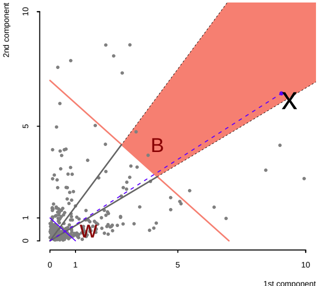
Polar decomposition and angular measure

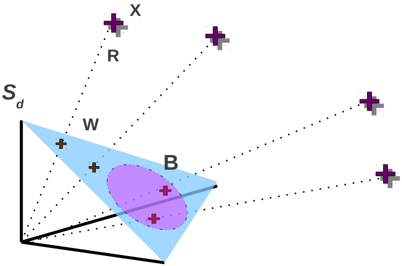
- Polar coordinates : $R = \sum_{j=1}^{d} X_j$ (L_1 norm) ; $\mathbf{W} = \frac{\mathbf{X}}{R}$.
- $\mathbf{W} \in$ simplex $\mathbf{S}_d = \{\mathbf{w} : \quad w_j \geq 0, \sum_j w_j = 1\}$.

Model above large radial threshold r_0 **Haan, Resnick, 77**

$$\mathbb{P}(R > r, \mathbf{W} \in B \mid R \geq r_0) \simeq \frac{r_0}{r} H(B)$$

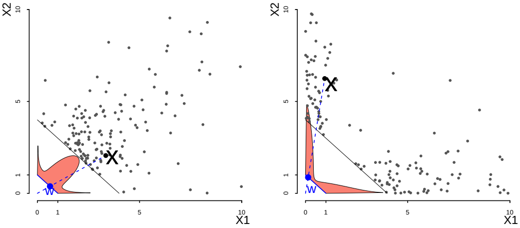
Angular measure H (+ margins)





Angular distribution

- H (+ margins) rules the joint distribution



- **Non parametric** family : Only one moment constraint on H,

$$\text{Center of mass} = \text{Center of the simplex}$$

- Statistician's goal : estimate H
  (if possible, together with margins)

Estimating the angular measure (assume margins known)

- **Non parametric estimation** :empirical likelihood, Einmahl *et al.*, 2001, Einmahl, Segers, 2009, Guillotte *et al*, 2011.

  Issues : asymptotic variance , Bayesian inference with $d > 2$, censored data

- Restriction to **parametric family** : Gumbel, logistic, pairwise Beta . . . Coles & Tawn, 91, Cooley *et al.*, 2010, Ballani & Schlather, 2011 : **Model uncertainty ?**

- Compromise : **Mixture** of countably many parametric models $\rightarrow$ Infinite-dimensional model

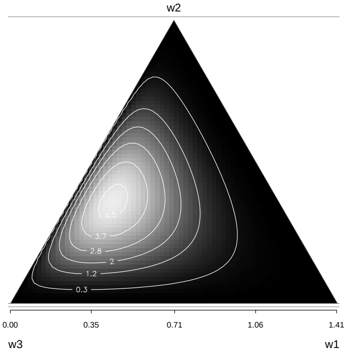
Dirichlet mixture model

(Boldi, Davison, 2007 ; S. , Naveau, 2013)

Dirichlet distribution ("multivariate Beta")

$$\forall \mathbf{w} \in \mathring{\mathbf{S}}_d,\ \mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}, \nu) = \frac{\Gamma(\nu)}{\prod_{i=1}^{d} \Gamma(\nu\mu_i)} \prod_{i=1}^{d} w_i^{\nu\mu_i - 1}\,.$$

- $\boldsymbol{\mu} \in \mathring{\mathbf{S}}_d$: location parameter (point on the simplex) : 'center' ;
- $\nu > 0$: concentration parameter.



Dirichlet mixture model

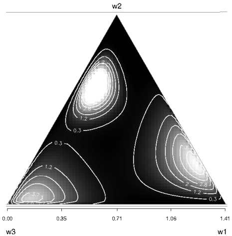
Boldi, Davison, 2007

- $\boldsymbol{\mu} = \boldsymbol{\mu}_{\cdot\,,1:k}$, $\boldsymbol{\nu} = \nu_{1:k}$, $\mathbf{p} = p_{1:k}$,

$$h(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{p}) = \sum_{m=1}^{k} p_m \,\mathrm{diri}(\mathbf{w} \mid \boldsymbol{\mu}_{\cdot\,,\,m}, \nu_m)$$

- Moments constraint $\rightarrow$ on $(\boldsymbol{\mu}, p)$:

$$\sum_{m=1}^{k} p_m \,\boldsymbol{\mu}_{.,m} = (\frac{1}{d}, \ldots, \frac{1}{d})\,.$$



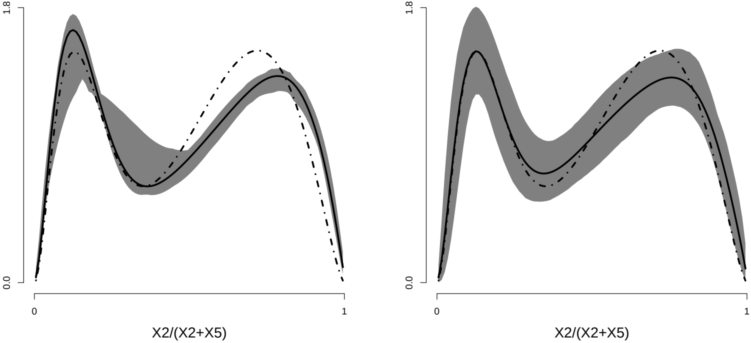
Weakly dense family ($k \in \mathbb{N}$) in the space of admissible angular measures

Bayesian inference

- Moments constraints (Boldi, Davison, 2007) $\rightarrow$ difficult to handle in a Bayesian setting in dimension $d > 2$
- Re-parametrization S. , Naveau (13) : work with unconstrained parameter in a product space
  - Weak posterior consistency
  - MCMC with reversible jumps manageable in moderate dimension ($\simeq 5$).
- Inference with censored data S. , 2015, JMVA

examples of results : mixing properties of the MCMC algorithm

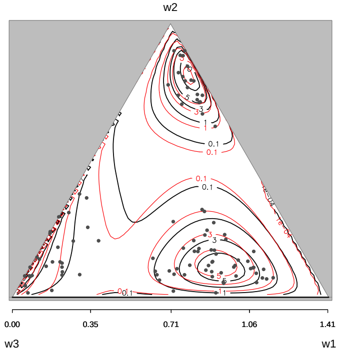
Simulated data, dimension 5,
showing 2D predictive angular measure)



original algo (Boldi, Davison, 07) reparametrized (S., Naveau, 2014)

predictive angular density

Simulated data, dimension 3,
showing true / predictive angular density level sets)



Connection with Point process/max-stable models

Similar to the 1-D case

$$t\mathbb{P}(\frac{X}{t} \in \cdot\,) \xrightarrow[n\to\infty]{} \mu(\,\cdot\,) \qquad \text{RV}$$

$$\Longleftrightarrow$$

$$\left[\left(\|X\|, \frac{X}{\|X\|}\right) \,\middle|\, \|X\| > r\right] \xrightarrow[n\to\infty]{d} d\frac{\mathrm{d}r}{r^2}\,\mathrm{d}H \qquad \text{(POT CV)}$$

$$\Longleftrightarrow$$

$$N_n(\,\cdot\,) := \sum_1^n \delta_{\frac{i}{n}, \frac{X_i}{n}} \xrightarrow[n\to\infty]{d} PP(d\frac{\mathrm{d}r}{r^2}\,\mathrm{d}H(w)) \qquad \text{(PP CV)}$$

$$\Longleftrightarrow$$

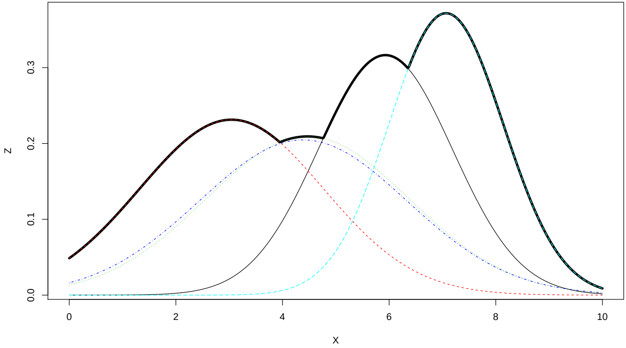
$$\frac{\bigvee_{i=1}^n X_i}{n} \xrightarrow[n\to\infty]{d} G \qquad \text{max CV}$$

where

$$G(x_1, \ldots, x_n) = \exp\left(-d\int_{\mathbf{S}_d} \bigvee_{j=1}^{d} \frac{w_i}{x_i}\,\mathrm{d}H(w)\right) \quad \text{max-stable distribution}$$

Outline

Pointwise maxima of continuous processes



possible limits (in distribution) of pointwise paxima

- A continuous stochastic process Z on a domain $\mathbb{D}$ is *max-stable* if $\exists$ continuous normalizing functions $(\alpha_n) > 0$ and β_n s.t.

$$Z \overset{d}{=} \bigvee_{i=1}^{n} \frac{Z_i - \beta_n}{\alpha_n}.$$

- N.B : equality in distribution : determined by finite-dimensional distributions $[Z_{s_1}, \ldots, Z_{s_n}]$.
- If (X_i) are i.i.d. continuous processes and $\exists a_n > 0$, b_n, s.t.

$$\bigvee_{i=1}^{n} \frac{X_i - b_n}{a_n} \xrightarrow[d]{n \to \infty} Z$$

then Z is a max-stable process.

de Haan, 84, A spectral representation for max-stable processes.

- Idea for statistical modeling : same as before : use a max-stable family (or its equivalent for peaks-over-thresholds) as a model for maxima/ excesses.

Spectral representation

- Standardization to unit fréchet margins (probability integral transform) $\to \mathbb{P}(Z(s) \leq x) = e^{-1/x}$, $s \in \mathbb{D}$. Then Z is a *simple max-stable* process.

Theorem (de Haan, 84, Penrose, 92)

Any non degenerate continuous, simple max-stable process $\{Z(s) : in\mathbb{D}\}$ defined on a compact set $\mathbb{D} \subset \mathbb{R}^d$, satisfies

$$Z(x) \stackrel{d}{=} \vee_{i \geq 1} \zeta_i f_i(s)$$

where $\{\zeta_i, f_i, i \geq 1\}$ points of a Poisson process on $(0, \infty) \times \mathcal{C}$ with intensity $\zeta^{-2}\,\mathrm{d}\zeta\,\mathrm{d}\nu(f)$, for some locally finite measure ν on the space $\mathcal{C}$ of continuous, ≥ 0 functions on $\mathbb{D}$ such that $\int_{\mathcal{C}} f(s)\,\mathrm{d}\nu(f) = 1$, $s \in \mathbb{D}$.

Intuition : $f =$ infinite-dim generalisation of an "angle".

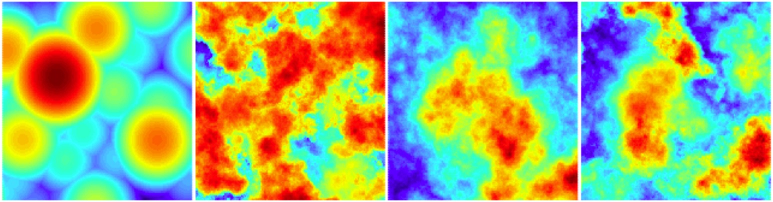
- f_i : rainstorm profile
- ζ_i : rainstorm intensity

Models for spatial extremes Ribatet, Dombry, Oesting,

Spatial extremes and max-stable processes (To appear)

model for spectral functions $f_i \to$ model for Z

- f_i : random gaussian density (mean = center of the storm, variance : inverse width : *Smith model* (Smith, 90)
- $f_i(s) = \max(0, W_i(s))$, W_i : stationary Gaussian process : Schlather model (Schlather, 2002)
- More flexible (and difficult to simulate until recently (Kabluchko et. al. , 2009) : Brown-Resnick process, Extremal-t process (Opitz, 2013)



Smith, Schlather, Brown-Resnick, Extremal-t

Inference for max-stable processes

- c.d.f. for a finite number of location : $F_{s_1,\dots,s_d}(x_1,\dots,x_d) = e^{-V(x_1,\dots,x_d)} \rightarrow$ likelihood expression is a d^{th} derivative, huge number of terms !
  $\rightarrow$ use
  - composite likelihood, Padoan *et.al*, 2010, Likelihood-based inference for max-stable processes, can be Bayesian (Cooley, Davison, Ribatet, 11)
  - concurrent extremes : conditioning on the underlying spectral function (hitting scenario), Dombry, Eyi-Minko 2013
- Implementation of standard methods in package `spatialExtremes` (Ribatet, 15)
- Bayesian hierarchical model : Reich, Shaby, 2013.
- Peaks-over-Threshold framework (Thibaud *et.al.*, 2013)

Outline

If no spatial structure but large number of features

- Example : collection of air pollutants, network data (features of the connection requests), blood toxines, ...

Issues in large (≥ 10) dimension for sandard multivariate models :

- MCMC convergence would take ages.

- Implicit assumption in many model : "All variables must be concomitantly large", (or some pre-specified susbets, as in the logistic model) :
  not reasonable in a spatial context (localized storms, affecting only some subsets of variables) :

End of this talk

joint work with **Nicolas Goix** and Stephan Clémençon

Dimension reduction in multivariate extremes

Exhibit sparsity ?

Anomaly detection in 'extreme' data

'Extremes' = points located in the tail of the distribution.

What does 'normal' mean among extremes ?

End of this talk joint work with **Nicolas Goix** and Stephan Clémençon

Dimension reduction in multivariate extremes

Exhibit sparsity ?

Error ('uncertainty') assessment

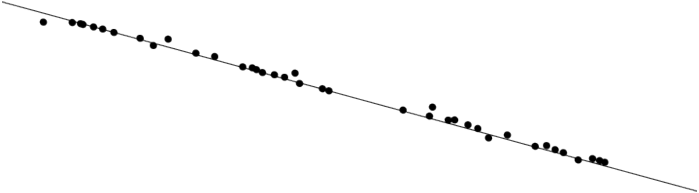
Finite sample Bounds on the error ? (not Bayesian ...)

Anomaly detection in 'extreme' data

'Extremes' = points located in the tail of the distribution.

What does 'normal' mean among extremes ?

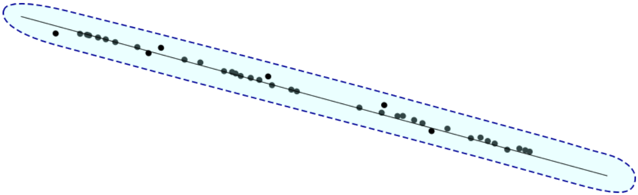
What is Anomaly Detection (AD) ?



- **Training step 1 : Learn a profile** characterizing 'normal' behavior, *e.g.* approximate support.

Applications : Public health, network intrusions, finance, surveillance

What is Anomaly Detection (AD) ?

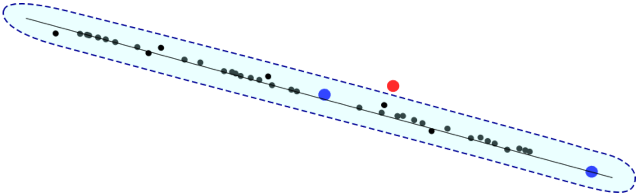


- **Training step 1 : Learn a profile** characterizing 'normal' behavior, *e.g.* approximate support.

- **Training step 2 :** Build a decision function
  $\rightarrow$ **'normal' region** around the profile.

Applications : Public health, network intrusions, finance, surveillance

What is Anomaly Detection (AD) ?



- **Training step 1 : Learn a profile** characterizing 'normal' behavior, *e.g.* approximate support.

- **Training step 2 :** Build a decision function
  $\rightarrow$ **'normal' region** around the profile.

- **Step 3 :** with new data :
  Anomalies = points outside the 'normal region'

Applications : Public health, network intrusions, finance, surveillance

Multivariate EVT for Anomaly detection

- If 'normal' data are heavy tailed, there may be **extreme** normal data.

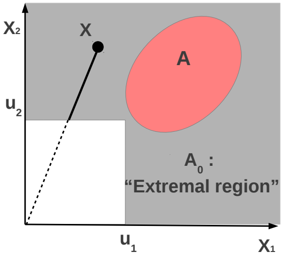
How to distinguish between large anomalies and normal extremes ?

- Yet : no multivariate AD algorithm has a specific treatment for multivariate extreme data

- **Our goal** (from an AD point of view) : Improve performance of standard AD algorithms on extremal regions using MEVT.
  $\rightarrow$ reduce # false positives

Recall Multivariate extremes

- Random vectors $\mathbf{Y} = (Y_1, \dots, Y_{d,})$; $\quad Y_j \geq 0$
- Margins : $Y_j \sim F_j$, $1 \leq j \leq d$ (continuous).
- **Preliminary step : Standardization** $X_j = \frac{1}{1-F_j(Y_j))}$, $\mathbb{P}(X_j > x) = \frac{1}{x}$.
- Goal : $\mathbb{P}(\mathbf{X} \in A)$, A 'far from 0' ?



Fundamental assumption and consequences

de Haan, Resnick, 70's, 80's

Intuitively : $\mathbb{P}(\mathbf{X} \in tA) \simeq \frac{1}{t}\mathbb{P}(\mathbf{X} \in A)$

Multivariate regular variation

$$0 \notin \bar{A} : \qquad t\,\mathbb{P}\left(\frac{\mathbf{X}}{t} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \qquad \mu : \text{ Exponent measure}$$

necessarily : $\mu(tA) = t^{-1}\mu(A)$ (Radial homogeneity)

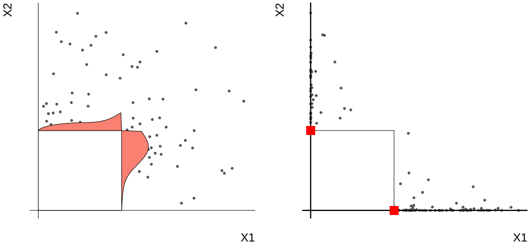
$\to$ **angular measure** on the sphere : $\Phi(B) = \mu\{tB, t \geq 1\}$

General model for extremes

$$\mathbb{P}\left(\|\mathbf{X}\| \geq r \; ; \quad \frac{\mathbf{X}}{\|\mathbf{X}\|} \in B \right) \simeq r^{-1}\,\Phi(B)$$

Angular measure

- Φ rules the joint distribution of extremes



- Asymptotic dependence : (X_1, X_2) may be large together.

  vs

- Asymptotic independence : only X_1 *or* X_2 may be large.

No assumption on Φ : non-parametric framework.

Multivariate extremes in large dimension ?

- Flexible multivariate models for **moderate dimension** ($d \simeq 5$)

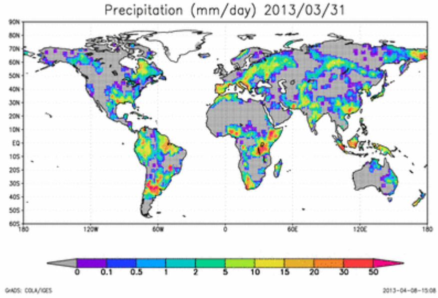
  Dirichlet Mixtures (Boldi,Davison 07 ; S., Naveau 12), Logistic family (Stephenson 09, Fougères *et.al*, 13), Pairwise Beta (Cooley *et.al*) ...

- Theory for angular measure (dependence) estimation : **asymptotic, $\mathbf{d = 2}$**, rates under **second order conditions**

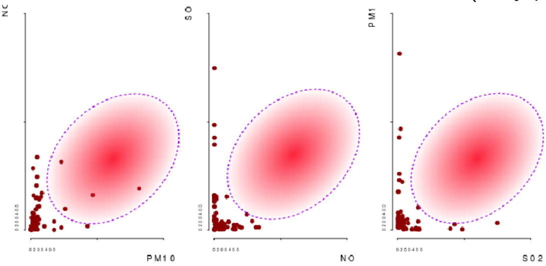
  (Einmahl, 01) Empirical likelihood (Einmahl, Segers 09)

- **High dimension ?** ($d \gg 1$) :
  - Spatial $\to$ max-stable models (parametric)
  - **Non spatial $\to$? ?**
    (multiple air pollutants, assets, features for AD ...)
  - Theory for integrated versions (tail dependence function)
    Asymptotic normality (Einmahl *et. al.*, 12, 15) (parametric case),
    Finite sample bounds (Goix *et. al*, 15)
    $\not\to$ structure of extremes (which components may be large together)

It cannot rain everywhere at the same time



(daily precipitation)



(air pollutants)

Towards high dimension

- Reasonable hope : only a moderate number of X_j's may be simultaneously large $\rightarrow$ **sparse angular measure**

- **Our goal** from a MEVT point of view :

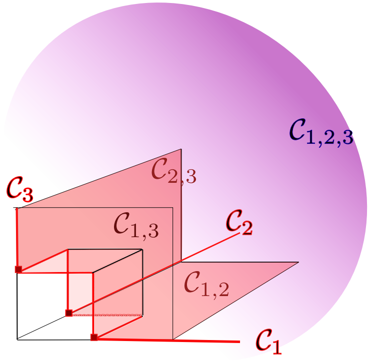
Estimate the (sparse) support of the angular measure
(*i.e.* the dependence structure).

Which components may be large together, while the other are small ?

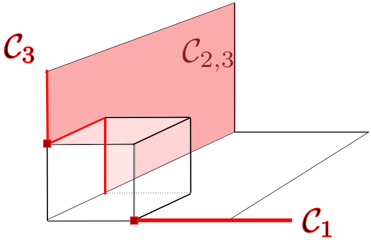
- For MEVT modeling : recover the asymptotically dependent groups of components $\rightarrow$ use simplified model.

- for AD : support $=$ normal profile
$\rightarrow$ anomalies $=$ points 'far away' from the support.

Sparse angular support



Full support :
anything may happen

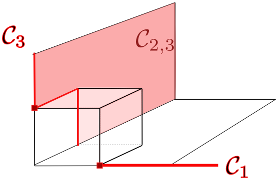


Sparse support
(X_1 not large if X_2 or X_3 large)

Where is the mass ?

Subcones of $\mathbb{R}_+^d$: $\boldsymbol{\mathcal{C}_\alpha} = \{x \succeq 0, x_i \geq 0 \ (i \in \alpha), \ x_j = 0 \ (j \notin \alpha), \|x\| \geq 1\}$
$\alpha \subset \{1, \ldots, d\}$.

Support recovery + representation



- $\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}$: partition of the unit sphere
- $\{\mathcal{C}_\alpha, \alpha \subset \{1, \ldots, d\}$: corresponding partition of $\{x : \|x\| \geq 1\}$
- μ-mass of subcone $\mathcal{C}_\alpha$: $\mathcal{M}(\alpha)$ (unknown)
- **Goal :** learn the $2^d - 1$-dimensional representation (potentially sparse)

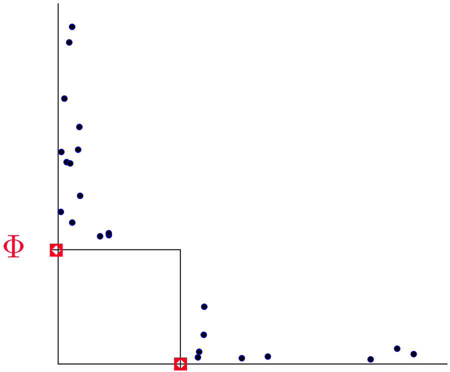
$$\mathcal{M} = \Big(\mathcal{M}(\alpha)\Big)_{\alpha \subset \{1,\ldots,d\}, \alpha \neq \emptyset}$$

- $\mathcal{M}(\alpha) > 0 \iff$
  features $j \in \alpha$ may be large together while the others are small.

Identifying non empty edges

Issue : real data = non-asymptotic : $X_j > 0$.

Cannot just count data on each edge :
Only the largest-dimensional sphere has empirical mass !



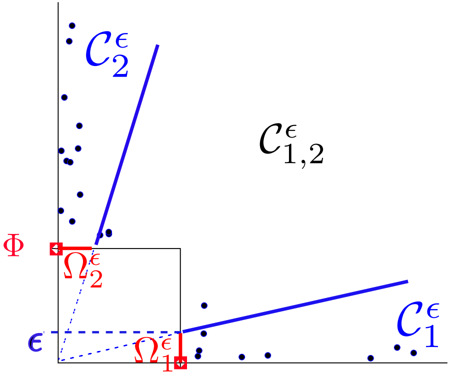
Identifying non empty edges

Fix $\varepsilon > 0$. Affect data ε-close to an edge, to that edge.

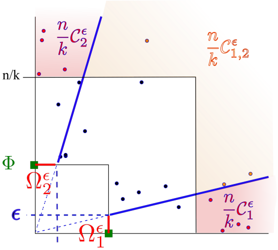
$$\Omega_\alpha \to \Omega_\alpha^\varepsilon = \{w : \quad w_i > \varepsilon \,(i \in \alpha), \quad w_j < \varepsilon \; (j \notin \alpha)\}.$$

$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\varepsilon = \{t\, \Omega_\alpha^\varepsilon, t \geq 1\}.$$

$\to$ New partition of the input space, compatible with non asymptotic data.



Empirical estimator : Counts the standardized points in $\mathcal{C}_\alpha^\varepsilon$, far from 0.



Algorithm

$$\text{data : } \mathbf{Y}_i, i = 1, \ldots, n, \ \mathbf{Y}_i = (X_{i,1}, \ldots, Y_{i,d}).$$

- Standardize : $\hat{X}_i = \frac{1}{1-\hat{F}_j(Y_{i,j})}$, with $\hat{F}_j(Y_{i,j}) = \frac{rank(Y_{i,j})-1}{n}$
- Natural estimator

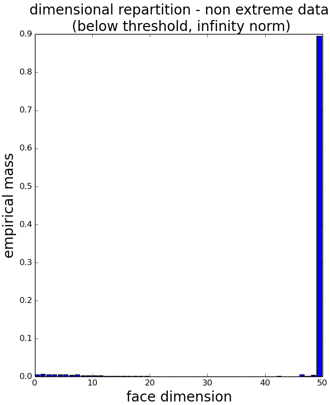
$$\hat{\Phi}_n(\Omega_\alpha) = \mu_n(\mathcal{C}_\alpha^\varepsilon) = \frac{n}{k}\mathbb{P}_n(\hat{\mathbf{X}} \in \frac{n}{k}\mathcal{C}_\alpha^\varepsilon).$$

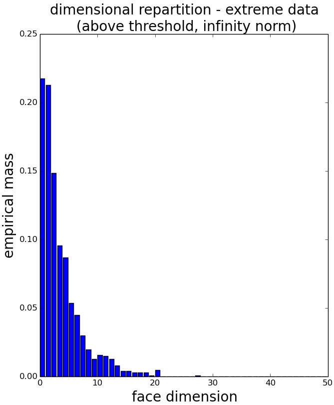
$$\longrightarrow \hat{\mathcal{M}} = (\hat{\Phi}_n(\Omega_\alpha), \alpha \subset \{1, \ldots, d\})$$

Sparsity in real datasets

Data=50 wave direction from buoys in North sea.

(Shell Research, thanks J. Wadsworth)





	Non-extreme data	Extreme Data
nb of faces with positive mass	2761	782
nb of faces with positive mass after thresholding	21	76
nb of faces with positive mass after 2nd thresholding	1	26

Finite sample error bound

VC-bound adapted to low probability regions (see Goix et. al. , 2015)

Theorem

If the margins F_j are continuous and if the density of the angular measure is bounded by $M > 0$ on each subface,
There is a constant C s.t. for any n, d, k, $\delta \leq e^{-k}$, $\varepsilon \leq 1/4$,
with probability $\geq 1 - \delta$,

$$\max_{\alpha} |\hat{\Phi}_n(\Omega_\alpha) - \Phi(\Omega_\alpha)| \leq Cd \left(\sqrt{\frac{1}{k\varepsilon} \log \frac{d}{\delta}} + Md\varepsilon \right) + \mathrm{Bias}_{\frac{n}{k},\varepsilon}(F, \mu).$$

Bias : using non asymptotic data to learn about an asymptotic quantity

$$\text{Regular variation} \iff \mathrm{Bias}_{t,\varepsilon} \xrightarrow[t \to \infty]{} 0$$

- Existing litterature ($\mathbf{d = 2}$) : $1/\sqrt{k}$.
- **Here** : $1/\sqrt{k\varepsilon} + Md\varepsilon$. Price to pay for biasing estimator with ε.
  OK if $\varepsilon\, k \to \infty$, $\varepsilon \to 0$.
  Choice of ε : cross-validation or '$\varepsilon = 0.1$'

Tools for the proof

1. VC inequality for small probability classes (Goix *et.al.*, 2015)

$$\rightarrow \text{max deviations} \leq \sqrt{p} \times \text{(usual bound)}$$

2. Apply it on VC-class of rectangles $\{\frac{k}{n} R(x, z, \alpha),\ x, z \succ \varepsilon\}$

$$\rightarrow p \leq d \frac{k}{\varepsilon n}$$

3. Approach $\mathcal{C}_\alpha^\varepsilon$ with such rectangles $\rightarrow$ error $\leq d\sqrt{\varepsilon}$

4. Approach $\mu(\mathcal{C}_\alpha)$ with $\mu(\mathcal{C}_\alpha^\varepsilon)$ $\rightarrow$ error $\leq d\varepsilon$ (bounded angular density).

Results : support recovery

- Asymmetric logistic, $d = 10$, dependence parameter $\alpha = 0.1$
  $\rightarrow$ Non asymptotic data (not exactly Generalized Pareto)
- K randomly chosen (asymptotically) non-empty faces.
- parameters : $k = \sqrt{n}$, $\epsilon = 0.1$
- Additional (heuristic) step : eliminate faces supporting less than 1% of total mass.

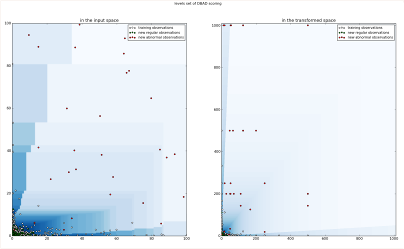
# sub-cones K	10	15	20	30	35	40	45	50
Aver. # errors (n=5e4)	0.01	0.09	0.39	1.82	3.59	6.59	8.06	11.21
Aver. # errors (n=15e4)	0.06	0.02	0.14	0.98	1.85	3.14	5.23	7.87

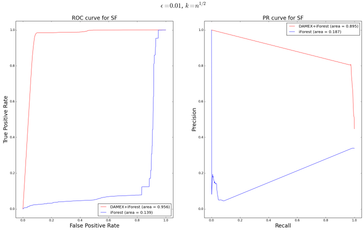
Algorithm DAMEX (Detecting Anomalies with Multivariate Extremes)

Anomaly = new observation ‘violating the sparsity pattern’ :
observed in empty or light subcone.

Scoring function : for x such that $\hat{v} \in \mathcal{C}_\alpha^\varepsilon$,

$$s_n(x) = \frac{1}{\|\hat{v}\|} \hat{\phi}_n(\Omega_\alpha^\epsilon) \quad \simeq_{x \text{ large}} \quad \mathbb{P}\left(X \in \mathcal{C}_\alpha, \|X\| > x\right)$$





Conclusion

- Adequate notion of **'sparsity' for MEVT** : sparse **angular measure**
- **Empirical estimation** ($\rightarrow$ algorithm) to learn this sparse asymptotic support **from non-asymptotic, non sparse data**.
- **Finite sample error bounds** (tools from statistical learning theory)
- **Applications :**
  - Immediate application to AD
  - View towards multivariate extreme (or spatial ?) modeling :

use sparsity information to build a simplified model, need to do clustering ?
(ongoing work, Maël Chiapino)

- **Question :** can we detect sparsity in a Bayesian framework ? ideas welcome. . .

Some references

- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection : A survey, 2009
- E. Chautru. Dimension reduction in multivariate extreme value analysis, 2015
- J. H. J. Einmahl , J. Segers. Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution, 2009.
- J. H. J. Einmahl, A. Krajina, J. Segers. An M-estimator for tail dependence in arbitrary dimensions, 2012.
- J.H.J Einmahl, A. Kiriliouk, A. Krajina, J. Segers. An M-estimator of spatial tail dependence, 2015
- N. Goix, A. Sabourin, S. Clémençon. Learning the dependence structure of rare events : a non-asymptotic study, 2015
- FT Liu, Kai Ming Ting, Zhi-Hua Zhou. Isolation forest, 2008
- Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics, 1997
- S.J. Roberts. Novelty detection using extreme value statistics, 1999

Some references

- M. O. Boldi, A. Davison. A mixture model for multivariate extremes.
- J. H. J. Einmahl, A. Krajina, J. Segers. *An m-estimator for tail dependence in arbitrary dimensions*, 2012.
- N. Goix, A. Sabourin, S. Clémençon. *Learning the dependence structure of rare events : a non-asymptotic study*, to appear.
- S. Resnick. *Extreme Values, Regular Variation, Point Processes*, 1987
- A. Sabourin, P. Naveau. *Bayesian Dirichlet mixture model for multivariate extremes : A re-parametrization*, 2014.
- A. Sabourin. *Semi-parametric modeling of excesses above high multivariate thresholds with censored data*, 2015
- A. Sabourin, B. Renard, *Combining regional estimation and historical floods : a multivariate semi-parametric peaks-over-threshold model with censored data*, 2015
- J. Tawn. *Modelling multivariate extreme value distributions*, 1990.
- D. Van Dyk, X.Meng. *The art of data augmentation*, 2001