Discovering patterns in high-dimensional extremes

Maël Chiapino

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0035

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Maël Chiapino

le 28 juin 2018

Apprentissage de structures dans les valeurs extrêmes en grande dimension

Discovering Patterns in High-dimensional Extremes

Directeur de thèse : **François Roueff**
Co-encadrement de la thèse : **Anne Sabourin**

Jury

Mathilde Mougeot, Professeur, ENSIIE, Université Paris Diderot	Rapporteur
Patrice Bertail, Professeur, MODAL'X, Université Paris Ouest (Nanterre)	Rapporteur
Anne Sabourin, Maitre de conférence, LTCI, Télécom ParisTech	Co-directeur
François Roueff, Professeur, LTCI, Télécom ParisTech	Directeur
Maud Thomas, Maitre de conférence, LPSM, Université Pierre et Marie Curie	Examinateur
Jessica Tressou, Chargée de recherche, INRA, AgroParisTech	Examinateur
Florence d'Alché, Professeur, LTCI, Télécom ParisTech	Examinateur

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Contents

List of Figures

List of Tables

Apprentissage de structures dans les valeurs extrêmes en grande dimension

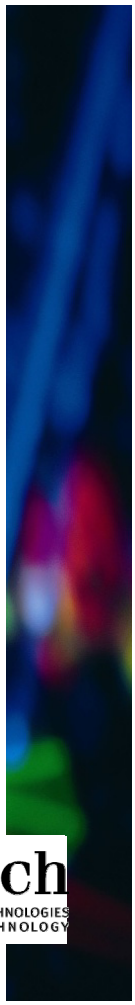
Maël Chiapino

RESUME : Nous présentons et étudions des méthodes d'apprentissage non-supervisé de phénomènes extrêmes multivariés en grande dimension. Dans le cas où chacune des distributions marginales d'un vecteur aléatoire est à queue lourde, l'étude de son comportement dans les régions extrêmes (i.e. loin de l'origine) ne peut plus se faire via les méthodes usuelles qui supposent une moyenne et une variance finies. La théorie des valeurs extrêmes offre alors un cadre adapté à cette étude, en donnant notamment une base théorique à la réduction de dimension à travers la mesure angulaire. La thèse s'articule autour de deux grandes étapes : - Réduire la dimension du problème en trouvant un résumé de la structure de dépendance dans les régions extrêmes. Cette étape vise en particulier à trouver les sous-groupes de composantes étant susceptible de dépasser un seuil élevé de façon simultané. - Modéliser la mesure angulaire par une densité de mélange qui suit une structure de dépendance déterminée à l'avance. Ces deux étapes permettent notamment de développer des méthodes de classification non-supervisée à travers la construction d'une matrice de similarité pour les points extrêmes.

MOTS-CLEFS : Théorie des valeurs extrêmes, apprentissage non-supervisé, réduction de dimension.

ABSTRACT : We present and study unsupervised learning methods of multivariate extreme phenomena in high-dimension. Considering a random vector on which each marginal is heavy-tailed, the study of its behavior in extreme regions is no longer possible via usual methods that involve finite means and variances. Multivariate extreme value theory provides an adapted framework to this study. In particular it gives theoretical basis to dimension reduction through the angular measure.

KEY-WORDS : Extreme value theory, unsupervised learning, dimension reduction, clustering.



1 Résumé

1.1 Introduction

La théorie des valeurs extrêmes répond à la nécessité de construire un modèle d'extrapolation de phénomènes hors des limites au sein desquelles ils sont habituellement observés. Étant donné un certain nombre d'observations, l'objectif est de prévoir dans quelle mesure les futures réalisations du phénomène sont susceptibles de franchir des seuils du même ordre de grandeurs que les valeurs maximales précédemment atteintes. Et plus particulièrement, de caractériser leur comportement au-delà de ces limites. Ces considérations ne prennent pleinement leur légitimité que lorsque la distribution des quantitées étudiées est à queue lourde, c'est à dire qu'elle ne concentre pas toute sa masse autour d'un point central et donc que la probabilité qu'une valeur soit extrême est non négligeable. Lorsque le processus considéré est univarié, la théorie des valeurs extrêmes peut se résumer à l'étude des valeurs maximales que prend ce processus. Le comportement asymptotique du maximum d'une telle variable aléatoire est bien connu et est entièrement décrit par une famille de distributions paramétriques, la loi de valeurs extrêmes généralisée (De Haan (1970)). Dans un contexte multivarié, contrairement à $\mathbb{R}$, l'espace ne possède plus de relation d'ordre et il n'est donc plus possible d'ordonner les points du processus de sorte à pouvoir en définir les points maximaux. Cependant, via les dépassements de seuils, le maximum composante par composante ou plus généralement la norme du vecteur aléatoire, il est possible de donner un sens au caractère extrémal de certains points.

Plusieurs difficultés propres aux extrêmes surviennent avec la dimensionnalité du problème. Les méthodes habituelles de réduction de la dimension, comme l'analyse en composantes principales, se basent sur l'étude de la matrice de covariance afin d'en extraire un modèle dont la structure est simplifiée. Dans le cas de distributions à queue lourde la covariance des composantes n'est plus définie, il n'est donc plus possible d'utiliser de tels outils. S'ajoute à cela un problème d'ordre statistique inhérent à l'objectif poursuivi d'extrapoler au-delà des limites déjà observées du phénomène. Plus précisémment, le caractère extrémal du processus que nous voulons estimer nous oblige à ne considérer, pour l'inférence statistique, qu'une partie des observations données, *i.e.* les plus extrêmes d'entre elles. Et cela ne peut représenter, par nature, qu'une faible proportion du jeu de données.

Ce résumé est articulé de la manière suivante. Tout d'abord dans la Section

1.2, le cadre général de la théorie des extrêmes est présenté dans le cas univarié 1.2.1 puis multivarié 1.2.2. En particulier y est introduite la *mesure angulaire* Φ, qui caractérise la structure de dépendance de toute loi d'extrêmes multivariés, après une certaine normalisation des distributions marginales. Une représentation parcimonieuse de cette mesure est décrite dans la Section 1.3. Sur ces bases, une methode générale pour trouver le support de Φ est détaillé dans la section 1.4. En particulier, l'heuristique CLEF (CLustering of Extreme Features) est présentée dans 1.4.2 suivi de versions alternatives de l'algorithme basées sur une série de tests statistiques 1.4.3. Une modélisation paramétrique de la mesure angulaire est enfin proposée dans la Section 1.5.

1

1.2 Théorie des valeurs extrêmes

1.2.1 Théorie des valeurs extrêmes univariées

Dans le cas univarié, la théorie des valeurs extrêmes caractérise le comportement asymptotique du maximum de variables aléatoires à valeurs dans $\mathbb{R}$. Soient $X_1, \ldots, X_n$, n copies indépendantes et identiquement distribuées ($i.i.d.$) d'une variable aléatoire $X \in \mathbb{R}$ de distribution F, telle que $0 < F(x) < 1$ pour tout $x \in \mathbb{R}$. Admettons que l'on se donne pour but d'estimer de larges quantiles situés dans la queue de distribution, *i.e.* $\mathbb{P}[X > x_p] = p$ avec x_p un seuil élevé. Dans le cas où aucune des données observées X_i ne se trouve dans la région $[x_p, \infty)$, la seule solution est de construire un modèle d'extrapolation. L'approche classique porte donc sur l'étude du maximum $M_n := \max\{X_1, \ldots, X_n\}$ et la problématique se traduit par la limite suivante, soit $x \in \mathbb{R}$:

$$\mathbb{P}[M_n \leq x] = F^n(x) \underset{n\to\infty}{\longrightarrow} 0, \tag{1.1}$$

de telle sorte que la distribution limite est nécessairement dégénérée. Il est clair que les méthodes usuelles qui portent sur le comportement 'moyen' de la variable d'étude ainsi que toute approche proprement empirique ne peuvent ici être appliquées. Afin de contourner (1.1) il est nécessaire d'appliquer une certaine normalisation au maximum M_n permettant l'émergence d'une limite non-dégénérée.

La propriété suivante, fondatrice de la théorie des extrêmes, est en quelque sorte l'analogue du théorème central limite (TCL) et de la limite de la somme normalisée de variables aléatoires indépendantes et identiquement distribuées. Bien que l'existence d'une telle limite non-dégénérée ne soit pas garantie, à supposer qu'elle existe, sa forme est quant à elle entièrement caractérisée. C'est une différence majeure avec le TCL qui stipule l'existence de la limite pour toute variable dont les deux premiers moments sont finis. Supposons l'existence de deux suites $(a_n)_{n\geq 1}$ et $(b_n)_{n\geq 1}$ avec $a_i > 0$ telles que:

$$\lim_{n\to\infty} \mathbb{P}\Big[\frac{M_n - b_n}{a_n} \leq x\Big] = G(x), \tag{1.2}$$

où G est une distribution non-dégénérée, alors cette distribution limite a nécessairement la forme suivante:

$$G(x) = exp\Big\{ - \Big[1 + \xi \frac{x-\mu}{\sigma}\Big]_+^{-\frac{1}{\xi}} \Big\}. \tag{1.3}$$

On dit que F appartient au domaine d'attraction de la loi de valeurs extrêmes G, noté $F \in DA(G)$. Lorsque $\xi = 0$, G est une loi extremum de type I (Gumbel), $\xi > 0$ correspond à une loi extremum de type II (Fréchet) et $\xi < 0$ une loi extremum de type III (Weibull).

Le comportement asymptotique du maximum de telles variables aléatoires étant entièrement caractérisé au sein d'une famille de distributions paramétriques, les problématiques liées à l'inférence de tels modèles concernent principalement la proportion de points extrémaux du jeu de données à prendre en compte.

1.2.2 Théorie des valeurs extrêmes multivariées

Dans un contexte multivarié, la définition d'un point extrême n'est plus aussi naturelle. En effet, dès lors que $d \geq 2$, $\mathbb{R}^d$ n'est plus un ensemble ordonné, il n'est plus possible de classer les points du plus au moins grand par une relation d'ordre. Néanmoins, considérant $\boldsymbol{x} \in \mathbb{R}^d$ il est approprié de définir un tel point 'extrême' lorsque $\max_{j\in\{1,\dots,d\}} x_j > t$ ou plus généralement $\|\boldsymbol{x}\| > t$ pour $t > 0$ un seuil élevé et $\|\cdot\|$ une norme quelconque. De façon analogue au cas univarié, la théorie des valeurs extrêmes multivariées se fonde sur la distribution limite du maximum composante par composante normalisé.

Soient $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, n variables aléatoires $i.i.d.$ dans $\mathbb{R}^d$ de distribution F. Les distributions marginales de F sont notées F_j pour $j = 1, \dots, d$. Soit $\boldsymbol{M}_n := (\max_{i=1,\dots,n} X_{i,1}, \dots, \max_{i=1,\dots,n} X_{i,d})$ le n-ème maximum composante par composante. Supposons l'existence de deux suites $(\boldsymbol{a}_n)_{n\geq 1}$ dans $(0,\infty)^d$ et $(\boldsymbol{b}_n)_{n\geq 1}$ dans $\mathbb{R}^d$, telles que:

$$\lim_{n\to\infty} \mathbb{P}\Big[\frac{\boldsymbol{M}_{n,j} - \boldsymbol{b}_{n,j}}{\boldsymbol{a}_{n,j}} \leq \boldsymbol{x}_j,\ j = 1, \dots, d\Big] = G_0(\boldsymbol{x}), \tag{1.4}$$

où G_0 est une distribution multivariée dont les distributions marginales $G_{0,j}$, $j = 1, \dots, d$ sont non-dégénérées. Alors G_0 est caractérisée par l'expression suivante (Resnick (1987)), pour un certain $\boldsymbol{x}_0 \in \mathbb{R}^d$:

$$G_0(\boldsymbol{x}) = \begin{cases} \exp\Big[-\mu_0\Big([\boldsymbol{x}_0, \boldsymbol{x}]^c\Big)\Big], & \text{pour } \boldsymbol{x} \geq \boldsymbol{x}_0, \\ 0 \text{ sinon}, \end{cases} \tag{1.5}$$

où μ_0 est la dite *mesure exposant*, une mesure de Radon définie sur $[\boldsymbol{x}_0, \boldsymbol{\infty}] \setminus \{\boldsymbol{x}_0\}$ et où $[\boldsymbol{x}_0, \boldsymbol{x}]^c = [\boldsymbol{x}_0, \boldsymbol{\infty}] \setminus [\boldsymbol{x}_0, \boldsymbol{x}]$. Chaque marginale $G_{0,j}$ est une loi de valeurs extrêmes univariée et donc pour tout $j \in \{1, \dots, d\}$, $F_j \in DA(G_{0,j})$. Cependant,

la mesure exposant μ_0, et donc G_0, ne peuvent être caractérisées par une famille de distributions paramétriques. C'est une différence majeure avec le cas univarié.

Soit $\boldsymbol{X} = (X_1, \ldots, X_d)$ une variable aléatoire à valeurs dans $\mathbb{R}^d$ de distribution F. L'étude de la queue de distribution de F se sépare en deux parties distinctes. D'une part, la caractérisation des distributions marginales, qui relève du cas univarié. D'autre part, l'étude de la structure de dépendance, qui constitue l'essentiel du problème. De façon similaire à l'étude des copules, une étape préliminaire naturelle est de standardiser chaque marginale de sorte à ce qu'elles soient de distributions identiques. Bien que le choix de la distribution soit arbitraire, il est commun de transformer chaque marginale en Pareto unitaire (Resnick (1987)); une standardisation que l'on obtient par le procédé suivant:

$$V_j := (1 - F_j(X_j))^{-1}, \text{ pour } j = 1, \ldots, d. \tag{1.6}$$

Remarque 1. Soit $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, n copies *i.i.d.* de $\boldsymbol{X}$. Dans la pratique, les distributions marginales F_j n'étant pas connues, la transformation (1.6) est effectuée par le biais des distributions empiriques $\hat{F}_j(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_{i,j}<x\}}$ où $x \in [x_{0,j}, \infty)$. Notons $rang(X_{i',j}) = n - \sum_{i=1}^n \mathbb{1}_{\{X_{i,j}<X_{i',j}\}}$. La transformation devient donc, pour tout $j \in \{1, \ldots, d\}$:

$$\begin{aligned}\hat{V}_{i,j} &= (1 - \hat{F}_j(X_{i,j}))^{-1}) \\ &= \frac{n}{rang(X_{i,j})}.\end{aligned}$$

Nous travaillons désormais avec le vecteur transformé $\boldsymbol{V} = (V_1, \ldots, V_d)$, à valeurs dans $(1, \infty)^d$, et de distributions marginales Pareto unitaire. L'existence de la distribution asymptotique G_0 dans (1.4) peut alors se reformuler de façon approprié à travers le concept de variation régulière. La variable aléatoire $\boldsymbol{V}$ est dite à variation régulière d'indice -1 sur $(1, \infty)^d$ et de mesure limite μ définie sur $[0, \infty]^d \setminus \{\boldsymbol{0}\}$ si (voir par ex. Resnick (2013)):

$$t\mathbb{P}\left(t^{-1}\boldsymbol{V} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \tag{1.7}$$

pour tout borélien A dans $[0, \infty]^d \setminus \{\boldsymbol{0}\}$ tel que $\mu(\partial A) = 0$ et $\boldsymbol{0} \notin \partial A$. Autrement dit, μ est la *mesure exposant* associée à la loi de valeurs extrêmes G de $\boldsymbol{V}$ de telle sorte que pour tout $\boldsymbol{v} \in [0, \infty)^d \setminus \{\boldsymbol{0}\}$:

$$G(\boldsymbol{v}) = \exp[-\mu([\boldsymbol{0}, \boldsymbol{v}]^c)]. \tag{1.8}$$

Le lien entre les distributions asymptotiques G_0 et G associées respectivement à $\boldsymbol{X}$ et $\boldsymbol{V}$ se traduit par:

$$G(\boldsymbol{v}) = G_0(G_{0,1}^{\leftarrow}(e^{-1/v_1}), \ldots, G_{0,d}^{\leftarrow}(e^{-1/v_d})),$$

où $G_{0,i}^{\leftarrow}$ est l'inverse généralisé de $G_{0,i}$.

Remarque 2. Soit $t > 0$ un seuil élevé, sachant que pour une coordonnée particulière $j_0 \in \{1, \dots, d\}$, V_{j_0} est extrême, *i.e.* $V_{j_0} > t$, la fonction $\boldsymbol{x} \mapsto \mu\Big([\mathbf{0}, \boldsymbol{x}]^c\Big)$ est approximativement, pour un $\boldsymbol{v} \in [0, \infty)^d \setminus \{\mathbf{0}\}$ donné, la probabilité conditionnelle suivante:

$$\begin{aligned}\mathbb{P}\Big[V_1 > tv_1 \text{ or } \dots \text{ or } V_d > tv_d | V_{j_0} > t\Big] &= \mathbb{P}\Big[\boldsymbol{V} \in t[\mathbf{0}, \boldsymbol{v}]^c | V_{j_0} > t\Big] \\ &= t\mathbb{P}\Big[\boldsymbol{V} \in t[\mathbf{0}, \boldsymbol{v}]^c\Big] \\ &\approx \mu\Big([\mathbf{0}, \boldsymbol{v}]^c\Big) = -\log G(\boldsymbol{v})\end{aligned}$$

L'intérêt de la transformation (1.6) ainsi que de l'hypothèse de variation régulière (1.7) est bien l'accent porté sur la structure de dépendance asymptotique caractérisée par la mesure μ. Il convient de rappeler qu'il n'existe pas de famille paramétrique englobant complètement la classe de telles mesures. Néanmoins, en conséquence de (1.6) et du caractère max-stable de toute loi de valeurs extrêmes généralisée (Resnick (1987)), la fonction μ est homogène de degré -1:

$$\mu(tA) = t^{-1}\mu(A), \tag{1.9}$$

pour tout $t > 0$ et pour tout borélien A tel que $\mathbf{0} \notin \partial A$. On peut déduire de (1.9) que les distributions marginales de G sont de loi Fréchet unitaire.

Remarque 3. Il suit de la propriété d'homogénéité (1.9) que les régions de $[0, \infty]^d \setminus \{\mathbf{0}\}$, minimales pour l'inclusion, sur lesquelles μ possède une masse non nulle sont nécessairement des cônes, *i.e.* $C \subset [0, \infty]^d \setminus \{\mathbf{0}\}$ tel que $x \in C \Rightarrow tx \in C$ pour tout $t > 0$.

Cette propriété est essentielle afin de mieux caractériser la classe des mesures exponents suite à la standardisation des marginales. En effet, (1.9) induit une décomposition pseudo-polaire de μ. Pour tout $\boldsymbol{v} \in [0, \infty)^d \setminus \{\mathbf{0}\}$, considérons l'application bijective:

$$T : \boldsymbol{v} \mapsto (\|\boldsymbol{v}\|, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}),$$

où $\|\cdot\|$ est une norme quelconque sur $\mathbb{R}^d$. Il est alors possible de définir une mesure Φ sur le quadrant positif de la sphère unité $\mathcal{S}_d = \{\boldsymbol{x} \geq \mathbf{0} : \|x\| = 1\}$, la dite *mesure angulaire*, caractérisant la structure de dépendance de $\boldsymbol{V}$:

$$\Phi(A) := \mu(\{\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in A\}), \tag{1.10}$$

pour tout $A \subset \mathcal{S}_d$. La propriété d'homogénéité de μ (1.9) se traduit alors par:

$$\begin{aligned}\mu(\{\boldsymbol{v} : T(\boldsymbol{v}) \in (t, \infty) \times A\}) &= \mu(\{\boldsymbol{v} : \|\boldsymbol{v}\| > t, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in A\}) \\ &= t^{-1}\Phi(A).\end{aligned}$$

Finalement, cette dernière équation ainsi que la bijectivité de T impliquent la factorisation de μ en deux mesures indépendantes (de Haan and Resnick (1977)):

$$\mu \circ T^{-1}(\mathrm{d}r, \mathrm{d}\boldsymbol{w}) = r^{-2}\mathrm{d}r\Phi(\mathrm{d}\boldsymbol{w}). \tag{1.11}$$

Remarque 4. Lorsque $\|\boldsymbol{V}\|$ est grand, la mesure angulaire Φ jauge la part relative que chacune des coordonnées V_j prend dans le caractère extrême de $\boldsymbol{V}$, autrement dit, la direction dans laquelle $\boldsymbol{V}$ sera susceptible d'être extrême. Posons $R = \|\boldsymbol{V}\|$ et $\boldsymbol{W} = R^{-1}\boldsymbol{V}$, une réécriture de la limite (1.7) en fonction de Φ est:

$$\mathbb{P}\left(\boldsymbol{W} \in A, R > rt \mid R > t\right) \xrightarrow[t\to\infty]{} r^{-1}\Phi(\mathcal{S}_d)^{-1}\Phi(A), \tag{1.12}$$

pour tout borélien A de $\mathcal{S}_d$ tel que $\Phi(\partial A) = 0$ et $r > 1$. En d'autres termes, lorsque la composante radiale R est grande, R et le pseudo-angle $\boldsymbol{W}$ sont approximativement indépendants. La distribution de $\boldsymbol{W}$ est alors approximativement (et à une normalisation près) la mesure angulaire et R suit approximativement une loi de Pareto unitaire. Une étude plus précise de ces caractéristiques est développée dans la section 1.3.

Tout comme pour la mesure exposant, il n'existe pas de famille paramétrique décrivant entièrement l'espace des mesures angulaires. La seule condition sur Φ, pour être une mesure angulaire valide, est la *contrainte des moments*:

$$\int_{\mathcal{S}_d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) = 1, \, \forall j \in \{1, \ldots, d\}. \tag{1.13}$$

C'est une conséquence de la transformation de chaque marginale en Pareto unitaire (1.6). En effet nous avons, pour tout $j \in \{1, \ldots, d\}$ et pour tout $x_j \in (0, \infty)$:

$$G(\infty, \ldots, x_j, \ldots, \infty) = e^{-\frac{1}{x_j}}.$$

De plus, en utilisant la décomposition (1.11), nous pouvons réécrire $\boldsymbol{x} \mapsto \mu([\boldsymbol{0}, \boldsymbol{x}]^c)$ en fonction de la mesure angulaire. Soit $\boldsymbol{x}$ dans $[0, \infty]^d \setminus \{\boldsymbol{0}\}$:

$$\begin{aligned}
\mu([\boldsymbol{0}, \boldsymbol{x}]^c) &= \int \mathbb{1}_{\{\exists j : u_j > x_j\}} \mu(\mathrm{d}u) \\
&= \int_{\mathcal{S}_d} \int_0^\infty \mathbb{1}_{\{r > \min_j \frac{x_j}{w_j}\}} r^{-2}\mathrm{d}r\Phi(\mathrm{w}) \\
&= \int_{\mathcal{S}_d} (\min_j \frac{x_j}{w_j})^{-1}\Phi(\mathrm{w}) \\
&= \int_{\mathcal{S}_d} \max_j \frac{w_j}{x_j}\Phi(\mathrm{w}).
\end{aligned}$$

Ainsi, pour $\boldsymbol{x} = (\infty, \ldots, x_j, \ldots, \infty)$, avec $x_j \in (0, \infty)$ nous avons:

$$\begin{aligned}
\frac{1}{x_j} &= -\log G(\boldsymbol{x}) \\
&= \mu([\boldsymbol{0}, \boldsymbol{x}]^c) \\
&= \int_{\mathcal{S}_d} \frac{w_j}{x_j} \, \Phi(\mathrm{d}\boldsymbol{w}).
\end{aligned}$$

La contrainte (1.13) est donc une condition nécessaire pour que Φ soit une mesure angulaire, et c'est aussi une condition suffisante (voir la proposition 5.11 dans Resnick (1987)). En conséquence directe de la contrainte des moments, dans le cas où $\|\cdot\|$ est la norme L_1, *i.e.* $\|\boldsymbol{x}\| = |x_1| + \ldots + |x_d|$, nous avons $\Phi(\mathcal{S}_d) = d$. En effet:

$$\begin{aligned}\Phi(\mathcal{S}_d) &= \int_{\mathcal{S}_d} \Phi(\mathrm{d}\boldsymbol{w}) \\ &= \int_{\mathcal{S}_d} \sum_{j=1}^{d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) \\ &= \sum_{j=1}^{d} \int_{\mathcal{S}_d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) \\ &= d.\end{aligned}$$

1

1.2.3 Travaux antérieurs et problèmes

D'un point de vue statistique, de nombreux modèles paramétriques ont été proposés pour la loi de valeurs extrêmes G à travers la modélisation de la fonction $\boldsymbol{x} \mapsto \mu[\mathbf{0}, \boldsymbol{x}]^c$, *i.e.* $G(\boldsymbol{x}) = \exp[-\mu[\mathbf{0}, \boldsymbol{x}]^c]$ (voir par ex. Coles and Tawn (1991)). Les deux stratégies principales pour l'inférence d'un tel modèle sont les méthodes de maximum composante par composante et des excès de seuils, l'étude porte alors respectivement sur l'estimation de G et sur l'estimation de $\boldsymbol{x} \mapsto \mu[\mathbf{0}, \boldsymbol{x}]^c$. Par le biais du maximum composante par composante, l'estimation bute sur le calcul de la vraisemblance lorsque la dimension augmente ($d > 10$). En effet, il est nécessaire de calculer:

$$\frac{\partial^d}{\partial x_1 \ldots \partial x_d} e^{-\mu[\mathbf{0}, \boldsymbol{x}]^c}$$

qui est une somme dont le nombre de termes explose en fonction de d, (voir la revue Huser et al. (2016)). Des simplifications ont été proposées (par ex. Wadsworth (2015), Stephenson and Tawn (2005)) mais le calcul reste infaisable lorsque $d > 10$.

Que ce soit pour le maximum composante par composante ou pour les excès de seuils, la plupart de ces approches ne considère que le cas de dépendance totale entre les variables, c'est-à-dire que le support de la mesure exposant μ est confiné à l'intérieur de $(0, \infty)^d$. L'estimation non-paramétrique n'échappe pas, quant à elle, au problème plus général du *fléau de la dimension*. Vient s'ajouter un problème inhérent aux extrêmes: le faible nombre de points disponibles pour effectuer l'estimation.

Estimation de la mesure angulaire Diverses approches paramétriques et non-paramétriques ont été proposées pour l'estimation de la mesure angulaire. (Boldi and Davison (2007b)) et (Sabourin et al. (2013)) considèrent la classe des mélanges de Dirichlet sur le simplexe. Cette classe de modèles est dense (au sens faible) dans l'espace des mesures angulaires. D'autres modèles non-paramétriques

ont été proposés (Guillotte et al. (2011), Fougeres et al. (2013)). Cependant les expérimentations ne vont pas au-delà de dimensions moyennes ($d \approx 5$). De récentes approches tentent de dépasser ce cap à travers des formes de *clustering* (Chautru (2015)) et un algorithme permettant de retrouver le support de la mesure angulaire en grande dimension ($d \approx 50$) est proposé dans (Goix et al. (2015a), Goix et al. (2016a)).

1.3 Mesure angulaire parcimonieuse

En grande dimension, lorsque $d \approx 100$, il est raisonnable de supposer qu'un phénomène extrême, décrit par une variable aléatoire $\boldsymbol{V} = (V_1, \ldots, V_d)$, ne soit pas dû à l'ensemble de ses composantes simultanément, *i.e.* $V_j > t$, $\forall j \in \{1, \ldots, d\}$ où $t > 0$ est un seuil élevé. Plus précisément, l'hypothèse fondamentale que nous faisons est qu'un tel phénomène ne prend son caractère extrémal que par le biais de certains sous-groupes précis de composantes. De telle sorte que lorsque $\|\boldsymbol{V}\|$ est grand, il existe un sous-groupe $\alpha \subset \{1, \ldots, d\}$ tel que $\|\boldsymbol{V}\| \approx \|\boldsymbol{V}_\alpha\|$, où $|\alpha| \ll d$ et $\boldsymbol{V}_\alpha = (V_j)_{j \in \alpha}$. Nous entendons mettre en avant le caractère multimodal du comportement extrême de $\boldsymbol{V}$ et donc la pluralité de tels α, qui identifient les sous-groupes de composantes dépendantes lors des différentes réalisations de $\|\boldsymbol{V}\|$ au delà d'un large seuil. Il est naturel de supposer que le nombre de sous-groupes de composantes pouvant vérifier la propriété précédente est bien moindre que le nombre total de sous-ensembles de $\{1, \ldots, d\}$, à savoir $2^d - 1$. De façon plus formelle, soit Φ la mesure angulaire associée à $\boldsymbol{V}$, la structure de dépendance asymptotique de $\boldsymbol{V}$ est caractérisée par la répartition de la masse de Φ sur l'orthant positif de la sphère unité. Dans le cas d'une structure de dépendance parcimonieuse, la masse de Φ n'est répartie que sur l'intérieur de certains bords de $\mathcal{S}_d$, correspondant aux différents groupes de composantes asymptotiquement dépendantes. En effet, supposons qu'au moins deux composantes V_i et V_j soient asymptotiquement indépendantes. Cela signifie:

$$t\mathbb{P}\left(V_i > t, V_j > t\right) = t\mathbb{P}\left(t^{-1}\boldsymbol{V} \in \{\boldsymbol{v} : v_i > 1, v_j > 1\}\right) \xrightarrow[t \to \infty]{} 0.$$

Or pour tout $\boldsymbol{w} \in \mathring{\mathcal{S}}_d$, il existe $\boldsymbol{v} \in [0, \infty)^d \setminus \{\boldsymbol{0}\}$ tel que $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \boldsymbol{w}$, $v_i > 1$ et $v_j > 1$, donc:

$$\begin{aligned}\Phi(\mathring{\mathcal{S}}_d) &= \mu\Big(\{\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in \mathring{\mathcal{S}}_d\}\Big) \\ &\leq \mu\Big(\{\boldsymbol{v} : v_i > 1, v_j > 1\}\Big) = 0.\end{aligned}$$

À travers cet exemple simple, nous voyons que l'indépendence asymptotique entre certaines des composantes implique que la masse de la mesure angulaire est confinée sur les bords de $\mathcal{S}_d$. Nous dénommons $\mathcal{S}_\alpha$ l'intérieur du bord de $\mathcal{S}_d$ associée aux coordonnées $\alpha \subset \{1, \ldots, d\}$:

$$\mathcal{S}_\alpha := \{w \in \mathcal{S}_d : \forall j \in \alpha\, w_j > 0, \forall j \notin \alpha\, w_j = 0\}, \tag{1.14}$$

et $\mathcal{M}$ l'ensemble des sous-groupes de composantes de $\{1, \ldots, d\}$ pour lesquels la mesure angulaire attribue une masse non nulle au sous-espace associé:

$$\mathcal{M} := \{\alpha \subset \{1, \ldots, d\} : \Phi(\mathcal{S}_\alpha) > 0\}. \tag{1.15}$$

L'hypothèse de parcimonie se traduit donc par la double inégalité suivante:

$$\begin{cases} \max_{\alpha \in \mathcal{M}} |\alpha| \ll d, \\ |\mathcal{M}| \ll 2^d - 1. \end{cases} \tag{1.16}$$

1.3.1 Problème de l'estimation de $\mathcal{M}$.

Soient $\boldsymbol{V}$ la variable aléatoire à valeurs dans $(1, \infty)^d$ de marginales Pareto unitaire et de mesure angulaire Φ, et $(R, \boldsymbol{W})$ la décomposition pseudo-polaire associée où $R = \|\boldsymbol{V}\|$ et $\boldsymbol{W} = \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}$. Il est primordial de rappeler que l'équation limite (1.12), où Φ est la mesure limite de $\boldsymbol{W}$ pour $R > t$ avec $t \to \infty$, n'est vérifiée que sur les boréliens A de $\mathcal{S}_d$ tels que $\Phi(\partial A) = 0$. En particulier, cette limite n'est valide que pour les sous-espaces de $\mathcal{S}_d$ de mesure de Lebesgue non nulle. Cela ne concerne donc pas les sous-espaces $\mathcal{S}_\alpha$ chargés, puisque dans $\mathbb{R}^d$, $\mathcal{S}_\alpha \subset \partial\mathcal{S}_\alpha$, et donc pour $\alpha \in \mathcal{M}$:

$$0 < \Phi(\mathcal{S}_\alpha) \leq \Phi(\partial\mathcal{S}_\alpha). \tag{1.17}$$

Par ailleurs, suivant l'hypothèse de parcimonie (1.16), la mesure angulaire attribue une masse nulle à l'intérieur de $\mathcal{S}_d$, *i.e.* $\Phi(\mathring{\mathcal{S}}_d) = 0$, or $\boldsymbol{W}$ est toujours à valeurs dans $\mathring{\mathcal{S}}_d$ *i.e.* $\mathbb{P}(\boldsymbol{W} \in \mathring{\mathcal{S}}_d) = 1$. Dans ce cadre, l'estimation empirique naïve de Φ à partir d'un échantillon *i.i.d.* $(\boldsymbol{W}_i)_{i=1,\ldots,n}$ n'est plus possible, comme le suggérait la remarque 4.

1.3.1.1 Parcimonie de la mesure angulaire pour le modèle logistique

Considérons le modèle logistique (Coles and Tawn (1991)), qui avec le modèle logistique asymmétrique, constitue une classe de modèles paramétriques très flexible en dimension $d > 1$ quelconque. Pour tout $\boldsymbol{v}$ dans $(0, \infty)^d$:

$$\mu_{lgtc,\theta}([\mathbf{0}, \boldsymbol{v}]^c) = \left(\sum_{j=1}^{d} v_j^{-\frac{1}{\theta}}\right)^\theta, \tag{1.18}$$

où $\theta > 0$. L'indépendance asymptotique est atteinte pour $\theta = 1$ et dans ce cas, la mesure angulaire associée place toute sa masse sur les axes, *i.e.* $\Phi(\boldsymbol{e}^j) = 1$ pour $j \in \{1, \ldots, d\}$ où $\boldsymbol{e}^j \in \mathbb{R}^d$ tel que $\boldsymbol{e}^j_i = 1$ pour $i = j$ et $\boldsymbol{e}^j_i = 0$ sinon. Pourtant, il est clair que tout point distribué selon la loi de valeurs extrêmes $G(\boldsymbol{v}) = \exp\Big(-\mu_{lgtc,1}([\mathbf{0}, \boldsymbol{v}]^c)\Big)$ est presque sûrement à valeurs dans $(0, \infty)^d$ et donc le pseudo-angle associé est, quant à lui, presque sûrement à valeurs dans l'intérieur de $\mathcal{S}_d$.

Remarque 5. Le lien entre la loi de valeurs extrêmes G et la mesure exposant μ n'existe qu'à travers l'équation (1.8), ainsi, le support de μ (et donc de la mesure angulaire Φ) n'est pas le support de G. Autrement dit, la mesure angulaire ne représente que l'angle **asymptotique** de $\boldsymbol{V}$, et le fait que le support de Φ ne soit inclus que dans certains sous-espaces $\mathcal{S}_\alpha$ ne signifie pas que le support de G est inclus dans des hyperplans de $\mathbb{R}^d$.

L'avantage de modéliser Φ plutôt que G est que cela permet d'appliquer naturellement un modèle parcimonieux à la structure de dépendance. Le revers, est que le support de la distribution que l'on modélise n'est jamais atteint dans l'espace au sein duquel les points sont observés (Remarque 5). Nous considérons par la suite deux voies principales afin de répondre à ce problème. La première, correspondant aux Chapitres 3 et 4, revient à considérer des sous-espaces proches des bords de $\mathcal{S}_d$ mais de mesure de Lebesgue non nulle, et d'estimer la mesure angulaire (ou la mesure exposant) de façon non-paramétrique par le biais d'une mesure de comptage sur ces sous-espaces. La deuxième approche est d'appliquer un modèle de mélange paramétrique pour Φ, où chaque composante du mélange correspond à un groupe de composantes asymptotiquement dépendantes, puis de projeter $\boldsymbol{V}$ sur les sous-espaces correspondants, et de considérer le résidu comme un bruit orthogonal. Cette dernière approche est développée dans le Chapitre (5).

1.4 Estimation du support de la mesure angulaire

D'un point de vue pratique, lorsqu'il s'agit simplement d'estimer le support de la mesure angulaire sur $\mathcal{S}_d$, et non de modélisation à proprement parler, il est plus naturel de se placer directement dans l'espace de définition de la mesure exposant μ, *i.e.* $[0,\infty]^d \setminus \{\mathbf{0}\}$. Par définition de Φ à l'équation (1.10), la contrepartie d'un sous-espace $\mathcal{S}_\alpha$ où $\alpha \subset \{1,\dots,d\}$, en terme de sous-espace de $[0,\infty]^d \setminus \{\mathbf{0}\}$, est le cône tronqué:

$$\mathcal{C}_\alpha = \{\boldsymbol{x} \geq 0 : \|\boldsymbol{x}\| > 1,\, \forall j \in \alpha\, x_j > 0,\, \forall j \notin \alpha\, x_j = 0\}. \tag{1.19}$$

En effet, pour tout α dans $\{1,\dots,d\}$:

$$\begin{aligned}\Phi(\mathcal{S}_\alpha) &= \mu(\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in \mathcal{S}_\alpha)\\ &= \mu(\mathcal{C}_\alpha).\end{aligned}$$

Pour la norme infinie, l'ensemble des cônes $\mathcal{C}_\alpha$ forme une partition de $[\mathbf{0},\mathbf{1}]^c$, et trouver le support de μ est équivalent à trouver celui de Φ. La figure (1.1) montre la partition de $\mathbb{R}^3_+ \setminus [0,1]^3$ par les cônes $\mathcal{C}_\alpha$.

Remarque 6. En considérant la partition $\bigcup_{\alpha\subset\{1,\dots,d\}} \mathcal{C}_\alpha$ nous parcourons l'ensemble des associations de dépendance asymptotique possibles des composantes $\{1,\dots,d\}$. En

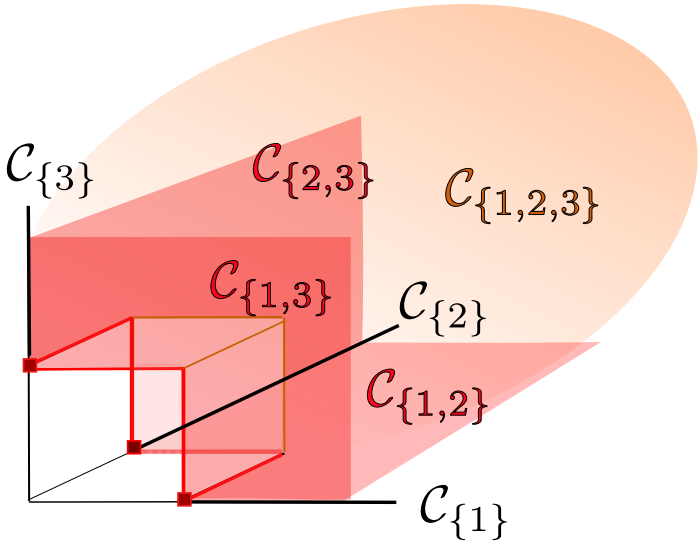


Figure 1.1: Cônes pour $d = 3$, $\|\boldsymbol{x}\| = \max_{j=1}^{d} x_j$

effet, soit $\alpha \subset \{1, \ldots, d\}$ un sous-groupe de composantes tel qu'une probabilité non-négligeable est attribué à l'évènement:

{ *Toutes les variables* $(V_j)_{j\in\alpha}$ *sont grandes alors que les variables complémentaires* $(V_j)_{j\notin\alpha}$ *sont petites.* }

Alors il existe un cône $C \subset \mathcal{C}_\alpha$ tel que $\mu(C) > 0$. Par extension, soit $C_{tot} \subset [0, \infty]^d \setminus \{\mathbf{0}\}$ le cône contenant l'ensemble de la masse de la mesure exposant, alors C_{tot} est nécessairement une union de sous-cônes, eux-mêmes inclus dans des cônes $\mathcal{C}_\alpha$ particuliers.

1.4.1 Inférence

Considérons à nouveau la variable aléatoire $\boldsymbol{V}$, à variation régulière dans $(1, \infty)^d$ de mesure exposant μ et de marginales Pareto unitaires. À l'instar de Φ pour le pseudo-angle $\frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}$ à valeurs dans l'intérieur de $\mathcal{S}_d$, $\boldsymbol{V}$ étant à valeurs $(1, \infty)^d$, pour tout $t > 0$ et $\alpha \subset \{1, \ldots, d\}$ un sous-ensemble strict:

$$\mathbb{P}\left(\frac{\boldsymbol{V}}{t} \in \mathcal{C}_\alpha\right) = 0. \tag{1.20}$$

De plus, les cônes $\mathcal{C}_\alpha$ étant d'intérieur vide nous avons $\mu(\partial\mathcal{C}_\alpha) > 0$, de façon analogue à (1.17) pour les sous-espaces $\mathcal{S}_\alpha$.

Pour estimer $\mu(\mathcal{C}_\alpha)$ à partir d'un échantillon $i.i.d.$ de copies de $\boldsymbol{V}$, il est nécessaire d'élargir les cônes tronqués $\mathcal{C}_\alpha$ dans l'intérieur de $[0, \infty]^d \setminus \{\mathbf{0}\}$. L'idée est de trouver vers quels bords de l'orthant positif (*i.e.* $\mathcal{C}_\alpha$), $\boldsymbol{V}$ se concentre lorsque $\|\boldsymbol{V}\|$ est grand. Considérons donc les cônes élargis suivant, pour $\epsilon > 0$ et $\alpha \subset \{1, \ldots, d\}$:

$$\mathcal{R}_\alpha^\epsilon = \{\boldsymbol{x} \geq \mathbf{0} : \|\boldsymbol{x}\| > 1, \forall j \in \alpha\, x_j > \epsilon, \forall j \notin \alpha\, x_j \leq \epsilon\}. \tag{1.21}$$

Remarque 7. Notons que pour $\epsilon \in (0,1)$ l'ensemble des $\mathcal{R}_\alpha^\epsilon$ forme toujours une partition de $[\mathbf{0},\mathbf{1}]^c$ (pour la norme infinie). De plus, supposons $\mu(\mathcal{C}_\alpha) > 0$, en vertu de la remarque 3 le sous-espace $C \subset \mathcal{C}_\alpha$, minimal pour l'inclusion et tel que $\mu(C) > 0$, est un cône. Or tout cône inclu dans $\mathcal{C}_\alpha$ est nécessairement d'intersection non vide avec $\mathcal{R}_\alpha^\epsilon$. De ce fait, les cônes élargis $\mathcal{R}_\alpha^\epsilon$ vérifient $\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \mu(\mathcal{R}_\alpha^\epsilon) \geq \mu(C) > 0$ et sont de mesure de Lebesgue non nulle.

De façon plus formelle, et conséquemment à la propriété de passage à la limite monotone décroissante de toute mesure:

$$\mu(\mathcal{C}_\alpha) = \mu(\cap_{\epsilon>0,\epsilon\in\mathbb{Q}}\mathcal{R}_\alpha^\epsilon) = \lim_{\epsilon\to 0}\mu(\mathcal{R}_\alpha^\epsilon).$$

1

Afin de pouvoir appliquer la limite (1.7) aux ensembles $\mathcal{R}_\alpha^\epsilon$, ces derniers doivent vérifier $\mu(\partial\mathcal{R}_\alpha^\epsilon) = 0$. Le bord de $\mathcal{R}_\alpha^\epsilon$ dans $[0,\infty]^d \setminus \{\mathbf{0}\}$ est composé d'hyperplans parallèles au axes et disjoints pour des valeurs différentes de $\epsilon \in (0,1)$. Il n'existe de ce fait qu'un nombre au plus dénombrable de ϵ tel que $\mu(\partial\mathcal{R}_\alpha^\epsilon) > 0$, sinon μ serait infinie sur une région compacte de $[0,\infty]^d \setminus \{\mathbf{0}\}$. Ainsi nous pouvons choisir un seuil $\epsilon > 0$, arbitrairement petit et tel que $\mu(\partial\mathcal{R}_\alpha^\epsilon) = 0$, pour estimer la masse des cônes $\mathcal{C}_\alpha$ via la limite suivante:

$$\mu(\mathcal{C}_\alpha) = \lim_{\epsilon\to 0}\lim_{t\to\infty} t\mathbb{P}\left(\boldsymbol{V} \in t\mathcal{R}_\alpha^\epsilon\right)$$

De récents travaux (Goix et al. (2016a)) proposent un algorithme nommé DAMEX permettant d'estimer la masse $\mu(\mathcal{C}_\alpha)$ par une méthode de comptage sur les sous-espaces $t\mathcal{R}_\alpha^\epsilon$ pour $\epsilon > 0$ petit et un large seuil $t > 0$. Plus précisément, soient $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$, n copies $i.i.d.$ de la variable $\boldsymbol{V}$, alors l'estimateur de $\mu(\mathcal{C}_\alpha)$ est:

$$\hat{\mu}_\alpha = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k}\mathcal{R}_\alpha^\epsilon\}}, \tag{1.22}$$

où l'on a remplacé le seuil t par $\frac{n}{k}$, avec $k \leq n$ un ordre de grandeur du nombre de points extrêmes parmi l'échantillon considéré. Un seuil minimal $\mu_{min} > 0$ est ensuite choisit de telle sorte que l'on décide:

$$\mu(\mathcal{C}_\alpha) > 0 \text{ si } \hat{\mu}_\alpha \geq \mu_{min}.$$

L'algorithme est testé sur des données réelles (données de directions de vagues dans la mer du nord fournies par Shell) et parvient à réduire nettement la dimension du problème en rassemblant l'essentiel de la masse de μ dans un nombre limité de sous-cônes $\mathcal{C}_\alpha$ de dimension moyenne. De la même façon, cette méthode montre de bons résultats sur des données simulées à partir d'un modèle paramétrique pour G, une version parcimonieuse du modèle logistique asymmétrique (Tawn (1990a)):

$$G(\boldsymbol{z}) = \exp\left[-\sum_{\alpha\in\mathcal{M}}\left\{\sum_{j\in\alpha}(|\mathcal{A}(j)|z_j)^{-1/\theta_\alpha}\right\}^{\theta_\alpha}\right], \tag{1.23}$$

où $\mathcal{A}(j) = \{\alpha \in \mathcal{M} : j \in \alpha\}$ et où $\theta_\alpha > 0$ est un paramètre de dépendance tel que $\theta_\alpha = 1$ correspond à l'indépendance asymptotique et $\theta_\alpha \downarrow 0$ à la dépendance totale. Néanmoins, les expériences sont faites pour $\theta_\alpha = 0.1$ (dépendance forte) et lorsque l'on augmente la dimension d, θ_α et le nombre $|\mathcal{M}|$ de sous-groupes de composantes dépendantes, la méthode ne parvient plus à retrouver $\mathcal{M}$. De façon intrinsèque, en raison de la partition de l'espace en un nombre considérable de sous-cônes — $(2^d - 1)$, le modèle montre une forte sensibilité à la moindre variation des sous-groupes de composantes impliquées dans le comportement extrême de $\boldsymbol{V}$.

En effet, soit α_0 tel que $\mu(\mathcal{C}_{\alpha_0}) > 0$, alors il est possible que pour un certain nombre de composantes $j \in \{1, \ldots, d\} \setminus \alpha_0$ les grandeurs estimées $\hat{\mu}_{\alpha_0 \cup \{j\}}$ soient non-négligeables:

$$\hat{\mu}_{\alpha_0 \cup \{j\}} = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k} \mathcal{R}^{\epsilon}_{\alpha_0 \cup \{j\}}\}} \geq \mu_{min}. \tag{1.24}$$

Il en résulte un éparpillement de la masse au sein de sous-groupes de cônes proches les uns des autres, rendant impossible l'émergence d'un modèle simplifié de la structure de dépendance. Dans le cas d'une dépendance asymptotique moyenne, voire faible (*i.e.* $\theta_\alpha \in [0.5, 1)$) entre les composantes ou plus généralement de la présence de bruit dans le jeu de données, il est nécessaire de contruire une méthode plus robuste pour estimer le support de μ.

Les figures suivantes (1.2) montrent la répartition des points extrêmes dans les cônes $\frac{n}{k}\mathcal{R}^{\epsilon}_{\alpha}$ où k est choisit de telle sorte que le nombre de points extrêmes $\sum_{i=1}^{n} \mathbb{1}_{\{\|\boldsymbol{V}_i\| > \frac{n}{k}\}}$ représente 5% des données. Pour chaque figure le coefficient de dépendance θ_α est fixé à une valeur particulière. Les points $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$, avec $n = 1e5$, sont générés à partir d'une distribution logistique asymmétrique en dimension $d = 50$, où l'on a généré au préalable 50 sous-groupes de composantes de façon aléatoire pour former $\mathcal{M}$. L'axe des abscisses représente les sous-groupes de composantes α pour lesquels l'estimateur $\hat{\mu}_\alpha$ est non nul, ordonné de façon décroissante. L'axe des ordonnés représente la valeurs des estimateurs $\hat{\mu}_\alpha$. En rouge sont représentés les éléments de $\mathcal{M}$ au nombre de 50. Ainsi, afin de retrouver $\mathcal{M}$, le seuil minimal $\mu_{min} > 0$ doit être tel que:

$$\begin{cases} \hat{\mu}_\alpha \geq \mu_{min}, & \text{pour tout } \alpha \in \mathcal{M} \\ \hat{\mu}_\alpha < \mu_{min}, & \text{sinon.} \end{cases} \tag{1.25}$$

Pour $\theta_\alpha = 0.1$ (1.2 en haut à gauche) l'écart entre les 'vrais' sous-groupes de composantes (*i.e.* appartenant à $\mathcal{M}$) et les autres est net, et il est possible de choisir un seuil μ_{min} tel que (1.25) soit vérifiée. Lorsque θ_α augmente, correspondant à une dépendence asymptotique plus faible, la répartition de la masse des points extrêmes s'éparpille et pour $\theta_\alpha = 0.7, 0.9$ (1.2 en bas gauche (resp. droite)), il n'est plus possible de retrouver $\mathcal{M}$.

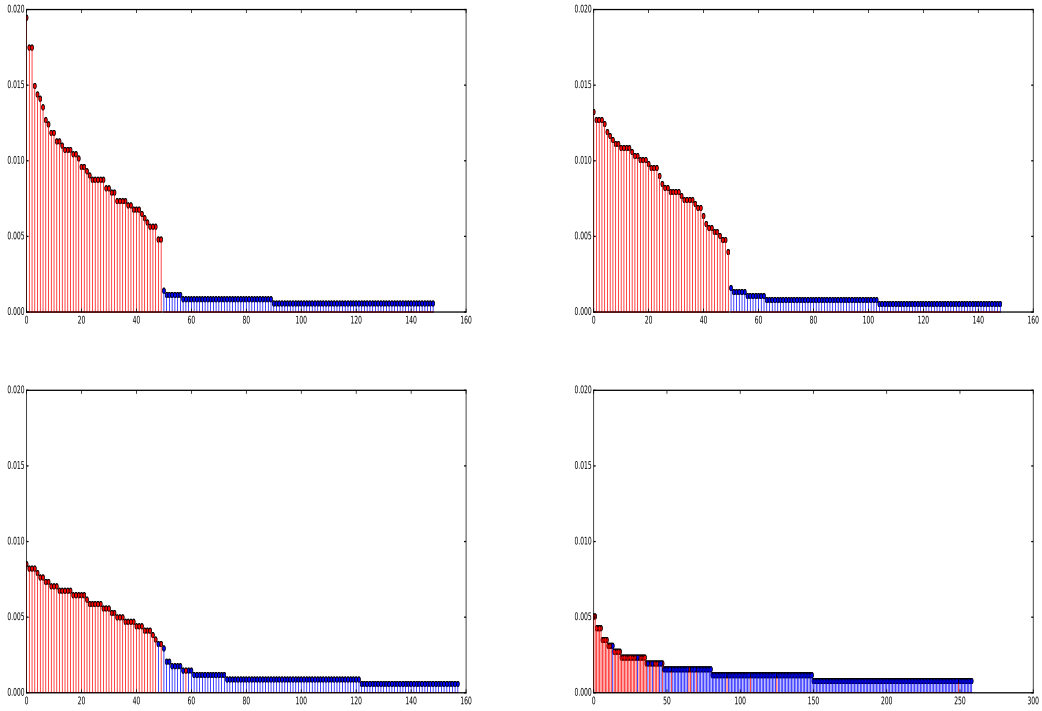


Figure 1.2: Répartition de la masse empirique de μ sur les cônes $\mathcal{C}_\alpha$ pour $\theta_\alpha \in \{0.1, 0.5, 0.7, 0.9\}$ (de haut en bas, de gauche à droite).

1.4.2 Algorithme CLustering Extreme Feature

Plutôt que de considérer la partition en sous-cônes $\mathcal{C}_\alpha$ qui conduit à différencier des événements extrêmes proches, l'idée est de rechercher seulement les sous-groupes $\alpha \subset \{1, \dots, d\}$ de taille $|\alpha| \geq 2$ tels que pour tout $j \in \alpha$, V_j soit grand, quelles que soient les contreparties $V_{j'}$ pour $j' \in \{1, \dots, d\} \setminus \alpha$. Nous définissons les sous-espaces imbriqués suivant:

$$\Gamma_\alpha = \{\boldsymbol{x} : \forall j \in \alpha \, x_j > 1\}. \tag{1.26}$$

De façon analogue au sous-cônes $\mathcal{R}^\epsilon_\alpha$, nous avons $\mu(\partial\Gamma_\alpha) = 0$, et il est possible d'estimer leur masse en utilisant la limite (1.7).

L'avantage de cette approche est qu'elle permet d'éviter la situation (1.24) d'éparpillement de la masse lors de l'estimation du support de μ. Pour tout $\alpha \subset \{1, \dots, d\}$, Γ_α correspond à l'évènement:

{ *Toutes les variables* $(V_j)_{j\in\alpha}$ *sont grandes (quelles que soient les variables complémentaires* $(V_j)_{j\notin\alpha}$*).* }

Soit $t > 0$ un seuil élevé, chaque rectangle $t\Gamma_\alpha$ rassemble les points extrêmes possédant de large coordonnées en α indépendamment de leurs coordonnées complémentaires $\{1, \dots, d\} \setminus \alpha$. De cette façon, les points appartenant à $t\Gamma_\alpha$, appartiennent nécessairement aux sous-rectangles $t\Gamma_\beta$ pour tout $\beta \subset \alpha$. L'objectif est alors de rechercher, de façon incrémentale, les sous-groupes α les plus grands tel que le nombre de points dans le rectangle $t\Gamma_\alpha$ soit non-négligeable.

En termes d'estimation du support de la mesure exposant μ, ce cadre est équivalent à la partition de l'espace en cônes $\mathcal{C}_\alpha$ lorsque l'on considère les sous-groupes $\alpha \subset \{1, \dots, d\}$ maximaux pour l'inclusion tels que $\mu(\mathcal{C}_\alpha) > 0$. En effet, admettons que μ donne une masse non nulle au cône $\mathcal{C}_\alpha$. Suivant la remarque 3, la région $C \subset \mathcal{C}_\alpha$, minimale pour l'inclusion, où se concentre la masse de μ, est elle-même un cône. Or tout cône inclus dans $\mathcal{C}_\alpha$ possède une intersection non vide avec Γ_α, dès lors $\mu(\mathcal{C}_\alpha) > 0$ implique $\mu(\Gamma_\alpha) > 0$. À l'inverse, supposons que $\mu(\Gamma_\alpha) > 0$, alors si pour tout β tel que $\alpha \subset \beta$, $\mu(\Gamma_\beta) = 0$, nous avons nécessairement $\mu(\mathcal{C}_\alpha) > 0$. Par contraposition, supposons $\mu(\Gamma_\alpha) > 0$ et $\mu(\mathcal{C}_\alpha) = 0$, alors il existe $j_C \in \{1, \dots, d\} \setminus \alpha$ et un cône C inclu dans $\Gamma_{\alpha\cup\{j_C\}}$ tel que $\mu(C) > 0$, autrement dit, il existe β tel que $\alpha \subset \beta$ et $\mu(\Gamma_\beta) > 0$.

De ce fait, les sous-groupes $\alpha \subset \{1, \dots, d\}$ maximaux pour l'inclusion tels que $\mu(\mathcal{C}_\alpha) > 0$, sont les sous-groupes maximaux tels que $\mu(\Gamma_\alpha) > 0$. Soit $\mathbb{M}$ l'ensemble des sous-groupes $\alpha \subset \{1, \dots, d\}$ maximaux tels que $\mu(\Gamma_\alpha) > 0$, alors nous avons la propriété suivante (Chiapino and Sabourin (2016)), pour tout $\alpha \subset \{1, \dots, d\}$:

$$\alpha \in \mathbb{M} \Leftrightarrow \alpha \text{ est maximal dans } \mathcal{M}. \tag{1.27}$$

Cette approche a une contrepartie: elle implique une explosion combinatoire. En effet, à la différence du nombre de cônes $\mathcal{C}_\alpha$ à estimer, qui est au maximum le nombre de points extrêmes du jeu de données, le nombre de rectangles imbriqués Γ_α qu'il

faut parcourir peut être beaucoup plus grand, *i.e.* $O(2^{d_0}-1)$ pour $d_0 \leq d$. C'est une conséquence du fait que l'union $\cup_{\alpha\subset\{1,\ldots,d\}}\Gamma_\alpha$ ne forme plus une partition de l'espace considéré.

1.4.2.1 Algorithme 'Apriori'

Plus généralement, cette problématique exploratoire se traduit de la façon suivante: nous voulons trouver tous les sous-groupes $\mathcal{I}$ d'un certain ensemble d'indices $\mathcal{T}$ tel que $\mathcal{I}$ vérifie une condition particulière $C(\mathcal{I})$. La propriété essentielle étant que pour tout $\mathcal{I}_1 \subset \mathcal{I}_2$ alors $C(\mathcal{I}_2)$ implique $C(\mathcal{I}_1)$. Ce problème est directement lié à l'algorithme 'Apriori' introduit dans (Agrawal et al. (1994)). L'algorithme permet en particulier de réduire l'exploration totale des sous-groupes de $\mathcal{T}$ par le biais d'une exploration incrémentale.

Appliquée à notre cadre, la procédure parcourt les sous-groupes $\alpha \subset \{1,\ldots,d\}, |\alpha| \geq 2$ de taille croissante en ne gardant que ceux qui vérifient un certain critère de décision $C(\alpha)$ équivalent à $\mu(\Gamma_\alpha) > 0$. La sortie de l'algorithme est donc:

$$\mathbb{M}_0 := \Big\{\alpha \subset \{1,\ldots,d\} : |\alpha| \geq 2,\, C(\alpha)\Big\}, \tag{1.28}$$

dont on ne garde que les éléments maximaux pour l'inclusion afin d'obtenir $\mathbb{M}$.

Cette procédure peut être visualisée à travers le diagramme de Hasse (1.3), que l'on explore de haut en bas, en coupant toutes les branches partant d'un noeud lorsque celui-ci ne vérifie pas la propriété C. Cette méthode réduit drastiquement le nombre de sous-groupes qu'il aurait potentiellement fallu tester.

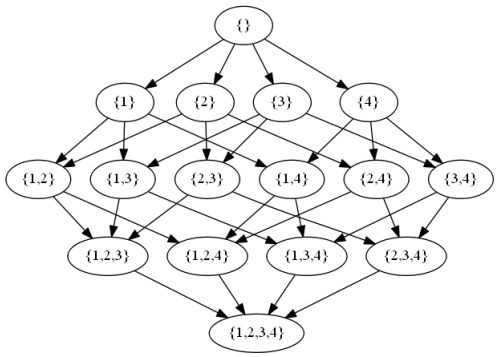


Figure 1.3: Diagramme de Hasse pour 4 composantes.

Pour le critère $\mu(\Gamma_\alpha) > 0$, la réduction du graphe s'effectue par le biais de la propriété suivante:

$$\text{Soit } \alpha \text{ tel que } \mu(\Gamma_\alpha) = 0, \text{ alors pour tout } \beta \supset \alpha,\ \mu(\Gamma_\beta) = 0. \tag{1.29}$$

À l'étape s, étant donné tous les sous-groupes α de taille s tels que $C(\alpha)$ est vérifiée, l'algorithme construit les sous-groupes $\beta \subset \{1,\ldots,d\}$ de taille $s+1$ susceptibles

de vérifier $C(\beta)$. Soit $\mathbb{M}_0^s = \Big\{\alpha \subset \{1,\dots,d\} : |\alpha| = s \text{ et } C(\alpha)\Big\}$ l'ensemble des sous-groupes de composantes de taille s et vérifiant le critère C. L'ensemble des sous-groupes candidats de taille $s+1$ pour $s \geq 3$ est donc:

$$\mathbb{A}^s = \Big\{\beta \subset \{1,\dots,d\} : |\beta| = s \text{ et } \beta \setminus \{j\} \in \mathbb{M}_0^{s-1} \text{ pour tout } j \in \beta\Big\}. \tag{1.30}$$

Pour $s = 2$, posons $\mathbb{A}^2 = \Big\{\alpha \subset \{1,\dots,d\} : |\alpha| = 2\Big\}$.

1.4.2.2 Critère d'arrêt

Soit $(\boldsymbol{V}_1,\dots,\boldsymbol{V}_n)$ un échantillon $i.i.d.$ de variables aléatoires à valeurs dans $(1,\infty)^d$ associées à la mesure angulaire Φ. Pour estimer $\mathbb{M}_0$ il apparait naturel, d'une façon analogue à l'algorithme DAMEX, d'avoir pour critère de décision sur $\alpha \subset \{1,\dots,d\}$ un seuil $\gamma_{min} > 0$ sur la proportion de points extrêmes appartenant à Γ_α:

$$\widehat{\gamma}_\alpha = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k}\Gamma_\alpha\}} \geq \gamma_{min}. \tag{1.31}$$

Cependant, du fait de l'imbrication des rectangles Γ_α entre eux, la valeur de $t\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] \approx \mu(\Gamma_\alpha)$, pour t grand, décroit nécessairement pour des tailles croissantes de α. En effet, soient $\alpha_1 \subset \alpha_2 \subset \{1,\dots,d\}$, alors $\Gamma_{\alpha_2} \subset \Gamma_{\alpha_1}$ et donc $t\mathbb{P}[\boldsymbol{V} \in t\Gamma_{\alpha_2}] \leq t\mathbb{P}[\boldsymbol{V} \in t\Gamma_{\alpha_1}]$ pour tout $t > 0$. Le choix du seuil $\gamma_{min} > 0$ devrait donc dépendre de la taille des sous-groupes α considérés.

De ce fait, un résumé numérique alternatif du degré de dépendance des composantes $\alpha \subset \{1,\dots,d\}$, dénommé κ_α, est proposé:

$$\kappa_\alpha := \frac{\mu(\Gamma_\alpha)}{\mu(\bigcup\limits_{\substack{\beta\subset\alpha \\ |\beta|=|\alpha|-1}} \Gamma_\beta)}. \tag{1.32}$$

La quantité κ_α ne décroit pas nécessairement avec la taille des α, et $\kappa_\alpha > 0$ si et seulement si $\mu(\Gamma_\alpha) > 0$. Il est donc possible de choisir un seuil $\kappa_{min} > 0$, indépendant de la taille de α, afin d'avoir un critère $\{\hat{\kappa}_\alpha \geq \kappa_{min}\}$ pour décider de $\{\kappa_\alpha > 0\}$. L'idée est de faire dépendre l'acceptation d'un sous-groupe α dans $\mathbb{M}_0$ en fonction des sous-groupes de taille inférieure le composant:

$$\kappa_\alpha = \mathbb{P}[\text{ Tous les } V_j \text{ avec } j \in \alpha \text{ sont grands } | \text{ Tous sont grands sauf au plus un }] \tag{1.33}$$

L'estimateur $\hat{\kappa}_\alpha$ de la quantité (1.32) est donc:

$$\hat{\kappa}_\alpha = \frac{\hat{\gamma}_\alpha}{\hat{\mu}(\bigcup\limits_{\substack{\beta\subset\alpha \\ |\beta|=|\alpha|-1}} \Gamma_\beta)}. \tag{1.34}$$

L'algorithme permettant l'estimation de $\mathbb{M}$ est décrit par la procédure (1).

Algorithm 1 CLEF (CLustering Extreme Features)

Entrée: Seuil $\kappa_{\min} > 0$.

PHASE 1: Construire l'ensemble $\widehat{\mathbb{M}}_0$ de tous les sous-groupes asymptotiquement dépendant.
Étape 1: $\widehat{\mathbb{M}}_0^2 = \left\{\alpha \subset \{1, \ldots, d\} : |\alpha| = 2,\ \hat{\kappa}_\alpha > \kappa_{\min}\right\}$, $S = 2$.
Étape $s = 3, \ldots, d$: Si $\widehat{\mathbb{M}}_0^{s-1} = \emptyset$, finir **PHASE 1**. Sinon:

- Générer les sous-groupes candidats de taille s:
  $\mathbb{A}^s = \{\alpha \subset \{1, \ldots, d\} : |\alpha| = s \text{ et } \alpha \setminus j \in \widehat{\mathbb{M}}_0^{s-1} \text{ pour tout } j \in \alpha\}$.
- Retenir les sous-groupes vérifiant le critère, $\widehat{\mathbb{M}}_0^s = \left\{\alpha \in \mathbb{A}^s : \hat{\kappa}_\alpha > \kappa_{\min}\right\}$.
- Si $\widehat{\mathbb{M}}_0^s \neq \emptyset$, $S = s$.

Sortie: $\widehat{\mathbb{M}}_0 = \emptyset$ si $S = 1$ et $\widehat{\mathbb{M}}_0 = \bigcup_{s=2}^{S} \widehat{\mathbb{M}}_0^s$ si $S \geq 2$.

PHASE 2: Ne retenir que les α maximaux.
Si $S = 1$, alors $\widehat{\mathbb{M}} = \emptyset$. Sinon:
Initialisation: $\widehat{\mathbb{M}} \leftarrow \widehat{\mathbb{M}}_0^s$.
pour $s = (S-1) : 2$,
    pour $\alpha \in \widehat{\mathbb{M}}_0^s$,
        S'il n'existe pas de $\beta \in \widehat{\mathbb{M}}$ tel que $\alpha \subset \beta$, alors $\widehat{\mathbb{M}} \leftarrow \widehat{\mathbb{M}} \cup \{\alpha\}$.
Sortie: $\widehat{\mathbb{M}}$

1.4.2.3 Résultats

Dans la pratique, le choix du résumé du degré de dépendance κ_α est d'autant plus justifié lorsque du bruit est observé sur les composantes. Afin de reproduire un jeu de données pouvant être assimilé à un jeu de données réelles bruité, nous reprenons le modèle logistique décrit dans la section 1.4.1, pour $d = 20$, $|\mathbb{M}| = 15$ et $\theta_\alpha = 0.5$, sur lequel nous perturbons artificiellement la structure de dépendance. Plus précisément, chacun des points $V_i \in (0,\infty)^d, i = 1,\dots,n$, du jeu de données est généré par le biais d'une version des sous-groupes de composantes dépendantes $\mathbb{M}^i$, légèrement perturbé par rapport à $\mathbb{M}$. Pour $i = 1,\dots,n$, chaque sous-groupe de composantes α_k^i dans $\mathbb{M}^i$ est modifié d'une composante par rapport à α_k, de telle sorte qu'il existe $j_0 \in \{1,\dots,d\}$ tel que:

$$\{j_0 \in \alpha_k \text{ et } j_0 \notin \alpha_k^i\} \text{ ou } \{j_0 \notin \alpha_k \text{ et } j_0 \in \alpha_k^i\}. \tag{1.35}$$

L'objectif est de retrouver $\mathbb{M}$, à partir de ce jeu de données perturbé. Il est donc nécessaire, à chaque étape $s = 2,\dots,S$, avec $S = \max_{\alpha\in\mathbb{M}} |\alpha|$, de discerner les sous-groupes de composantes suivants:

$$\begin{aligned}\hat{\kappa}_\alpha &\geq \kappa_{min} \text{ pour tout } \alpha \in \mathbb{M}_0^s\\ \hat{\kappa}_\alpha &< \kappa_{min} \text{ pour tout } \alpha \in \mathbb{A}^s \setminus \mathbb{M}_0^s.\end{aligned}$$

Nous dénommons $\mathbb{M}^F = \underset{s=1,\dots,S}{\cup} \mathbb{A}^s \setminus \mathbb{M}_0$ l'ensemble des sous-groupes de composantes générés comme candidats mais n'appartenant pas à $\mathbb{M}_0$. Notons que $\mathbb{M}_0 = \cup_{s=2}^S \mathbb{M}_0^s$. De cette façon, la condition permettant l'existence d'un seuil $\kappa_{min} > 0$ tel que $\widehat{\mathbb{M}} = \mathbb{M}$ est:

$$\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha < \min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha. \tag{1.36}$$

Si (1.36) est vérifié, il suffit de choisir un seuil κ_{min} appartenant à l'intervalle $(\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha, \min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha)$ pour retrouver $\mathbb{M}$. Les figures (1.4) et (1.5) montrent la valeur de la masse empirique $\hat{\gamma}_\alpha$ (1.31) et de $\hat{\kappa}_\alpha$(1.32) sur les ensembles $\mathbb{M}_0$ (en rouge) et $\mathbb{M}^F$ (en bleue). La ligne bleu représente $\max_{\alpha\in\mathbb{M}^F} \hat{\gamma}_\alpha$ (resp. $\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha$) et la ligne rouge représente $\min_{\alpha\in\mathbb{M}_0} \hat{\gamma}_\alpha$ (resp. $\min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha$). Les sous-groupes rouges sont classés par taille croissante, et l'on remarque que la masse empirique (1.31) décroit avec la taille des sous-groupes. De plus, comme $\max_{\alpha\in\mathbb{M}^F} \hat{\gamma}_\alpha > \min_{\alpha\in\mathbb{M}_0} \hat{\gamma}_\alpha$, il est impossible de trouver un seuil γ_{min} pour lequel on peut retrouver $\mathbb{M}$.

Nous comparons à présent l'algorithme DAMEX ainsi que l'algorithme CLEF pour les différents critères $C(\alpha) = \{\hat{\gamma}_\alpha > \gamma_{min}\}$ et $C(\alpha) = \{\hat{\kappa}_\alpha > \kappa_{min}\}$ sur des données journalières de débit fluvial. Le jeu de données provient de $d = 92$ stations réparties sur les différents fleuves de France, enregistré entre le 1er janvier 1969 et le 31 décembre 2008, et comprend $n = 14610$ vecteurs $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ à valeur dans $\mathbb{R}_+^d$. L'objectif est de trouver les sous-groupes de stations susceptibles de subir une

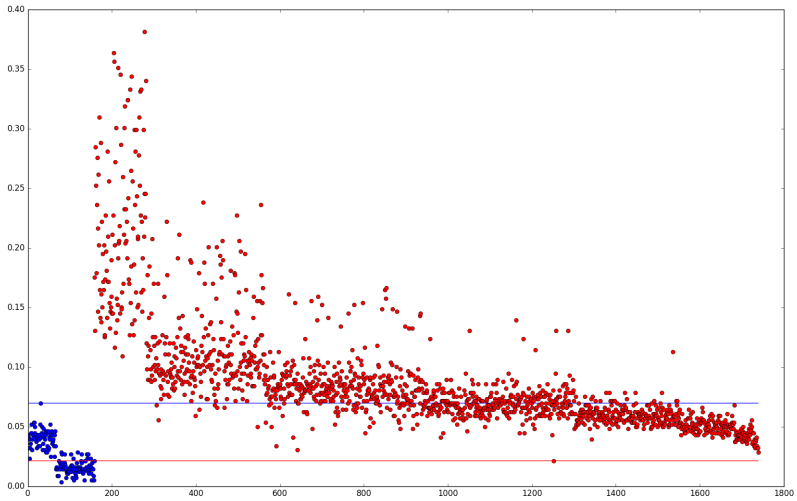


Figure 1.4: Valeurs de $\hat{\gamma}_\alpha$ pour les sous-groupes de composantes de $\mathbb{M}^F$ et de $\mathbb{M}_0$.

crue de façon concomitante. Après transformation (1.6), nous considérons les points extrêmes $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n_{extr}})$ où $n_{extr} = \sum_{i=1}^n \mathbb{1}_{\|\boldsymbol{V}_i\| > \frac{n}{k}}$ avec k choisit tel que $n_{extr} = n \cdot 5\%$. Les figures (1.6), (1.7) et (1.8) montrent respectivement les résultats de l'algorithme DAMEX, de l'algorithme CLEF avec la masse empirique (1.31) puis avec $\hat{\kappa}_\alpha$ (1.32), pour les seuils respectifs $\mu_{min} = 0.002$, $\gamma_{min} = 0.2$ et $\kappa_{min} = 0.2$. La figure (1.6) représente $\widehat{\mathcal{M}}$ et les figures (1.7) et (1.8) représente $\widehat{\mathbb{M}}$. Chaque sous-groupe de composantes est représenté par l'enveloppe convexe des stations qu'il comprend, et chaque couleur correspond au nombre spécifique de stations par sous-groupe. Nous voyons que contrairement à la masse empirique (1.31) et à l'algorithme DAMEX, le critère $\{\hat{\kappa}_\alpha \geq \kappa_{min}\}$ permet de discerner de large sous-groupes de composantes sans que le résultat soit perturbé par des sous-groupes de petites tailles comme le suggèrent les figures précédentes (1.4 et 1.5).

1.4.3 Coefficient de dépendance de queue

En vue de construire une série de tests statistiques qui ne reposent pas sur l'utilisation d'un seuil arbitraire, il est nécessaire d'introduire un nouveau critère, autre que $\{\kappa_\alpha > 0\}$, en prenant en compte la variance, afin de décider de l'hypothèse $\mu(\Gamma_\alpha) > 0$. En effet, lorsque $\mu(\Gamma_\alpha) = 0$ la distribution limite des statistiques $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha)$ est dégénérée en zéro. De ce fait, nous n'avons pas de contrôle sur le niveau asymptotique des tests basés sur ces statistiques sous l'hypothèse $H_0 : \kappa_\alpha = 0$. Considérons le coefficient de dépendance de queue de distribution $\eta_\alpha \in (0, 1]$, introduit dans le cas bivarié dans (Ledford and Tawn (1996)) et étendu en dimension $d \geq 3$ dans (De Haan and Zhou (2011)) et (Eastoe and Tawn (2012)). L'hypothèse

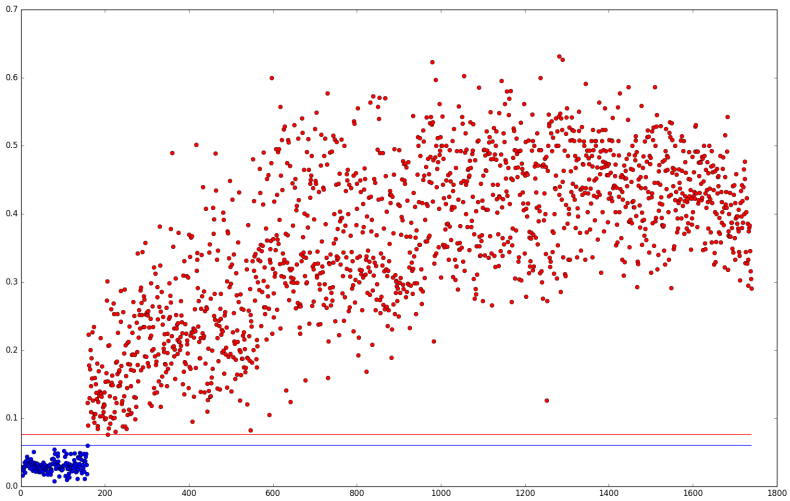


Figure 1.5: Valeurs de $\hat{\kappa}_\alpha$ pour les sous-groupes de composantes de $\mathbb{M}^F$ et de $\mathbb{M}_0$.

fondamentale, stipule l'existence d'un coefficient $\eta_\alpha \in (0,1]$ et d'une fonction à variation lente $\mathcal{L}_\alpha$ tels que:

$$\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] = t^{-1/\eta_\alpha}\mathcal{L}_\alpha(t). \tag{1.37}$$

Supposons que la limite $\lim_{t\to\infty} t\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] = \mu(\Gamma_\alpha)$ existe et que l'hypothèse précédente (1.37) est valide, alors $\mu(\Gamma_\alpha) > 0$ implique $\eta_\alpha = 1$. À l'inverse, supposons (1.37) et $\liminf_{t\to\infty} \mathcal{L}_\alpha(t) > 0$ alors $\eta_\alpha = 1$ implique $\mu(\Gamma_\alpha) > 0$. L'hypothèse nulle $\mu(\Gamma_\alpha) > 0$ correspond donc à l'hypothèse $\eta_\alpha = 1$, à une condition peu coûteuse sur $\mathcal{L}_\alpha$ près. Dès lors, si $\eta_\alpha = 1$ la limite $\sqrt{k}(\hat{\eta}_\alpha - \eta_\alpha)$ est non-dégénérée et il est ainsi possible de contrôler le niveau asymptotique du test associé.

Un nouveau critère $C(\alpha)$ pour l'algorithme CLEF est construit à partir de l'estimateur du coefficient de dépendance de queue η_α. Cette serie de tests statistiques est développée dans la partie 4 pour différents estimateurs non-paramétriques de η_α, à savoir une extension multivariée de l'estimateur de Peng (Peng (1999)) et l'estimateur de Hill (Draisma et al. (2001), Draisma et al. (2004)). Une version non-dégénérée du test $H_0 : \kappa_\alpha = 0$ est aussi développée, en considérant la variance de κ_α ainsi qu'un seuil $\kappa_{min} > 0$, donnant la nouvelle hypothèse $H_0 : \kappa_\alpha > \kappa_{min}$.

En reprenant la méthode employée pour comparer les résultats des critères $\{\hat{\kappa}_\alpha \geq \kappa_{min}\}$ et $\{\hat{\gamma}_\alpha \geq \gamma_{min}\}$ sur les données artificiellement bruitées (1.4.2.3), nous testons l'estimateur de Hill par le biais d'un nouveau critère d'acceptation de sous-groupe de composantes $\alpha \subset \{1, \ldots, d\}$, dont nous contrôlons le niveau asymptotique:

$$\hat{\eta}_\alpha > 1 - q_{1-\delta}\frac{\hat{\sigma}_\alpha}{\sqrt{k}},$$

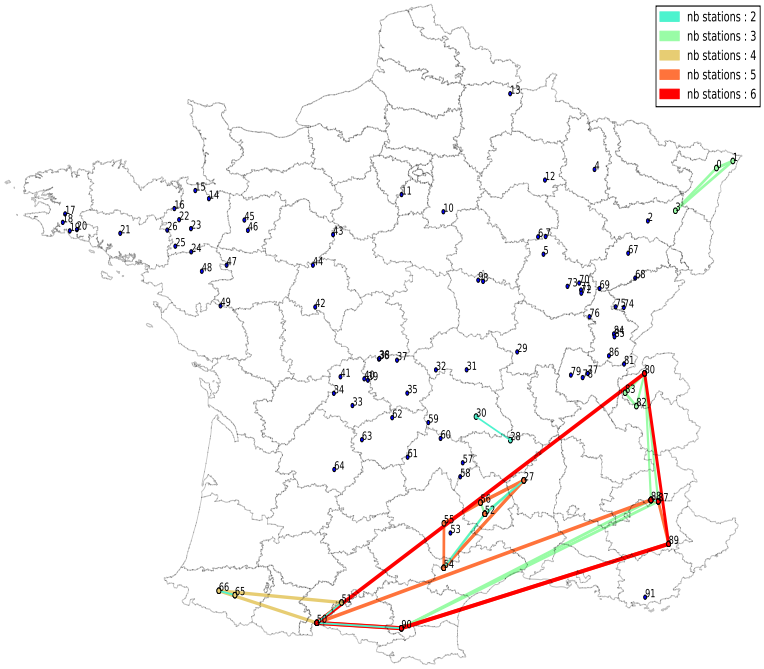


Figure 1.6: Représentation de $\widehat{\mathcal{M}}$ pour l'algorithme DAMEX avec $\mu_{min} = 0.002$.

où $q_{1-\delta}$ est le $(1-\delta)$-quantile de la loi normale centrée réduite, $\hat{\sigma}_\alpha$ la variance asymptotique de l'estimateur et k un ordre de grandeur du nombre de points extrêmes. La figure 1.9 montre les résultats du critère pour les mêmes paramètres initiaux $d = 20$, $|\mathbb{M}| = 15$ et $\theta_\alpha = 0.5$ que pour les figures 1.4 et 1.5 et pour une très faible valeur $\delta = 1e-7$. La valeur mesurée en ordonnée est $\hat{\eta}_\alpha - (1 - q_{1-\delta}\frac{\hat{\sigma}_\alpha}{\sqrt{k}})$ et est strictement positive pour tous les sous-groupes devant être acceptés (en rouge) et strictement négative pour tous les sous-groupes devant être rejetés (en bleu).

1.5 Modèle paramétrique pour la mesure angulaire

Cette dernière partie est motivée par un problème de détection d'anomalies sur des données aéronautiques fournit par Airbus. L'approche générale de la détection d'anomalies est de construire un modèle décrivant le comportement 'normal' d'un phénomène puis de définir comme 'anormaux' des évènements en fonction de l'écart qu'ils ont avec ce modèle. Dans le cadre de la théorie des extrêmes multivariés, nous proposons par une approche différente, d'assimiler les évènements 'anormaux' aux évènements extrêmes et de construire un modèle permettant la classification non supervisée de ces évènements.

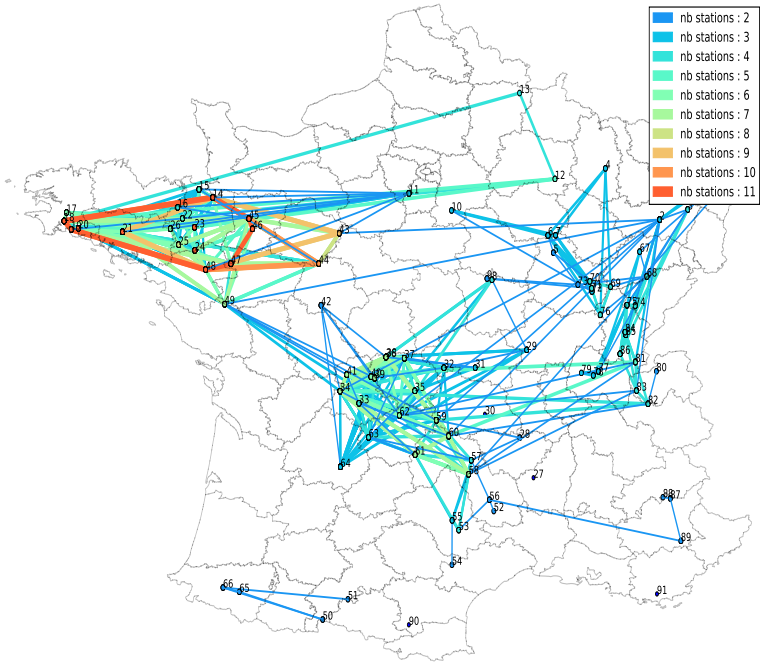


Figure 1.7: Représentation de $\widehat{\mathcal{M}}$ pour le critère $\{\hat{\gamma}_\alpha > \gamma_{min}\}$ avec $\gamma_{min} = 0.2$.

Pour ce faire, après avoir estimé le support de la structure de dépendance $\mathcal{M}$, un modèle de mélange paramétrique pour la mesure angulaire Φ est proposé. L'intérêt de cette modélisation est double. Premièrement, cela permet d'attribuer de manière probabiliste des points à des sous-groupes de composantes. Deuxièmement, il devient possible de construire une matrice de similarité $\boldsymbol{S} = (s_{i,j})_{i,j=1,\dots,n}$, où n est le nombre de points extrêmes, telle que:

$$s_{i,j} = \mathbb{P}[\text{ les points } i \text{ et } j \text{ sont attribués au même sous-groupe de composantes }] \tag{1.38}$$

L'intérêt de la matrice de similarité $\boldsymbol{S}$ est que l'on peut lui appliquer un algorithme de partitionnement spectral et ainsi faire du *clustering* sur les points extrêmes, non plus simplement en les attribuant à un sous-groupe de composante particulier.

Par souci de simplicité nous supposons ici $\min_{\alpha\in\mathcal{M}} |\alpha| > 1$ ansi que $\mathbb{M} = \mathcal{M}$, le cas général étant développé dans la partie 5. Supposons que la fonction $\boldsymbol{z} \mapsto \mu([\mathbf{0}, \boldsymbol{z}]^c)$ soit différentiable d fois sur $[0,\infty)^d \setminus \{0\}$, alors Φ possède des densités sur l'intérieur du simplexe $\mathcal{S}_d$ ($\|\cdot\|$ étant la norme L_1 dans cette partie) ainsi que sur chacun des sous-simplexes $\mathcal{S}_\alpha$ (voir le Théorème 1 dans Coles and Tawn (1991)).

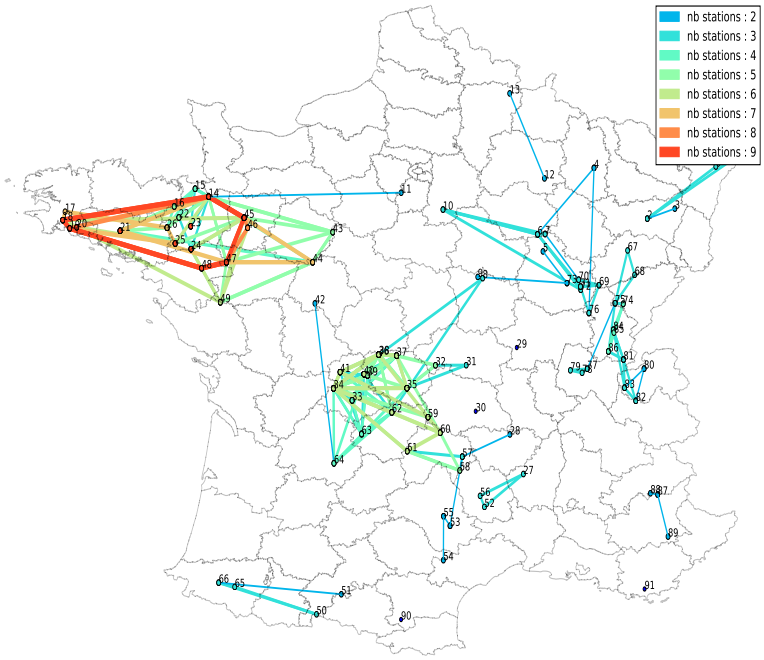


Figure 1.8: Représentation de $\widehat{\mathcal{M}}$ pour l'algorithme CLEF avec $\kappa_{min} = 0.2$.

Nous pouvons donc opérer la décomposition:

$$\frac{1}{\Phi(\mathcal{S}_d)}\Phi(\mathrm{d}\boldsymbol{w}) = \sum_{\alpha\in\mathbb{M}} \pi_\alpha \phi_\alpha(\boldsymbol{w}_\alpha)\mathrm{d}\boldsymbol{w}_\alpha \tag{1.39}$$

avec les poids $\pi_\alpha \in (0,1)$ tels que $\sum_\alpha \pi_\alpha = 1$ et ϕ_α une densité non nulle sur $\mathcal{S}_\alpha$. Le modèle de mélange, afin de correspondre à une mesure angulaire valide, doit vérifier la contrainte des moments (1.13). Les densités ϕ_α ainsi que les poids π_α doivent donc vérifier la contrainte:

$$\sum_{\substack{\alpha\in\mathbb{M}\\ j\in\alpha}} \pi_\alpha \int_{\mathcal{S}_\alpha} w_j\, \phi_\alpha(\boldsymbol{w_\alpha})\mathrm{d}\boldsymbol{w_\alpha} = \frac{1}{d},\ \forall j \in \{1,\dots,d\}. \tag{1.40}$$

La loi de Dirichlet étant une distribution naturelle sur le simplexe, nous proposons comme modèle paramétrique un mélange de Dirichlet. Cette loi peut être paramétrée par sa moyenne $\boldsymbol{m}_\alpha \in \mathcal{S}_\alpha$ et un paramètre de concentration $\nu_\alpha > 0$. Nous avons, pour tout $\boldsymbol{w} \in \mathcal{S}_\alpha$:

$$\phi_\alpha(\boldsymbol{w}|\boldsymbol{m}_\alpha, \nu_\alpha) = \frac{\Gamma(\nu_\alpha)}{\prod_{i\in\alpha}\Gamma(\nu_\alpha m_{\alpha,i})} \prod_{i\in\alpha} w_i^{\nu_\alpha m_{\alpha,i}-1}. \tag{1.41}$$

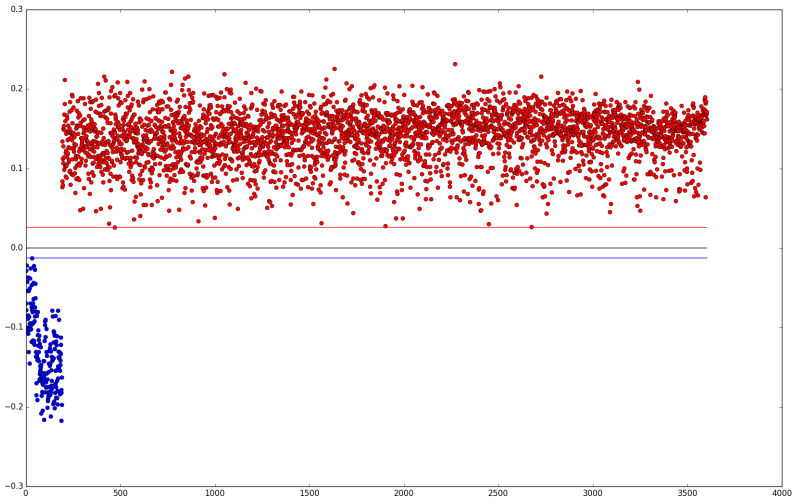


Figure 1.9: Valeurs de $\hat{\eta}_\alpha - (1 - q_{1-\delta}\frac{\hat{\sigma}_\alpha}{\sqrt{k}})$ pour les sous-groupes de composantes de $\mathbb{M}^F$ et de $\mathbb{M}_0$.

Dans ce cadre, étant donné que $\int_{\mathcal{S}_\alpha} \boldsymbol{w}\, \phi_\alpha(\boldsymbol{w}|\boldsymbol{m}_\alpha, \nu_\alpha)\mathrm{d}\boldsymbol{w} = \boldsymbol{m}_\alpha$, la contrainte des moments devient:

$$\sum_{\alpha\in\mathbb{M}} \pi_\alpha \boldsymbol{m}_{\alpha,j} = \frac{1}{d},\ \forall j \in \{1,\dots,d\}. \tag{1.42}$$

Deux difficultés principales émergent lors de l'inférence d'un tel modèle. Premièrement, les estimateurs $\widehat{\boldsymbol{m}}$ et $\widehat{\boldsymbol{\pi}}$ doivent vérifier les trois contraintes suivantes:

$$\begin{cases} \sum_{\alpha\in\mathbb{M}} \pi_\alpha m_{\alpha,j} = \frac{1}{d},\ \forall j \in \{1,\dots,d\} \\ \sum_{j\in\alpha} m_{\alpha,j} = 1,\ \forall \alpha \in \mathbb{M} \\ \sum_{\alpha\in\mathbb{M}} \pi_\alpha = 1 \end{cases} \tag{1.43}$$

et la maximisation de la log-vraisemblance devient de ce fait non convexe. Cette difficulté est atténuée par le changement de variable suivant, pour tout $\alpha \in \mathcal{M}$:

$$\rho_{\alpha,j} = \pi_j m_{\alpha,j},\ \forall j \in \alpha. \tag{1.44}$$

Ainsi, les trois contraintes précédentes sont réduites à:

$$\sum_{\alpha\in\mathcal{M}} \rho_{\alpha,j} = \frac{1}{d},\ \forall j \in \{1,\dots,d\}. \tag{1.45}$$

En effet, en posant $\pi_\alpha := \sum_{j\in\alpha} \rho_\alpha$ et $m_{\alpha,j} := \frac{\rho_{\alpha,j}}{\sum_{j\in\alpha}\rho_\alpha}$, pour tout $\alpha \in \mathcal{M}$ et pour tout $j \in \alpha$, il est aisé de voir que (1.45) implique $\sum_{j\in\alpha} m_{\alpha,j} = 1$, $\forall \alpha \in \mathbb{M}$ et $\sum_{\alpha\in\mathbb{M}} \pi_\alpha = 1$, et donc (1.45) est équivalent à (1.42).

La deuxième difficulté vient de la nature asymptotique du modèle à estimer, conséquence directe du problème explicité dans la section 1.3.1. Pour tout élément $\boldsymbol{V}$, obtenu après transformation du jeu de données initial (1.6), le pseudo-angle $\frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}$ appartient à l'intérieur du simplexe central $\mathcal{S}_d$. Pour palier ce problème, l'idée est de séparer $\boldsymbol{V}$ entre sa contribution en α, $\boldsymbol{V}_\alpha = (V_j)_{j\in\alpha}$ et ce que l'on pourrait nommer son résidu non-asymptotique, $\boldsymbol{\varepsilon}_{\alpha^c} = (V_j)_{j\notin\alpha}$ (figure (1.10)). L'estimation se fait via l'ajout d'une variable cachée $\mathbf{Z}$ à valeurs dans $\{0,1\}^K$ telle que $\pi_{\alpha_k} = \mathbb{P}(Z_k = 1)$ où $\alpha_k \in \mathbb{M} = \{\alpha_1, \ldots, \alpha_K\}$ et $\sum_{k=1}^K Z_k = 1$.

Un modèle sous-asymptotique est alors proposé, comprenant le modèle de mélange de Dirichlet auquel s'ajoute un modèle pour les résidus non-asymptotiques. Ainsi, pour $\alpha_k \in \mathbb{M}$ et pour $\boldsymbol{v} \in (1,\infty)^d$, la vraisemblance est donnée par:

$$p(\boldsymbol{v}|z_k = 1) = r_k^{-|\alpha_k|-1}\phi_{\alpha_k}(\boldsymbol{w}_k|\boldsymbol{m}_{\alpha_k}, \nu_{\alpha_k}) \prod_{j\notin\alpha} f_\varepsilon(v_j|\lambda_k), \tag{1.46}$$

où $f_\varepsilon(\cdot|\lambda_k)$ est une distribution exponentielle de paramètre λ_k, $r_k = \sum_{j\in\alpha_k} v_j$ et $\boldsymbol{w}_k = (\frac{v_j}{r_k})_{j\in\alpha_k}$.

L'inférence sur le modèle de mélange est alors opérée pour les pseudo-angles $\boldsymbol{W}_\alpha = \frac{\boldsymbol{V}_\alpha}{\|\boldsymbol{V}_\alpha\|}$ et les résidus $\boldsymbol{\varepsilon}_{\alpha^c}$ via un algorithme espérance-maximisation (EM) adapté.

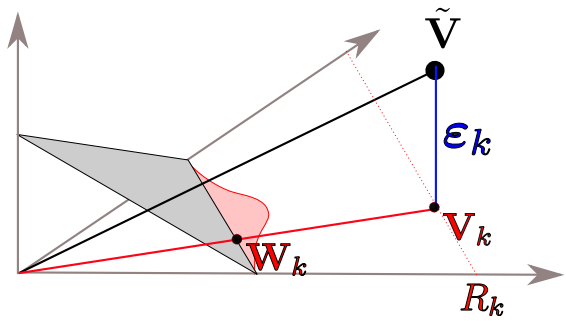


Figure 1.10: Pseudo-angle et résidu.

1.6 Contributions

Articles de conférence avec actes

- Feature clustering for extreme events analysis, with application to extreme stream-flow data. (ECML 2016)
  Authors: M. Chiapino, A. Sabourin.

- A multivariate extreme value theory approach to anomaly clustering and visualization. (Soumis)
  Authors: M. Chiapino, A. Sabourin, S. Clémençon, V. Feuillard.

Article de journal

- Identifying groups of variables with the potential of being large simultaneously. (Soumis)
  Authors: M. Chiapino, A. Sabourin, J. Segers.

1.7 Conclusion

Dans cette thèse différentes stratégies permettant l'exploration de données extrêmes en grande dimension ont été abordées. Pour l'étude de phénomènes extrêmes multivariés, la nécessité de réduire la dimension du problème est d'autant plus grande que la théorie est confrontée à deux difficultés majeures en matière d'inférence. La première est inhérente aux phénomènes étudiés: un évènement extrême est nécessairement rare, et de ce fait, le nombre de données dédié à l'estimation du modèle est limité. La deuxième difficulté vient du caractère limite du modèle d'étude. Après standardisation des distributions marginales, la mesure angulaire caractérise la structure de dépendance de la distribution jointe limite des points extrêmes. Mais cette mesure ne représente que la structure de dépendance **asymptotique** de la variable aléatoire étudiée, et de ce fait les données observées ne sont jamais distribuées selon la loi limite associée. Nous avons exploré deux pistes principales afin de répondre à ce dernier problème.

Dans un premier temps, une méthode d'estimation robuste, non paramétrique, du support de la mesure limite est développée. Le procédé considère séquentiellement des régions emboitées de l'espace de sorte que le plus petit de ces ensembles satisfaisant un certain critère correspond à un ensemble maximal de composantes asymptotiquement dépendantes. Suivant les principes de l'algorithme 'Apriori', nous répondons à l'explosion combinatoire induite par les régions emboitées, par l'algorithme CLEF (1.4.2) qui prouve son efficacité à estimer le support de la structure de dépendance sur des données bruitées et en grande dimension. L'algorithme ouvre une voie générale pour la réduction de dimension dans un contexte d'extrêmes multivariés et permet l'adaptation de différents résumés du degré de dépendance asymptotique (κ_α, η_α, etc.) (1.4.3) afin de déterminer la dépendance des différents groupes de composantes α.

Dans un deuxième temps, une modélisation paramétrique de la structure de dépendance parcimonieuse a été proposée. L'écart (en loi) entre les données observées et la mesure limite est modélisé par un bruit (à queue légère), ce qui permet d'appliquer des méthodes de vraisemblances classiques fondées sur les modèles de mélange. De multiples façons de modéliser cet écart sont possibles et leur étude constitue un vaste champ de recherche. Le choix de la décomposition $\boldsymbol{V}_{\text{observé}} = \boldsymbol{W}R + \textbf{ bruit}$, où $\boldsymbol{W}$ est utilisé pour estimer la mesure angulaire, est motivé par sa relative simplicité d'adaptation à l'algorithme EM, et prouve son efficacité pour la classification non supervisée de données extrêmes.

2 Introduction

2.1 Introduction

Extreme value theory meets the need to build a model for extrapolation over a particular kind of rare phenomena. Given a set of observations, the goal is to foresee how likely future realisations will exceed the ranges previoulsy recorded. More precisely, our goal is to characterize the behavior of random variables above some large thresholds. These considerations are all the more justified when the distribution is heavy-tailed, so that the probability for a point to be extreme is significant. In the univariate case, extreme value theory can be summed up to the study of the maximum values of a process. The asymptotic behavior of the maximum of a random variable is well-known and is characterized within a parametric familly of distributions, the generalized extreme value distribution (De Haan (1970)). In the multivariate case, unlike $\mathbb{R}$, the space is not ordered and therefore it is not possible to define the maximum points of a given dataset. However, by means of thresholds exceedances, componentwise maximum or more generally by using the norm of the vectors, it is possible to provide a meaning to the extreme aspect of some points.

Several difficulties specific to the extreme value theory arise when the dimensionality increases. Usual methods of dimension reduction like principal component analysis (PCA) rely on the study of the covariance matrix. Yet, the covariance between features of heavy-tailed distribution is not defined and therefore such tools are of no use. In addition, on a statistic point of view extreme value theory has an inherent weakness: extreme points are rare by nature thus only a small proportion of the data should be used for the inference.

This introduction is organized as follow. In Section 2.2, the general framework of extreme value theory is exposed, in the univariate 2.2.1 and multivariate cases 2.2.2. In particular, we introduce the *angular measure* Φ which characterizes the dependence structure of any extreme value distribution after standardization of the marginals. A sparse representation of this measure is described in Section 2.3. On this basis, a general method to find the support of Φ is detailed in Section 2.4. In particular, the algorithm CLEF (CLustering Extreme Feature) is developed in 2.4.2, followed by alternative versions of the algorithm based on test statistics 2.4.3. Finally, a parametric model for the angular measure is proposed in Section 2.5.

2.2 Extreme value theory

2.2.1 Univariate extreme value theory

In the univariate case, extreme value theory characterizes the asymptotic behavior of the maximum of random variables valued in $\mathbb{R}$. Let $X \in \mathbb{R}$ be a random variable distributed according to F and $X_1, \ldots, X_n$ $n \geq 1$ *i.i.d.* (independent and identically distributed) copies of X. Suppose that our goal is to estimate large quantiles in the distribution tail, *i.e.* $\mathbb{P}[X > x_p] = p$ with p small. In the case where none of the observed variables lie in $[x_p, \infty)$, the only solution is to build a model based on extrapolation. The standard approach relies on the study of the maximum $M_n := \max\{X_1, \ldots, X_n\}$. Given that the variable X is not bounded, the challenge is reflected by the following limit. For any $x \in \mathbb{R}$:

$$\mathbb{P}[M_n \leq x] = F^n(x) \underset{n\to\infty}{\longrightarrow} 0, \tag{2.1}$$

so the limiting distribution is necessarily degenerate. Usual methods based on the average behavior of the variable and naive empirical approaches are of no use. In order to tackle (2.1) one has to apply some normalization on the maximum M_n to make a non-degenerate limit arise.

The following property is the basis of extreme value theory. It applies to the normalized maximum in a similar way than the central limit theorem (CLT) does with the normalized sum of *i.i.d.* random variables. The main difference is that the existence of the limit in the CLT is guaranteed as soon as the first two moments of the variable are finite. However, assuming the existence of two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ with $a_i > 0$ such that:

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq x\right] = G(x), \tag{2.2}$$

with G a non-degenerate distribution, then the limiting distribution G has necessarily the form:

$$G(x) = \exp\left\{ -\left[1 + \xi\frac{x-\mu}{\sigma}\right]_+^{-\frac{1}{\xi}}\right\}. \tag{2.3}$$

F is said to be in the max-domain of attraction of the generalized extreme value (GEV) distribution G, written $F \in DA(G)$. When $\xi > 0$, G is a heavy-tailed Fréchet, $\xi = 0$ corresponds to a light-tailed Gumbel and $\xi < 0$ to a bounded-tail Weibull. The asymptotic behavior of the maximum is therefore entirely characterized within a parametric familly. Therefore, inferential issues on such models are mostly limited to the proportion of upper data points that has to be considered.

2.2.2 Multivariate extreme value theory

In the multivariate case, identifying extreme points lacks of a proper definition. Indeed when $d \geq 2$, $\mathbb{R}^d$ is not an ordered set. Nevertheless, let $\boldsymbol{x}$ be a vector valued

in $\mathbb{R}^d$ then one can consider $\boldsymbol{x}$ as extreme as soon as $\max_{j\in\{1,\dots,d\}} x_j > t$ or more generaly when $\|\boldsymbol{x}\| > t$ with $t > 0$ a large threshold and $\|\cdot\|$ a given norm. In a similar way than for the univariate case, multivariate extreme value theory is based on the limiting distribution of the standardized componentwise maximum of a random variable.

Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be $n \geq 1$ *i.i.d.* random variables valued in $\mathbb{R}^d$ distributed according to F. Marginal distributions of F are named F_j for $j = 1, \dots, d$ and are assumed to be continuous. Let $\boldsymbol{M}_n := (\max_{i=1,\dots,n} X_{i,1}, \dots, \max_{i=1,\dots,n} X_{i,d})$ be the n-th componentwise maximum. We assume that there exists two sequences $(\boldsymbol{a}_n)_{n\geq 1}$ in $(0,\infty)^d$ and $(\boldsymbol{b}_n)_{n\geq 1}$ in $\mathbb{R}^d$, such that:

$$\lim_{n\to\infty} \mathbb{P}\Big[\frac{\boldsymbol{M}_{n,j} - \boldsymbol{b}_{n,j}}{\boldsymbol{a}_{n,j}} \leq \boldsymbol{x}_j,\ j = 1, \dots, d\Big] = G_0(\boldsymbol{x}), \tag{2.4}$$

where G_0 is a multivariate distribution with non-degenerate marginals $G_{0,j}$, $j = 1, \dots, d$. Then G_0 is characterized by the following formula (Resnick (1987)), for some $\boldsymbol{x}_0 \in \mathbb{R}^d$:

$$G_0(\boldsymbol{x}) = \begin{cases} \exp\Big[-\mu_0\Big([\boldsymbol{x}_0, \boldsymbol{x}]^c\Big)\Big], & \text{for } \boldsymbol{x} \geq \boldsymbol{x}_0, \\ 0 \text{ otherwise}, \end{cases} \tag{2.5}$$

where μ_0 is the so-called *exponent measure*, a Radon measure defined on $[\boldsymbol{x}_0, \boldsymbol{\infty}] \setminus \{\boldsymbol{x}_0\}$ and with $[\boldsymbol{x}_0, \boldsymbol{x}]^c = \cup_{j=1}^d \Big\{\boldsymbol{y} \in [\boldsymbol{x}_0, \boldsymbol{\infty}] \setminus \{\boldsymbol{x}_0\} : y_j > x_j\Big\}$. The multivariate distribution F is said to be in the domain of attraction of G_0, *i.e.* $F \in DA(G_0)$. Also, each marginal $G_{0,j}$ is a univariate extreme value distribution, so that for all $j \in \{1, \dots, d\}$, $F_j \in DA(G_{0,j})$. However, the exponent measure μ_0 and therefore G_0 cannot be characterized within a unique familly of parametric distributions. This is a major difference with the univariate case.

Let $\boldsymbol{X} = (X_1, \dots, X_d)$ be a random variable valued in $\mathbb{R}^d$ according to F. The study of the tail of the distribution F splits into two separate parts. On the one hand, the characterization of the marginals which is related to the univariate case. On the other hand, the study of the dependence structure which is the main purpose of the multivariate case. In an analogous manner than the study of copulas, a natural preliminary step is to standardize each marginal to the same distribution. The choice of the distribution is somewhat arbitrary, though it is common to transform each marginal to a unit Pareto (Resnick (1987)). The standardization is given by the following transformation:

$$V_j := (1 - F_j(X_j))^{-1}, \text{ for } j = 1, \dots, d. \tag{2.6}$$

Remark 2.1. Let $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ be $n \geq 1$ *i.i.d.* copies of $\boldsymbol{X}$. In practice the marginals F_j are usually not known. Thefore, we apply the transformation through the empirical distributions $\hat{F}_j(x) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_{i,j}<x\}}$ where $x \in [x_{0,j}, \infty)$. Let us call

$rank(X_{i',j}) = n - \sum_{i=1}^{n} \mathbb{1}_{\{X_{i,j} < X_{i',j}\}}$. The transformation then becomes, for all $j \in \{1, \dots, d\}$:

$$\begin{aligned}\hat{V}_{i,j} &= (1 - \hat{F}_j(X_{i,j}))^{-1}) \\ &= \frac{n}{rank(X_{i,j})}.\end{aligned}$$

From now on, we work with the transformed random vector $\boldsymbol{V} = (V_1, \dots, V_d)$ valued in $(1, \infty)^d$ with unit Pareto marginals. The existence of the asymptotic distribution G_0 in (2.4) can be reformulated in a suitable way through regular variation. The random variable $\boldsymbol{V}$ is said to be regularly varying on $(1, \infty)^d$ with index -1 and with limiting measure μ defined on $[0, \infty]^d \setminus \{\mathbf{0}\}$ if (*e.g.*Resnick (2013)):

$$t\mathbb{P}\left(t^{-1}\boldsymbol{V} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \tag{2.7}$$

for any Borel set A in $[0, \infty]^d \setminus \{\mathbf{0}\}$ such that $\mu(\partial A) = 0$ and $\mathbf{0} \notin \partial A$. The distribution of $\boldsymbol{V}$ is in the domain of attraction of the multivariate extreme value distribution G, such that for all $\boldsymbol{v} \in [0, \infty]^d \setminus \{\mathbf{0}\}$:

$$G(\boldsymbol{v}) = \exp[-\mu([\mathbf{0}, \boldsymbol{v}]^c)]. \tag{2.8}$$

The link between the asymptotic distribution G_0 and G can be written:

$$G(\boldsymbol{v}) = G_0(G_{0,1}^{\leftarrow}(e^{-1/v_1}), \dots, G_{0,d}^{\leftarrow}(e^{-1/v_d})),$$

where $G_{0,i}^{\leftarrow}$ is the inverse function of $G_{0,i}$.

Remark 2.2. Let $t > 0$ be a large threshold, given that for some $j_0 \in \{1, \dots, d\}$, V_{j_0} is extreme, *i.e.* $V_{j_0} > t$, the function $\boldsymbol{x} \mapsto \mu\Big([\mathbf{0}, \boldsymbol{x}]^c\Big)$ is approximatively, for $\boldsymbol{v} \in [0, \infty)^d \setminus \{\mathbf{0}\}$, the following conditionnal probability:

$$\begin{aligned}\mathbb{P}\Big[V_1 > tv_1 \text{ or } \dots \text{ or } V_d > tv_d | V_{j_0} > t\Big] &= \mathbb{P}\Big[\boldsymbol{V} \in t[\mathbf{0}, \boldsymbol{v}]^c | V_{j_0} > t\Big] \\ &= t\mathbb{P}\Big[\boldsymbol{V} \in t[\mathbf{0}, \boldsymbol{v}]^c\Big] \\ &\approx \mu\Big([\mathbf{0}, \boldsymbol{v}]^c\Big) = -\log G(\boldsymbol{v})\end{aligned}$$

The benefit of the transformation (2.6) along with the regular variation assumption (2.7) is to focus on the dependence structure that is entirely characterized by the exponent measure μ. We recall that the class of such measures is not embedded within a parametric familly. Nevertheless, as a consequence of (2.6) and the max-stable nature of any extreme value distribution (Resnick (1987)), the measure μ is homogeneous of degree -1:

$$\mu(tA) = t^{-1}\mu(A), \tag{2.9}$$

for any $t > 0$ and for all Borel set A bounded away from the origin, *i.e.* $\mathbf{0} \notin \partial A$. One can deduce from (2.9) that the marginals of G are unit Fréchet. Indeed, for any $j \in \{1, \dots, d\}$ and $x_j \in (0, \infty)$:

$$\begin{aligned} G(\infty, \dots, x_j, \dots, \infty) &= \exp\Big[-\mu\Big(([0,\infty] \times \dots \times [0, x_j] \times \dots [0,\infty])^c\Big)\Big] \\ &= \exp\Big[-\frac{1}{x_j}\mu\Big(A\Big)\Big], \end{aligned}$$

with $A = ([0,\infty] \times \dots \times [0,1] \times \dots [0,\infty])^c$. And we get $\mu(A) = 1$ using the limit (2.7) along with the fact that $\mathbf{V}$ has unit Pareto marginals:

$$\begin{aligned} \mu(A) &= \lim_{t\to\infty} t\mathbb{P}\left(t^{-1}\boldsymbol{V} \in A\right) \\ &= \lim_{t\to\infty} t\mathbb{P}\left(V_j > t\right) \\ &= 1. \end{aligned}$$

Remark 2.3. It follows from the homogeneity property (2.9) that the inclusion-wise mininal sets of $[0,\infty]^d \setminus \{\mathbf{0}\}$ with non-zero μ-mass are necessarily cones, *i.e.* $C \subset [0,\infty]^d \setminus \{\mathbf{0}\}$ such that $x \in C \Rightarrow tx \in C$ for all $t > 0$.

The homogeneity property is fundamental in order to characterize the class of exponent measures resulting from the standardization of the marginals. Indeed (2.9) leads to a pseudo-polar decomposition of μ. For all $\boldsymbol{v} \in [0,\infty)^d \setminus \{\mathbf{0}\}$, let us consider the bijective function:

$$T : \boldsymbol{v} \mapsto (\|\boldsymbol{v}\|, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}),$$

where $\|\cdot\|$ is a given norm on $\mathbb{R}^d$. It is then possible to define a measure Φ on the positive orthant of the unit sphere $\mathcal{S}_d = \{\boldsymbol{x} \geq \mathbf{0} : \|x\| = 1\}$, the so-called *angular measure*, which characterizes the dependence structure of $\boldsymbol{V}$:

$$\Phi(A) := \mu(\{\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in A\}), \tag{2.10}$$

for any set $A \subset \mathcal{S}_d$. Considering the homogeneity property of μ (2.9) one can write:

$$\begin{aligned} \mu(\{\boldsymbol{v} : T(\boldsymbol{v}) \in (t,\infty) \times A\}) &= \mu(\{\boldsymbol{v} : \|\boldsymbol{v}\| > t, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in A\}) \\ &= t^{-1}\Phi(A). \end{aligned}$$

Finally, this last equation along with the bijectivity of T imply a factorization of μ in two independent measure (de Haan and Resnick (1977)):

$$\mu \circ T^{-1}(\mathrm{d}r, \mathrm{d}\boldsymbol{w}) = r^{-2}\mathrm{d}r\Phi(\mathrm{d}\boldsymbol{w}). \tag{2.11}$$

Remark 2.4. For large values of $\|\boldsymbol{V}\|$, the angular measure Φ quantifies the relative contribution of each coordinate V_j in the extremal behavior of $\boldsymbol{V}$. One can say that Φ measures the directions of $(1,\infty)^d$ in which $\boldsymbol{V}$ is likely to be extreme. Let us consider the pseudo-polar decomposition $R = \|\boldsymbol{V}\|$ and $\boldsymbol{W} = R^{-1}\boldsymbol{V}$. One can re-write the limit (2.7) depending on Φ:

$$\mathbb{P}\left(\boldsymbol{W} \in A, R > rt \mid R > t\right) \xrightarrow[t\to\infty]{} r^{-1}\Phi(\mathcal{S}_d)^{-1}\Phi(A), \tag{2.12}$$

for all Borel set A in $\mathcal{S}_d$ such that $\Phi(\partial A) = 0$ and $r > 1$. In other words, when the radial part R is large, then R and the pseudo-angle $\boldsymbol{W}$ are approximately independent. The distribution of $\boldsymbol{W}$ is approximately (and up to a normalization) the angular measure and R is approximately distributed according to a unit Pareto.

2

As for the exponent measure, there is no parametric familly that can entirely describe the class of angular measures. The only condition on Φ in order to be a proper angular measure, is the so-called *moment constraint*:

$$\int_{\mathcal{S}_d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) = 1, \, \forall j \in \{1, \ldots, d\}. \tag{2.13}$$

Indeed, using the decomposition (2.11), one can rewrite $\boldsymbol{x} \mapsto \mu([\mathbf{0}, \boldsymbol{x}]^c)$ depending on Φ. Let $\boldsymbol{x}$ be in $[0,\infty]^d \setminus \{\mathbf{0}\}$:

$$\begin{aligned}
\mu([\mathbf{0}, \boldsymbol{x}]^c) &= \int \mathbb{1}_{\{\exists j : u_j > x_j\}} \mu(\mathrm{d}u) \\
&= \int_{\mathcal{S}_d} \int_0^\infty \mathbb{1}_{\{r > \min_j \frac{x_j}{w_j}\}} r^{-2} \mathrm{d}r \Phi(\mathrm{w}) \\
&= \int_{\mathcal{S}_d} (\min_j \frac{x_j}{w_j})^{-1} \Phi(\mathrm{w}) \\
&= \int_{\mathcal{S}_d} \max_j \frac{w_j}{x_j} \Phi(\mathrm{w}).
\end{aligned}$$

Therefore, considering $\boldsymbol{x} = (\infty, \ldots, x_j, \ldots, \infty)$, with $x_j \in (0,\infty)$ we have:

$$\begin{aligned}
\frac{1}{x_j} &= -\log G(\boldsymbol{x}) \\
&= \mu([\mathbf{0}, \boldsymbol{x}]^c) \\
&= \int_{\mathcal{S}_d} \frac{w_j}{x_j} \Phi(\mathrm{d}\boldsymbol{w}).
\end{aligned}$$

The constraint (2.13) is thus a necessary condition on Φ to be a proper angular measure. The proposition 5.11 in Resnick (1987) demonstrates that it is also a sufficient condition. As a straight consequence of the moment constraint (2.13), in the case where the L_1 norm is chosen for $\|\cdot\|$, *i.e.* $\|\boldsymbol{x}\| = |x_1| + \ldots + |x_d|$, we have

$\Phi(\mathcal{S}_d) = d$. Indeed:

$$\begin{aligned}\Phi(\mathcal{S}_d) &= \int_{\mathcal{S}_d} \Phi(\mathrm{d}\boldsymbol{w}) \\ &= \int_{\mathcal{S}_d} \sum_{j=1}^{d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) \\ &= \sum_{j=1}^{d} \int_{\mathcal{S}_d} w_j \, \Phi(\mathrm{d}\boldsymbol{w}) \\ &= d.\end{aligned}$$

2.2.3 Previous works and issues

On a statistic point of view, various models have been proposed for the extreme value distribution G through the modelization of the function $\boldsymbol{x} \mapsto \mu[\mathbf{0}, \boldsymbol{x}]^c$, *i.e.* $G(\boldsymbol{x}) = \exp[-\mu[\mathbf{0}, \boldsymbol{x}]^c]$ (see *e.g.* Coles and Tawn (1991)). Two main strategies arise to infer such models. Methods based on the componentwise maximum are used to estimate G while threshold exceedence methods are used to estimate μ. Componentwise maximum methods face a major issue to estimate the likelihood when the dimensionality increase ($d > 10$). Indeed, one has to compute:

$$\frac{\partial^d}{\partial x_1 \dots \partial x_d} e^{-\mu([\mathbf{0}, \boldsymbol{x}]^c)},$$

which is a sum whose number of terms explodes with the dimension (see *e.g.*Huser et al. (2016)). Some simplifications have been proposed (*e.g.* Wadsworth (2015), Stephenson and Tawn (2005)), but the computation remains infeasible when $d > 10$.

Either for componentwise maximum methods or for threshold exceedence, most of these approaches only consider the case of total dependence between the variables. In other words, the case where the support of μ is infinite on any hyperplane of $[0, \infty]^d \setminus \{\mathbf{0}\}$. Also, in high dimension, non-parametric estimations of these measures are not spared by the general issue of the *curse of dimentionality* which is all the more burdened by the small proportion of points (*i.e.* the extreme points) that should be used for inference. The modelization of the angular measure, which is in theory equivalent to the modelization of the exponent measure, lends itself well to methods aiming at reducing the dimension (see Section 2.3).

Estimation of the angular measure Several parametric and non-parametric approaches have been proposed for the estimation of the angular measure. The class of Dirichlet mixture distributions on the simplex has been proposed in (Boldi and Davison (2007b)) and (Sabourin et al. (2013)). This class of models is dense (in the weak sense) in the class of angular measures. Non-parametric models have been proposed in (Guillotte et al. (2011), Fougeres et al. (2013)). However, experiments are only made in moderate dimensions ($d \approx 5$). Recently proposed works intend to

go through this issue with adapted clustering methods (Chautru (2015)). Finally, an algorithm developped in (Goix et al. (2015a), Goix et al. (2016a)) is used to bring out a sparse support of the exponent measure (and equivalently of the angular measure) in greater dimension ($d \approx 50$).

2.3 Sparse angular measure

In high dimension ($d \approx 100$), it is reasonable to assume that an extreme phenomenon described by a random variable $\boldsymbol{V} = (V_1, \ldots, V_d)$ is not due to all its features simultaneously, *i.e.* $V_j > t$ for all $j \in \{1, \ldots, d\}$ where $t > 0$ is a large threshold. More precisely, we make the assumption that the extreme nature of the phenomenon is only due to some particular subgroups of coordinates. Let us assume that $\|\boldsymbol{V}\|$ is large, then there exists a subgroup $\alpha \subset \{1, \ldots, d\}$ such that $\|\boldsymbol{V}\| \approx \|\boldsymbol{V}_\alpha\|$, where $|\alpha| \ll d$ and $\boldsymbol{V}_\alpha = (V_j)_{j \in \alpha}$. The aim is to bring to the fore the multimodal nature of the extremal behavior of $\boldsymbol{V}$ and thus the plurality of subgroups α that identify the asymptotically dependent coordinates. It is also reasonable to assume that the number of such subgroups is small compared to the total number of subsets of $\{1, \ldots, d\}$, that is $2^d - 1$. More formally, let Φ be the angular measure associated with $\boldsymbol{V}$, then the dependence structure of $\boldsymbol{V}$ is characterized through the distribution of the mass of Φ over the positive orthant of the unit sphere. In the case of a sparse dependence structure, the mass of Φ is only distributed over some particular subspaces of the boundary of $\mathcal{S}_d$, which correspond to asymptotic dependent subgroups of features. Indeed, let us assume that at least two coordinates V_i and V_j are asymptotically independent, which write:

$$t\mathbb{P}\left(V_i > t, V_j > t\right) = t\mathbb{P}\left(t^{-1}\boldsymbol{V} \in \{\boldsymbol{v} : v_i > 1, v_j > 1\}\right) \xrightarrow[t \to \infty]{} 0.$$

Now for all $\boldsymbol{w} \in \mathring{\mathcal{S}}_d$, there exists $\boldsymbol{v} \in [0, \infty)^d \setminus \{\boldsymbol{0}\}$ such that $\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} = \boldsymbol{w}$, $v_i > 1$ and $v_j > 1$, therefore:

$$\begin{aligned}\Phi(\mathring{\mathcal{S}}_d) &= \mu\Big(\{\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in \mathring{\mathcal{S}}_d\}\Big) \\ &\leq \mu\Big(\{\boldsymbol{v} : v_i > 1, v_j > 1\}\Big) = 0.\end{aligned}$$

This toy example shows that asymptotical independence between variables implies that the mass of the angular measure is confined on the boundary of $\mathcal{S}_d$. Let us call $\mathcal{S}_\alpha \subset \mathcal{S}_d$ the subspace associated with the coordinates $\alpha \subset \{1, \ldots, d\}$:

$$\mathcal{S}_\alpha := \{w \in \mathcal{S}_d : \forall j \in \alpha\, w_j > 0, \forall j \notin \alpha\, w_j = 0\}, \tag{2.14}$$

and $\mathcal{M}$ the set of all subgroups of coordinates of $\{1, \ldots, d\}$ that correspond to subspaces with non-zero Φ-mass:

$$\mathcal{M} := \{\alpha \subset \{1, \ldots, d\} : \Phi(\mathcal{S}_\alpha) > 0\}. \tag{2.15}$$

The sparsity assumption can then be stated through the double inequality:

$$\begin{cases} \max_{\alpha\in\mathcal{M}} |\alpha| \ll d, \\ |\mathcal{M}| \ll 2^d - 1. \end{cases} \tag{2.16}$$

2.3.1 Issue on the estimation of $\mathcal{M}$.

Let $\boldsymbol{V}$ be the random variable valued in $(1,\infty)^d$ with unit Pareto marginals and associated with the angular measure Φ. We apply the pseudo-polar decomposition $(R, \boldsymbol{W})$ where $R = \|\boldsymbol{V}\|$ and $\boldsymbol{W} = \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}$. It is essential to point out that the limit (2.12), where Φ is the limiting measure of $\boldsymbol{W}$ for $R > t$ with $t \to \infty$, is only satisfied on continuity sets of Φ, *i.e.* any Borel set A of $\mathcal{S}_d$ such that $\Phi(\partial A) = 0$. In particular, the limit is only established on subspace of $\mathcal{S}_d$ that has a non-zero Lebesgue measure. Therefore it does not apply on subspaces $\mathcal{S}_\alpha$ such that $\Phi(\mathcal{S}_\alpha) > 0$. Indeed, in $\mathcal{S}_d$ we have $\mathcal{S}_\alpha \subset \partial\mathcal{S}_\alpha$, and thus for $\alpha \in \mathcal{M}$:

$$0 < \Phi(\mathcal{S}_\alpha) \leq \Phi(\partial\mathcal{S}_\alpha). \tag{2.17}$$

Furthermore, the sparcity assumption (2.16) implies that the angular measure assigns no mass on the interior of $\mathcal{S}_d$, *i.e.* $\Phi(\mathring{\mathcal{S}}_d) = 0$. On the other side, $\boldsymbol{W}$ is always valued in $\mathring{\mathcal{S}}_d$ *i.e.* $\mathbb{P}(\boldsymbol{W} \in \mathring{\mathcal{S}}_d) = 1$. In this situation, a naive empirical estimation of Φ using an *i.i.d.* sample $(\boldsymbol{W}_i)_{i=1,\dots,n}$ is no longer possible as it was suggested by the remark 2.4.

2.3.1.1 Sparse angular measure for the logistic model

Let us consider the logistic model (Coles and Tawn (1991)) which constitutes, along with the asymmetric logistic model, a very flexible class of parametric extreme value distributions for general dimensions. For all $\boldsymbol{v}$ in $(0,\infty)^d$:

$$\mu_{lgtc,\theta}([\mathbf{0},\boldsymbol{v}]^c) = \left(\sum_{j=1}^{d} v_j^{-\frac{1}{\theta}}\right)^{\theta}, \tag{2.18}$$

where $\theta \in (0,1]$. Asymptotic independence is reached for $\theta = 1$ and in this case the mass of the angular measure is only located on the axes, *i.e.* $\Phi(\boldsymbol{e}^j) = 1$ for $j \in \{1,\dots,d\}$ where $\boldsymbol{e}^j \in \mathbb{R}^d$ such that $\boldsymbol{e}^j_i = 1$ for $i = j$ and $\boldsymbol{e}^j_i = 0$ otherwise. Nonetheless, it is obvious that any point distributed according to the extreme value distribution $G(\boldsymbol{v}) = \exp\Big(-\mu_{lgtc,1}([\mathbf{0},\boldsymbol{v}]^c)\Big)$ is almost surely valued in $(0,\infty)^d$ and, as well, the pseudo-angle almost surely lies in the interior of $\mathcal{S}_d$.

Remark 2.5. The extreme value distribution G and the exponent measure μ are only linked through the equation (2.8), so the support of μ (as well as the support of the angular measure) is not the support of G. Therefore, the fact that the support

of Φ is included in sub-dimensional spaces of $\mathcal{S}_d$ does not mean that the support of G is included in hyperplans of $[0,\infty]^d \setminus \{\mathbf{0}\}$. In other words, even if $\boldsymbol{V}$ was directly distributed according to G, the angular measure only appears as a limit of the conditional probability given that $\|\boldsymbol{V}\|$ goes to infinity.

Modeling Φ instead of G has the advantage to exploit a natural representation of sparse dependence structures. On the other hand, the support we aim to estimate is never reached in the space of the observations (Remark 2.5). Two main ways are considered in the following to tackle this issue. In Chapter 3 and Chapter 4, we approach the sub-parts of interest of the boundary with subspaces of non-zero Lebesgue measure, and we apply a non-parametric estimation of the support of the angular measure (or equivalently of the exponent measure) with a counting measure on these subspaces. In Chapter 5, we propose a parametric mixture model for the angular measure in which each mode corresponds to a subgroup of features asymptotically dependent. In order to infer such a model, we project the observations on the corresponding subspaces and we consider the residual as an orthogonal noise.

2.4 Estimation of the support of the angular measure

In practice, when the only matter is to estimate the support of the angular measure on $\mathcal{S}_d$, it is more suitable to work directly in the definition space of the exponent measure μ, *i.e.* $[0,\infty]^d \setminus \{\mathbf{0}\}$. By definition of Φ in equation (2.10), the counterpart of the subspace $\mathcal{S}_\alpha$ where $\alpha \subset \{1,\dots,d\}$, expressed in terms of subspaces of $[0,\infty]^d \setminus \{\mathbf{0}\}$, is the truncated cone:

$$\mathcal{C}_\alpha = \{\boldsymbol{x} \geq 0 : \|\boldsymbol{x}\| > 1,\ \forall j \in \alpha\, x_j > 0,\ \forall j \notin \alpha\, x_j = 0\}. \tag{2.19}$$

Indeed, for all α in $\{1,\dots,d\}$:

$$\begin{aligned}\Phi(\mathcal{S}_\alpha) &= \mu(\boldsymbol{v} : \|\boldsymbol{v}\| > 1, \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \in \mathcal{S}_\alpha)\\ &= \mu(\mathcal{C}_\alpha).\end{aligned}$$

So that finding the support of Φ amounts to find the support of μ. Note that the set of all the cones $\mathcal{C}_\alpha$ forms a partition of $\{\boldsymbol{v} \in \mathbb{R}_+^d : \|\boldsymbol{v}\| > 1\}$. The figure (2.1) shows a partition of $\mathbb{R}_+^3 \setminus [0,1]^3$ by the cones $\mathcal{C}_\alpha$.

Remark 2.6. By considering the partition $\bigcup_{\alpha \subset \{1,\dots,d\}} \mathcal{C}_\alpha$ we browse over all possible associations of asymptotically dependent coordinates of $\{1,\dots,d\}$. Indeed, let $\alpha \subset \{1,\dots,d\}$ be a subgroup such that a non-null probability is attributed to the event:

{ *All the variables* $(V_j)_{j\in\alpha}$ *are large while* ***all*** *the complementary variables* $(V_j)_{j\notin\alpha}$ *are small.* }

Then there exists a cone $C \subset \mathcal{C}_\alpha$ such that $\mu(C) > 0$. By extension, let $C_{tot} \subset [0,\infty]^d \setminus \{\mathbf{0}\}$ be the cone containing the whole mass of the exponent measure, then C_{tot} is necessarily a union of sub-cones themselves included in some cones $\mathcal{C}_\alpha$.

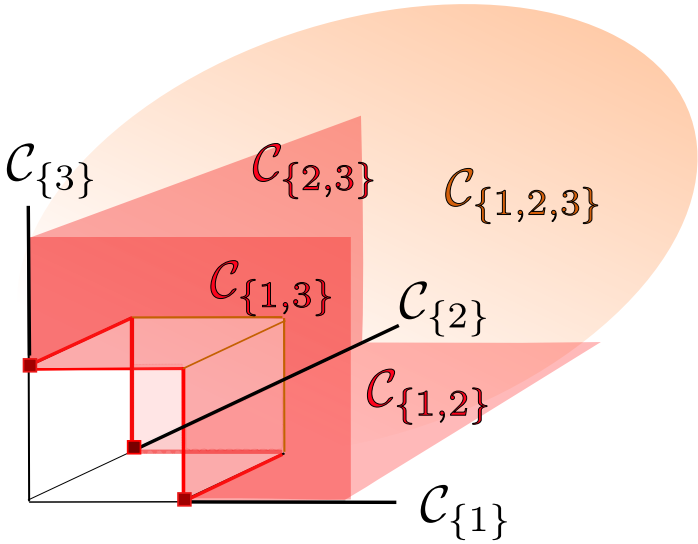


Figure 2.1: Cones for $d = 3$, $\|\boldsymbol{x}\| = \max_{j=1}^{d} x_j$

2.4.1 Inference

The inferential issue over the mass of Φ, explained in Section 2.3.1, transposes to the estimation of the mass of μ on the subspaces of $[0,\infty]^d \setminus \{\mathbf{0}\}$. We consider again the random variable $\boldsymbol{V}$ regularly varying in $(1,\infty)^d$ with limiting exponent measure μ and unit Pareto marginals. As $\boldsymbol{V}$ is almost surely valued in $(1,\infty)^d$, for all $t > 0$ and a strict subset $\alpha \subset \{1,\dots,d\}$:

$$\mathbb{P}\left(\frac{\boldsymbol{V}}{t} \in \mathcal{C}_\alpha\right) = 0. \tag{2.20}$$

Moreover, the cones $\mathcal{C}_\alpha$ are of empty interior and therefore $\mu(\partial\mathcal{C}_\alpha) > 0$. In order to estimate $\mu(\mathcal{C}_\alpha)$ using *i.i.d.* copies of $\boldsymbol{V}$, it is necessary to thicken the truncated cones $\mathcal{C}_\alpha$ in the interior of $[0,\infty]^d \setminus \{\mathbf{0}\}$. When $\|\boldsymbol{V}\|$ is large, the idea is to locate the subspaces of the positive orthant (*i.e.* $\mathcal{C}_\alpha$) toward which $\boldsymbol{V}$ settle down. We therefore consider the following thickened cones, for $\epsilon > 0$ and $\alpha \subset \{1,\dots,d\}$:

$$\mathcal{R}_\alpha^\epsilon = \{\boldsymbol{x} \geq \mathbf{0} : \|\boldsymbol{x}\| > 1, \forall j \in \alpha\, x_j > \epsilon, \forall j \notin \alpha\, x_j \leq \epsilon\}. \tag{2.21}$$

Remark 2.7. Note that for $\epsilon \in (0,1)$ the set of all cones $\mathcal{R}_\alpha^\epsilon$ remains a partition of $[\mathbf{0},\mathbf{1}]^c$ (for the L_∞-norm). Assuming that $\mu(\mathcal{C}_\alpha) > 0$, the inclusion-wise minimal set $C \subset \mathcal{C}_\alpha$ such that $\mu(C) > 0$, is a cone (Remark 2.3). Yet, any cone included in $\mathcal{C}_\alpha$ is necessarily of non-empty intersection with $\mathcal{R}_\alpha^\epsilon$. Therefore, the thickened cone $\mathcal{R}_\alpha^\epsilon$ satisfies $\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \mu(\mathcal{R}_\alpha^\epsilon) \geq \mu(C) > 0$ and is furthermore of non-zero Lebesgue measure.

More formally, as a consequence of the continuity from above of any measure:

$$\mu(\mathcal{C}_\alpha) = \mu(\cap_{\epsilon>0,\epsilon\in\mathbb{Q}}\mathcal{R}_\alpha^\epsilon) = \lim_{\epsilon\to 0}\mu(\mathcal{R}_\alpha^\epsilon).$$

However, in order to apply the limit (2.7) to the sets $\mathcal{R}_\alpha^\epsilon$, they must verify $\mu(\partial\mathcal{R}_\alpha^\epsilon) = 0$. The boundary of $\mathcal{R}_\alpha^\epsilon$ in $[0,\infty]^d \setminus \{\mathbf{0}\}$ is composed of hyperplans parallel to the axes. For different values of $\epsilon \in (0,1)$ the boundaries are disjoint. Therefore, there is at most a countable number of ϵ such that $\mu(\partial\mathcal{R}_\alpha^\epsilon) > 0$, otherwise μ would be infinite on a compact set of $[0,\infty]^d \setminus \{\mathbf{0}\}$. It is thus possible to choose a threshold $\epsilon > 0$ arbitrarily close to zero such that $\mu(\partial\mathcal{R}_\alpha^\epsilon) = 0$, and estimate the quantity $\mu(\mathcal{C}_\alpha)$ through the limit:

$$\mu(\mathcal{C}_\alpha) = \lim_{\epsilon\to 0}\lim_{t\to\infty} t\mathbb{P}\left(\boldsymbol{V} \in t\mathcal{R}_\alpha^\epsilon\right).$$

An algorithm named DAMEX (Goix et al. (2016a)) have been developed to estimate the mass $\mu(\mathcal{C}_\alpha)$ by a counting method on the subspaces $t\mathcal{R}_\alpha^\epsilon$ for a small $\epsilon > 0$ and a large threshold $t > 0$. More precisely, let $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ be $n \geq 1$ *i.i.d.* copies of the random variable $\boldsymbol{V}$. The estimator of $\mu(\mathcal{C}_\alpha)$ is defined as:

$$\hat{\mu}_\alpha = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k}\mathcal{R}_\alpha^\epsilon\}}, \tag{2.22}$$

where t is replaced by $\frac{n}{k}$, with $k \leq n$ an order of magnitude of the number of extreme points considered in the sample. A low threshold $\mu_{min} > 0$ is then chosen in order to decide:

$$\mu(\mathcal{C}_\alpha) > 0 \text{ if } \hat{\mu}_\alpha \geq \mu_{min}.$$

The algorithm is tested on real data (a wave directions dataset from the North Sea provided by Shell) and achieves to reduce drastically the dimensionality of the problem by gathering most of the mass of μ in a limited number of cones $\mathcal{C}_\alpha$ of low dimensions. Similarly, the method shows good results on simulated data. A dataset is generated from a sparse version of the asymmetric logistic model (Tawn (1990a)):

$$G(\boldsymbol{z}) = \exp\left[-\sum_{\alpha\in\mathcal{M}}\left\{\sum_{j\in\alpha}(|\mathcal{A}(j)|z_j)^{-1/\theta_\alpha}\right\}^{\theta_\alpha}\right], \tag{2.23}$$

where $\mathcal{A}(j) = \{\alpha \in \mathcal{M} : j \in \alpha\}$ and with $\theta_\alpha \in (0,1]$ a dependence parameter such that $\theta_\alpha = 1$ corresponds to asymptotic independence and $\theta_\alpha \downarrow 0$ to total dependence. The algorithm easily recovers the support $\mathcal{M}$ of the dependence structure for strong asymptotic dependence, *i.e.* $\theta_\alpha = 0.1$. However, when the dimension d, θ_α and the number $|\mathcal{M}|$ of dependent subgroups increase, the method is no longer able to recover $\mathcal{M}$. Due to the partition of the space in a huge number of subcones $(2^d - 1)$, the model is inherently sensible to any variability of the subgroups of coordinates involved in the extremal behavior of $\boldsymbol{V}$.

Indeed, let α_0 be a subgroup of coordinates such that $\mu(\mathcal{C}_{\alpha_0}) > 0$. Then it is possible that for several coordinates $j \in \{1,\ldots,d\} \setminus \alpha_0$ the quantity $\hat{\mu}_{\alpha_0\cup\{j\}}$ is significant:

$$\hat{\mu}_{\alpha_0\cup\{j\}} = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k}\mathcal{R}_{\alpha_0\cup\{j\}}^\epsilon\}} \geq \mu_{min}. \tag{2.24}$$

As a result, the mass is scattered into clusters of sub-cones that are close to each other and it is no longer possible to recover the sparse dependence structure. It is therefore necessary to build a more robust method to estimate the support of μ in the cases of moderate and low dependence between the components (*i.e.* $\theta_\alpha \in [0.5, 1)$) or more generaly in order to tackle noisy datasets. In order to exhibit this scattering of the mass, we generate several dataset $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ with $n = 1e5$, from an asymmetric logistic distribution for different levels of dependency θ_α. We set $d = 50$ and we randomly generate 50 subgroups of coordinates to form $\mathcal{M}$. The following figures (2.2) correspond to the different datasets and show the repartition of extreme points in the cones $\frac{n}{k}\mathcal{R}_\alpha^\epsilon$ where k is chosen such that the proportion of extreme points $\sum_{i=1}^n \mathbb{1}_{\{\|\boldsymbol{V}_i\| > \frac{n}{k}\}}$ represents 5% of the dataset. The x-axis represents the subgroups of coordinates α such that $\hat{\mu}_\alpha > 0$. The y-axis represents the values of the estimators $\hat{\mu}_\alpha$. The elements of $\mathcal{M}$ are represented in red. In order to recover $\mathcal{M}$ the threshold $\mu_{min} > 0$ must verify:

$$\begin{cases} \hat{\mu}_\alpha \geq \mu_{min}, & \text{for all } \alpha \in \mathcal{M} \\ \hat{\mu}_\alpha < \mu_{min}, & \text{otherwise.} \end{cases} \tag{2.25}$$

For $\theta_\alpha = 0.1$ (2.2 up-left) there is a big gap between the 'true' subgroups (*i.e.* in $\mathcal{M}$) and the others. Thus, it is possible to chose $\mu_{min} > 0$ such that (2.25) is verified. When θ_α increases, corresponding to lower asymptotic dependences, we observe a scattering of the extreme points. For $\theta_\alpha = 0.7, 0.9$ (2.2 bottom-left (resp. right)) it is no longer possible to recover $\mathcal{M}$.

2.4.2 CLustering Extreme Feature Algorithm

The inner weakness of the partition of the space into cones is that it leads to differenciate close extreme events. Instead, we decide to only look for subgroups $\alpha \subset \{1, \ldots, d\}$ such that V_j is large for all $j \in \alpha$, regardless of the counterparts $V_{j'}$ for $j' \in \{1, \ldots, d\} \setminus \alpha$. We define the nested rectangles:

$$\Gamma_\alpha = \{\boldsymbol{x} : \forall j \in \alpha,\, x_j > 1\}. \tag{2.26}$$

One can show that $\mu(\partial\Gamma_\alpha) = 0$, just as for the cones $\mathcal{R}_\alpha^\epsilon$, and thus we can use the limit (4.1) to estimate their mass. The advantage of this approach is to avoid the scattering of the empirical mass of μ as in (2.24). For all $\alpha \subset \{1, \ldots, d\}$, Γ_α corresponds to the event:

{ *All the variables* $(V_j)_{j\in\alpha}$ *are large (**regardless** of their counterpart* $(V_j)_{j\notin\alpha}$*).* }

Let $t > 0$ be a high threshold, each rectangle $t\Gamma_\alpha$ gathers extreme points with large coordinates in α regardless of the complementary coordinates $\{1, \ldots, d\} \setminus \alpha$. One simple but important property is that any point in $t\Gamma_\alpha$ also belongs to $t\Gamma_\beta$ for all $\beta \subset \alpha$. The goal is to find, in an incremental way, the larger subgroups α such that the number of points in $t\Gamma_\alpha$ is still significant.

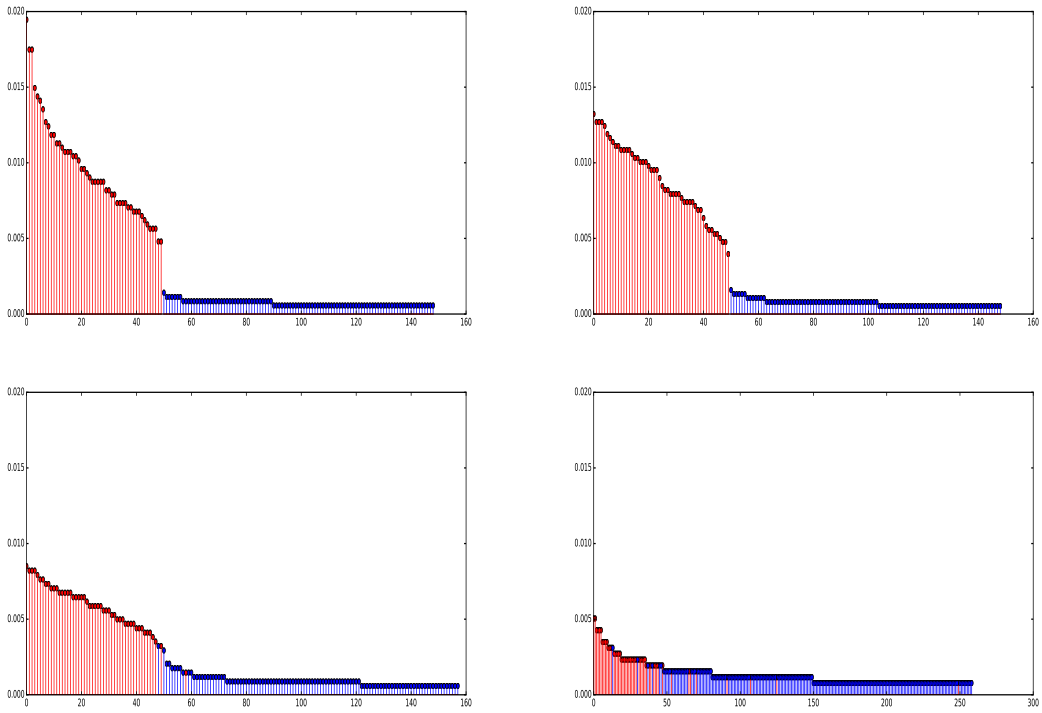


Figure 2.2: Distribution of the empirical mass of μ on the cones $\mathcal{C}_\alpha$ for $\theta_\alpha \in \{0.1, 0.5, 0.7, 0.9\}$ (top-down, left-right).

When we consider the inclusion-wise maximal subgroups $\alpha \subset \{1, \ldots, d\}$ such that $\mu(\mathcal{C}_\alpha) > 0$, the new framework becomes equivalent to the partition of the space into cones. On the one hand, let us assume that μ has a non-zero mass on $\mathcal{C}_\alpha$. The inclusion-wise minimal region $C \subset \mathcal{C}_\alpha$ such that $\mu(C) > 0$ is a cone (2.3). As any cone included in $\mathcal{C}_\alpha$ has a non-empty intersection with Γ_α, $\mu(\mathcal{C}_\alpha) > 0$ implies $\mu(\Gamma_\alpha) > 0$. On the other hand, let us assume $\mu(\Gamma_\alpha) > 0$. If for all β such that $\alpha \subset \beta$, $\mu(\Gamma_\beta) = 0$ then necessarily $\mu(\mathcal{C}_\alpha) > 0$. By contraposition, if we assume $\mu(\Gamma_\alpha) > 0$ and $\mu(\mathcal{C}_\alpha) = 0$, then there exists $j_C \in \{1, \ldots, d\} \setminus \alpha$ and a cone C included in $\Gamma_{\alpha \cup \{j_C\}}$ such that $\mu(C) > 0$. In other words, there exists β such that $\alpha \subset \beta$ and $\mu(\Gamma_\beta) > 0$.

Hence, maximal subgroups $\alpha \subset \{1, \ldots, d\}$ such that $\mu(\mathcal{C}_\alpha) > 0$, are also maximal subgroups such that $\mu(\Gamma_\alpha) > 0$. Let $\mathbb{M}$ be the set of maximal subgroups $\alpha \subset \{1, \ldots, d\}$ such that $\mu(\Gamma_\alpha) > 0$, then we have the following property:

$$\alpha \in \mathbb{M} \Leftrightarrow \alpha \text{ is maximal in } \mathcal{M}. \tag{2.27}$$

The counterpart of this approach is the implied combinatorial explosion due to the nested rectangles. Indeed, the maximal number of cones $\mathcal{C}_\alpha$ to estimate is at most the number of extreme points. On the contrary, the potential number of rectangles Γ_α that might be browsed is much bigger, *i.e.* $\mathrm{O}(2^{d_0} - 1)$ for $d_0 \leq d$. This issue arises from the fact that $\bigcup_{\alpha \subset \{1,\ldots,d\}} \Gamma_\alpha$ is no longer a partition of the space.

2.4.2.1 'Apriori' Algorithm

More generaly this exploratory issue can be reformulate in the following manner: we want to find all subgroups $\mathcal{I}$ of a set of indexes $\mathcal{T}$ that verify a condition $C(\mathcal{I})$. The essential property is that for all $\mathcal{I}_1 \subset \mathcal{I}_2$ then $C(\mathcal{I}_2)$ implies $C(\mathcal{I}_1)$. This problem is directly linked to the Apriori algorithm introduced in (Agrawal et al. (1994)). The algorithm reduces drastically the browse of all subset of $\mathcal{T}$ with an incremental procedure.

Applied to our framework, the procedure browses subsets $\alpha \subset \{1, \ldots, d\}, |\alpha| \geq 2$, of increasing sizes and only keeps those that verify a criterion $C(\alpha)$ equivalent to $\mu(\Gamma_\alpha) > 0$. The output of the algorithm is thus:

$$\mathbb{M}_0 := \Big\{\alpha \subset \{1, \ldots, d\} : |\alpha| \geq 2,\, C(\alpha)\Big\}. \tag{2.28}$$

In order to get $\mathbb{M}$, a postprocessing step is done to only keep the maximal elements of $\mathbb{M}_0$ for inclusion. This process can be visualized through the Hasse diagram (2.3). We explore the diagram top to bottom and prune all the edges of a node as soon as the node does not verify the criterion C. This method reduces notably the potential number of subgroups to be tested. For the criterion $\mu(\Gamma_\alpha) > 0$, the pruning of the graph follows the property:

$$\text{Let } \alpha \text{ be such that } \mu(\Gamma_\alpha) = 0, \text{ then for all } \beta \supset \alpha,\ \mu(\Gamma_\beta) = 0. \tag{2.29}$$

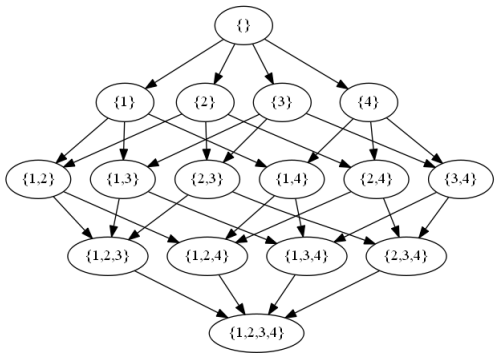


Figure 2.3: Hasse diagram with 4 components.

At step $s \geq 3$, given all subgroups α of size $s-1$ such that $C(\alpha)$ is verified, the algorithm builds the subgroups $\beta \subset \{1,\dots,d\}$ of size s that are likely to verify $C(\beta)$. Let $\mathbb{M}_0^{s-1} = \Big\{\alpha \subset \{1,\dots,d\} : |\alpha| = s-1 \text{ and } C(\alpha)\Big\}$ be the set of subgroups of size $s-1$ that verify C. The set of the candidate subgroups of size s for $s \geq 3$ is:

$$\mathbb{A}^s = \Big\{\beta \subset \{1,\dots,d\} : |\beta| = s \text{ and } \beta \setminus \{j\} \in \mathbb{M}_0^{s-1} \text{ for all } j \in \beta\Big\}. \tag{2.30}$$

For $s = 2$, we put $\mathbb{A}^2 = \Big\{\alpha \subset \{1,\dots,d\} : |\alpha| = 2\Big\}$.

2.4.2.2 Stopping criteria

Let $(\boldsymbol{V}_1,\dots,\boldsymbol{V}_n)$ be an *i.i.d.* sample of random variables valued in $(1,\infty)^d$ with limiting angular measure Φ. In order to decide whether or not $\alpha \subset \{1,\dots,d\}$ belongs to $\widehat{\mathbb{M}}_0$, one natural criterion would be:

$$\widehat{\gamma}_\alpha = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}_{\{\boldsymbol{V}_i \in \frac{n}{k}\Gamma_\alpha\}} \geq \gamma_{min}, \tag{2.31}$$

with $\gamma_{min} > 0$ a small threshold and $k \leq n$ an order of magnitude of the number of extreme points considered.

However, as the rectangles are nested, the value of $t\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] \approx \mu(\Gamma_\alpha)$, with $t > 0$, is necessarily decreasing for increasing sizes of α. Let us consider $\alpha_1 \subset \alpha_2 \subset \{1,\dots,d\}$, so that $\Gamma_{\alpha_2} \subset \Gamma_{\alpha_1}$ and thus $t\mathbb{P}[\boldsymbol{V} \in t\Gamma_{\alpha_2}] \leq t\mathbb{P}[\boldsymbol{V} \in t\Gamma_{\alpha_1}]$ for all $t > 0$. The choice of the threshold should therefore depends on the size of the subsets considered. To tackle this issue, an alternative summary of the dependency degree between the coordinates $\alpha \subset \{1,\dots,d\}$ is proposed:

$$\kappa_\alpha := \frac{\mu(\Gamma_\alpha)}{\mu(\bigcup\limits_{\substack{\beta \subset \alpha \\ |\beta| = |\alpha|-1}} \Gamma_\beta)}. \tag{2.32}$$

The quantity κ_α does not necessarily decrease with the size of α, and we have $\kappa_\alpha > 0$ if and only if $\mu(\Gamma_\alpha) > 0$. It is then possible to chose a threshold $\kappa_{min} > 0$, that

does not depend on the size of α, in order to get a criterion $\{\hat{\kappa}_\alpha \geq \kappa_{min}\}$ to decide $\{\kappa_\alpha > 0\}$. The idea is to make the acceptation of a group α in $\mathbb{M}_0$ dependent on the subgroups of α:

$$\begin{aligned}\kappa_\alpha = \mathbb{P}[\ &\text{All the variables } V_j \text{ for } j \in \alpha \text{ are large } | \\ &\text{All variables are large except at most one }]\end{aligned}$$

The estimator $\hat{\kappa}_\alpha$ of the quantity (2.32) is then:

$$\hat{\kappa}_\alpha = \frac{\hat{\gamma}_\alpha}{\hat{\mu}(\bigcup_{\substack{\beta \subset \alpha \\ |\beta| = |\alpha|-1}} \Gamma_\beta)}. \tag{2.33}$$

The algorithm that estimates $\mathbb{M}$ is described in (2).

Algorithm 2 CLEF (CLustering Extreme Features)

Input: Threshold $\kappa_{\min} > 0$.

STAGE 1: Construct the set $\widehat{\mathbb{M}}_0$ of tail-dependent groups.
Step 1: $\widehat{\mathbb{M}}_0^2 = \left\{\alpha \subset \{1, \ldots, d\} : |\alpha| = 2,\ \hat{\kappa}_\alpha > \kappa_{\min}\right\}$, $S = 2$.
Step $s = 3, \ldots, d$: If $\widehat{\mathbb{M}}_0^{s-1} = \emptyset$, end **PHASE 1**. Otherwise:

- Generate candidate of size s:
  $\mathbb{A}^s = \{\alpha \subset \{1, \ldots, d\} : |\alpha| = s \text{ and } \alpha \setminus j \in \widehat{\mathbb{M}}_0^{s-1} \text{ for all } j \in \alpha\}$.
- Keep all subgroups that verify the criterion, $\widehat{\mathbb{M}}_0^s = \left\{\alpha \in \mathbb{A}^s : \hat{\kappa}_\alpha > \kappa_{\min}\right\}$.
- Si $\widehat{\mathbb{M}}_0^s \neq \emptyset$, $S = s$.

Sortie: $\widehat{\mathbb{M}}_0 = \emptyset$ if $S = 1$ and $\widehat{\mathbb{M}}_0 = \bigcup_{s=2}^{S} \widehat{\mathbb{M}}_0^s$ if $S \geq 2$.

STAGE 2: Only keep maximal α.
If $S = 1$, then $\widehat{\mathbb{M}} = \emptyset$. Otherwise:
Initialization: $\widehat{\mathbb{M}} \leftarrow \widehat{\mathbb{M}}_0^s$.
for $s = (S-1) : 2$,
  for $\alpha \in \widehat{\mathbb{M}}_0^s$,
    If there is no $\beta \in \widehat{\mathbb{M}}$ such that $\alpha \subset \beta$, then $\widehat{\mathbb{M}} \leftarrow \widehat{\mathbb{M}} \cup \{\alpha\}$.
Output: $\widehat{\mathbb{M}}$

2.4.2.3 Results

In practice the choice of a criterion based on κ_α is all the more justified when noise is observed on the components. In order to reproduce a dataset that can be assimilated to a real-world noisy dataset, we consider again the asymmetric logistic model described in 2.4.1, for $d = 20$, $|\mathbb{M}| = 15$ and $\theta_\alpha = 0.5$, on which we artificially disturbe

the dependence structure. More precisely, each points $V_i \in (0,\infty)^d, i = 1,\dots,n$, is generated with a slightly modified version $\mathbb{M}^i$ of the subgroups of dependent coordinates $\mathbb{M}$. For $i = 1,\dots,n$, each subgroup of coordinates α_k^i in $\mathbb{M}^i$ is randomly altered from the original α_k, so that there exists $j_0 \in \{1,\dots,d\}$ such that:

$$\{j_0 \in \alpha_k \text{ and } j_0 \notin \alpha_k^i\} \text{ or } \{j_0 \notin \alpha_k \text{ and } j_0 \in \alpha_k^i\}. \tag{2.34}$$

The goal is to recover $\mathbb{M}$ from the disturbed dataset. At each step $s = 2,\dots,S$, with $S = \max_{\alpha\in\mathbb{M}} |\alpha|$, it is therefore necessary to distinguish the following subgroups, using a threshold $\kappa_{min} > 0$:

$$\hat{\kappa}_\alpha \geq \kappa_{min} \text{ for all } \alpha \in \mathbb{M}_0^s$$
$$\hat{\kappa}_\alpha < \kappa_{min} \text{ for all } \alpha \in \mathbb{A}^s \setminus \mathbb{M}_0^s.$$

We call $\mathbb{M}^F = \underset{s=1,\dots,S}{\cup} \mathbb{A}^s \setminus \mathbb{M}_0$ the set of all candidates that will not verify the criterion. Note also that $\mathbb{M}_0 = \cup_{s=2}^S \mathbb{M}_0^s$. The condition for the existence of a threshold $\kappa_{min} > 0$ such that $\widehat{\mathbb{M}} = \mathbb{M}$ is therefore:

$$\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha < \min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha. \tag{2.35}$$

If (2.35) is verified, it is possible to choose κ_{min} in the interval $(\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha, \min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha)$ to recover $\mathbb{M}$. Figures (2.4) (resp. (2.5)) displays the value of the empirical mass $\hat{\gamma}_\alpha$ (resp. $\hat{\kappa}_\alpha$) on the elements of $\mathbb{M}_0$ (in red) and $\mathbb{M}^F$ (in blue). The blue line represents $\max_{\alpha\in\mathbb{M}^F} \hat{\gamma}_\alpha$ (resp. $\max_{\alpha\in\mathbb{M}^F} \hat{\kappa}_\alpha$) while the red line represents $\min_{\alpha\in\mathbb{M}_0} \hat{\gamma}_\alpha$ (resp. $\min_{\alpha\in\mathbb{M}_0} \hat{\kappa}_\alpha$). The red subgroups are ranked by increasing sizes and it is clear that the empirical mass (2.31) decreases with the size of the subsets. Moreover, as $\max_{\alpha\in\mathbb{M}^F} \hat{\gamma}_\alpha > \min_{\alpha\in\mathbb{M}_0} \hat{\gamma}_\alpha$, it is not possible to find a threshold $\gamma_{min} > 0$ whereby we can recover $\mathbb{M}$.

We then compare the algorithm DAMEX along with the algorithm CLEF for the different criterions $C(\alpha) = \{\hat{\gamma}_\alpha > \gamma_{min}\}$ and $C(\alpha) = \{\hat{\kappa}_\alpha > \kappa_{min}\}$ on a dataset of daily recorded river flows. The dataset comes from $d = 92$ stations spread accross the different rivers of France, recorded from january 1st 1969 to december 31st 2008. It is composed of $n = 14610$ vectors $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ valued in $\mathbb{R}_+^d$. The goal is to find the subgroups of stations that are likely to be concomitantly flooded. After the transformation (2.6), we consider the extreme points $(\hat{\boldsymbol{V}}_1,\dots,\hat{\boldsymbol{V}}_{n_{extr}})$ where $n_{extr} = \sum_{i=1}^n \mathbb{1}_{\|\hat{\boldsymbol{V}}_i\|>\frac{n}{k}}$ with k chosen such that $n_{extr} = n \cdot 5\%$. Figures (2.6), (2.7) and (2.8) respectively show the results of the DAMEX algorithm, and the CLEF algorithm with the empirical mass (2.31) and with $\hat{\kappa}_\alpha$ (2.32), for the respective thresholds $\mu_{min} = 0.002$, $\gamma_{min} = 0.2$ and $\kappa_{min} = 0.2$.

Figure (2.6) displays $\widehat{\mathcal{M}}$ and the figures (2.7) and (2.8) display $\widehat{\mathbb{M}}$. Each subgroup of dependent coordinates is represented by the convex hull of the corresponding stations. Each color corresponds to a specific size of subgroup. We see that unlike the empirical mass (2.31) and the DAMEX algorithm, the criterion $\{\hat{\kappa}_\alpha \geq \kappa_{min}\}$ allows to reach large subgroups of stations without being disturbed by small subgroups of stations.

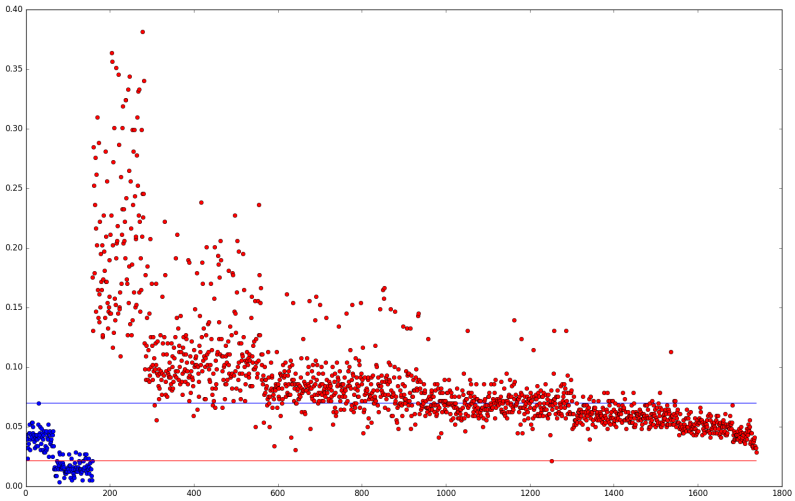


Figure 2.4: Value of $\hat{\gamma}_\alpha$ on the elements of $\mathbb{M}^F$ and $\mathbb{M}_0$.

2.4.3 Coefficient of tail dependence

The purpose of this section is to build a serie of test statistics based on the hypothesis $\mu(\Gamma_\alpha) > 0$. Using the criterion $\kappa_\alpha > 0$ raises an issue. Indeed, as soon as $\mu(\Gamma_\alpha) = 0$ the limiting distribution of the statistics $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha)$ is degenerate. Thus, we have no control of the asymptotic levels of the tests. Let us consider the tail dependence coefficient $\eta_\alpha \in (0, 1]$ introduced in the bivariate case in (Ledford and Tawn (1996)) and extended to general dimensions $d \geq 3$ in (De Haan and Zhou (2011)) et (Eastoe and Tawn (2012)). The fundamental assumption stipulates the existence of $\eta_\alpha \in (0, 1]$ and a slowly varying function $\mathcal{L}_\alpha$ such that:

$$\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] = t^{-1/\eta_\alpha}\mathcal{L}_\alpha(t). \tag{2.36}$$

Under the assumption that both the limit $\lim_{t\to\infty} t\mathbb{P}[\boldsymbol{V} \in t\Gamma_\alpha] = \mu(\Gamma_\alpha)$ exists and (2.36) is verified, $\mu(\Gamma_\alpha) > 0$ implies $\eta_\alpha = 1$. On the contrary, suppose (2.36) and $\liminf_{t\to\infty} \mathcal{L}_\alpha(t) > 0$ then $\eta_\alpha = 1$ implies $\mu(\Gamma_\alpha) > 0$. In other words, the null hypothesis $\mu(\Gamma_\alpha) > 0$ corresponds to the hypothesis $\eta_\alpha = 1$ under mild conditions on $\mathcal{L}_\alpha$. Hence, if $\eta_\alpha = 1$ the limit $\sqrt{k}(\widehat{\eta}_\alpha - \eta_\alpha)$ is non-degenerate and it is possible to control the asymptotic levels of the associated test.

A new criterion $C(\alpha)$ based on the estimator of the coefficient of tail dependence η_α is used in the algorithm CLEF. This series of test statistics is developed in Chapter 4 for different non-parametric estimators of η_α, namely a multivariate extention of the Peng estimator (Peng (1999)) along with the Hill estimator (Draisma et al. (2001), Draisma et al. (2004)). A non-degenerate version of the test $H_0 : \kappa_\alpha = 0$ is also developed by adding a threshold $\kappa_{min} > 0$, to get the new hypothesis $H_0 : \kappa_\alpha > \kappa_{min}$.

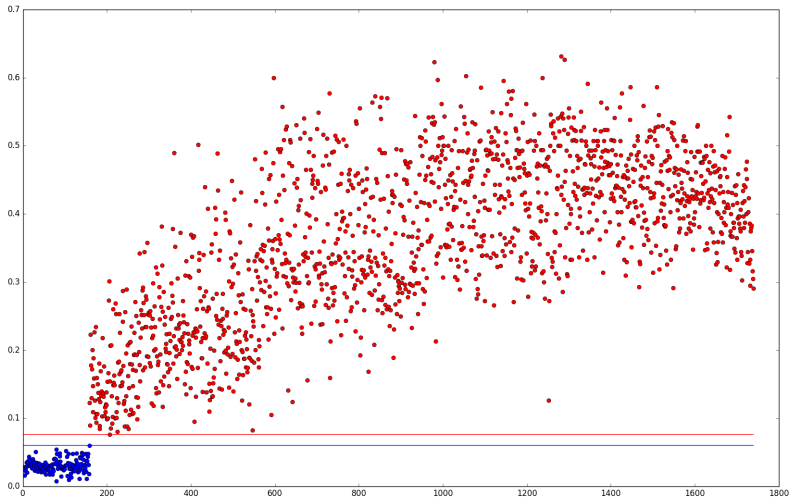


Figure 2.5: Value of $\hat{\kappa}_\alpha$ on the elements of $\mathbb{M}^F$ and $\mathbb{M}_0$.

Similar experiments than in (2.4.2.3) are made. A dataset is generated from the asymmetric logistic model with artificial noise, with the same initial parameters $d = 20$, $|\mathbb{M}| = 15$ and $\theta_\alpha = 0.5$ than for Figures 2.4 and 2.5. The Hill estimator is tested through a new criterion of acceptance for a subgroup $\alpha \subset \{1, \dots, d\}$, on which we have a control of the asymptotic level:

$$\hat{\eta}_\alpha > 1 - q_{1-\delta} \frac{\hat{\sigma}_\alpha}{\sqrt{k}},$$

where $q_{1-\delta}$ is the $(1-\delta)$-quantile of the normal distribution, $\hat{\sigma}_\alpha$ is the variance of the estimator and k is an order of magnitude of the number of extreme points.

Figure 2.9 displays the values of the quantity $\hat{\eta}_\alpha - (1 - q_{1-\delta}\frac{\hat{\sigma}_\alpha}{\sqrt{k}})$ over elements of the sets $\mathbb{M}^F$ and $\mathbb{M}_0$ for a low value $\delta = 1e-7$. We see that the quantity of interest is strictly positive for any elements of $\mathbb{M}_0$ (in red) and strictly negative on the elements of $\mathbb{M}^F$ (in blue).

2.5 Parametric modeling of the angular measure

The motivation of this part is to detect and classify anomalies of a flight dataset provided by Airbus. The general approach of anomaly detection is to build a model that describes the 'normal' behavior of a phenomenon. On this basis, some events are defined as 'anomalies' based on the deviation they have with the estimated model of the 'normal' behavior. We propose a different approach, using the extreme value theory framework, by assimilating the 'anomalies' with the extreme events. This method allows to apply clustering on the 'anomalies'.

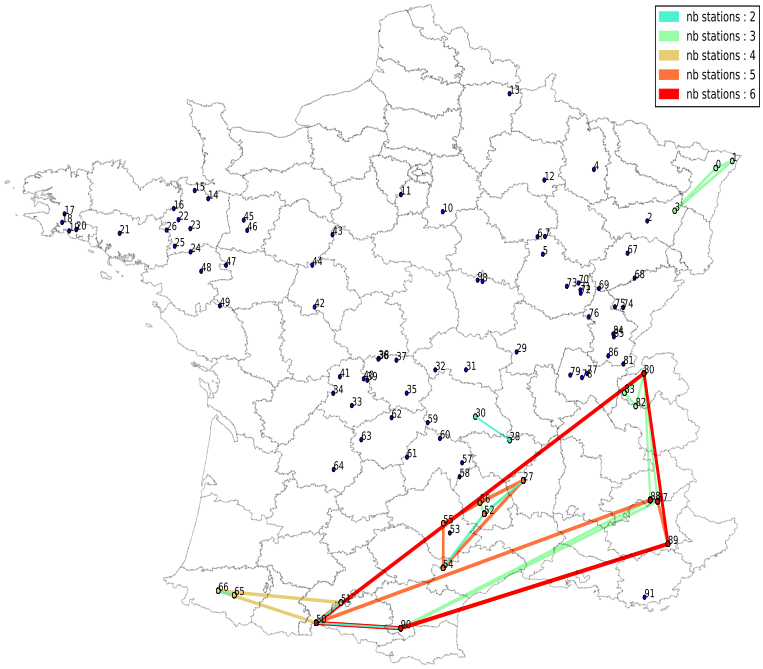


Figure 2.6: Representation of $\widehat{\mathcal{M}}$ for the algorithm DAMEX with $\mu_{min} = 0.002$.

To that end, once we have estimated the support of the dependence structure $\mathcal{M}$, a parametric mixture model is proposed for the angular measure Φ. This modelization has two benefits. First, it allows to assign the points to subgroups of features in a probabilistic manner. Secondly, it is possible to build a similarity matrix $\boldsymbol{S} = (s_{i,j})_{i,j=1,\dots,n}$, where n is the number of extreme points, such that:

$$s_{i,j} = \mathbb{P}[\text{ Points } i \text{ and } j \text{ are assign to the same subgroup of features. }] \tag{2.37}$$

Methods of spectral clustering are then applied on the similarity matrix $\boldsymbol{S}$ in order to cluster the extreme points.

For the sake of simplicity we assume here that $\min_{\alpha\in\mathcal{M}} |\alpha| > 1$ along with $\mathbb{M} = \mathcal{M}$, the general case being developed in Section 5. Suppose that $\boldsymbol{z} \mapsto \mu([\mathbf{0}, \boldsymbol{z}]^c)$ is d-times differentiable on $[0,\infty)^d \setminus \{0\}$, then Φ has densities on the interior of the simplex $\mathcal{S}_d$ ($\|\cdot\|$ being the L_1 in this part) and on each sub-simplex $\mathcal{S}_\alpha$ (see Theorem 1 in Coles and Tawn (1991)). We can thus apply the decomposition:

$$\frac{1}{\Phi(\mathcal{S}_d)}\Phi(\mathrm{d}\boldsymbol{w}) = \sum_{\alpha\in\mathbb{M}} \pi_\alpha \phi_\alpha(\boldsymbol{w}_\alpha)\mathrm{d}\boldsymbol{w}_\alpha \tag{2.38}$$

where $\pi_\alpha \in (0,1)$ is a weight, such that $\sum_\alpha \pi_\alpha = 1$, and ϕ_α is a non zero density on $\mathcal{S}_\alpha$.

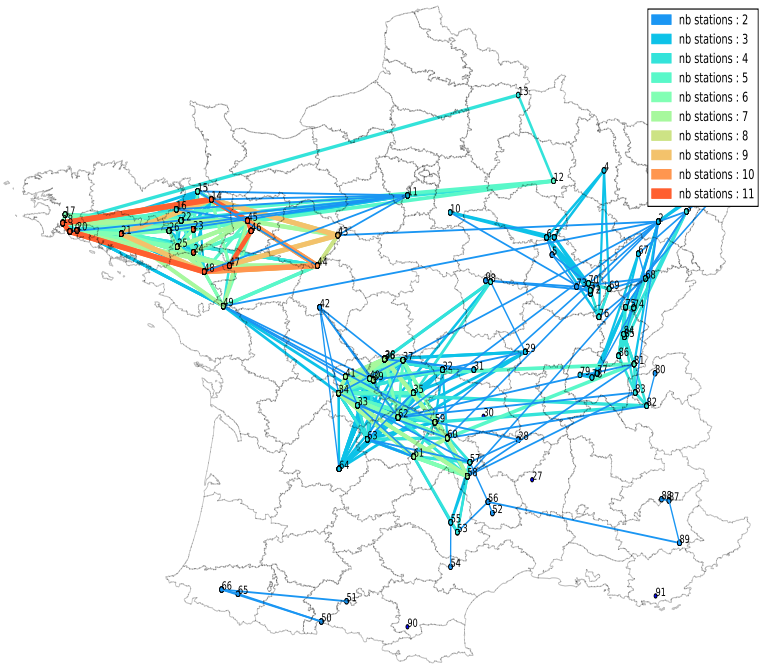


Figure 2.7: Representation of $\widehat{\mathcal{M}}$ for the criterion $\{\hat{\gamma}_\alpha > \gamma_{min}\}$ with $\gamma_{min} = 0.2$.

In order to correspond to a proper angular measure the mixture model has to verify the moment constraint (5.8). The densities ϕ_α along with the weights π_α have to verify:

$$\sum_{\substack{\alpha \in \mathbb{M} \\ j \in \alpha}} \pi_\alpha \int_{\mathcal{S}_\alpha} w_j \, \phi_\alpha(\boldsymbol{w_\alpha}) \mathrm{d}\boldsymbol{w_\alpha} = \frac{1}{d}, \, \forall j \in \{1, \ldots, d\}. \tag{2.39}$$

As the Dirichlet distribution is a natural distribution on the simplex we propose a Dirichlet mixture for the parametric modelling of Φ. The distribution can be parametrized by a mean parameter $\boldsymbol{m}_\alpha \in \mathcal{S}_\alpha$ and a concentration parameter $\nu_\alpha > 0$. We have, for all $\boldsymbol{w} \in \mathcal{S}_\alpha$:

$$\phi_\alpha(\boldsymbol{w} | \boldsymbol{m}_\alpha, \nu_\alpha) = \frac{\Gamma(\nu_\alpha)}{\prod_{i \in \alpha} \Gamma(\nu_\alpha m_{\alpha,i})} \prod_{i \in \alpha} w_i^{\nu_\alpha m_{\alpha,i} - 1}. \tag{2.40}$$

Therefore, given that $\int_{\mathcal{S}_\alpha} \boldsymbol{w} \, \phi_\alpha(\boldsymbol{w} | \boldsymbol{m}_\alpha, \nu_\alpha) \mathrm{d}\boldsymbol{w} = \boldsymbol{m}_\alpha$, the moment constraint becomes:

$$\sum_{\alpha \in \mathbb{M}} \pi_\alpha \boldsymbol{m}_{\alpha,j} = \frac{1}{d}, \, \forall j \in \{1, \ldots, d\}. \tag{2.41}$$

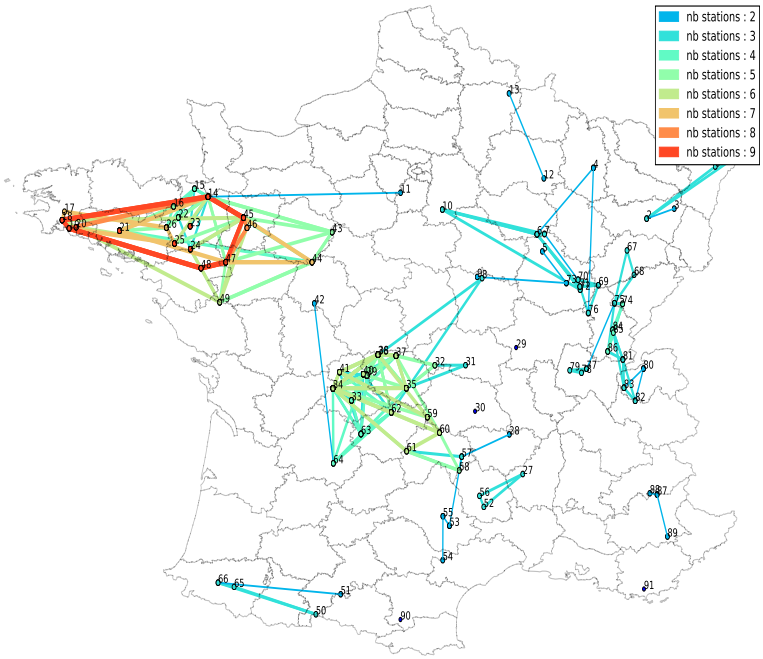


Figure 2.8: Representation of $\widehat{\mathcal{M}}$ for the algorithm CLEF with $\kappa_{min} = 0.2$.

Two main issues arise with the inference of the model. First, the estimators $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{\pi}}$ have to verify the constraints:

$$\begin{cases} \sum_{\alpha \in \mathbb{M}} \pi_\alpha m_{\alpha,j} = \frac{1}{d}, \forall j \in \{1, \dots, d\} \\ \sum_{j \in \alpha} m_{\alpha,j} = 1, \forall \alpha \in \mathbb{M} \\ \sum_{\alpha \in \mathbb{M}} \pi_\alpha = 1 \end{cases} \tag{2.42}$$

and therefore the log-likelihood maximization is non-convex. This difficulty is eased by the following substitution, for all $\alpha \in \mathcal{M}$:

$$\rho_{\alpha,j} = \pi_j m_{\alpha,j}, \forall j \in \alpha. \tag{2.43}$$

Thus, the three constraints become:

$$\sum_{\alpha \in \mathcal{M}} \rho_{\alpha,j} = \frac{1}{d}, \forall j \in \{1, \dots, d\}. \tag{2.44}$$

Indeed, considering $\pi_\alpha := \sum_{j \in \alpha} \rho_\alpha$ and $m_{\alpha,j} := \frac{\rho_{\alpha,j}}{\sum_{j \in \alpha} \rho_\alpha}$, for all $\alpha \in \mathcal{M}$ and for all $j \in \alpha$, it is straightforward that (2.44) implies $\sum_{j \in \alpha} m_{\alpha,j} = 1, \forall \alpha \in \mathbb{M}$ and $\sum_{\alpha \in \mathbb{M}} \pi_\alpha = 1$, so that (2.44) is equivalent to (2.41).

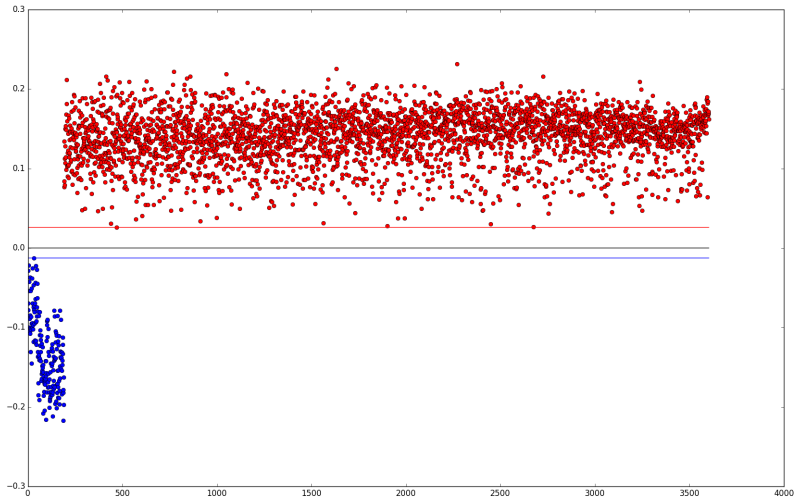


Figure 2.9: Values of $\hat{\eta}_\alpha - (1 - q_{1-\delta}\frac{\hat{\sigma}_\alpha}{\sqrt{k}})$ on the elements of $\mathbb{M}^F$ and $\mathbb{M}_0$.

The second difficulty comes from the asymptotic nature of the model. This is a straight consequence of the issue explained in the section 2.3.1. For all $\boldsymbol{V}$, obtained after transformation of the initial dataset (2.6), the pseudo-angle $\frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}$ lies in the interior of the central simplex $\mathcal{S}_d$. In order to tackle this issue, the idea is to split $\boldsymbol{V}$ between its contribution in α, $\boldsymbol{V}_\alpha = (V_j)_{j\in\alpha}$ and its non-asymptotic residual $\boldsymbol{\varepsilon}_{\alpha^c} = (V_j)_{j\notin\alpha}$ (figure (2.10)). A hidden variable $\mathbf{Z}$ valued in $\{0,1\}^K$ is added for the inference such that $\pi_{\alpha_k} = \mathbb{P}(Z_k = 1)$, where $\alpha_k \in \mathbb{M} = \{\alpha_1, \ldots, \alpha_K\}$ and $\sum_{k=1}^K Z_k = 1$.

A sub-asymptotic model is then proposed, that comprehends the Dirichlet mixture, on which is added a model for the non-asymptotic residuals. Thus, for $\alpha_k \in \mathbb{M}$ and for $\boldsymbol{v} \in (1,\infty)^d$, the likelihood is given by:

$$p(\boldsymbol{v}|z_k = 1) = r_k^{-|\alpha_k|-1}\phi_{\alpha_k}(\boldsymbol{w}_k|\boldsymbol{m}_{\alpha_k}, \nu_{\alpha_k}) \prod_{j\notin\alpha} f_\varepsilon(v_j|\lambda_k), \tag{2.45}$$

where $f_\varepsilon(\cdot|\lambda_k)$ is an exponential law of parameter λ_k, $r_k = \sum_{j\in\alpha_k} v_j$ and $\boldsymbol{w}_k = (\frac{v_j}{r_k})_{j\in\alpha_k}$. The inference on the model is then made by using the pseudo-angles $\boldsymbol{W}_\alpha = \frac{\boldsymbol{V}_\alpha}{\|\boldsymbol{V}_\alpha\|}$ and the residuals $\boldsymbol{\varepsilon}_{\alpha^c}$ through an adapted Expectation-Maximization (EM) algorithm.

2.6 Contributions

Conference articles with proceedings

- Feature clustering for extreme events analysis, with application to extreme stream-flow data. (ECML 2016)

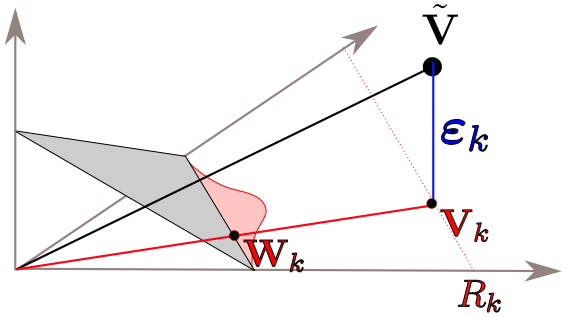


Figure 2.10: Pseudo-angle and residual.

Authors: M. Chiapino, A. Sabourin.

- A multivariate extreme value theory approach to anomaly clustering and visualization. (submitted)
  Authors: M. Chiapino, A. Sabourin, S. Clémençon, V. Feuillard.

Journal article

- Identifying groups of variables with the potential of being large simultaneously. (submitted)
  Authors: M. Chiapino, A. Sabourin, J. Segers.

2.7 Open problems

In this thesis, we proposed new methods for infering the dependence structure of high dimensional extreme phenomena. As usual methods for dimension reduction do not apply in this context, we found a way to tackle high dimension through a sparse version of the angular measure Φ. Nevertheless, as Φ is a limit measure, the observations cannot be used straightforwardly for inference. In particular, we have to deal with the different supports between the law of the observed data and the limit model. Under some sparsity assumption, the asymptotic tests carried out in Chapter 4 allow to recover the asymptotic dependence structure in the extremes, in relation to the support of Φ. The decomposition $\boldsymbol{V}_{observed} = \boldsymbol{W}R + \boldsymbol{\varepsilon}_{residual}$ is then used to infer the model (Chapter 5). The consistency of the model comes from the fact that $\lim_{R\to\infty} \boldsymbol{\varepsilon}_{residual} = 0$. However, we did not investigate the rate of convergence of such a limit, and this should constitute new studies. Also, other decompositions are possible such as $\boldsymbol{V}_{observed} = (\boldsymbol{W} + \boldsymbol{\varepsilon}_{residual})R$, that would lead to new models and algorithms.

3 Clustering Extreme Features

Abstract The dependence structure of extreme events of multivariate nature plays a special role for risk management applications, in particular in hydrology (flood risk). In a high dimensional context ($d > 50$), a natural first step is dimension reduction. Analyzing the tails of a dataset requires specific approaches: earlier works have proposed a definition of sparsity adapted for extremes, together with an algorithm detecting such a pattern under strong sparsity assumptions. Given a dataset that exhibits no clear sparsity pattern we propose a clustering algorithm allowing to group together the features that are 'dependent at extreme level', *i.e.*that are likely to take extreme values simultaneously. To bypass the computational issues that arise when it comes to dealing with possibly $O(2^d)$ subsets of features, our algorithm exploits the graphical structure stemming from the definition of the clusters, similarly to the Apriori algorithm, which reduces drastically the number of subsets to be screened. Results on simulated and real data show that our method allows a fast recovery of a meaningful summary of the dependence structure of extremes.

3.1 Introduction

Extreme value analysis is of primarily interest in many contexts. One example is the machine learning problem of anomaly detection, where one needs to control the false positive rate in the most remote regions of the sample space (Clifton et al. (2011a); Lee and Roberts (2008b); Goix et al. (2015a, 2016b)). Another example is the field of environmental sciences, where extreme events (floods, droughts, heavy rainfall, ...) are of particular concern to risk management, considering the disastrous impact these events may have. Using Extreme Value Theory (EVT) as a general setting to understand or predict extreme events has a long history (Katz et al. (2002)). In spatial problems, exhibiting areas (groups of weather stations) which may be concomitantly impacted by severe events is of direct interest for risk management policies. Identifying these groups may also serve as a preliminary dimensionality reduction step before more precise modeling. Before proceeding further, we emphasize that standard dimension reduction techniques such as PCA do not apply to extremes as these methods essentially focus on the data around the mean by analyzing their covariance structure, which does not characterize the behavior of extremes (*i.e.* data

far away in the tails of the distribution). In the present paper, the quantity of interest is river water-flow recorded at several locations of the French river system. The features of the experiment are thus the stream-flow records at different gauging stations, and the goal is to recover maximal groups of stations where extreme discharge may occur simultaneously. Our dataset consists of daily stream-flow recorded at 92 gauging stations scattered over the French river system, from 1969, January 1st to 2008, December 31st. It is the same dataset as in Giuntoli et al. (2013), up to 220 gauging stations presenting missing or censored records, which have been removed from our analysis, which results in $n = 14610$ vectors $X_1, .., X_n$ in $\mathbb{R}^d$, with $d = 92$ the number of stations. The reader is referred to Giuntoli et al. (2013) for more details.

Related work. Dimensionality reduction for extreme value analysis has emerged very recently in the literature. As far as we know, the seminal contribution is Chautru (2015) and is restricted to moderate dimensional settings ($d \leq 20$, see Section 3.3.1 for more details). The methodology proposed by Chautru (2015) allows to recover groups of components (features) which may take large values simultaneously, *while the other features stay small.* For the purpose of anomaly detection, Goix et al. (2015a, 2016b) proposed an alternative algorithm to do so with a reduced computational complexity of order $O(nd \log n)$. To the best of our knowledge, these are currently the only available examples in the literature to handle the recovery of groups of features which are representative of the *extremal dependence structure.* (See Section 3.2.2 for a precise definition of the latter). In Goix et al. (2015a, 2016b), the extremal dependence structure is called *sparse* if the number of such groups is small compared with $2^d - 1$, the total number of groups. The output of Goix et al. (2015a, 2016b)'s DAMEX algorithm is a (hopefully sparse) vector $\hat{\mathcal{M}} = (\hat{\mu}_\alpha, \alpha \subset \{1, \ldots, d\})$ of size $2^d - 1$, where $\hat{\mu}_\alpha$ is a summary of the dependence strength at extreme levels between features $j \in \alpha$. The fact that $\hat{\mu}_\alpha$ is positive means that the probability that all features in α be large while all others stay small, is not negligible. Various datasets have been analyzed in Goix et al. (2015a, 2016b) (wave data from the north sea, standard anomaly detection datasets, simulated data) for which the DAMEX algorithm does exhibit a sparsity pattern, thus pointing to a relatively small number of groups of features α (each being of relatively small size $|\alpha|$ compared to the original dimension of the problem) which could be jointly extreme. However, DAMEX becomes unusable in situations where the subsets of features impacted by extreme events vary from one event to another: DAMEX then finds a very large number of subsets to be dependent, but not significantly so, ($i.e. 0 < \hat{\mu}_\alpha \ll 1$), so that no sparsity pattern emerges. This is precisely the case with the river flow dataset analyzed in the present paper (see Section 3.5).

Contributions. One remarkable aspect of the preliminary analysis of the river flow dataset using DAMEX is the tendency of those many subsets α's such that $\hat{\mu}_\alpha > 0$, to form *clusters*, whose members differ from each other by a single or two features only. In practice, this means that several distinct events have impacted 'almost' the same group (cluster) of stations. The aim of this paper is to propose a methodol-

ogy enabling to gather together such 'close-by' feature subsets into feature clusters. This is done by relaxing the constraint that 'features not in α take small values' when constructing the representation of the dependence structure. The output of the CLEF algorithm (CLustering Extreme Features) proposed in the present work (Section 3.4) is an alternative representation which remains usable in this 'weakly sparse' context. This representation can still be explained and understood in the multivariate EVT framework (Section 3.3), as in Chautru (2015); Goix et al. (2015a, 2016b). We emphasize that the scope of CLEF algorithm concerns situations similar to the hydrological problem considered here, where the DAMEX algorithm does not yield a readable output. In the opposite case (*e.g.* with the wave dataset or the anomaly detection datasets analyzed in Goix et al. (2016b, 2015a)), DAMEX remains a better option than CLEF in view of its computational simplicity.

Relationships with Apriori. The dimension reduction problem considered here (determining for which subgroups of features concomitant large values are frequent) is closely related to the problem of frequent itemsets mining, specifically to the well known Apriori algorithm introduced by Agrawal et al. (1994), see also Gunopulos et al. (2003). Indeed, the present problem can be recast as follows: encoding as a '1' any value above a specified threshold and as a '0' any value below this threshold, one obtains a binary dataset. The goal is now to recover the groups items (features) for which concomitant '1' values are frequent, which is precisely the frequent itemsets mining problem. The combinatorial issue that arises with possibly $2^d - 1$ subsets is circumvented in Apriori by considering subsets of increasing sizes, letting a subset 'grow' until its frequency in the database is not significant anymore. This incremental principle is also related to a subset clustering method proposed in Agrawal et al. (2005). CLEF proceeds in a similar way to Apriori, the main difference being that CLEF comes with a natural interpretation in terms of multivariate EVT. Also, in practice, the stopping criterion used to decide whether incrementing a feature subset is different in CLEF and in Apriori, allowing CLEF to detect larger groups, as discussed in Sections 3.3 and 4.1.

The paper is organized as follows. Section 3.2 sets up the extremal feature clustering problem and establishes connections with multivariate EVT. The dimension reduction method that we promote is explained in Section 3.3: Section 3.3.1 recalls existing work and points out some limitations, Section 3.3.2 makes explicit the links between the considered problem and the Apriori algorithm. The CLEF algorithm is described in Section 3.4. Section 3.5 gathers results: the output of CLEF is compared with that of DAMEX and Apriori. Section 3.6 concludes. The `Python` code for CLEF, the scripts and the dataset used for our hydrological case study are available at `https://bitbucket.org/mchiapino/clef_algo`.

3.2 Problem statement and multivariate EVT viewpoint

3.2.1 Formal statement of the problem

Consider a random vector $X = (X^1, \ldots, X^d)$ in $\mathbb{R}^d$ (here, X^j is the water discharge recorded at location j). The first step when it comes to learning dependence properties of X is to standardize the features, in the same spirit as in the copula framework, which allows one to focus only on the *dependence* structure of X. One popular standardization choice in multivariate EVT is the probability integral transform: Denote by F the joint cumulative distribution function (*c.d.f.*) of X and by F^j the marginal *c.d.f.* of X^j. For simplicity, let us assume that each F^j is continuous (no point masses), so that with probability one, $0 < F^j(X^j) < 1$. The standardized variable used for dependence analysis are $V^j = (1 - F^j(X^j))^{-1}$, $j = 1, \ldots, d$ and $V = (V^1, \ldots, V^d)$. Doing so, the V^j's are identically distributed according to standard Pareto distribution, $\mathbb{P}(V^j > t) = 1/t, t \geq 1$.

3

Our goal here is to recover all the maximal subsets of features (stations) $\alpha \subset \{1, \ldots, d\}$ which 'may be large together' with non negligible probability. In more formal terms, define the *extremal joint excess coefficient*,

$$\rho_\alpha := \lim_{t\to\infty} t\mathbb{P}\left(\forall j \in \alpha, V^j > t\right) = \lim_{t\to\infty} \mathbb{P}\left(\forall j \in \alpha, V^j > t \mid V^{\alpha_1} > t\right) \in [0, 1]. \quad (3.1)$$

The variable t plays the role of a high threshold above which the standardized feature V^j is considered as extreme. In practice, estimation will be done by fixing a large t and assuming that the limit in (3.1) is approximately reached. An advantage of the standardization procedure is that a single threshold t is needed to define an extreme event, not d thresholds, since all the features share the same scale. The limit in (3.1) exists under the regularity property (3.3) in the next paragraph. Notice already that the second equality also comes from our standardization choice ensuring that for any $j \leq d$, $t^{-1} = \mathbb{P}(V^j > t) = \mathbb{P}(V^{\alpha_1} > t)$, which justifies the scaling factor t in the definition. The coefficient $\rho_\alpha \in [0, 1]$ may be seen as a 'correlation' coefficient for the features $X^j, j \in \alpha$ at extreme levels. We say that the features $\{V^j, j \in \alpha\}$ 'may be large together' if $\rho_\alpha > 0$. One relevant summary of the dependence structure of extremes is thus the set of subgroups

$$\mathbb{M} = \{\alpha \subset \{1, \ldots, d\} \ : \quad \rho_\alpha > 0\}. \quad (3.2)$$

More precisely, we would like to recover those subgroups $\alpha \in \mathbb{M}$ which are maximal for inclusion in $\mathbb{M}$, *i.e.* $\forall \beta$ such that $\alpha \subsetneq \beta$, $\beta \notin \mathbb{M}$. A maximal set of features $\alpha \in \mathbb{M}$ may be viewed as a *cluster*, in the sense that every subset $\beta \subset \alpha$ is dependent at extreme level (*i.e.*$\rho_\beta > 0$), and that α 'gathers' all of them together. In this paper, a 'cluster' of features is understood as a maximal element $\alpha \in \mathbb{M}$.

3.2.2 Connections with multivariate EVT

The working hypothesis in EVT is that, up to marginal standardization, the distribution of X is 'approximately homogeneous' on extreme regions. As pointed out

above, if the margins F^j are continuous, then the V^j's have the homogeneity property: $t\mathbb{P}\left(\frac{V^j}{t} \geq x\right) = 1/x$, for $1 \leq j \leq d$, $t > 1$, $x > 0$. The key assumption is that the latter property holds *jointly* at extreme levels, *i.e.* that V is jointly *regularly varying* (see *e.g.* Resnick (2013)), which writes

$$t\mathbb{P}\left(\frac{V}{t} \in A\right) \xrightarrow[t\to\infty]{} \mu(A), \tag{3.3}$$

where μ is the so-called *exponent measure* and where A is any set in $\mathbb{R}^d$ which is bounded away from 0 and such that $\mu(\partial A) = 0$. The exponent measure is finite on any such set A and satisfies, for $t > 0, A \subset \mathbb{R}^d_+$, $t\mu(tA) = \mu(A)$, where $tA = \{tx : x \in A\}$. Notice that many commonly used textbook multivariate distributions (*e.g.* multivariate Gaussian or Student distributions) satisfy (3.3), after standardization to V variables. The measure μ characterizes the distribution of V at extreme levels, since for t large enough (so that the region tA is an 'extreme region' of interest), one may use the approximation $\mathbb{P}(V \in tA) \simeq t^{-1}\mu(A)$. The connection between μ and the ρ_α's is as follows: consider the 'rectangle'

$$\Gamma_\alpha := \{x \in \mathbb{R}^d_+ : \quad \forall j \in \alpha,\ x^j > 1\} \tag{3.4}$$

From the definitions (3.1) and (3.3), it follows that $\rho_\alpha = \mu(\Gamma_\alpha)$. Thus the family of subset $\mathbb{M}$ in (3.2) writes $\mathbb{M} = \{\alpha : \mu(\Gamma_\alpha) > 0\}$.

Non parametric estimation. In a word, non parametric estimation of extremal characteristics based on *i.i.d.* data $X_1, \ldots, X_n$ (distributed as X) is performed by replacing probability distributions with their empirical counterparts, and by proceeding as if the limit in (3.3) were reached above some large threshold t. Since the F^j's are unknown, set $\hat{V}_i^j = 1/(1 - \hat{F}^j(X_i^j)),\ = 1, \ldots, n,\ j = 1, \ldots, d$, where $\hat{F}^j(x) = n^{-1}\sum_{i=1}^n \mathbb{1}\{X_i^j < x\}$. Thus $\hat{V}_i^j \in \{1, n/(n-1), n/(n-2), \ldots, n/2, n\}$ and for each fixed j, and $t \leq n$, the number of examples i such that $\hat{V}_i^j > t$ is equal to $\lceil n/t \rceil$. This suggests that t should be chosen as a function of the sample size and indeed, theoretical guarantees on the estimators are obtained for $t = o(n)$ and $t \to \infty$, *e.g.* $t \sim \sqrt{n}$, see Beirlant et al. (2006), Chapter 3 for more details. After this data preprocessing step, the exponent measure μ of any region $A \subset \mathbb{R}^d_+ \setminus \{0\}$ is approximated by

$$\mu_n(A) = t\hat{P}_n(tA), \qquad \text{where } \hat{P}_n(A) = n^{-1}\sum_{i=1}^n \delta_{\hat{V}_i}(A), \tag{3.5}$$

where δ denotes the Dirac mass. Statistical properties of μ_n (or of other functional summaries of it) have been investigated by many authors, see *e.g.* Qi (1997); Einmahl and Segers (2009); Fougeres et al. (2015) for the asymptotic behavior, Goix et al. (2015b) for finite sample error bounds.

3.3 Dimension reduction for multivariate extremes

3.3.1 Existing work

Numerous modeling strategies for low dimensional multivariate extremes (say $d \leq 10$) have been proposed, see *e.g.* Coles and Tawn (1991); Cooley et al. (2010); Sabourin et al. (2013) for parametric modeling, Boldi and Davison (2007a); Guillotte et al. (2011); Sabourin and Naveau (2014); Fougeres et al. (2013) for semi- or non-parametric ones. For higher dimensional problems, to this date, the only available dimensionality reduction methods are (to our best knowledge) the recent works Chautru (2015); Goix et al. (2015a, 2016b). These three references share the common idea of recovering the sub-cones of $\mathbb{R}_+^d$ on which the exponent measure μ concentrates. The seminal paper on this subject appears to be Chautru (2015). It relies on principle nested spheres and spherical k-means and is designed for moderate dimensional problems only ($d \leq 20$ in their simulation experiments and $d = 4$ in their case study) with a relatively simple dependence structure (at most 4 groups of features with extremal dependence, only two for $d = 20$ in their simulation experiments). The computational burden significantly increases for larger dimensions or more elaborate dependence structures, as discussed in Section 4.4 of the cited reference. In particular the dimensionality of the problem considered in the present paper ($d = 92$ with up to 53 dependent groups of features) is outside the scope of Chautru (2015)'s algorithm.

The present work is mainly related to Goix et al. (2015a, 2016b) insofar as it relies on a simple counting procedure on rectangular regions as in (3.4). As a comparison, Goix et al. (2015a, 2016b) consider the truncated cones

$$\mathcal{C}_\alpha = \{x : \|x\|_\infty \geq 1, x_j > 0 \text{ for } j \in \alpha \; ; \; x_j = 0 \text{ for } j \notin \alpha\}. \tag{3.6}$$

The importance of such cones in the analysis comes from the homogeneity property of μ. More precisely, a subset of features α may take large values together while the others take small values, if and only if μ assigns a positive mass to $\mathcal{C}_\alpha$. The approach proposed in Goix et al. (2016b) consists in 'thickening' the cones $\mathcal{C}_\alpha$, *i.e.* defining for some small $\epsilon > 0$ (typically, $\epsilon = 0.1$),

$$\mathcal{C}_{\alpha,\epsilon} = \{x \in \mathbb{R}_+^d : \; \|x\|_\infty \geq 1 \; ; \; \|x\|_\infty^{-1} x_j > \epsilon \text{ for } j \in \alpha \; ; \; \|x\|_\infty^{-1} x_j \leq \epsilon \text{ for } j \notin \alpha\}. \tag{3.7}$$

The quantity $\mu_\alpha := \mu(\mathcal{C}_\alpha)$ is approximated by its empirical counterpart on $\mathcal{C}_{\alpha,\epsilon}$, $\hat{\mu}(\mathcal{C}_\alpha) = \mu_n(\mathcal{C}_{\alpha,\epsilon})$, where μ_n is the empirical estimator defined in (3.5). In practice a tolerance parameter $\mu_{\min}$ has to be chosen: for any α such that $\mu_n(\mathcal{C}_{\alpha,\epsilon}) < \mu_{\min}$, one sets $\hat{\mu}(\mathcal{C}_\alpha) = 0$. The final output of Goix et al. (2016b)'s DAMEX algorithm is the potentially sparse $2^d - 1$-vector $\hat{\mathcal{M}} = (\hat{\mu}_\alpha)_{\alpha \subset \{1,\ldots,d\}}$ mentioned in the introduction, with $\hat{\mu}_\alpha := \hat{\mu}(\mathcal{C}_\alpha)$.

One shortcoming of DAMEX is that no sparsity pattern is produced in case of 'noise'. Here, noise is understood as a small variability affecting the groups of features concomitantly impacted by an extreme event. As an example, for the hydrological dataset considered here, geophysics determines the main underlying dependence patterns, *i.e.* the groups of stations where floods *tend to* occur simultaneously (such

as, say, group $\alpha_0 = \{1, 2, 3, 4\}$); however due to meteorological variability, the actual observed floods sometimes affect some neighboring stations $5, 6$, so that in the dataset, the observed groups would be *e.g.* $\big\{\{1, 2, 3, 4, 5\}, \{1, 2, 3, 4, 6\}, \{1, 2, 3, 4\}\big\}$. In such a case, the empirical mass is scattered over many sub-cones $\mathcal{C}_{\alpha,\epsilon}$ (three instead of one). This example suggests an alternative approach allowing to gather together those $\mathcal{C}_{\alpha,\epsilon}$'s that are 'close', as detailed next.

3.3.2 Gathering together 'close-by' cones, incremental strategy

One way to gather different $\mathcal{C}_{\alpha,\epsilon}$'s together is to relax the condition that 'all the features V^j for $j \notin \alpha$ take small values' in the definition of $\mathcal{C}_{\alpha,\epsilon}$. This yields the rectangular region Γ_α defined in (3.4). Unlike the regions $\mathcal{C}_{\alpha,\epsilon}$'s, the Γ_α's do not form a partition of the positive orthant of $\mathbb{R}^d$, and indeed the fact that a point V_i belongs to Γ_α does not tell anything about its features V_i^j for $j \notin \alpha$. The problem addressed in Goix et al. (2016b) (recovering $\mathcal{M} := \{\alpha : \mu(\mathcal{C}_\alpha) > 0\}$) and the relaxed problem considered here (recovering $\mathbb{M} := \{\alpha : \rho_\alpha > 0\} = \{\alpha : \mu(\Gamma_\alpha) > 0\}$) are different but however related through the maximal elements of $\mathcal{M}$ and $\mathbb{M}$, as stated in the following lemma. Recall that α is said to be maximal in $\mathbb{M}$ (*resp.* $\mathcal{M}$) if there is no superset $\alpha' \supsetneq \alpha$ in $\mathbb{M}$ (*resp.* $\mathcal{M}$).

Lemma 3.1. *For $\alpha \subset \{1, \dots, d\}$,*

$$\alpha \text{ is maximal in } \mathbb{M} \Leftrightarrow \alpha \text{ is maximal in } \mathcal{M}. \tag{3.8}$$

The proof is deferred to the Appendix.
Another important property from an algorithmic perspective is the following:

Lemma 3.2. *For $\alpha \subset \{1, \dots, d\}$, if $\rho_\alpha = 0$ then also for all $\alpha' \supset \alpha$, $\rho_{\alpha'} = 0$.*

The proof is immediate: remind that $\rho_\alpha = \mu(\Gamma_\alpha)$ and notice that for $\alpha \subset \alpha'$, $\Gamma_{\alpha'} \subset \Gamma_\alpha$.

Apriori-like incremental strategy Lemma 3.2 suggests searching for α's satisfying $\rho_\alpha > 0$ following the Hasse diagram, among α's of increasing size, and stopping the search along a given path as soon as $\rho_\alpha = 0$ for some α. This incremental strategy is also the main ingredient of the Apriori algorithm (Agrawal et al. (1994)), which we recall for convenience: Let $I = \{\text{item}_1, \dots, \text{item}_d\}$ be set of items and let $T = \{t_1, \dots, t_n\}$ be a set of transactions with $t_i \subset I, \forall i \in \{1, \dots, n\}$. The frequency of occurrence of the list of items (itemset) $\alpha \subset I$ is defined as $f_\alpha := \frac{1}{n} \sum_{1 \le i \le n} \mathbb{1}_{\alpha \subset t_i}$. Apriori returns the set $\{\alpha : f_\alpha > f_{min}\}$ with $f_{min} > 0$. It begins with pairs of items and then increments the size of the itemsets at each step. Indeed if $f_\alpha \le f_{min}$ then all supersets $\alpha' \supset \alpha$ verify $f_{\alpha'} \le f_{min}$ as well, which reduces drastically the number of subsets to be tested.

The CLEF algorithm described next proceeds similarly: a concomitant occurrence of threshold excesses $\{V_i^j > t$ for features $j \in \alpha\}$ can be identified with a

transaction and the dependence parameter ρ_α can be seen as a (rescaled) theoretical frequency. The main difference between CLEF and Apriori concerns the stopping criterion used by CLEF, which involves a *ratio* between frequencies for a group α and for subgroups $\beta \subset \alpha$. The idea behind is to allow detection of larger groups, as described in the following section.

3.4 Empirical criterion and implementation

3.4.1 Conditional criterion for extremal dependence

Considering the relaxed framework where the goal is to recover the set $\mathbb{M}$ defined in (3.2), one needs an empirical criterion for testing the condition '$\rho_\alpha(= \mu(\Gamma_\alpha)) > 0$'. One option would be to consider the empirical estimator $\hat{\rho}_\alpha = \mu_n(\Gamma_\alpha)$ where μ_n is defined in (3.5) which would be the (rescaled) counterpart of the empirical frequency f_α used in Apriori. Then the stopping criterion would be '$\hat{\rho}_\alpha \leq \rho_{\min}$', with $\rho_{\min}$ a user-defined tolerance level. However, since the Γ_α's (for increasing α's) are nested, the ρ_α's can only decrease with increasing sizes of α. In other words larger groups tend to be less frequent than smaller groups, even dependent ones. Thus in principle, detecting larger groups as well as smaller ones would require the tolerance level $\rho_{\min}$ to depend on the size $|\alpha|$ of the considered subgroup, which would result in $d-1$ tuning parameters instead of one.

The alternative chosen in the present paper is to consider a *conditional* frequency, the conditioning event for a group α of size s being such that at least $s-1$ features are large among the s considered ones. Now, there is no reason why the conditional frequency of occurrence should decrease with $|\alpha|$, so that a single tuning parameter needs to be chosen, without preventing the detection of large groups. In practice, computing conditional frequencies amounts to compare $\mu_n(\Gamma_\alpha)$ with $\mu_n(\Gamma_\beta)$, with $\beta \subset \alpha$. More precisely, let $\alpha \subset \{1, \ldots, d\}$ be such that for some $j \in \alpha$, $\rho_{\alpha\setminus\{j\}} > 0$. Consider the probability that all the features in α be large given that *all of them but at most one are large* and call κ_α the limiting conditional probability, namely

$$\kappa_\alpha = \lim_{t\to\infty} \frac{\mathbb{P}\Big(\forall j \in \alpha, V_i^j > t\Big)}{\mathbb{P}\Big(\text{for all but at most one } j \in \alpha, V_i^j > t\Big)}. \tag{3.9}$$

In the sequel, κ_α is referred to as the *conditional dependence coefficient* of α. Notice that the limit in (3.9) does exist: indeed, let

$$\Delta_\alpha = \cup_{j\in\alpha}\Gamma_{\alpha\setminus\{j\}} = \{x \in \mathbb{R}_+^d : \ \|x\|_\infty > 1, \ \textstyle\sum_{j\in\alpha} \mathbb{1}_{x_j \geq 1} \geq |\alpha| - 1\}.$$

Since by assumption on α, for some j $\mu(\Gamma_{\alpha\setminus\{j\}}) = \rho_{\alpha\setminus\{j\}} > 0$, in view of (3.3), we have

$$\mu(\Delta_\alpha) = \lim_{t\to\infty} t\mathbb{P}(\text{ for all but at most one } j \in \alpha, V_i^j > t) > 0,$$

so that an equivalent definition of κ_α is

$$\kappa_\alpha = \frac{\lim_{t\to\infty} t\mathbb{P}\Big(\forall j \in \alpha, V_i^j > t\Big)}{\lim_{t\to\infty} t\mathbb{P}\Big(\text{for all but at most one } j \in \alpha, V_i^j > t\Big)} = \frac{\mu(\Gamma_\alpha)}{\mu(\Delta_\alpha)}. \tag{3.10}$$

The idea is now to compare empirical counterparts of κ_α –using μ_n instead of μ, see (3.5)– with a single fixed tolerance parameter $\kappa_{\min} > 0$. This amounts to decide that $\mu_n(\Gamma_\alpha)$ results from noise if $\mu_n(\Gamma_\alpha) \ll \mu_n(\Delta_\alpha)$. Notice that $\Gamma_\alpha \subset \Delta_\alpha$, so that the empirical version of κ_α is again a conditional probability and thus belongs to $[0, 1]$ whenever $\mu_n(\Gamma_\beta) > 0$ for some $\beta \subset \alpha$ such that $|\alpha \setminus \beta| = 1$, which is another argument in favor of an incremental strategy.

3

3.4.2 Algorithm

CLEF (summarized in Algorithm 3) uses the empirical counterpart of the conditional criterion κ_α, which depends on a (high) threshold t as in (3.5):

$$\hat{\kappa}_{\alpha,t} := \frac{\mu_n(\Gamma_\alpha)}{\mu_n(\Delta_\alpha)} = \frac{\sum_{i=1}^n \mathbb{1}_{\{\#\{j\in\alpha:\, \hat{V}_i^j > t\} = |\alpha|\}}}{\sum_{i=1}^n \mathbb{1}_{\{\#\{j\in\alpha:\, \hat{V}_i^j > t\} \geq |\alpha|-1\}}}. \tag{3.11}$$

For $s \geq 2$, families $\hat{\mathcal{A}}_s$ of subsets α of size s are constructed in an incremental way, among a set of candidates $\mathcal{A}'_s$, as follows: Set $\hat{\mathcal{A}}_1 = \{\{1\}, \ldots, \{d\}\}$, then

$$\begin{aligned} \mathcal{A}'_s &= \Big\{\alpha \subset \{1, \ldots, d\} : |\alpha| = s, \forall \beta \subset \alpha \text{ s.t. } |\beta| = s - 1 : \beta \in \hat{\mathcal{A}}_{s-1}\} \\ \hat{\mathcal{A}}_s &= \{\alpha \in \mathcal{A}'_s : \hat{\kappa}_{\alpha,t} > \kappa_{\min}\Big\}. \end{aligned} \tag{3.12}$$

The procedure stops at step $S \leq d-1$ if $\hat{\mathcal{A}}_{S+1} = \emptyset$, at which point our estimator of the family $\mathbb{M}$ of dependent subsets is $\hat{\mathbb{M}} = \cup_{s=1}^S \hat{\mathcal{A}}_s$. Notice that restricting the search to the set of candidates $\mathcal{A}'_s$ ensures that the 'empirical counterpart' of Lemma 3.2 is satisfied, namely $\alpha \notin \hat{\mathbb{M}} \Rightarrow \forall \beta \supset \alpha, \beta \notin \hat{\mathbb{M}}$. It also avoids division by zero when computing (3.12). The final output of CLEF is the set $\hat{\mathbb{M}}_{\max}$ of maximal elements of $\hat{\mathbb{M}}$.

Remark 3.3 (Choice of the parameters t and $\kappa_{\min}$). The choice of t is a classical bias/variance trade-off: according to standard good practice in EVT (see *e.g.* *Coles (2001)*), t is chosen in a 'stability region' of relevant summaries of the output. Here we consider the cardinal of $\hat{\mathbb{M}}$ and the mean cardinal of maximal subsets $\alpha \in \hat{\mathbb{M}}$. When t is too small, the observed data may not have reached there ultimate regime (the extremal dependence structure characterized by μ in (3.3)), so that the bias

Algorithm 3 CLEF (CLustering Extreme Features)

INPUT: High threshold t, tolerance parameter $\kappa_{\min} > 0$.
STAGE 1: constructing the $\hat{\mathcal{A}}_s$'s .
Initialization: set $S = d$.
Step 1: Construct the family of extremal-dependent pairs:
set $\hat{\mathcal{A}}_2 = \Big\{\{i,j\} \subset \{1,\dots,d\},$ such that $\hat{\kappa}_{\{i,j\}} > \kappa_{min}\Big\}$.
Step 2: If $\hat{\mathcal{A}}_2 = \emptyset$, set $S = 2$; end **STAGE 1**. Otherwise

- generate candidate triplets $\mathcal{A}'_3 = \{i,j,k\} \subset \{1,\dots,d\}$ s.t $\{i,j\},\{i,k\},\{j,k\} \in \hat{\mathcal{A}}_2\}$,
- set $\hat{\mathcal{A}}_3 = \Big\{\alpha \in \mathcal{A}'_3$ s.t. $\hat{\kappa}_\alpha > \kappa_{min}\Big\}$.

$\vdots$

Step $s(s \leq d)$: If $\hat{\mathcal{A}}_s = \emptyset$, set $S = s$; end **STAGE 1**. Otherwise

- generate candidates of size $s+1$:
  $\mathcal{A}'_{s+1} = \{\alpha \subset \{1,\dots,d\}, |\alpha| = s+1, \alpha \setminus \{j\} \in \hat{\mathcal{A}}_s$ for all $j \in \alpha\}$,
- set $\hat{\mathcal{A}}_{s+1} = \Big\{\alpha \in \mathcal{A}'_{s+1}$ such that $\hat{\kappa}_\alpha > \kappa_{min}\Big\}$.

Output: $\hat{\mathbb{M}} = \cup_{s=1}^{S} \hat{\mathcal{A}}_s$.
STAGE 2: pruning (keeping maximal α's only)
Initialization: $\hat{\mathbb{M}}_{\max} \leftarrow \hat{\mathcal{A}}_S$.
for $s = (S-1):2$, for $\alpha \in \hat{\mathcal{A}}_s$,
If there is no $\beta \in \hat{\mathbb{M}}_{\max}$ such that $\alpha \subset \beta$, $\hat{\mathbb{M}}_{\max} \leftarrow \hat{\mathbb{M}}_{\max} \cup \{\alpha\}$.
Output: $\hat{\mathbb{M}}_{\max}$

of $\hat{\kappa}_{\alpha,t}$ may be large. In contrast, for too large values of t, very few excesses are observed so that the sample size of the data used to compute $\hat{\kappa}_{\alpha,t}$ is very small and the variance becomes too large. To wit, due to our standardization choice it holds that $\mathbb{P}(V_i^j > t) = 1/t$. Thus for each $j \in \{1,\dots,d\}$, $\Big|\{i : \hat{V}_i^j > t\}\Big| \simeq 1/t$, so that the total number of data points for which at least one feature exceeds t is approximately within the interval $[n/t, dn/t]$. Results on real and simulated data (Section 3.5.3) bring out such a stability region for the above mentioned output summaries. It is empirically verified on simulated data that this region corresponds to near optimal values of t.

As for the tolerance parameter $\kappa_{\min}$, it should be chosen according to the context, keeping in mind that $\hat{\kappa}_{\alpha,t}$ is an empirical conditional probability of a joint threshold excess of all features $j \in \alpha$ (given that at least $|\alpha| - 1$ excesses have occurred). $\kappa_{\min}$ is the level above which this probability is considered as non negligible. The higher $\kappa_{\min}$, the more stringent the condition, the smaller and fewer the discovered groups

α. In this work, we set $\kappa_{\min} = 0.25$.

Remark 3.4. [Construction of the candidates $\mathcal{A}'_{s+1}$] The graphical structure of the groups of features is exploited to construct candidate incremented groups of features. Namely, members of $\mathcal{A}'_{s+1}$ are the maximal cliques of size s in the graph $(\mathcal{A}_s, \mathcal{E}_s)$, where $\mathcal{E}_s = \{(\alpha, \alpha') \in \mathcal{A}_s \times \mathcal{A}_s : |\alpha \cap \alpha'| = s - 1\}$. The maximal clique problem is typically attacked *via* the max-clique algorithm (Xie and Philip (2010)). In the present work, clique extraction is performed using the function `find_clique` of the `Python` package `NetworkX`, which uses the Bron & Kerbosch (Bron and Kerbosch (1973),Tomita et al. (2006)) algorithm.

3.5 Results

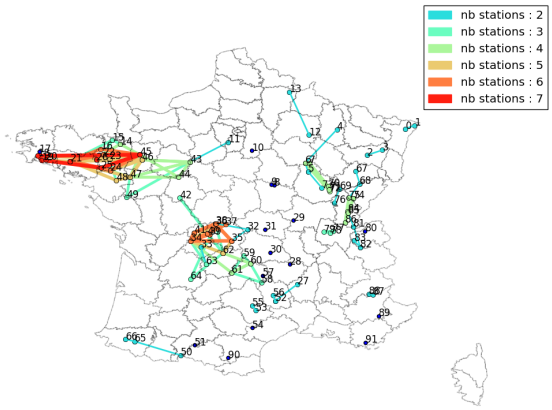
The aim of our experiments is threefold. First, CLEF's output on the hydrological data is illustrated and compared with DAMEX's (Section 3.5.1). Second, the respective performances of CLEF, DAMEX and Apriori are compared quantitatively on simulated data (Section 3.5.2). Finally (Section 3.5.3), the question of the threshold choice is investigated: the goal is to verify whether a stability region such as the one mentioned in Remark 3.3 exists and whether it corresponds to optimal performances of CLEF.

3.5.1 Stream-flow data

The output of CLEF for the stream-flow data may be visualized in Figure 3.1 (Execution time: 0.09 s on a recent 4 cores laptop computer). Following the heuristic mentioned in Remark 3.3, the extremal threshold t was fixed to 320, yielding $k = 1186$ extreme events (time indexes i such that $\|\hat{V}_i\|_\infty \geq t$). The parameter $\kappa_{\min}$ was fixed to 0.25. A total number of 53 clusters (elements of $\hat{\mathbb{M}}_{\max}$) are returned by the CLEF algorithm, the size of which varies between 2 and 7. At first inspection, Figure 3.1 agrees with general climatologic facts: in the north-western part of France, the climate is driven by large scale oceanographic perturbations, so that extreme floods tend to impact a large number of gauging stations simultaneously. The south-eastern part of France is rather subject to localized events (*e.g.* the so-called 'orages Cévenols' in the vicinity of the Mediterranean coast). This yields smaller clusters, both in terms of number of stations and of spatial extent.

As a comparison, Table 3.1 shows the outcome of Goix et al. (2016b)'s DAMEX algorithm with the stream-flow data. These results show that no matter the choice of the thickening parameter ϵ in (3.7), the data do not concentrate on 'a few' thickened cones $\mathcal{C}_{\alpha,\epsilon}$, instead most of the empirical mass is spread onto many of them. In other words, there are too many subcones with positive mass, but not in a significant way.

Figure 3.1: Output of CLEF for the stream-flow dataset: Maximal groups of stations $\alpha \in \hat{\mathbb{M}}$ that are likely to be jointly impacted by an extreme event.



Clusters of stations are marked by colored edges between their members, the color scale indicates the number of stations forming the cluster.

3.5.2 Simulation experiments

In order to quantify the relative performances of CLEF, DAMEX and Apriori, we generate d-dimensional datasets under a model such that the exponent measure μ concentrates on p specified cones $(\mathcal{C}_{\alpha_1}, \ldots, \mathcal{C}_{\alpha_p})$. Notice that $p, (\alpha_1, \ldots, \alpha_p)$ only determine the generating model, they are not used as inputs of any of the three algorithms compared here. The generated data are 'realistic' in the sense that all the features are positive (the points lie in the interior cone $\mathcal{C}_{\{1,\ldots,d\}}$), even though the furthest points in the tails concentrate near the subcones $\mathcal{C}_{\alpha_k}$'s. Namely, we use the *asymmetric logistic* extreme value model (Tawn (1990b)), from which data is simulated using Algorithm 2.2 in Stephenson (2003). 20 datasets of size $n = 100.10^3, d = 100$, are generated. For each dataset, p subsets $\alpha_1, \ldots, \alpha_p$ of $\{1, \ldots, d\}$ are randomly chosen, which sizes follow a truncated geometric distribution (the maximum cluster size is 8). We aim at reproducing the fact that different events associated with a single α usually impact a group of stations which differs from α by a few stations (the impacted area is not deterministic). To this end, for each step $i = 1, \ldots, n$, and each subset α_j, $j = 1, \ldots, p$, one randomly chosen 'noisy' feature $l_{i,j} \in \{1, \ldots, d\} \setminus \alpha_j$ is added to α_j. For CLEF, DAMEX and Apriori algorithms, the extreme threshold parameter t is chosen so that $\frac{\#\{i \leq n : \|\hat{V}_i\|_\infty \geq t\}}{n} \approx 5\%$. Table 3.2 summarizes the average performance of the three algorithms, for $p = 40, 50, 60, 70$. In these experiments, the CLEF algorithm recovers most of the charged p subsets $\alpha_1, \ldots, \alpha_p$ in average, and significantly more than Apriori. In contrast, DAMEX does not recover the sparse structure of the data. It should be noted that in situations

Table 3.1: Output of Goix et al. (2016b)'s DAMEX algorithm with the hydrological dataset.

ϵ	$\#\ \{\alpha : \mu_n(\mathcal{C}_{\alpha,\epsilon}) > 0\}$	$\%\Big\{\alpha : \frac{\#\{i : t^{-1}V_i \in \mathcal{C}_{\alpha,\epsilon}\}}{\#\{i : \|V_i\| \geq t\}} < 1\%\Big\}$
0.01	740	100%
0.05	688	98%
0.1	639	94%
0.2	559	88%

Columns 1 and 2 indicate respectively the number of thickened cones $\mathcal{C}_{\alpha,\epsilon}$ with non zero empirical mass, and the percentage of cones (among those such that $\mu_n(\mathcal{C}_{\alpha,\epsilon}) > 0$) containing less than 1% of the 'extreme data', that is of $\#\{i : \|\hat{V}_i\|_\infty > t\}$.

Table 3.2: Average number of errors (non recovered and falsely discovered clusters) of CLEF, Apriori and DAMEX with simulated, noisy data.

p	# errors CLEF	# errors Apriori	# errors DAMEX
40	1.2	6.4	72.2
50	3.5	10.9	91.0
60	6.3	14.6	112.4
70	10.1	25.8	134.0

(not reported here) where no noisy feature is added, Apriori and DAMEX perform as well as CLEF.

3.5.3 Influence of the threshold choice

The high threshold t plays a decisive role in our framework as it determines which standardized features V_i^j are considered as extreme. Recall that the estimate $\hat{V}_i^j$ is discrete (see Section 3.2.2). A more convenient way to evaluate the influence of the threshold t is thus to consider instead $k := \#\Big\{i \in \{1, \ldots, n\} : ||\hat{V}_i||_\infty > t\Big\}$ the total number of extreme points. Two significant summaries of CLEF output which are the number of clusters $|\hat{\mathbb{M}}|$ and their average sizes $\frac{1}{|\hat{\mathbb{M}}|}\sum_{\alpha\in\hat{\mathbb{M}}}|\alpha|$, are plotted as a function of k. Figure 3.2 (hydrological data) and the first two panels of Figure 3.3 (simulated data) confirm the existence of stability regions (vertical red lines). The simulation experiments show that choosing the parameter in such regions ensures an optimal performance for CLEF, since both match exactly the one of lowest errors. The large width of the stability region for the simulated data (Figure 3.3 compared to Figure 3.2) may be explained by the fact that the generative model is a classical parametric extreme value model for which the asymptotic regime is nearly reached even for small thresholds, leading to large stability regions.

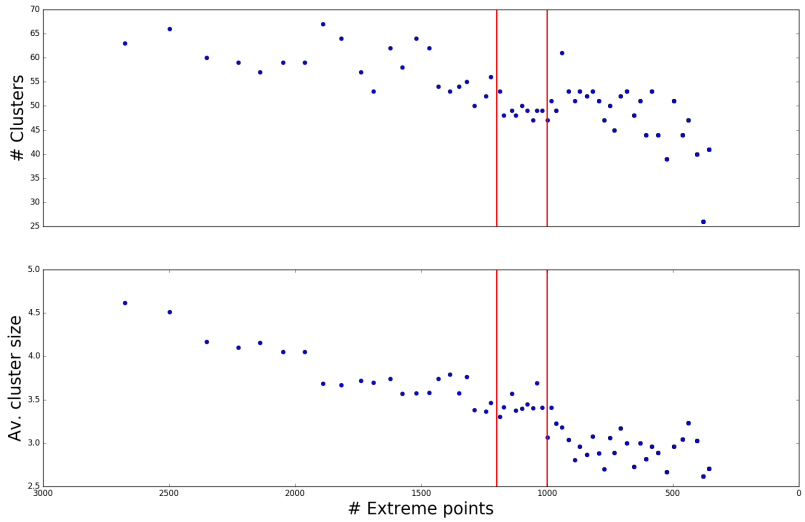


Figure 3.2: Stability region for k (number of extreme points) on the stream-flow data.

Upper panel: number of detected clusters, lower panel: average cluster size. Vertical red lines ($k \in \{1000, \ldots, 1200\}$ / $t \in [320, 400]$): stability region.

3.6 Conclusion

We propose a novel dimension reduction method for the analysis of extremes of multivariate datasets *via* feature clustering. This is done in adequacy with the framework of multivariate extreme value theory. The proposed algorithm makes use of the graphical structure of the problem, scanning the multiple possible subsets of features in a time efficient way. Results on a hydrological stream-flow data and on simulated data demonstrate the relevance of this approach on datasets which would not exhibit any sufficiently sparse structure when analyzed with existing algorithms. Future work will focus on the statistical properties of the empirical criteria $\hat{\kappa}_{\alpha,t}$ involved in the algorithm, which would allow to analyze the output as a statistical test for independence at extreme levels.

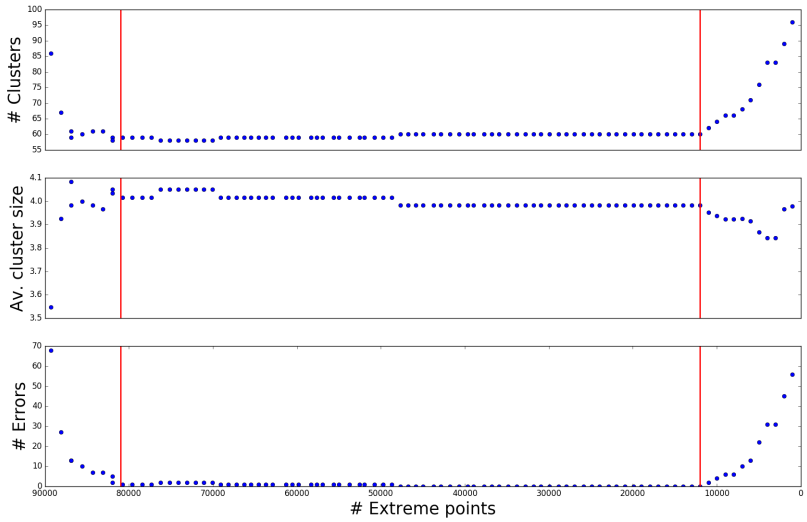


Figure 3.3: Stability region for k (number of extreme points) on simulated data.

Upper panel: number of detected clusters, middle panel: average cluster size, lower panel: number of errors of CLEF (as in table 3.2). Vertical red lines ($k \in \{12000, \ldots, 81000\}$): stability region.

3.7 Appendix: Proof of Lemma 3.1

Step 1. As a first step we show that $\mathcal{M} \subset \mathbb{M}$, *i.e.* $\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \mu(\Gamma_\alpha) > 0$.

Proof. Write $\mathcal{C}_\alpha = \bigcup_{\epsilon>0,\epsilon\in\mathbb{Q}} R_{\alpha,\epsilon}$, where $R_{\alpha,\epsilon} = \{x \in \mathbb{R}_+^d : \|x\|_\infty \geq 1; \quad x_j > \epsilon \ (j \in \alpha); \quad x_i = 0 \ (i \notin \alpha)\}$. Assume $\mu(\mathcal{C}_\alpha) > 0$. Since $\mu(\mathcal{C}_\alpha) < \infty$, by the monotonous limit property of the measure μ, we have $\mu(\mathcal{C}_\alpha) = \lim_{\epsilon\to 0} \mu(R_{\alpha,\epsilon})$. Also, from the definitions, $R_{\alpha,\epsilon} \subset \epsilon\Gamma_\alpha$. Thus,

$$\begin{aligned}\mu(\mathcal{C}_\alpha) > 0 \Rightarrow \exists \epsilon \in (0,1) : \mu(R_{\alpha,\epsilon}) > 0 \quad &\Rightarrow \mu(\epsilon\Gamma_\alpha) > 0 \\ \Rightarrow \rho_\alpha = \mu(\Gamma_\alpha) = \epsilon\mu(\epsilon\Gamma_\alpha) > 0.\end{aligned}$$

□

Step 2. We now prove the reverse inclusion for maximal elements of $\mathbb{M}$, *i.e.*

$$\alpha \text{ is maximal in } \mathbb{M} \quad \Rightarrow \alpha \in \mathcal{M}. \tag{3.13}$$

Proof. Consider, for $i \notin \alpha$, the set $\Delta_{i,\epsilon} = \Gamma_\alpha \cap \{x \in \mathbb{R}_+^d : \quad x_i > \epsilon\}$, so that $\Gamma_\alpha = \left\{ \bigcup_{\substack{i\in\{1,\ldots,d\}\setminus\alpha \\ \epsilon\in\mathbb{Q}\cap(0,1)}} \Delta_{i,\epsilon} \right\} \cup R_{\alpha,1}$. Thus,

$$\alpha \in \mathbb{M} \quad \Rightarrow \quad \mu(\Gamma_\alpha) > 0 \quad \Rightarrow \quad \Big(\exists i, \mu(\Delta_{i,\epsilon}) > 0 \quad \text{or} \quad \mu(R_{\alpha,1}) > 0\Big) \tag{3.14}$$

To prove (3.13), it is enough to show that

$$\alpha \in \mathbb{M} \quad \Rightarrow \quad \text{for } i \notin \alpha,\ \mu(\Delta_{i,\epsilon}) = 0. \tag{3.15}$$

Indeed if (3.15) is true, and if $\alpha \in \mathbb{M}$, then (3.14) implies that $\mu(R_{\alpha,1}) > 0$, and the result follows from the inclusion $R_{\alpha,1} \subset \mathcal{C}_\alpha$. We show (3.15) by contradiction. If $\mu(\Delta_{i,\epsilon}) > 0$ for some $i \notin \alpha$, then

$$\frac{1}{\epsilon}\Delta_{i,\epsilon} = \left(\frac{1}{\epsilon}\Gamma_\alpha\right) \cap \{x \in \mathbb{R}_+^d : x_i > 1\} \subset \Gamma_{\alpha\cup\{i\}},$$

thus $\mu(\Gamma_{\alpha\cup\{i\}}) > 0$, which contradicts the maximality of α in $\mathbb{M}$. □

Step 3. From (3.13), if α is maximal in $\mathbb{M}$ then $\alpha \in \mathcal{M}$. Now if α is maximal in $\mathbb{M}$ but not in $\mathcal{M}$, there exists $\beta \supsetneq \alpha$ in $\mathcal{M}$. Thus from Step 1, $\beta \in \mathbb{M}$, a contradiction. Hence α is also maximal in $\mathcal{M}$. Conversely, if α is maximal in $\mathcal{M}$ then (Step 1) $\alpha \in \mathbb{M}$. If α was not maximal in $\mathbb{M}$, there would exist $\beta \supsetneq \alpha$ maximal in $\mathbb{M}$, and from (3.13), $\beta \in \mathcal{M}$, contradicting the maximality of α in $\mathcal{M}$.

Acknowledgments

Part of this work has been funded by the the 'LabEx Mathématiques Hadamard' (LMH) project, by the 'AGREED' project from the PEPS JCJC program (INS2I, CNRS) and by the chair 'Machine Learning for Big Data' from Télécom ParisTech. The authors would like to thank Benjamin Renard for interesting discussions about the hydrological use case and for sharing the data.

4 Asymptotic Tests on the Coefficient of Tail Dependence

Abstract Identifying groups of variables that may be large simultaneously amounts to finding out which joint tail dependence coefficients of a multivariate distribution are positive. The asymptotic distribution of a vector of nonparametric, rank-based estimators of these coefficients justifies a stopping criterion in an algorithm that searches the collection of all possible groups of variables in a systematic way, from smaller groups to larger ones. The issue that the tolerance level in the stopping criterion should depend on the size of the groups is circumvented by the use of a conditional tail dependence coefficient. Alternatively, such stopping criteria can be based on limit distributions of rank-based estimators of the coefficient of tail dependence, quantifying the speed of decay of joint survival functions. Numerical experiments indicate that the algorithm's effectiveness for detecting tail-dependent groups of variables is highest when paired with a criterion based on a Hill-type estimator of the coefficient of tail dependence.

4.1 Introduction

A question that often arises when monitoring several variables is which groups of variables are prone to be large simultaneously. In food risk management, for instance, the variables under consideration may be the concentrations of different contaminants in blood samples of consumers. In environmental applications, one may be interested in several physical variables such as wind speed and precipitation recorded at several locations, with the purpose of setting off a regional warning when several of these variables exceed a high threshold. In the context of semi-supervised anomaly detection, when the training sample is mostly made of normal instances, identifying the groups of variables which are likely to be large together allows to label certain new instances as abnormal.

The latter use case is the motivation behind the DAMEX algorithm Goix et al. (2016b, 2017). In a regular variation framework, identifying those groups among d variables that may be large simultaneously amounts to identifying the support of the exponent measure. The algorithm returns the list of groups of features $\alpha \subset \{1, \ldots, d\}$ such that the mass of the empirical exponent measure on certain cones exceeds a user-defined threshold. However, when the empirical version of the exponent measure is scattered over a large number of such cones, the DAMEX

algorithm does not discover a clear-cut structure. Chiapino and Sabourin (2016) encounter this difficulty for extreme streamflow data recorded at several locations of the French river system.

To overcome this issue, the same authors come up with the CLEF (CLustering Extreme Features) algorithm. Instead of partitioning the sample space, CLEF considers nested regions corresponding to increasing subsets of components. A group of variables is enlarged until there is no longer enough evidence that all features in it may be large together. In this respect, CLEF resembles the Apriori algorithm Agrawal et al. (1994), which is a data-mining tool for discovering maximal sets of items among d available items that are frequently bought together by consumers. Apriori considers increasing itemsets that are made to grow until their frequency falls below a user-defined threshold. In CLEF, the stopping criterion concerns the relative frequency of simultaneous occurrences of large values of all components in a considered subset compared to the frequency of simultaneous occurrences of larges values of all but one component in this subset. Chiapino and Sabourin (2016) find the method to work well on real and simulated data but do not investigate the asymptotic properties of the statistic underlying the stopping criterion.

Our contributions are three-fold. First, we investigate the asymptotic behavior of the statistic underlying CLEF. In this way, the informal stopping criterion can be turned into a proper hypothesis test with controllable level. A second issue concerns the specification of the null hypothesis in the CLEF stopping criterion. Originally, a certain conditional tail dependence coefficient, κ_α, related to a given group of variables $\alpha \subset \{1, \ldots, d\}$ is supposed to be above a strictly positive, user-defined and therefore somewhat arbitrary threshold. We propose instead to base the stopping criterion on the hypothesis that a multivariate version of the coefficient of Ledford and Tawn (1996) and Ramos and Ledford (2009) is equal to one. The test is based on the limit distributions of multivariate extensions of nonparametric estimators in Peng (1999) and Draisma et al. (2001, 2004). Third, we conduct a numerical experiment to compare the finite-sample performance of the DAMEX algorithm and the CLEF algorithm with the various stopping criteria. We find that overall, the multivariate extension of the Hill-type estimator in Draisma et al. (2004) yields the most reliable procedure to detect maximal groups of asymptotically dependent variables.

Section 4.2 casts the problem in the language of regular variation and introduces the tail dependence coefficients upon which the CLEF stopping criteria will be based. Necessary background on empirical tail dependence functions and processes is reviewed in Section 4.3, including a new result for the empirical joint tail function. In Section 4.4, we derive the asymptotic distribution of the statistic used in CLEF and turn the heuristic stopping criterion implemented in Chiapino and Sabourin (2016) into a statistical test with asymptotically controllable level. Two alternative tests based on the asymptotic distributions of estimators of the Ledford–Tawn–Ramos coefficient of tail dependence are constructed in Sections 4.5 and 4.6. We report the results of our simulation experiments in Section 4.7. Section 4.8 concludes. Proofs

are gathered in Appendix 4.9 while the pseudo-code for the CLEF algorithm and variations is provided in Appendix 4.10.

4.2 Regular variation and tail dependence coefficients

Bold letters denote vectors and binary operations between vectors are understood componentwise. The indicator function of a set A is denoted by $\mathbb{1}_A$. For $t \in \mathbb{R} \cup \{\infty\}$, we let $\mathbf{t}_\alpha$ denote the constant vector of $(\mathbb{R} \cup \{\infty\})^\alpha$ with all coordinates equal to t. In the special case $\alpha = \{1, \ldots, d\}$, the index α is usually omitted for brevity when clear from the context: for instance, $\mathbf{0} = \mathbf{0}_{\{1,\ldots,d\}} = (0, \ldots, 0) \in \mathbb{R}^d$.

Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a random vector in $\mathbb{R}^d$ with cumulative distribution function F, whose margins $F_1, \ldots, F_d$ are continuous. We assume that the transformed vector $\mathbf{V} = (V_1, \ldots, V_d)$ with $V_j = 1/\{1 - F_j(X_j)\}$ for all $j \in \{1, \ldots, d\}$ is regularly varying on the cone $[0, \infty]^d \setminus \{\mathbf{0}\}$ with (nonzero) limit or exponent measure μ. This means that μ is finite on Borel sets of $[0, \infty]^d \setminus \{\mathbf{0}\}$ bounded away from the origin and that

$$\lim_{t \to \infty} t\mathbb{P}[\mathbf{V} \in tA] = \mu(A), \tag{4.1}$$

for all Borel sets $A \subset [0, \infty]^d \setminus \{\mathbf{0}\}$ such that $\mathbf{0} \notin \partial A$ and $\mu(\partial A) = 0$. The measure μ is homogeneous, i.e., $\mu(s\,\cdot) = s^{-1}\mu(\,\cdot\,)$ for all $0 < s < \infty$, and therefore assigns no mass to hyperplanes parallel to the coordinate axes. As a consequence, (4.1) applies to finite and infinite rectangles that are bounded away from the origin and whose sides are parallel to the coordinate axes. The measure μ characterizes the extremal dependence structure of $\mathbf{X}$. The reader is referred to Resnick (2007a, 2013) for an introduction to regular variation.

Let $\varnothing \neq \alpha \subset \{1, \ldots, d\}$. Particular instances of (4.1) include the extremal coefficient λ_α Schlather and Tawn (2003) and the joint tail coefficient ρ_α:

$$\lambda_\alpha = \lim_{t \to \infty} t\mathbb{P}[\exists j \in \alpha : V_j > t] = \mu(\{\mathbf{u} \in [0, \infty)^d \mid \exists j \in \alpha : u_j > 1\}), \tag{4.2}$$

$$\rho_\alpha = \lim_{t \to \infty} t\mathbb{P}[\forall j \in \alpha : V_j > t] = \mu(\{\mathbf{u} \in [0, \infty)^d \mid \forall j \in \alpha : u_j > 1\}). \tag{4.3}$$

In the bivariate case, $|\alpha| = 2$, and with our choice of Pareto margins, we have $\rho_\alpha = \lim_{t \to \infty} \mathbb{P}(V_{\alpha_1} > t \mid V_{\alpha_2} > t)$, the upper tail dependence coefficient denoted by χ in Coles et al. (1999).

Our general objective is to propose statistically sound procedures to recover maximal subgroups α of components that are likely to be concomitantly large. Our aim can thus be phrased as recovering the maximal subsets $\alpha \subset \{1, \ldots, d\}$ such that $\rho_\alpha > 0$.

Since $\rho_\alpha \leq \rho_\beta$ as soon as $\alpha \supset \beta$, any positive tolerance level with which we would like to compare an estimate of ρ_α should depend on α and in particular be decreasing as a function of the cardinality $|\alpha|$. To circumvent this issue, Chiapino and Sabourin (2016) consider for α such that $|\alpha| \geq 2$ the conditional tail dependence

coefficient

$$\kappa_\alpha = \lim_{t\to\infty} \mathbb{P}\left[\forall j \in \alpha : V_j > t \;\middle|\; \textstyle\sum_{j\in\alpha} \mathbb{1}\{V_j > t\} \geq |\alpha| - 1\right], \tag{4.4}$$

which is the limiting conditional probability that all variables in α exceed a large threshold given that all but at most one already do. In contrast to ρ_α, the coefficient κ_α has no particular reason to decrease as a function of $|\alpha|$. Note that $\rho_\alpha = \mu(\Gamma_\alpha)$ while $\kappa_\alpha = \mu(\Gamma_\alpha)/\mu(\Delta_\alpha) = \rho_\alpha/\mu(\Delta_\alpha)$ where $\Gamma_\alpha = \{\mathbf{x} \in [0,\infty)^d \mid \forall j \in \alpha : x_j > 1\}$ and $\Delta_\alpha = \{\mathbf{x} \in [0,\infty)^d \mid \sum_{j\in\alpha} \mathbb{1}_{\{x_j \geq 1\}} \geq |\alpha| - 1\}$, provided $|\alpha| \geq 2$. If $\mu(\Delta_\alpha) = 0$, then $\mu(\Gamma_\beta) = 0$ for all $\beta \subset \alpha$ with $|\beta| = |\alpha| - 1$; in that case, we define $\kappa_\alpha = 0$.

In the CLEF algorithm (Chiapino and Sabourin, 2016), the criterion to decide whether $\rho_\alpha > 0$ or not is that $\widehat{\kappa}_\alpha \geq C$, where C is a user-defined tolerance level, $\widehat{\kappa}_\alpha = \widehat{\mu}(\Gamma_\alpha)/\widehat{\mu}(\Delta_\alpha)$, and $\widehat{\mu}$ is the empirical exponent measure in (4.8) below. The level C can be chosen independently of α. Still, its choice is somewhat arbitrary, and in particular, the user has no control of false positives. In Section 4.4, we will provide the asymptotic distribution of $\widehat{\kappa}_\alpha$ and propose a test statistic with a guaranteed asymptotic level.

If $\rho_\alpha = 0$ (or $\kappa_\alpha = 0$), the limiting distributions of the statistics $\sqrt{k}(\widehat{\rho}_\alpha - \rho_\alpha)$ and $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha)$ are degenerate at zero. We therefore have no control on the asymptotic levels of tests based on those statistics under $H_0 : \kappa_0 = 0$. This is why will have to define a CLEF stopping criterion in terms of a test of $H_0 : \kappa_\alpha \geq \kappa_{\min}$ versus $H_1 : \kappa_\alpha < \kappa_{\min}$ instead, in terms of a user-defined level $\kappa_{\min} > 0$. The choice of $\kappa_{\min}$ is somewhat arbitrary; in the simulation experiments (Section 4.7), we choose $\kappa_{\min} = 0.08$.

In Sections 4.5 and 4.6, we consider alternative CLEF stopping criteria based on estimators of the coefficient of tail dependence $\eta_\alpha \in (0,1]$. For bivariate distributions, the coefficient has been introduced by Ledford and Tawn (1996) and extended by Ramos and Ledford (2009) in order to model situations in between asymptotic dependence ($\rho_{\{1,2\}} > 0$) and full independence of X_1 and X_2. De Haan and Zhou (2011) and Eastoe and Tawn (2012) proposed and studied a multivariate extension of η_α for $|\alpha| \geq 3$. The model assumption is that there exist $\eta_\alpha \in (0,1]$ and a slowly varying function $\mathcal{L}_\alpha$ such that

$$\mathbb{P}[\forall j \in \alpha : V_j > t] = t^{-1/\eta_\alpha} \mathcal{L}_\alpha(t). \tag{4.5}$$

Suppose that the limit ρ_α in (4.3) exists and that (4.5) holds. Then $\rho_\alpha > 0$ implies $\eta_\alpha = 1$. The converse is true as well, provided $\liminf_{t\to\infty} \mathcal{L}_\alpha(t) > 0$. Modulo this side condition, which we will take for granted, the null hypothesis $\rho_\alpha > 0$ corresponds to the simple hypothesis $\eta_\alpha = 1$.

We will test the null hypothesis $\eta_\alpha = 1$ via multivariate extensions of nonparametric estimators of η_α in Peng (1999) and Draisma et al. (2004). The null limit of the test statistic is non-degenerate, so that the asymptotic level of the test can be controlled, with no need to introduce an additional tolerance parameter $\kappa_{\min}$. The

estimators that we will study are related to the Pickands estimator and the Hill estimator for the extreme value index of $T_\alpha = \min_{j\in\alpha} V_j$, respectively. The maximum likelihood estimator, also considered in Draisma et al. (2004), is less suitable to our context due to its relative computational complexity, since the test is destined to be performed on a large number of subsets of $\{1,\dots,d\}$. See also the review Bacro and Toulemonde (2013) and the references therein.

Remark 4.1. The DAMEX algorithm (Goix et al., 2017) is designed to recover the family $\mathcal{M}$ of non-empty subsets α of $\{1,\dots,d\}$ with the property that

$$\mu\Big(\Big\{\mathbf{x}\in[0,\infty)^d \;\Big|\; \|\mathbf{x}\|_\infty \geq 1;\ \forall j\in\alpha\,, x_j>0 \ \text{ and } \forall j\notin\alpha,\ x_j=0\Big\}\Big) > 0.$$

In contrast, our focus is on $\mathbb{M} = \{\alpha \mid \rho_\alpha > 0\} = \{\alpha \mid \kappa_\alpha > 0\}$. Still, the maximal elements of $\mathbb{M}$ for the inclusion order are also the maximal elements of $\mathcal{M}$ (Chiapino and Sabourin, 2016, Lemma 1). The two problems of finding the maximal elements of $\mathbb{M}$ or $\mathcal{M}$ are thus equivalent.

4

4.3 Empirical tail dependence functions and processes

To find the asymptotic distribution of nonparametric estimators of the various dependence coefficients, we rely on empirical tail processes. Let the random vector $\mathbf{X} \sim F$ be as in Section 4.2; in particular, assume regular variation as in (4.1) with exponent measure μ. Let Λ be the push-forward measure of μ on $[0,\infty]^d \setminus \{\infty\}$ induced by the transformation $\mathbf{x} \mapsto 1/\mathbf{x} = (1/x_1,\dots,1/x_d)$, i.e., $\Lambda(\,\cdot\,) = \mu(\{\mathbf{x}\in[0,\infty]^d\setminus\{\mathbf{0}\} \mid 1/\mathbf{x}\in\,\cdot\,\})$.

For $\varnothing \neq \alpha \subset \{1,\dots,d\}$, consider the stable tail dependence function $\ell_\alpha : [0,\infty)^\alpha \to [0,\infty)$ and the joint tail dependence function $r_\alpha : [0,\infty]^\alpha\setminus\{\infty_\alpha\} \to [0,\infty)$ given by

$$\begin{aligned}
\ell_\alpha(\mathbf{x}) &= \lim_{t\to 0} t^{-1}\mathbb{P}[\exists j\in\alpha : F_j(X_j) > 1 - tx_j] = \Lambda(\{\mathbf{y} \mid \exists j\in\alpha : y_j < x_j\}),\\
r_\alpha(\mathbf{x}) &= \lim_{t\to 0} t^{-1}\mathbb{P}[\forall j\in\alpha : F_j(X_j) > 1 - tx_j] = \Lambda(\{\mathbf{y} \mid \forall j\in\alpha : y_j < x_j\}).
\end{aligned} \tag{4.6}$$

From (4.2) and (4.3), clearly $\lambda_\alpha = \ell_\alpha(\mathbf{1}_\alpha)$ and $\rho_\alpha = r_\alpha(\mathbf{1}_\alpha)$. For brevity, we write $\ell = \ell_{\{1,\dots,d\}}$ and $r = r_{\{1,\dots,d\}}$. Note that $\ell_\alpha(\boldsymbol{x}) = \ell(\boldsymbol{x}\mathbf{e}_\alpha)$ for $\boldsymbol{x}\in[0,\infty)^\alpha$, where $\mathbf{e}_\alpha \in \{0,1\}^d$ has components $\mathbf{e}_{\alpha,j} = \mathbb{1}_\alpha(j)$. Similarly, $r_\alpha(\boldsymbol{x}) = r(\boldsymbol{x}\boldsymbol{\iota}_\alpha)$ for $\boldsymbol{x} \in [0,\infty]^\alpha \setminus \{\boldsymbol{\infty}_\alpha\}$, where $\boldsymbol{\iota}_\alpha \in \{1,\infty\}^d$ denotes the vector such that $\boldsymbol{\iota}_{\alpha,j} = 1$ if $j\in\alpha$ and $\boldsymbol{\iota}_{\alpha,j} = +\infty$ otherwise. By the inclusion–exclusion formula, for $\boldsymbol{x}\in[0,\infty)^\alpha$, writing $\boldsymbol{x}_\beta = (x_j)_{j\in\beta}$, we have

$$r_\alpha(\mathbf{x}) = \sum_{\varnothing\neq\beta\subset\alpha} (-1)^{|\beta|+1}\ell_\beta(\mathbf{x}_\beta), \qquad \ell_\alpha(\mathbf{x}) = \sum_{\emptyset\neq\beta\subset\alpha} (-1)^{|\beta|+1} r_\beta(\mathbf{x}_\beta). \tag{4.7}$$

Let $\mathbf{X}_i = (X_{i,1},\dots,X_{i,d})$, for $i\in\{1,\dots,n\}$, be an independent random sample from F, having continuous margins and satisfying (4.1). Let $k = k(n) \to \infty$ as

$n \to \infty$, while $k(n) = \mathrm{o}(n)$. Following for instance Einmahl et al. (2012); Goix et al. (2017); Qi (1997), we rely on ranks to obtain an approximately Pareto-distributed sample $\widehat{\mathbf{V}}_i = (\widehat{V}_{i,1}, \ldots, \widehat{V}_{i,d})$. Let $\widehat{F}_j(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{\{X_{i,j} < x\}}$ be the (left-continuous) empirical distribution function of component $j \in \{1, \ldots, d\}$ and put $\widehat{V}_{i,j} = 1/\{1 - \widehat{F}_j(X_{i,j})\} = n/(n + 1 - R_{i,j})$, where $R_{i,j}$ is the rank of $X_{i,j}$ among $X_{1,j}, \ldots, X_{n,j}$. The empirical counterparts to μ and Λ are

$$\widehat{\mu}(\,\cdot\,) = \frac{1}{k} \sum_{i=1}^n \delta_{(k/n)\widehat{\mathbf{V}}_i}(\,\cdot\,), \qquad \widehat{\Lambda}(\,\cdot\,) = \frac{1}{k} \sum_{i=1}^n \delta_{(n/k)/\widehat{\mathbf{V}}_i}(\,\cdot\,), \tag{4.8}$$

respectively, with δ_a the Dirac measure at the point a. Replacing Λ by $\widehat{\Lambda}$ in the definition of ℓ_α and r_α produces the empirical tail dependence function

$$\begin{aligned} \widehat{\ell}_\alpha(\boldsymbol{x}) &= k^{-1} \textstyle\sum_{i=1}^n \mathbb{1}\{\exists j \in \alpha : n + 1 - R_{i,j} \le \lfloor k x_j \rfloor\} \\ &= k^{-1} \textstyle\sum_{i=1}^n \mathbb{1}\{\exists j \in \alpha : X_{i,j} \ge X_{(n - \lfloor k x_j \rfloor + 1), j}\} \end{aligned}$$

and the empirical joint tail function

$$\begin{aligned} \widehat{r}_\alpha(\boldsymbol{x}) &= k^{-1} \textstyle\sum_{i=1}^n \mathbb{1}\{\forall j \in \alpha : n + 1 - R_{i,j} \le \lfloor k x_j \rfloor\} \\ &= k^{-1} \textstyle\sum_{i=1}^n \mathbb{1}\{\forall j \in \alpha : X_{i,j} \ge X_{(n - \lfloor k x_j \rfloor + 1), j}\}, \end{aligned} \tag{4.9}$$

where $X_{(1),j} \le \ldots \le X_{(n),j}$ are the ascending order statistics of $X_{1,j}, \ldots, X_{n,j}$ and $\lfloor \cdot \rfloor$ is the floor function. The identities (4.7) hold for $\widehat{\ell}_\alpha$ and $\widehat{r}_\alpha$ as well.

Einmahl et al. (2012, Theorem 4.6) find the weak limit of the empirical process $\sqrt{k}(\widehat{\ell} - \ell)$ on $[0, T]^d$ for any $T > 0$. We leverage their theorem to show a similar result for $\sqrt{k}(\widehat{r}_\alpha - r_\alpha)$, jointly in α. The following conditions stem from the cited article.

Condition 1 (Uniform tail convergence). There exists $\gamma > 0$ such that, uniformly in $\mathbf{x} \in [0, 1]^d$ with $\sum_{j=1}^d x_j = 1$, we have

$$t^{-1} \mathbb{P}[\exists j = 1, \ldots, d : \, F_j(X_j) > t x_j] - \ell(\mathbf{x}) = \mathrm{O}(t^\gamma), \qquad t \to \infty.$$

Condition 2 (Moderate k). The sequence $k = k(n)$ satisfies $k = \mathrm{o}(n^{2\gamma/(1+2\gamma)})$ as $n \to \infty$, with $\gamma > 0$ as in Condition 1.

Condition 3 (Smoothness). For all $j \in \{1, \ldots, d\}$, the partial derivative $\partial_j \ell = \partial \ell / \partial x_j$ exists and is continuous on the set $\{\mathbf{x} \in [0, \infty)^d \mid x_j > 0\}$.

Since ℓ is convex, it is continuously differentiable Lebesgue almost everywhere (Rockafellar, 1970, Theorem 25.5). Condition 3 is satisfied for many popular max-stable models (logistic, asymmetric logistic, Brown–Resnick) but fails for max-linear models. Under Condition 3, the partial derivative $\partial_j r_\alpha = \partial r_\alpha / \partial x_j$ ($j \in \alpha$) exists and is continuous on $\{\mathbf{x} \in [0, \infty)^\alpha \mid x_j > 0\}$ and satisfies $\partial_j r_\alpha(\mathbf{x}) = \sum_{\beta : j \in \beta \subset \alpha} (-1)^{|\beta|+1} \partial_j \ell_\beta(\mathbf{x}_\beta)$, where $\mathbf{x}_\beta = (x_s)_{s \in \beta}$.

Einmahl (1997) and Einmahl et al. (2012) consider a centered Gaussian process W indexed by the Borel sets of $[0,\infty]^d \setminus \{\infty\}$ bounded away from ∞ with covariance function

$$\mathbb{E}[W(A)\,W(B)] = \Lambda(A \cap B). \tag{4.10}$$

Note that $W(\emptyset) = 0$ almost surely. For $\emptyset \neq \alpha \subset \{1,\ldots,d\}$ and $\boldsymbol{x} \in [0,\infty)^\alpha$, write

$$W_\alpha(\boldsymbol{x}) = W(\{\mathbf{y} \in [0,\infty]^d \mid \forall j \in \alpha : y_j < x_j\}).$$

We consider weak convergence as in van der Vaart (1998); van der Vaart and Wellner (1996); notation $\rightsquigarrow$. We work in the metric space $\ell^\infty(S)$ of bounded, real functions f on an arbitrary set S, the metric being the one induced by the supremum norm, $\|f\|_\infty = \sup_{x\in S}|f(x)|$; the double use of the symbol ℓ should not give rise to any confusion. The proof of the following proposition and of other results in the paper is deferred to Appendix 4.9.

Proposition 4.2. *Let $\mathbf{X}_i = (X_{i,1},\ldots,X_{i,d})$, for $i \in \{1,\ldots,n\}$, be an independent random sample from F, having continuous margins and satisfying (4.1). Let $k = k(n) \to \infty$ as $n \to \infty$, while $k(n) = \mathrm{o}(n)$. If Conditions 1, 2 and 3 hold, then, for $T > 0$, in the product space $\prod_{\varnothing\neq\alpha\subset\{1,\ldots,d\}} \ell^\infty([0,T]^\alpha)$, we have, as $n \to \infty$, the weak convergence*

$$\sqrt{k}\,\{\widehat{r}_\alpha(\mathbf{x}) - r_\alpha(\mathbf{x})\} \rightsquigarrow W_\alpha(\mathbf{x}) - \sum_{j\in\alpha} \partial_j r_\alpha(\mathbf{x})\, W_{\{j\}}(x_j) = Z_\alpha(\mathbf{x}). \tag{4.11}$$

4.4 Estimating the conditional tail dependence coefficient

This section investigates the asymptotic distribution of the empirical conditional dependence coefficient $\widehat{\kappa}_\alpha$ based on the empirical exponent measure $\widehat{\mu}$. This is achieved by re-writing $\widehat{\kappa}_\alpha$ as a function of the empirical joint tail coefficients $\widehat{\rho}_\alpha$, the distribution of which follows from Proposition 4.2. We also propose consistent estimators of the asymptotic variance of $\widehat{\kappa}_\alpha$. Combining the two yields a test for the null hypothesis $\kappa_\alpha \geq \kappa_{\min}$ where $\kappa_{\min} \in (0,1)$ is a tolerance level fixed by the user, to be seen as the minimal limiting conditional probability that all components in a random vector exceed a threshold, given that all of them but at most one already do.

Let $\varnothing \neq \alpha \subset \{1,\ldots,d\}$ and recall the sets $\Gamma_\alpha = \{\mathbf{x} \in [0,\infty)^d \mid \forall j \in \alpha : x_j > 1\}$ and, provided α has at least two elements, $\Delta_\alpha = \{\mathbf{x} \in [0,\infty)^d \mid \sum_{j\in\alpha} \mathbb{1}_{\{x_j\geq 1\}} \geq |\alpha| - 1\}$. Write $\alpha \setminus j = \alpha \setminus \{j\}$ for $j \in \alpha$. Since Δ_α is the disjoint union of the sets $\Gamma_{\alpha\setminus j} \setminus \Gamma_\alpha$ and Γ_α, where $j \in \alpha$, we find, for every Borel measure ν, the equality

$$\nu(\Delta_\alpha) = \sum_{j\in\alpha} \nu(\Gamma_{\alpha\setminus j}) - (|\alpha| - 1)\,\nu(\Gamma_\alpha). \tag{4.12}$$

Recall $\rho_\alpha = \mu(\Gamma_\alpha)$ and $\kappa_\alpha = \mu(\Gamma_\alpha)/\mu(\Delta_\alpha)$ in (4.4). By (4.12) applied to $\nu = \mu$, we have

$$\kappa_\alpha = \frac{\rho_\alpha}{\sum_{j\in\alpha} \rho_{\alpha\setminus j} - (|\alpha| - 1)\rho_\alpha}. \tag{4.13}$$

Recall the joint tail function r_α and its nonparametric estimator $\widehat{r}_\alpha$ in (4.6) and (4.9), respectively. Since $\rho_\alpha = r_\alpha(\mathbf{1}_\alpha)$, we define the estimators $\widehat{\rho}_\alpha = \widehat{\mu}(\Gamma_\alpha) = \widehat{r}_\alpha(\mathbf{1}_\alpha)$ and, provided $|\alpha| \geq 2$,

$$\widehat{\kappa}_\alpha = \frac{\widehat{\mu}(\Gamma_\alpha)}{\widehat{\mu}(\Delta_\alpha)} = \frac{\widehat{\rho}_\alpha}{\sum_{j\in\alpha} \widehat{\rho}_{\alpha\setminus j} - (|\alpha| - 1)\widehat{\rho}_\alpha}.$$

The asymptotic distribution of the vector of empirical joint tail coefficients follows immediately from Proposition 4.2. Write $\dot{\rho}_{\alpha,j} = \partial_j r_\alpha(\mathbf{1}_\alpha)$.

Corollary 4.3. In the setting of Proposition 4.2, we have, jointly in $\varnothing \neq \alpha \subset \{1, \ldots, d\}$, the weak convergence

$$\sqrt{k_n}\,(\widehat{\rho}_\alpha - \rho_\alpha) \rightsquigarrow Z_\alpha(\mathbf{1}_\alpha) = G_\alpha, \qquad n \to \infty. \tag{4.14}$$

The limit distribution is centered Gaussian with covariance matrix

$$\mathbb{E}[G_\alpha G_{\alpha'}] = \rho_{\alpha\cup\alpha'} - \sum_{j\in\alpha} \dot{\rho}_{j,\alpha}\rho_{\alpha'\cup\{j\}} - \sum_{j'\in\alpha'} \dot{\rho}_{j',\alpha'}\rho_{\alpha\cup\{j'\}} + \sum_{j\in\alpha}\sum_{j'\in\alpha'} \dot{\rho}_{j,\alpha}\,\dot{\rho}_{j',\alpha'}\,\rho_{\{j,j'\}}. \tag{4.15}$$

The asymptotic distribution of $\widehat{\kappa}_\alpha$ follows from the one of $(\widehat{\rho}_\beta)_\beta$ via the delta method. The asymptotic variance involves the partial derivative $\partial_j\kappa_\alpha = \partial\kappa_\alpha/\partial x_j$ of the function

$$\kappa_\alpha(\mathbf{x}) = \frac{r_\alpha(\mathbf{x})}{\sum_{j\in\alpha} r_{\alpha\setminus j}(\mathbf{x}_{\alpha\setminus j}) - (|\alpha| - 1)r_\alpha(\mathbf{x})} \tag{4.16}$$

for $\mathbf{x} \in [0, \infty)^\alpha$. Note that $\kappa_\alpha(\mathbf{1}_\alpha) = \kappa_\alpha$. Write $\dot{\kappa}_{j,\alpha} = \partial_j\kappa_\alpha(\mathbf{1}_\alpha)$.

Proposition 4.4. *In the setting of Corollary 4.3, we have, as $n \to \infty$ and jointly in $\alpha \subset \{1, \ldots, d\}$ such that $|\alpha| \geq 2$ and $\mu(\Delta_\alpha) > 0$, the weak convergence*

$$\sqrt{k}\,(\widehat{\kappa}_\alpha - \kappa_\alpha) \rightsquigarrow \mu(\Delta_\alpha)^{-2}\left\{\left(\textstyle\sum_{j\in\alpha} \rho_{\alpha\setminus j}\right) G_\alpha - \rho_\alpha \textstyle\sum_{j\in\alpha} G_{\alpha\setminus j}\right\}. \tag{4.17}$$

For a fixed such α, the limit distribution is $\mathcal{N}(0, \sigma^2_{\kappa,\alpha})$ with

$$\sigma^2_{\kappa,\alpha} = \left(1 - \kappa_\alpha\right)\kappa_\alpha \left\{\mu(\Delta_\alpha)^{-1} - \textstyle\sum_{j\in\alpha} \dot{\kappa}_{j,\alpha}\right\} + \sum_{i\in\alpha}\sum_{j\in\alpha} \dot{\kappa}_{i,\alpha}\dot{\kappa}_{j,\alpha}\rho_{\{i,j\}} + \kappa_\alpha \sum_{j\in\alpha} \dot{\kappa}_{j,\alpha}\left\{1 - \mu(\Delta_\alpha)^{-1}\rho_{\alpha\setminus j}\right\}. \tag{4.18}$$

Following Peng (1999), the asymptotic variance $\sigma^2_{\kappa,\alpha}$ in (4.18) can be estimated consistently by estimating the partial derivatives $\dot{\kappa}_{i,\alpha}$ via finite differencing applied

to the empirical version of $\kappa_\alpha(\mathbf{x})$ in (4.16) obtained by replacing r_α and $r_{\alpha\setminus j}$ by $\widehat{r}_\alpha$ and $\widehat{r}_{\alpha\setminus j}$, respectively:

$$\widehat{\kappa}_\alpha(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}\{\forall j \in \alpha : X_{i,j} \geq X_{(n-\lfloor kx_j \rfloor+1),j}\}}{\sum_{i=1}^n \mathbb{1}\{\exists m \in \alpha : \forall j \in \alpha \setminus m : X_{i,j} \geq X_{(n-\lfloor kx_j \rfloor+1),j}\}}$$

Define

$$\dot{\kappa}_{j,\alpha,n} = \frac{1}{2k^{-1/4}} \left\{ \widehat{\kappa}_\alpha(\mathbf{1}_\alpha + k^{-1/4}\mathbf{e}_j) - \widehat{\kappa}_\alpha(\mathbf{1}_\alpha - k^{-1/4}\mathbf{e}_j) \right\}, \tag{4.19}$$

with $\mathbf{e}_j$ the canonical unit vector of $\mathbb{R}^\alpha$ pointing in direction $j \in \alpha$, and put

$$\begin{aligned}\widehat{\sigma}^2_{\kappa,\alpha} = \left(1 - \widehat{\kappa}_\alpha\right)\widehat{\kappa}_\alpha \left\{ \widehat{\mu}(\Delta_\alpha)^{-1} - \textstyle\sum_{j\in\alpha} \dot{\kappa}_{j,\alpha,n} \right\} + \sum_{i,j\in\alpha} \dot{\kappa}_{i,\alpha,n}\dot{\kappa}_{j,\alpha,n}\widehat{\rho}_{\{i,j\}} \\ + \widehat{\kappa}_\alpha \sum_{j\in\alpha} \dot{\kappa}_{j,\alpha,n} \left\{ 1 - \widehat{\mu}(\Delta_\alpha)^{-1}\widehat{\rho}_{\alpha\setminus j} \right\}.\end{aligned} \tag{4.20}$$

Proposition 4.5. *Under the conditions of Proposition 4.4, we have* $\widehat{\sigma}^2_{\kappa,\alpha} = \sigma^2_{\kappa,\alpha} + \mathrm{o}_{\mathbb{P}}(1)$ *as* $n \to \infty$*, so that* $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha)/\widehat{\sigma}_{\kappa,\alpha} \rightsquigarrow \mathcal{N}(0,1)$*, provided* $\sigma^2_{\kappa,\alpha} > 0$*.*

The proof relies on the weak convergence of the empirical process $\sqrt{k}\{\widehat{\kappa}_\alpha(\,\cdot\,) - \kappa_\alpha(\,\cdot\,)\}$ on $[0,T]^\alpha$ for any $T > 0$. This property follows in turn from Proposition 4.2 and the functional delta method.

We consider a tolerance level $\kappa_{\min} \in (0,1)$ under which the tail dependence between components $j \in \alpha$ is deemed negligible compared to the one between components $j \in \beta \subsetneq \alpha$. In other words, we aim at testing $H_0 : \kappa_\alpha \geq \kappa_{\min}$. Since $\kappa_\alpha = \rho_\alpha/\mu(\Delta_\alpha)$, the null hypothesis is that ρ_α is greater than some level depending on α. Let $0 < \delta < 1$ be a (small) probability, and consider the test

$$\tau_{\alpha,n} = \mathbb{1}\left\{ \widehat{\kappa}_\alpha < \kappa_{\min} + q_\delta k^{-1/2}\widehat{\sigma}_{\kappa,\alpha} \right\} \tag{4.21}$$

where q_δ is the δ-quantile of the standard normal distribution. By Proposition 4.5, if $\sigma_{\kappa,\alpha} > 0$, the test in (4.21) has asymptotic level δ for H_0 against $H_1 : \kappa_\alpha < \kappa_{\min}$.

If $\rho_\alpha = 0$, then, in Proposition 4.3, we have $\sqrt{k}(\widehat{\rho}_\alpha - \rho_\alpha) = \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$: indeed, on the one hand, we have $\sqrt{k}(\widehat{\rho}_\alpha - \rho_\alpha) = \sqrt{k}\widehat{\rho}_\alpha \geq 0$, and on the other hand, its limit distribution is centered Gaussian. Likewise, we have $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha) = \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$ in Proposition 4.4 if $\kappa_\alpha = 0$. As a consequence, under the simple hypothesis $H_0 : \rho_\alpha = 0$, the asymptotic level of a test based on the asymptotic distribution of $\sqrt{k}(\widehat{\rho}_\alpha - \rho_\alpha)$ or $\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha)$ cannot be controlled. This is why the test in (4.21) concerns the null hypothesis $H_0 : \kappa_\alpha \geq \kappa_{\min}$ for some $\kappa_{\min} > 0$ instead. Alternatively, we propose tests based on estimators of the coefficient of tail dependence η_α in (4.5). In Sections 4.5 and 4.6, we consider two such estimators, extending the ones of Peng (1999) and Draisma et al. (2004), respectively, to the multivariate setting.

4.5 Coefficient of tail dependence: Peng's estimator

For bivariate distributions, Peng's (Peng, 1999) estimator of the coefficient of tail dependence $\eta = \eta_{\{1,2\}}$ is based on the property that the curve $t \mapsto (\log t, \log \mathbb{P}[V_1 > t, V_2 > t])$ has an affine asymptote with slope $-1/\eta$. A similar idea motivates Pickands' (Pickands III, 1975) estimator for the extreme value index. Estimating the ordinate of the curve at $t = n/k$ and $t = n/(2k)$ allows to estimate that slope. Under a second-order regular variation condition, Peng (1999) shows that his estimator is asymptotically normal, both if $\eta = 1$ and if $\eta < 1$. In the former case, the asymptotic variance depends on the tail dependence function and its partial derivatives, which are unknown but may be estimated consistently, thus leading to tests whose asymptotic levels can be controlled.

Let $\alpha \subset \{1, \dots, d\}$ have at least two elements. Recall the empirical joint tail function $\widehat{r}_\alpha$ in (4.9). We define the multivariate extension of Peng's (Peng, 1999) estimator of η_α in (4.5) as

$$\widehat{\eta}_\alpha^P = \log(2)/\log\{\widehat{r}_\alpha(\mathbf{2}_\alpha)/\widehat{r}_\alpha(\mathbf{1}_\alpha)\}. \tag{4.22}$$

The asymptotic normality of $\widehat{\eta}_\alpha^P$ follows from Proposition 4.2 and the delta method.

Proposition 4.6. *In the setting of Proposition 4.2, we have, as $n \to \infty$ and jointly in $\alpha \subset \{1, \dots, d\}$ such that $|\alpha| \geq 2$ and $\rho_\alpha > 0$, the weak convergence*

$$\sqrt{k}(\widehat{\eta}_\alpha^P - 1) \rightsquigarrow \frac{-1}{2\rho_\alpha \log 2} \left\{ Z_\alpha(\mathbf{2}_\alpha) - 2Z_\alpha(\mathbf{1}_\alpha) \right\}.$$

The right-hand side is a $\mathcal{N}(0, \sigma^2_{\alpha,P})$ random variable with variance

$$\sigma^2_{\alpha,P} = \frac{1}{2(\rho_\alpha \log 2)^2} \Bigg[\rho_\alpha - 4\rho_\alpha^2 + 2 \sum_{j \in \alpha} \dot{\rho}_{j,\alpha} r_\alpha(\mathbf{2}_\alpha \wedge \boldsymbol{\iota}_j) + \sum_{j \in \alpha} \sum_{j' \in \alpha} \dot{\rho}_{j,\alpha} \dot{\rho}_{j',\alpha} \left\{ 3\rho_{\{j,j'\}} - 2r_{\{j,j'\}}(2,1) \right\} \Bigg], \tag{4.23}$$

where $\rho_{\{j,j'\}} = r_{\{j,j'\}}(2,1) = 1$ if $j = j'$ and where $\boldsymbol{\iota}_j \in \{1, \infty\}^\alpha$ is the vector which all coordinates equal to 1 except for the j-th one which equals ∞, so that $(\mathbf{2}_\alpha \wedge \boldsymbol{\iota}_j)_m = 1$ if $m \in \alpha \setminus j$ and $(\mathbf{2}_\alpha \wedge \boldsymbol{\iota}_j)_m = 2$ if $m = j$.

By extending the proof of (Peng, 1999, Theorem 2.1), it is also possible to obtain asymptotic normality of $\widehat{\eta}_\alpha^P$ in the case $\rho_\alpha = 0$ and $\eta_\alpha < 1$ in (4.5). This would require a multivariate extension of the second-order regular variation condition in Peng (1999) in the style of Condition 4 below. For the application as a stopping criterion in the CLEF algorithm, we are only interested in the asymptotic distribution of $\widehat{\eta}_\alpha^P$ under the hypothesis $\rho_\alpha > 0$, so we do not pursue this idea any further.

As in Proposition 4.4, the asymptotic variance $\sigma^2_{\alpha,P}$ in (4.23) involves unknown quantities, all of which we can estimate consistently. For $\alpha \subset \{1,\dots,d\}$ and $j \in \alpha$, define

$$\dot{\rho}_{j,\alpha,n} = \frac{1}{2k^{-1/4}} \left\{ \widehat{r}_\alpha(\mathbf{1}_\alpha + k^{-1/4}\mathbf{e}_j) - \widehat{r}_\alpha(\mathbf{1}_\alpha - k^{-1/4}\mathbf{e}_j) \right\}, \tag{4.24}$$

where $\mathbf{e}_j$ is the canonical unit vector in $\mathbb{R}^\alpha$ pointing in dimension j. Define

$$\widehat{\sigma}^2_{\alpha,P} = \frac{1}{2(\widehat{\rho}_\alpha \log 2)^2} \Bigg[\widehat{\rho}_\alpha + \sum_{j\in\alpha} \dot{\rho}_{j,\alpha,n} \{-4\widehat{\rho}_\alpha + 2\widehat{r}_\alpha(\mathbf{2}_\alpha \wedge \boldsymbol{\iota}_j)\} + \sum_{j\in\alpha}\sum_{j'\in\alpha} \dot{\rho}_{j,\alpha,n}\dot{\rho}_{j',\alpha,n} \left\{ 3\widehat{\rho}_{\{j,j'\}} - 2\widehat{r}_{\{j,j'\}}(2,1) \right\} \Bigg]. \tag{4.25}$$

Proposition 4.7. *In the setting of Proposition 4.2, we have $\widehat{\sigma}^2_{\alpha,P} = \sigma^2_{\alpha,P} + \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$, where $\alpha \subset \{1,\dots,d\}$ is such that $|\alpha| \geq 2$ and $\rho_\alpha > 0$. If $\sigma^2_{\alpha,P} > 0$, then $\sqrt{k}(\widehat{\eta}^P_\alpha - 1)/\widehat{\sigma}_{\alpha,P} \rightsquigarrow \mathcal{N}(0,1)$ as $n \to \infty$.*

The proof parallels the one of Proposition 4.5 and is omitted for brevity. The main step is to verify that $\dot{\rho}_{j,\alpha,n} = \dot{\rho}_{j,\alpha} + \mathrm{o}_{\mathbb{P}}(1)$ as $n \to \infty$, which follows from Proposition 4.2.

To test the hypothesis $H_0 : \rho_\alpha > 0$ at significance level $\delta \in (0,1)$, we propose

$$\tau_{\alpha,\eta^P,n} = \mathbb{1}\left\{ \widehat{\eta}^P_\alpha < 1 - q_{1-\delta} k^{-1/2} \widehat{\sigma}_{\alpha,P} \right\}, \tag{4.26}$$

where $q_{1-\delta}$ is the $(1-\delta)$-quantile of the standard normal distribution. In the setting of Proposition 4.7, the test in (4.26) has asymptotic level δ for H_0 against $H_1 : \eta_\alpha < 1$.

4.6 Coefficient of tail dependence: Hill estimator

The coefficient of tail dependence η_α in (4.5) is the tail index of the random variable $T_\alpha = \min_{j\in\alpha} V_j$: the function $t \mapsto \mathbb{P}[T_\alpha > t]$ is regularly varying at infinity with index $-1/\eta_\alpha$. A tractable alternative to Peng's estimator for η_α is a Hill-type estimator as in Draisma et al. (2001, 2004). Replacing the unobservable Pareto variables $V_{i,j}$ by the rank-based versions $\widehat{V}_{i,j} = n/(n+1-R_{ij})$ in Section 4.3 yields an approximate sample

$$\widehat{T}_{i,\alpha} = \min_{j\in\alpha} \widehat{V}_{i,j}, \qquad i = 1,\dots,n,$$

from the distribution of T_α. Let $\widehat{T}_{(1),\alpha} \leq \dots \leq \widehat{T}_{(n),\alpha}$ denote the order statistics of $\widehat{T}_{1,\alpha},\dots,\widehat{T}_{n,\alpha}$. The Hill estimator for η_α is defined as

$$\widehat{\eta}^H_\alpha = \frac{1}{k}\sum_{i=1}^{k} \log \frac{\widehat{T}_{(n-i+1),\alpha}}{\widehat{T}_{(n-k),\alpha}}. \tag{4.27}$$

Under the second-order regular variation conditions stated below, the asymptotic normality of $\widehat{\eta}_\alpha^H$ follows from (Draisma et al., 2004, proof of Theorem 2.1). The results in the cited reference cover the bivariate case only. In this section, we verify that they remain valid in any dimension $d \geq 2$, and we provide the general expression for the asymptotic variance. Put $E_\alpha = [0,\infty]^\alpha \setminus \{\boldsymbol{\infty}_\alpha\}$.

Condition 4. For each $\alpha \subset \{1,\ldots,d\}$ with $|\alpha| \geq 2$, there exist functions $c_\alpha, c_{1,\alpha} : E_\alpha \to [0,\infty)$ such that $c_{1,\alpha}$ is neither constant nor a multiple of c_α, and there exists $q_{1,\alpha} : (0,\infty) \to (0,\infty)$, with $q_{1,\alpha}(t) \to 0$ as $t \to 0$, such that, for all $\mathbf{x} \in E_\alpha$, we have

$$\lim_{t\to 0} \left\{ \frac{\mathbb{P}[\forall j \in \alpha : 1 - F_j(X_j) \leq t x_j]}{\mathbb{P}[\forall j \in \alpha : 1 - F_j(X_j) \leq t]} - c_\alpha(\mathbf{x}) \right\} \Big/ q_{1,\alpha}(t) = c_{1,\alpha}(\mathbf{x}).$$

Under Condition 4, the function $q_\alpha(t) = \mathbb{P}[\forall j \in \alpha : 1 - F_j(X_j) \leq t]$ is regularly varying at 0 with some index $1/\eta_\alpha$. Condition 4 implies that the first-order condition (4.5) holds with the same index $1/\eta_\alpha$. In addition, $c_\alpha(\mathbf{1}_\alpha) = 1$ and c_α is homogeneous of order $1/\eta_\alpha$, i.e., $c_\alpha(t\mathbf{x}) = t^{1/\eta_\alpha} c_\alpha(\mathbf{x})$ for $t > 0$, see Draisma et al. (2001, 2004). Under the regular variation assumption (4.1), we have $\rho_\alpha = \lim_{t\to 0} q_\alpha(t)/t$, so that, under Condition 4, $\rho_\alpha > 0$ implies $\eta_\alpha = 1$, as in Draisma et al. (2004) for the bivariate case. Finally, if $\rho_\alpha > 0$, then $c_\alpha(\mathbf{x}) = r_\alpha(\mathbf{x})/r_\alpha(\mathbf{1}_\alpha) = r_\alpha(\mathbf{x})/\rho_\alpha$. Note that in Draisma et al. (2004), our ρ_α is denoted by l for $\alpha = \{1,2\}$.

The asymptotic variance of the Hill estimator (4.27) involves a Gaussian process whose distribution depends on whether $\rho_\alpha = 0$ or $\rho_\alpha > 0$. As in Draisma et al. (2004), introduce a centered Gaussian process W_1 on E_α with covariance function $\mathbb{E}[W_1(\mathbf{x})\, W_1(\mathbf{y})] = c_\alpha(\mathbf{x} \wedge \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in E_\alpha$. Recall the stochastic process Z_α in (4.11) and the random variable $G_\alpha = Z_\alpha(\mathbf{1}_\alpha)$ in (4.14).

Proposition 4.8. *Let $\mathbf{X}_i = (X_{i,1},\ldots,X_{i,d})$, for $i \in \{1,\ldots,n\}$, be an independent random sample from F, having continuous margins and satisfying (4.1). Let $k = k(n) \to \infty$ as $n \to \infty$, while $k(n) = \mathrm{o}(n)$. If Conditions 1, 2, 3, and 4 hold, then, as $n \to \infty$,*

$$\sqrt{k}\left(\widehat{\eta}_\alpha^H - \eta_\alpha\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\alpha,H}),$$

with $\sigma^2_{\alpha,H} = \eta_\alpha^2 \operatorname{Var}\{\tilde{W}(\mathbf{1}_\alpha)\}$, where $\tilde{W}(\mathbf{x}) = W_1(\mathbf{x})$ if $\rho_\alpha = 0$ and $\tilde{W}(\mathbf{x}) = \rho_\alpha^{-1/2} Z_\alpha(\mathbf{x})$ if $\rho_\alpha > 0$. In particular, if $\rho_\alpha > 0$, we have

$$\sigma^2_{\alpha,H} = \rho_\alpha^{-1} \operatorname{Var}(G_\alpha) = 1 - 2\rho_\alpha + \rho_\alpha^{-1} \sum_{j\in\alpha} \sum_{j'\in\alpha} \dot{\rho}_{j,\alpha} \dot{\rho}_{j',\alpha} \rho_{\{j,j'\}}. \tag{4.28}$$

The proof of Proposition 4.8 is based on the arguments developed in the proofs of (Draisma et al., 2004, Theorem 2.1), (Drees, 1998b, Theorem 3.2), and (Drees, 1998a, Example 3.1), which we gather in Appendix 4.9.

Again, the unknown terms in (4.28) may be replaced by their empirical counterparts, leading to an asymptotically consistent test. Recall $\dot{\rho}_{j,\alpha,n}$ in (4.24) and define

$$\widehat{\sigma}^2_{\alpha,H} = 1 - 2\widehat{\rho}_\alpha + \widehat{\rho}_\alpha^{-1} \sum_{j\in\alpha} \sum_{j'\in\alpha} \dot{\rho}_{j,\alpha,n} \dot{\rho}_{j',\alpha,n} \widehat{\rho}_{\{j,j'\}}.$$

The proof of the consistency of the variance estimator follows the same lines as the proofs of Propositions 4.5 and 4.7 and is omitted.

Corollary 4.9. Under the conditions of Proposition 4.8, if $\rho_\alpha > 0$, we have $\widehat{\sigma}^2_{\alpha,H} = \sigma^2_{\alpha,H} + o_{\mathbb{P}}(1)$ as $n \to \infty$ and thus $\sqrt{k}(\widehat{\eta}^H_\alpha - 1)/\widehat{\sigma}_{\alpha,P} \rightsquigarrow \mathcal{N}(0,1)$, provided $\sigma^2_{\alpha,H} > 0$.

We may exploit Corollary 4.9 to test $H_0 : \rho_\alpha > 0$ in the same way as we did by using Peng's estimator in (4.26): at significance level $\delta \in (0,1)$, the null hypothesis is rejected in favour of $H_1 : \eta_\alpha < 1$ when $\widehat{\eta}^H_\alpha < 1 - q_{1-\delta} k^{-1/2} \widehat{\sigma}_{\alpha,H}$.

Remark 4.10. The condition $\sigma^2_{\alpha,H} > 0$ in Corollary 4.9 is satisfied whenever $0 < \rho_\alpha < 1$. Indeed, in (4.28), we have $\rho_{\{j,j'\}} \geq \rho_\alpha$ and $\dot{\rho}_{j,\alpha}\dot{\rho}_{j',\alpha} \geq 0$, whence $\sigma^2_{\alpha,H} \geq 1 - 2\rho_\alpha + \sum_{(j,j')\in\alpha^2} \dot{\rho}_{j,\alpha}\dot{\rho}_{j',\alpha} = 1 - 2\rho_\alpha + \rho^2_\alpha = (1-\rho_\alpha)^2$.

4.7 Simulation study

Our aim is to compare the finite sample performance of the various tests proposed in Sections 4.4, 4.5 and 4.6 within the framework of the CLEF algorithm, the pseudo-code of which is given in Appendix 4.10. Three variants of the CLEF algorithm are obtained by varying the criterion according to which a subset α is declared as tail-dependent: $\widehat{\kappa}_\alpha > \kappa_{\min} - q_\delta \widehat{\sigma}_{\kappa,\alpha}/\sqrt{k}$ for CLEF-asymptotic; $\widehat{\eta}_{\alpha,P} > 1 - q_\delta \widehat{\sigma}_{\alpha,P}/\sqrt{k}$ for CLEF-Peng; and $\widehat{\eta}_{\alpha,H} > 1 - q_\delta \widehat{\sigma}_{\alpha,H}/\sqrt{k}$ for CLEF-Hill. The original CLEF criterion was $\widehat{\kappa}_\alpha > C$ for some constant C chosen by the user. For completeness, the output of the DAMEX algorithm Goix et al. (2016b) is included in the comparison.

In practice, the dependence tests based on the tail dependence coefficient should not be carried out to the letter when the test statistic is not defined or when its estimated variance is infinite. Thus, in our experiments, CLEF-Peng and CLEF-Hill are modified so as to take into account additional, common-sense stopping criteria. A subset α will *not* be part of the list returned by the algorithms under the following conditions:

1. Concerning CLEF-Hill, when $\widehat{\rho}_\alpha = 0$, that is, no extreme record impacts all coordinates in α, the estimated variance of the Hill estimator of η_α is infinite. Therefore, $\widehat{\rho}_\alpha = 0$ is considered as a stopping criterion in CLEF-Hill.

2. Concerning CLEF-Peng, when $\widehat{r}_\alpha(\mathbf{2}_\alpha) = \widehat{r}_\alpha(\mathbf{1}_\alpha)$, the Peng estimator (4.22) is ill-defined. Such a case arises when there are very few points in the joint tail within the subspace generated by α. When the estimated derivatives $\dot{\rho}_{j,\alpha,n}$ are close to zero, and when $\widehat{\rho}_\alpha \ll 1$, the estimated variance $\widehat{\sigma}^2_{\alpha,P}$ in (4.25) becomes large, preventing rejection of the null hypothesis. To prevent these issues, each of the conditions $\widehat{\rho}_\alpha < 0.05$ and $\widehat{r}_\alpha(\mathbf{2}_\alpha) = \widehat{r}_\alpha(\mathbf{1}_\alpha)$ are declared as a stopping criterion in CLEF-Peng.

Experimental setting. : CLEF Chiapino and Sabourin (2016) is designed to face situations where DAMEX (Goix et al., 2016b) fails to exhibit a clear-cut dependence structure. A major issue reported in Chiapino and Sabourin (2016) for certain

hydrological data is the high variability of the groups of features for which large values occur simultaneously. Because of this, the empirical exponent measure $\widehat{\mu}$ assigns low mass to any sub-region partitioning the sample space, see Remark 4.1. The empirical finding motivating the latter work is that the various subsets α involved in simultaneous extreme records could nevertheless be clustered, meaning that many of them have a significant intersection, whereas many symmetric differences comprise just a single or at most a few features.

A natural assumption in this context is that a 'true' list of dependent subsets $\mathcal{M} = \{\alpha_1, \ldots, \alpha_K\}$ exists such that $\mu(\mathcal{C}_\alpha) > 0$ for $\alpha \in \mathcal{M}$ and that noisy features are involved in each extreme event. Observed large records then concern groups of the kind $\alpha' = \alpha \cup \{j\}$, where $\alpha \in \mathcal{M}$ and $j \in \{1, \ldots, d\} \setminus \alpha$.

In our experiments, datasets are generated as follows: The dimension is fixed to $d = 100$. A family of 'true' dependent subsets $\mathcal{M} = \{\alpha_1, \ldots, \alpha_K\}$ of cardinality $K = 80$ is randomly chosen: the subset sizes $|\alpha|$ follow a truncated geometric distribution, with a maximum subset size set to 8. For simplicity, we forbid nested subsets, so $\alpha_j \not\subset \alpha_k$ whenever $j \neq k$. The maximal elements of $\mathbb{M} = \{\alpha \subset \{1, \ldots, d\} \mid \rho_\alpha > 0\}$ are then precisely the elements of $\mathcal{M}$, as explained in Remark 4.1. Finally, two different subsets may have at most two features in common.

Once the dependence structure $\mathcal{M}$ has been fixed, the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are sampled independently from d-dimensional asymmetric logistic distributions Tawn (1990b), using Algorithm 2.2 in Stephenson (2003). The underlying 'true' distribution function is

$$G(\mathbf{x}) = \exp\left[-\sum_{m=1}^{K}\left\{\sum_{j\in\alpha_m}(|\mathcal{A}(j)|x_j)^{-1/w_{\alpha_m}}\right\}^{w_{\alpha_m}}\right], \tag{4.29}$$

where $\mathcal{A}(j) = \{\alpha \in \mathcal{M} \mid j \in \alpha\}$ and w_{α_m} is a dependence parameter which is set to 0.1 in our simulations. Actually, to mimic the noisy situation described above, each point $\mathbf{X}_i$ is simulated according to a slightly different version, G_i, of G. For each $i = 1, \ldots, n$ and $k = 1, \ldots, K$, we randomly select an additional 'noisy feature' $j_{i,k} \in \{1, \ldots, d\} \setminus \alpha_k$ and set $\alpha'_{i,k} = \alpha_k \cup \{j_{i,k}\}$. Then $\mathcal{M}'_i = \{\alpha'_{i,1}, \ldots, \alpha'_{i,K}\}$ is the collection of 'noisy subsets' for $\mathbf{X}_i$ and $G_i(\mathbf{x})$ is as in (4.29) with $\mathcal{A}(j)$ replaced by $\mathcal{A}'_i(j) = \{\alpha' \in \mathcal{M}'_i \mid j \in \alpha'\}$.

Results. : We generate datasets of size $n = 5e4$ and $n = 1e5$. For each sample size, 50 independent datasets are simulated according to the procedure summarized in the preceding paragraph. We compare the average performance of the three proposed versions of CLEF, together with the original CLEF and DAMEX algorithms, for different choices of k and confidence level δ.

Tables 4.1 and 4.2 gather the results for a confidence level δ equal to 0.001 and 0.0001, respectively. In both tables, the results obtained with the original version of CLEF and DAMEX are included in the comparison with an identical choice of tuning parameters, so that the last two lines of the two tables are the same. In CLEF, the threshold C was chosen by trial and error in the interval $(0, \kappa_{\min})$, namely $C = 0.05$. Imposing that $C < \kappa_{\min}$ is intended to reproduce the effect of the variance term

Table 4.1: Average number of recovered clusters and errors of CLEF-asymptotic ($\kappa_{\min} = 0.08$), CLEF-Peng, CLEF-Hill, CLEF and DAMEX on 50 datasets.

$n = 5e4$	k/n	recovered	subset errors	superset errors	other errors
CLEF-asymptotic	0.003	71.1 (3.0)	7.4 (4.7)	5.1 (2.1)	28.0 (13.3)
	0.005	73.0 (3.7)	8.0 (6.3)	2.4 (1.7)	14.6 (8.9)
CLEF-Peng	0.003	**79.70 (0.7)**	**1.00 (2.5)**	**0. (0.)**	3.9 (2.7)
	0.005	**79.98 (0.1)**	**0.06 (0.4)**	**0. (0.)**	0.9 (0.9)
CLEF-Hill	0.003	79.0 (1.4)	2.4 (3.5)	0.04 (0.2)	17.9 (7.0)
	0.005	75.7 (2.4)	9.2 (6.8)	**0. (0.)**	**0. (0.)**
CLEF	0.003	69.9 (4.4)	16.2 (8.1)	0.5 (0.6)	**2.3 (2.2)**
	0.005	75.0 (3.6)	8.1 (6.4)	0.2 (0.5)	0.9 (1.2)
DAMEX	0.003	0.6 (0.2)	1.7 (1.4)	32.9 (5.6)	45.4 (5.9)
	0.005	0.1 (0.4)	2.4 (1.5)	18.3 (5.5)	59.1 (5.9)
$n = 1e5$					
CLEF-asymptotic	0.003	73.2 (3.7)	9.5 (6.7)	0.9 (0.8)	4.7 (2.7)
	0.005	72.6 (4.4)	11.7 (7.6)	0.1 (0.4)	0.5 (0.9)
CLEF-Peng	0.003	**79.9 (0.2)**	**0.2 (1.0)**	**0. (0.)**	0.1 (0.4)
	0.005	**80.0 (0.)**	**0. (0.)**	**0. (0.)**	**0. (0.)**
CLEF-Hill	0.003	77.0 (2.0)	6.1 (4.6)	**0. (0.)**	**0. (0.)**
	0.005	67.2 (4.8)	22.8 (10.4)	**0. (0.)**	**0. (0.)**
CLEF	0.003	75.2 (3.2)	7.5 (5.9)	0.0 (0.2)	0.2 (0.5)
	0.005	77.9 (2.3)	3.2 (3.9)	0.02 (0.1)	0.02 (0.1)
DAMEX	0.003	0.04 (0.2)	1.3 (1.0)	24.4 (6.7)	54.2 (7.0)
	0.005	0.1 (0.3)	1.9 (1.6)	10.3 (3.7)	67.6 (4.7)

Confidence level for the tests: $\delta = 0.001$. Standard deviations over the 50 samples in brackets. Bold face indicates the best performing algorithm on average for a given n and a given choice of k/n, the proportion of extreme data used.

upon the stopping criterion in CLEF-asymptotic. In DAMEX, the 80 subsets with highest empirical mass are retained and the subspace thickening parameter ϵ is set to the default value of 0.1, following the guidelines of the authors.

Each algorithm produces a list, $\widehat{\mathbb{M}}$, of groups of features $\alpha \in \{1, \ldots, d\}$. This list is to be compared with the one of $K = 80$ 'true' subsets $\mathcal{M}$. The performance of each algorithm is measured in terms of two criteria: the number of 'true' subsets $\alpha \in \mathcal{M}$ that appear in $\widehat{\mathbb{M}}$ (third column of Tables 4.1 and 4.2); the number of 'errors', that is, the subsets $\alpha \in \widehat{\mathbb{M}}$ that do not belong to $\mathcal{M}$. These can be understood as 'false positives'. Among these errors, we make the distinction between those which are respectively proper subsets (fourth column of Tables 4.1 and 4.2) or proper

Table 4.2: Same setting as Table 4.1 with $\delta = 0.0001$

$n = 5e4$	k/n	recovered	subset errors	superset errors	other errors
CLEF-asymptotic	0.003	71.8 (2.4)	2.3 (2.5)	7.8 (2.8)	41.9 (19.3)
	0.005	73.5 (2.8)	3.7 (3.8)	4.8 (2.5)	25.8 (12.2)
CLEF-Peng	0.003	**79.7 (0.7)**	**1.0 (2.5)**	**0. (0.)**	3.9 (2.7)
	0.005	**80.0 (0.1)**	**0.1 (0.4)**	**0. (0.)**	0.9 (0.9)
CLEF-Hill	0.003	79.5 (0.8)	**0.3 (1.1)**	0.5 (0.8)	142.2 (33.2)
	0.005	79.2 (1.0)	1.6 (2.3)	**0. (0.)**	**0.2 (0.5)**
CLEF	0.003	69.9 (4.4)	16.2 (8.1)	0.5 (0.6)	2.3 (2.2)
	0.005	75.0 (3.6)	8.1 (6.4)	0.2 (0.5)	0.9 (1.2)
DAMEX	0.003	0.6 (0.2)	1.7 (1.4)	32.9 (5.6)	45.4 (5.9)
	0.005	0.1 (0.4)	2.4 (1.5)	18.3 (5.5)	59.1 (5.9)
$n = 1e5$					
CLEF-asymptotic	0.003	75.7 (2.8)	3.7 (3.8)	2.0 (1.4)	11.0 (5.5)
	0.005	76.0 (2.9)	5.6 (4.5)	0.4 (0.7)	1.9 (1.9)
CLEF-Peng	0.003	**79.9 (0.2)**	**0.2 (1.0)**	**0. (0.)**	0.1 (0.4)
	0.005	**80. (0.)**	**0. (0.)**	**0. (0.)**	**0. (0.)**
CLEF-Hill	0.003	79.5 (1.0)	1.2 (2.3)	**0. (0.)**	**0.1 (0.2)**
	0.005	75.4 (2.8)	8.7 (5.2)	**0. (0.)**	**0. (0.)**
CLEF	0.003	75.2 (3.2)	7.5 (5.9)	0.0 (0.2)	0.2 (0.5)
	0.005	77.9 (2.3)	3.2 (3.9)	0.02 (0.1)	0.02 (0.1)
DAMEX	0.003	0.04 (0.2)	1.3 (1.0)	24.4 (6.7)	54.2 (7.0)
	0.005	0.1 (0.3)	1.9 (1.6)	10.3 (3.7)	67.6 (4.7)

supersets (fifth column) of some true $\beta \in \mathcal{M}$, and the other errors (sixth column).

CLEF-Peng obtains the best overall scores for both values of δ, but as explained above, a special treatment is reserved for the case $\widehat{\rho}_\alpha \leq 0.05$, and this threshold constitutes an arbitrary tuning parameter, which can impact the performance significantly. On the other hand, CLEF-Hill does not require any other adjustment than for the special case $\widehat{\rho}_\alpha = 0$ and performs nearly as well as CLEF-Peng with $\delta = 0.0001$ and $k/n = 0.005$. In addition, CLEF-Hill outperforms all the other methods. In particular, CLEF-asymptotic is globally less accurate than CLEF-Peng and CLEF-Hill. This reflects the fact that the null hypothesis in this algorithm involves an arbitrary $\kappa_{\min} > 0$ fixed by the user. Our own choice $\kappa_{\min} = 0.08$ was fixed by trial and error, which is straightforward with synthetic data and could also be achieved by cross-validation in a real use case. Finally, as expected, DAMEX obtains very low scores, because it is not designed to handle the addition of noisy features, as explained earlier.

4.8 Conclusion

In this work, we propose three variants of the CLEF algorithm (Chiapino and Sabourin, 2016), replacing the heuristic criterion in the original version with a formal test for asymptotic dependence, and this for all possible subsets of features among $\{1, \ldots, d\}$. As in the original CLEF implementation, only a small proportion of all $2^d - 1$ subsets has to be examined, while the computational complexity for each such subset is low. Experimental results indicate that the CLEF algorithm is most effective when based on a test constructed from an extension of the Hill estimator (Draisma et al., 2004) of the multivariate coefficient of tail dependence.

The procedure we propose is nonparametric and rank-based. Parametric approaches, based for instance on the nested asymmetric logistic distribution (Tawn, 1990b), could have a greater sensitivity, at the cost of increased model risk and greater computational complexity. We have also assumed that the observations are serially independent; in the contrary case, the asymptotic variances of the various estimator need to be estimated by some form of bootstrap, which, in high dimensions, poses important theoretical and computational challenges; see (Bücher and Dette, 2013) for the bivariate and serially independent case.

4.9 Proofs

Proof of Proposition 4.2. For $\emptyset \neq \alpha \subset \{1, \ldots, d\}$ and $\boldsymbol{x} \in [0, \infty)^\alpha$, put

$$L_\alpha(\boldsymbol{x}) = \{\boldsymbol{y} \in [0, \infty]^d \mid \exists j \in \alpha : y_j < x_j\},$$
$$R_\alpha(\boldsymbol{x}) = \{\boldsymbol{y} \in [0, \infty]^d \mid \forall j \in \alpha : y_j < x_j\}.$$

If $\alpha = \{1, \ldots, d\}$, then just write L rather than $L_{\{1,\ldots,d\}}$. Note that $L_\alpha(\boldsymbol{x}) = L(\boldsymbol{x}\boldsymbol{e}_\alpha)$ with $\boldsymbol{e}_\alpha = (\mathbb{1}_\alpha(j))_{j=1}^d$ and that $L_{\{j\}}(x_j) = R_{\{j\}}(x_j)$ and thus $W(L_{\{j\}}(x_j)) = W_{\{j\}}(x_j)$. Einmahl et al. (2012, Theorem 4.6) show that, in the space $\ell^\infty([0, T]^d)$ and under Conditions 1, 2 and 3, we have weak convergence

$$\sqrt{k}\{\widehat{\ell}(\boldsymbol{x}) - \ell(\boldsymbol{x})\} \rightsquigarrow W(L(\boldsymbol{x})) - \sum_{j=1}^d \partial\ell_j(\boldsymbol{x}) W_{\{j\}}(x_j)$$

as $n \to \infty$. Here, we have taken a version of the Gaussian process W such that the trajectories $\mathbf{x} \mapsto W(L(\boldsymbol{x}))$ are continuous almost surely.

As in (4.7), we have, for $\emptyset \neq \alpha \subset \{1, \ldots, d\}$ and $\boldsymbol{x} \in [0, \infty)^\alpha$, the identity

$$\widehat{r}_\alpha(\boldsymbol{x}) = \sum_{\emptyset \neq \beta \subset \alpha} (-1)^{|\beta|+1} \widehat{\ell}(\boldsymbol{x}_\beta \boldsymbol{e}_\beta)$$

where $\boldsymbol{x}_\beta = (x_j)_{j \in \beta}$. Hence, we can view the vector $(\sqrt{k}(\widehat{r}_\alpha - r_\alpha))_{\emptyset \neq \alpha \subset \{1,\ldots,d\}}$ as the result of the application to $\sqrt{k}(\widehat{\ell} - \ell)$ of a bounded linear map from the space

$\ell^\infty([0,T]^d)$ to the product space $\prod_{\emptyset\neq\alpha\in\{1,\dots,d\}} \ell^\infty([0,T]^\alpha)$. By the continuous mapping theorem, we obtain, in the latter space, the weak convergence

$$\sqrt{k}\,\{\widehat{r}_\alpha(\mathbf{x}) - r_\alpha(\mathbf{x})\} \rightsquigarrow \sum_{\emptyset\neq\beta\subset\alpha} (-1)^{|\beta|+1} \left\{W(L_\beta(\boldsymbol{x}_\beta)) - \textstyle\sum_{j=1}^d \partial_j \ell_\beta(\boldsymbol{x}_\beta) W_{\{j\}}(x_j \mathbb{1}_\beta(j))\right\}.$$

Here we used $\ell(\boldsymbol{x}_\beta \boldsymbol{e}_\beta) = \ell_\beta(\boldsymbol{x}_\beta)$.

The set-indexed process W satisfies the remarkable property that $W(A\cup B) = W(A) + W(B)$ almost surely whenever A and B are disjoint Borel sets of $[0,\infty]^d \setminus \{\boldsymbol{\infty}\}$ that are bounded away from $\boldsymbol{\infty}$: indeed, (4.10) implies $\mathbb{E}[\{W(A\cup B) - W(A) - W(B)\}^2] = 0$. It follows that the trajectories of W obey the inclusion-exclusion formula, so that, for $\emptyset \neq \alpha \subset \{1,\dots,d\}$ and $\boldsymbol{x} \in [0,\infty)^\alpha$, we have, almost surely,

$$\begin{aligned}\sum_{\emptyset\neq\beta\subset\alpha} (-1)^{|\beta|+1} W(L_\beta(\boldsymbol{x}_\beta)) &= \sum_{\emptyset\neq\beta\subset\alpha} (-1)^{|\beta|+1} W\left(\textstyle\bigcup_{j\in\beta} R_{\{j\}}(x_j)\right)\\ &= W\left(\textstyle\bigcap_{j\in\alpha} R_{\{j\}}(x_j)\right) = W(R_\alpha(\boldsymbol{x})) = W_\alpha(\boldsymbol{x}).\end{aligned}$$

We can make this hold true almost surely jointly for all such α and $\boldsymbol{x}$: first, consider points $\boldsymbol{x}$ with rational coordinates only and then consider a version of W by extending W_α to points $\boldsymbol{x}$ with general coordinates via continuity. Similarly, since $W_{\{j\}}(0) = W(\emptyset) = 0$ almost surely, we have

$$\begin{aligned}\sum_{\emptyset\neq\beta\subset\alpha} (-1)^{|\beta|+1} \sum_{j=1}^d \partial_j \ell_\beta(\boldsymbol{x}_\beta) W_{\{j\}}(x_j \mathbb{1}_\beta(j)) &= \sum_{j\in\alpha} \sum_{\beta: j\in\beta\subset\alpha} \partial_j \ell_\beta(\boldsymbol{x}_\beta) W_{\{j\}}(x_j)\\ &= \sum_{j\in\alpha} \partial r_j(\boldsymbol{x}) W_{\{j\}}(x_j).\end{aligned}$$

We have thus shown weak convergence as stated in (4.11). □

Proof of Corollary 4.3. The weak convergence statement (4.14) is a special case of (4.11): set $\boldsymbol{x} = \mathbf{1}_\alpha$. The covariance formula (4.15) follows from the fact that

$$\begin{aligned}\mathbb{E}[W_\alpha(\mathbf{1}_\alpha) W_{\alpha'}(\mathbf{1}_{\alpha'})] &= \Lambda(\{\boldsymbol{y} \in [0,\infty]^d \mid \forall i \in \alpha\cup\alpha' : y_i < 1\})\\ &= \mu(\{\boldsymbol{u} \in [0,\infty)^d \mid \forall i \in \alpha\cup\alpha' : u_i > 1\}) = \rho_{\alpha\cup\alpha'};\end{aligned}$$

the first equality follows from (4.10) and the last one from (4.3). We obtain (4.15) by expanding $G_\alpha = Z_\alpha(\mathbf{1}_\alpha)$ using (4.11) and working out $\mathbb{E}[G_\alpha G_{\alpha'}]$ with the above identity. □

Proof of Proposition 4.4. Let $\alpha = \{\alpha_1,\dots,\alpha_S\} \subset \{1,\dots,d\}$ with $S = |\alpha| \geq 2$ and such that $\mu(\Delta_\alpha) > 0$. In view of (4.13), we have $\kappa_\alpha = g_\alpha(\theta_\alpha)$ and $\widehat{\kappa}_\alpha = g_\alpha(\widehat{\theta}_\alpha)$ where $\theta_\alpha = (\rho_\alpha, \rho_{\alpha\setminus\alpha_1}, \dots, \rho_{\alpha\setminus\alpha_S})$, $\widehat{\theta}_\alpha = (\widehat{\rho}_\alpha, \widehat{\rho}_{\alpha\setminus\alpha_1}, \dots, \widehat{\rho}_{\alpha\setminus\alpha_S})$, and

$$g_\alpha(x_0, x_1, \dots, x_S) = \frac{x_0}{\sum_{j=1}^S x_j - (S-1)x_0}, \qquad x \in [0,\infty)^{1+S}. \tag{4.30}$$

Let $\nabla g_\alpha(x)$ denote the gradient vector of g_α evaluated x and let $\langle \cdot, \cdot \rangle$ denote the scalar product in Euclidean space. Proposition 4.3 combined with the delta method as in (van der Vaart, 1998, Theorem 3.1) gives, as $n \to \infty$,

$$\begin{aligned}\sqrt{k}(\widehat{\kappa}_\alpha - \kappa_\alpha) = \sqrt{k}\{g_\alpha(\widehat{\theta}_\alpha) - g_\alpha(\theta_\alpha)\} &= \Big\langle \nabla g_\alpha(\theta_\alpha),\, \sqrt{k}(\widehat{\theta}_\alpha - \theta_\alpha)\Big\rangle + \mathrm{o}_{\mathbb{P}}(1) \\ &\rightsquigarrow \Big\langle \nabla g_\alpha(\theta_\alpha),\, (G_\alpha, G_{\alpha\setminus\alpha_1}, \ldots, G_{\alpha\setminus\alpha_S})\Big\rangle,\end{aligned}$$

the weak convergence holding jointly in α by Slutsky's lemma and Proposition 4.3. The partial derivatives of g_α are

$$\begin{aligned}\frac{\partial g}{\partial x_0}(x) &= \frac{\sum_{j=1}^S x_j}{\{\sum_{j=1}^S x_j - (S-1)x_0\}^2}, \\ \frac{\partial g}{\partial x_j}(x) &= \frac{-x_0}{\{\sum_{j=1}^S x_j - (S-1)x_0\}^2}, \qquad j = 1, \ldots, S.\end{aligned}$$

Evaluating these at $x = \theta_\alpha$ and using $\sum_{j\in\alpha} \rho_{\alpha\setminus j} - (S-1)\rho_\alpha = \mu(\Delta_\alpha)$ as in (4.12) and (4.13), we find that

$$\Big\langle \nabla g_\alpha(\theta_\alpha),\, (G_\alpha, G_{\alpha\setminus\alpha_1}, \ldots, G_{\alpha\setminus\alpha_S})\Big\rangle = \mu(\Delta_\alpha)^{-2}\Big\{\Big(\textstyle\sum_{j\in\alpha} \rho_{\alpha\setminus j}\Big) G_\alpha - \rho_\alpha \textstyle\sum_{j\in\alpha} G_{\alpha\setminus j}\Big\},$$

in accordance to the right-hand side in (4.17).

To calculate the asymptotic variance $\sigma^2_{\kappa,\alpha}$, we introduce a few abbreviations: we write $R_\beta = R_\beta(\mathbf{1}_\beta)$ and $W^\cap_\beta = W_\beta(\mathbf{1}_\beta) = W(R_\beta)$ for $\emptyset \neq \beta \in \{1, \ldots, d\}$ and we put $W_j = W_{\{j\}}(1)$ for $j = 1, \ldots, d$, so that $G_\alpha = W^\cap_\alpha - \sum_{j\in\alpha} \dot{\rho}_{j,\alpha} W_j$. We find

$$\begin{aligned}H_\alpha &= \Big(\sum_{i\in\alpha} \rho_{\alpha\setminus i}\Big) G_\alpha - \rho_\alpha \sum_{i\in\alpha} G_{\alpha\setminus i} \\ &= \Big(\sum_{i\in\alpha} \rho_{\alpha\setminus i}\Big)\Big(W^\cap_\alpha - \sum_{j\in\alpha} \dot{\rho}_{j,\alpha} W_j\Big) - \rho_\alpha \sum_{i\in\alpha}\Big(W^\cap_{\alpha\setminus i} - \sum_{j\in\alpha\setminus i} \dot{\rho}_{j,\alpha\setminus i} W_j\Big).\end{aligned}$$

From the proof of Proposition 4.2, recall that $W(A \cup B) = W(A) + W(B)$ almost surely for disjoint Borel sets A and B of $[0, \infty]^d \setminus \{\boldsymbol{\infty}\}$ bounded away from $\boldsymbol{\infty}$; moreover, for such A and B, the variables $W(A)$ and $W(B)$ are uncorrelated. Since $R_{\alpha\setminus i}$ is the disjoint union of R_α and $R_{\alpha\setminus i} \setminus R_\alpha$, we have therefore $W^\cap_{\alpha\setminus i} = W^\cap_\alpha + W(R_{\alpha\setminus i} \setminus R_\alpha)$ almost surely. In addition, $\sum_{i\in\alpha} \rho_{\alpha\setminus i} = \mu(\Delta_\alpha) + (S-1)\rho_\alpha$ by (4.12) applied to $\nu = \mu$. As a consequence,

$$H_\alpha = \{\mu(\Delta_\alpha) - \rho_\alpha\} W^\cap_\alpha - \rho_\alpha \sum_{j\in\alpha} W(R_{\alpha\setminus j} \setminus R_\alpha) + \sum_{j\in\alpha} K_{\alpha,j} W_j$$

where

$$K_{\alpha,j} = \rho_\alpha\Big(\sum_{i\in\alpha\setminus j} \dot{\rho}_{j,\alpha\setminus i}\Big) - \Big(\sum_{i\in\alpha} \rho_{\alpha\setminus i}\Big)\dot{\rho}_{j,\alpha}, \qquad j \in \alpha.$$

The $S+1$ variables $W^\cap_\alpha = W(R_\alpha)$ and $W(R_{\alpha\setminus j} \setminus R_\alpha)$, $j \in \alpha$, are all uncorrelated, since they involve evaluating W at disjoint sets; $W_j = W(R_{\{j\}})$ is uncorrelated with

$W(R_{\alpha\setminus j}\setminus R_\alpha)$, for the same reason. Moreover, $\mathbb{E}[W_\alpha^{\cap}W_j] = \Lambda(R_\alpha\cap R_{\{j\}}) = \Lambda(R_\alpha) = \rho_\alpha$ and similarly $\mathbb{E}[W(R_{\alpha\setminus i}\setminus R_\alpha)W_j] = \Lambda(R_{\alpha\setminus i}\setminus R_\alpha) = \rho_{\alpha\setminus i} - \rho_\alpha$ if $i, j \in \alpha$ and $i \neq j$. Hence

$$\begin{aligned}\mathrm{Var}(H_\alpha) = \{\mu(\Delta_\alpha) - \rho_\alpha\}^2\rho_\alpha + \rho_\alpha^2\sum_{j\in\alpha}(\rho_{\alpha\setminus j} - \rho_\alpha) + \sum_{i,j\in\alpha} K_{\alpha,i}K_{\alpha,j}\rho_{\{i,j\}} \\ + \{\mu(\Delta_\alpha) - \rho_\alpha\}\rho_\alpha\sum_{j\in\alpha}K_{\alpha,j} - \rho_\alpha\sum_{j\in\alpha}K_{\alpha,j}\sum_{i\in\alpha\setminus j}(\rho_{\alpha\setminus i} - \rho_\alpha).\end{aligned}$$

As $\sum_{j\in\alpha}(\rho_{\alpha\setminus j} - \rho_\alpha) = \mu(\Delta_\alpha) - \rho_\alpha$ and $\sum_{i\in\alpha\setminus j}(\rho_{\alpha\setminus i} - \rho_\alpha) = \mu(\Delta_\alpha) - \rho_{\alpha,j}$, we get

$$\begin{aligned}\mathrm{Var}(H_\alpha) = \{\mu(\Delta_\alpha) - \rho_\alpha\}\rho_\alpha\Big\{\mu(\Delta_\alpha) + \sum_{j\in\alpha}K_{\alpha,j}\Big\} + \sum_{i,j\in\alpha}K_{\alpha,i}K_{\alpha,j}\rho_{\{i,j\}} \\ - \rho_\alpha\sum_{j\in\alpha}K_{\alpha,j}\{\mu(\Delta_\alpha) - \rho_{\alpha\setminus j}\}. \quad (4.31)\end{aligned}$$

Recall $\kappa_\alpha(\boldsymbol{x})$ in (4.16). We have

$$\begin{aligned}\frac{\partial}{\partial x_j}\left(\frac{1}{\kappa_\alpha(\boldsymbol{x})}\right)_{\boldsymbol{x}=\mathbf{1}_\alpha} &= \frac{\partial}{\partial x_j}\left(\frac{\sum_{i\in\alpha}r_{\alpha\setminus i}(\boldsymbol{x}_{\alpha\setminus i})}{r_\alpha(\boldsymbol{x})}\right) \\ &= \rho_\alpha^{-2}\left(\rho_\alpha\sum_{i\in\alpha\setminus j}\dot\rho_{j,\alpha\setminus i} - \dot\rho_{j,\alpha}\sum_{i\in\alpha}\rho_{\alpha\setminus i}\right) = \rho_\alpha^{-2}K_{\alpha,j}.\end{aligned}$$

It follows that $\dot\kappa_{j,\alpha} = -\rho_\alpha^{-2}K_{\alpha,j}/(1/\kappa_\alpha)^2 = -K_{\alpha,j}/\mu(\Delta_\alpha)^2$. By (4.31), we find that $\sigma_{\kappa,\alpha}^2 = \mu(\Delta_\alpha)^{-4}\,\mathrm{Var}(H_\alpha)$ is equal to the right-hand side of (4.18). □

Proof of Proposition 4.5. We only need to prove that $\widehat\sigma_{\kappa,\alpha}^2 = \sigma_{\kappa,\alpha}^2 + \mathrm{o}_{\mathbb{P}}(1)$ as $n\to\infty$. In view of the expressions (4.18) and (4.20) for $\sigma_{\kappa,\alpha}^2$ and $\widehat\sigma_{\kappa,\alpha}$, it is enough to show that $\dot\kappa_{j,\alpha,n} = \dot\kappa_{\alpha,j} + \mathrm{o}_{\mathbb{P}}(1)$, with $\dot\kappa_{j,\alpha,n}$ in (4.19); indeed, Corollary 4.3 already gives consistency of $\widehat\mu(\Delta_\alpha)$ and $\widehat\rho_\beta$. Now since $2^{-1}k^{1/4}\{\kappa_\alpha(\mathbf{1}_\alpha + k^{-1/4}\mathbf{e}_j) - \kappa_\alpha(\mathbf{1}_\alpha - k^{-1/4}\mathbf{e}_j)\} \to \dot\kappa_{\alpha,j}$ as $n\to\infty$, a sufficient condition is that for some $\epsilon > 0$,

$$\sup_{[1-\epsilon,2+\epsilon]^\alpha} k^{1/4}\Big|\widehat\kappa_\alpha(\mathbf{x}) - \kappa_\alpha(\mathbf{x})\Big| = \mathrm{o}_{\mathbb{P}}(1), \qquad n\to\infty. \quad (4.32)$$

In turn, (4.32) follows from weak convergence of $k^{1/2}(\widehat\kappa_\alpha - \kappa_\alpha)$ as $n\to\infty$ in the space $\ell^\infty([1-\varepsilon, 1+\varepsilon]^\alpha)$. In light of the expressions of $\widehat\kappa_\alpha$ and κ_α in terms of the (empirical) joint tail dependence functions $\widehat r_\beta$ and r_β, respectively, weak convergence of $k^{1/2}(\widehat\kappa_\alpha - \kappa_\alpha)$ follows from Proposition 4.2 and the functional delta method (van der Vaart, 1998, Theorem 20.8). The calculations are similar to the ones for the Euclidean case in the proof of Proposition 4.4; an extra point to be noted is that if α is such that $\mu(\Delta_\alpha) > 0$, then the denominator in the definition of $\kappa_\alpha(\boldsymbol{x})$ in (4.16) is positive for all $\boldsymbol{x}$ in a neighbourhood of $\mathbf{1}_\alpha$. □

Proof of Proposition 4.6. Proposition 4.2 implies, as $n \to \infty$, the weak convergence

$$\Big(\sqrt{k}\{\widehat{r}_\alpha(\mathbf{2}_\alpha) - r_\alpha(\mathbf{2}_\alpha\}, \sqrt{k}\{\widehat{r}_\alpha(\mathbf{1}_\alpha) - r_\alpha(\mathbf{1}_\alpha\}\Big) \rightsquigarrow \Big(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)\Big).$$

Now $\widehat{\eta}_\alpha^P = g(\widehat{r}_\alpha(\mathbf{2}_\alpha), \widehat{r}_\alpha(\mathbf{1}_\alpha))$ and $\eta_\alpha = 1 = g(r_\alpha(\mathbf{2}_\alpha), r_\alpha(\mathbf{1}_\alpha)) = g(2\rho_\alpha, \rho_\alpha)$, with $g(x,y) = \log(2)/\log(x/y)$; note that the function r_α is homogeneous. Since the gradient of g is $\nabla g(x,y) = \log(2)(\log(x/y))^{-2}(-x^{-1}, y^{-1})$, the delta method gives

$$\begin{aligned}\sqrt{k}(\widehat{\eta}^P - 1) &\rightsquigarrow \Big\langle \nabla g(2\rho_\alpha, \rho_\alpha), \Big(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)\Big)\Big\rangle \\ &= \frac{1}{\rho_\alpha \log 2}\Big\langle (-1/2, 1), \Big(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)\Big)\Big\rangle \\ &= \frac{-1}{2\rho_\alpha \log 2}\{Z_\alpha(\mathbf{2}_\alpha) - 2Z_\alpha(\mathbf{1}_\alpha)\}.\end{aligned}$$

The first part of the assertion follows. As for the variance,

$$\mathrm{Var}(Z_\alpha(\mathbf{2}_\alpha) - 2Z_\alpha(\mathbf{1}_\alpha)) = \mathrm{Var}(Z_\alpha(\mathbf{2}_\alpha)) + 4\,\mathrm{Var}(Z_\alpha(\mathbf{1}_\alpha)) - 4\,\mathbb{C}\mathrm{ov}(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)),$$

The function r_α is homogeneous of order 1, so that $\partial_j r_\alpha$ is constant along rays, that is, the function $0 < t \mapsto \partial_j r_\alpha(t\,\mathbf{x})$ is constant. Moreover, the measure Λ is homogeneous of order 1 too. In view of (4.10) and (4.11), it follows that $\mathrm{Var}(Z_\alpha(t\boldsymbol{x})) = t\,\mathrm{Var}(Z_\alpha(\boldsymbol{x}))$ for $t > 0$; in particular $\mathrm{Var}(Z_\alpha(\mathbf{2}_\alpha) = 2\,\mathrm{Var}(Z_\alpha(\mathbf{1}_\alpha)$. Further, $\rho_\alpha = (\mathrm{d}r_\alpha(t,\ldots,t)/\mathrm{d}t)_{t=1} = \sum_{j\in\alpha}\dot\rho_{j,\alpha}$ and thus

$$\begin{aligned}\mathrm{Var}(Z_\alpha(\mathbf{1}_\alpha)) &= \rho_\alpha - 2\sum_{j\in\alpha}\dot\rho_{j,\alpha}\rho_\alpha + \sum_{j\in\alpha}\sum_{j'\in\alpha}\dot\rho_{j,\alpha}\dot\rho_{j',\alpha}\rho_{\{j,j'\}} \\ &= \rho_\alpha - 2\rho_\alpha^2 + \sum_{j\in\alpha}\sum_{j'\in\alpha}\dot\rho_{j,\alpha}\dot\rho_{j',\alpha}\rho_{\{j,j'\}}.\end{aligned}$$

The covariance term is

$$\begin{aligned}\mathbb{C}\mathrm{ov}(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)) = \rho_\alpha - \sum_{j\in\alpha}\dot\rho_{j,\alpha}\rho_\alpha - \sum_{j\in\alpha}\dot\rho_{j,\alpha}r_\alpha(\mathbf{2}_\alpha\wedge\boldsymbol{\iota}_j) \\ + \sum_{j\in\alpha}\sum_{j'\in\alpha}\dot\rho_{j,\alpha}\dot\rho_{j',\alpha}r_{\{j,j'\}}(2,1),\end{aligned}$$

with $\mathbf{2}_\alpha \wedge \boldsymbol{\iota}_j$ as explained in the statement of the proposition. Since $\sum_{j\in\alpha}\dot\rho_{j,\alpha} = \rho_\alpha$, we can simplify and find

$$\begin{aligned}\mathrm{Var}(Z_\alpha(\mathbf{2}_\alpha) - 2Z_\alpha(\mathbf{1}_\alpha)) &= 6\,\mathrm{Var}(Z_\alpha(\mathbf{1}_\alpha)) - 4\,\mathbb{C}\mathrm{ov}(Z_\alpha(\mathbf{2}_\alpha), Z_\alpha(\mathbf{1}_\alpha)) \\ &= 2\rho_\alpha - 8\rho_\alpha^2 + 4\sum_{j\in\alpha}\dot\rho_{j,\alpha}r_\alpha(\mathbf{2}_\alpha\wedge\boldsymbol{\iota}_j) \\ &\quad + \sum_{j\in\alpha}\sum_{j'\in\alpha}\dot\rho_{j,\alpha}\dot\rho_{j',\alpha}\Big[6\rho_{\{j,j'\}} - 4r_{\{j,j'\}}(2,1)\Big].\end{aligned}$$

Divide the right-hand side by $(2\rho_\alpha \log 2)^2$ to obtain (4.23). □

Proof of Proposition 4.8. To alleviate notations, $\emptyset \neq \alpha \subset \{1, \ldots, d\}$ is fixed and the subscript α is omitted throughout the proof. Introduce the tail empirical process $Q_n(t) = \widehat{T}_{(n-\lfloor kt \rfloor)}$ for $0 < t < n/k$. The key is to represent the Hill estimator as a statistical tail functional (Drees, 1998a, Example 3.1) of Q_n, i.e., $\widehat{\eta}^H = \Theta(Q_n)$, where Θ is the map defined for any measurable function $z : (0, 1] \to \mathbb{R}$ as $\Theta(z) = \int_0^1 \log^+\{z(t)/z(1)\}\mathrm{d}t$ when the integral is finite and $\Theta(z) = 0$ otherwise. Let $z_\eta : t \in (0, 1] \mapsto t^{-\eta}$ denote the quantile function of a standard Pareto distribution with index $1/\eta$; it holds that $\Theta(z_\eta) = \eta$. The map Θ is scale invariant, i.e., $\Theta(tz) = \Theta(z), t > 0$.

The proof consists of three steps:

1. Introduce a function space $D_{\eta,h}$ allowing to control $Q_n(t)$ and $z_\eta(t)$ as $t \to 0$. In this space and up to rescaling, $Q_n - z_\eta$ converges weakly to a Gaussian process.

2. Show that the map Θ is Hadamard differentiable at z_η tangentially to some well chosen subspace of $D_{\eta,h}$.

3. Apply the functional delta method to show that $\eta^H = \Theta(Q_n)$ is asymptotically normal and compute its asymptotic variance via the Hadamard derivative of Θ.

Step 1. : Let $\epsilon > 0$ and $h(t) = t^{1/2+\epsilon}$, $t \in [0, 1]$. Then $h \in \mathcal{H}$, where

$$\mathcal{H} = \{z : [0, 1] \to \mathbb{R} \mid z \text{ continuous}, \lim_{t\to 0} z(t)t^{-1/2}(\log\log(1/t))^{1/2} = 0\}.$$

Introduce the function space

$$D_{\eta,h} = \{z : [0, 1] \to \mathbb{R} \mid \lim_{t\to 0} t^\eta h(t)z(t) = 0\,;\; t \mapsto t^\eta h(t)z(t) \in D[0, 1]\},$$

where $D[0, 1]$ is the space of càdlàg functions. Notice that $z_\eta \in D_{\eta,h}$. Equip $D_{\eta,h}$ with the seminorm $\|z\|_{\eta,h} = \sup_{t\in(0,1]} |t^\eta h(t)z(t)|$. Let $m = \lceil nq^{\leftarrow}(k/n)\rceil$, with $\lceil \cdot \rceil$ the ceil function, so that $k/m \to \rho$; for self-consistency of the present paper, the roles of k and m are reversed compared to the notation in Draisma et al. (2004). From (Draisma et al., 2004, Lemma 6.2), we have, for all $t_0 > 0$, in the space $D_{\eta,h}$, the weak convergence

$$\sqrt{k}\left(\frac{m}{n}Q_n - z_\eta\right) \rightsquigarrow \left(\eta t^{-(\eta+1)}\bar{W}(t)\right)_{t\in[0,t_0]} \tag{4.33}$$

where $\bar{W}(t) = \tilde{W}(\mathbf{t}_\alpha)$, and $\tilde{W}$ is defined as in the statement of Proposition 4.8. Indeed, the process $\bar{W}$ in the statement from (Draisma et al., 2004, Lemmata 6.1 and 6.2) has same distribution as $W_1(\mathbf{t}_\alpha)$ in the case $\rho = 0$; recall that our ρ is denoted by l in Draisma et al. (2004). Put $U_{i,j} = 1 - F_j(X_{i,j})$, and let $U_{(1),j} \leq \ldots \leq U_{(d),j}$ be the order statistics of $U_{1,j}, \ldots, U_{n,j}$. In the case $\rho > 0$, $\bar{W}$ equals in

distribution $W_{\text{dra}}(\mathbf{t}_\alpha)$ where W_{dra} appears in Lemma 6.1 in the cited reference as the limit in distribution (for $\alpha = \{1,2\}$), for $\mathbf{x} \in E_\alpha$, of

$$\begin{aligned}\Delta_{n,k,m}(\mathbf{x}) &= \sqrt{k}\left[\frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\{\forall j \in \alpha : U_{i,j} \le U_{(\lfloor m x_j \rfloor),j}\} - c(\mathbf{x})\right] \\ &= \underbrace{\sqrt{\frac{m}{k}}}_{\to \rho^{-1/2}} \sqrt{m}\left[\underbrace{\frac{1}{m}\sum_{i=1}^{n} \mathbb{1}\{\forall j \in \alpha : U_{i,j} \le U_{(\lfloor m x_j \rfloor),j}\}}_{r_n(\mathbf{x}) \text{ with } k \text{ replaced by } m} - r(\mathbf{x})\underbrace{\frac{k}{m\rho}}_{\to 1}\right].\end{aligned}$$

From Proposition 4.2 and Slutsky's Lemma, we have $\Delta_{n,k,m} \rightsquigarrow \rho^{-1/2} Z_\alpha$ in $\ell^\infty([0,1]^\alpha)$. Therefore, $W_{\text{dra}} = \rho^{-1/2} Z_\alpha$, as claimed.

Step 2. : The right-hand side of (4.33) belongs to $\mathcal{C}_{h,\eta} = \{z \in D_{\eta,h} \mid z \text{ is continuous}\}$. To apply the functional delta-method (van der Vaart, 1998, Theorem 20.8), we must verify that the restriction of Θ to $\bar{D}_{\eta,h}$ is Hadamard-differentiable tangentially to $\mathcal{C}_{\eta,h}$, with derivative Θ', where $\bar{D}_{\eta,h}$ is a subspace of $D_{\eta,h}$ such that $\mathbb{P}(Q_n \in \bar{D}_{\eta,h}) \to 1$ as $n \to \infty$; see the remark following Condition 3 in Drees (1998a). Then it will follow from the scale invariance of Θ, the identities $\Theta(Q_n) = \widehat{\eta}^H$ and $\Theta(z_\eta) = \eta$, and the weak convergence in (4.33) that

$$\sqrt{k}\left(\widehat{\eta}^H - \eta\right) = \sqrt{k}\left(\Theta(\frac{m}{n}Q_n) - \Theta(z_\eta)\right) \rightsquigarrow \Theta'\left[\left(\eta t^{-(\eta+1)}\bar{W}(t)\right)_{t\in[0,1]}\right] \tag{4.34}$$

as $n \to \infty$. From (Drees, 1998a, Example 3.1), the restriction of Θ to $\bar{D}_{\eta,h}$, the subset of functions on $D_{\eta,h}$ which are positive and non increasing, is indeed Hadamard differentiable; letting ν denote the measure $\mathrm{d}\nu(t) = t^\eta \mathrm{d}t + \mathrm{d}\epsilon_1(t)$, with ϵ_1 a point mass at 1, the derivative is

$$\Theta'(z) = \int_0^1 t^\eta z(t)\mathrm{d}t - y(1) = \int_{[0,1]} z(t)\mathrm{d}\nu(t).$$

Step 3. : The weak limit in (4.34) is thus equal to $\int_{[0,1]} \eta t^{-(\eta+1)}\bar{W}(t)\mathrm{d}\nu(t)$. From (Shorack and Wellner, 2009, Proposition 2.2.1), the latter random variable is centered Gaussian with variance

$$\sigma^2 = \iint_{[0,1]^2} \eta^2 (st)^{-(\eta+1)}\, \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(t))\mathrm{d}\nu(s)\mathrm{d}\nu(t).$$

By definition of ν and by symmetry of the covariance,

$$\begin{aligned}\sigma^2/\eta^2 = 2\underbrace{\int_{s=0}^1 \int_{t=0}^s (st)^{-1}\, \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(t))\mathrm{d}t\mathrm{d}s}_{A} \\ - 2\underbrace{\int_{s=0}^1 \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(1))s^{-1}\mathrm{d}s}_{B} + \mathrm{Var}(\bar{W}(1)).\end{aligned}$$

For any $s \in (0,1)$,

$$\int_{t=0}^{s} \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(t))(st)^{-1}\mathrm{d}t = \int_{u=0}^{1} \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(us))(su)^{-1}\mathrm{d}u$$
$$= \int_{u=0}^{1} \mathbb{C}\mathrm{ov}(\bar{W}(1), \bar{W}(u))(u)^{-1}\mathrm{d}u = B.$$

The penultimate equality follows from $\mathbb{C}\mathrm{ov}(\bar{W}(\lambda s), \bar{W}(\lambda t)) = \lambda\, \mathbb{C}\mathrm{ov}(\bar{W}(s), \bar{W}(t))$ for $\lambda > 0$ and $s, t \in (0,1]$. Therefore $A = B$ and $\sigma^2 = \eta^2 \operatorname{Var}(\bar{W}(1))$, as required. □

4.10 CLEF ALGORITHM AND VARIANTS

The CLEF algorithm is described at length in Chiapino and Sabourin (2016). For completeness, its pseudo-code is provided below. The underlying idea is to iteratively construct pairs, triplets, quadruplets... of features that are declared 'dependent' whenever $\widehat{\kappa}_\alpha \geq C$ for some user-defined tolerance level $C > 0$. Varying this criterion produces three variants of the original algorithm, namely CLEF-Asymptotic, CLEF-Peng, and CLEF-Hill. The pruning stage of the algorithm is the same for all three variants.

Algorithm 4 CLEF (CLustering Extreme Features)

Input: Tolerance parameter $\kappa_{\min} > 0$.

STAGE 1: constructing the collection $\widehat{\mathbb{M}}$ of tail-dependent groups.
Step 1: Put $\hat{\mathcal{A}}_1 = \{\{1\}, \ldots, \{d\}\}$ and $S = 1$.
Step $s = 2, \ldots, d$: If $\hat{\mathcal{A}}_{s-1} = \emptyset$, end **STAGE 1**. Otherwise:

- Generate candidates of size s:
  $\mathcal{A}'_s = \{\alpha \subset \{1, \ldots, d\} : |\alpha| = s \text{ and } \alpha \setminus j \in \hat{\mathcal{A}}_{s-1} \text{ for all } j \in \alpha\}$.
- Put $\hat{\mathcal{A}}_s = \left\{\alpha \in \mathcal{A}'_s : \hat{\kappa}_\alpha > \kappa_{\min}\right\}$.
- If $\hat{\mathcal{A}}_s \neq \emptyset$, put $S = s$.

Output: $\widehat{\mathbb{M}} = \emptyset$ if $S = 1$ and $\widehat{\mathbb{M}} = \bigcup_{s=2}^{S} \hat{\mathcal{A}}_s$ if $S \geq 2$.

STAGE 2: pruning, keeping maximal groups α only.
If $S = 1$, then $\widehat{\mathbb{M}}_{\max} = \emptyset$. Otherwise:
Initialization: $\widehat{\mathbb{M}}_{\max} \leftarrow \hat{\mathcal{A}}_S$.
for $s = (S-1) : 2$,
  for $\alpha \in \hat{\mathcal{A}}_s$,
    If there is no $\beta \in \widehat{\mathbb{M}}_{\max}$ such that $\alpha \subset \beta$, then $\widehat{\mathbb{M}}_{\max} \leftarrow \widehat{\mathbb{M}}_{\max} \cup \{\alpha\}$.
Output: $\widehat{\mathbb{M}}_{\max}$

5 Clustering of Extreme points and Visualization

Abstract In a wide variety of situations, anomalies in the behaviour of a complex system, whose health is monitored through the observation of a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ valued in $\mathbb{R}^d$, correspond to the simultaneous occurence of extreme values for certain subgroups $\alpha \subset \{1, \ldots, d\}$ of variables X_j. Under the heavy-tail assumption, which is precisely appropriate for modeling these phenomena, a statistical method for identifying such events/subgroups has been recently developed in Goix et al. (2016a), relying on the concept of angular measure in multivariate extreme value theory, which characterizes the dependence structure of the X_j's in the extremes. It is the purpose of this paper to exploit this approach further, by means of a mixture model that permits to describe the distribution of extremal observations and where the anomaly type α is viewed as a latent variable. In particular, the model enables to assign to any such point $\mathbf{X}$ a posterior probability for each anomaly type α, defining implicitely a similarity measure between anomalies. A procedure based on the EM algorithm is also proposed here to infer the parameters of the mixture model from a (truncated) training dataset and it is explained at length how the corresponding posterior similarity measure estimates permit to obtain an informative planar representation of anomalies using standard graph-mining tools. The relevance and usefulness of the 2-d visual display thus designed is illustrated on real datasets, in the aeronautics application domain.

5.1 Introduction

Motivated by a wide variety of applications ranging from fraud detection to aviation safety management through the health monitoring of complex networks, data center infrastructure management or food risk analysis, unsupervised anomaly detection is now the subject of much attention in the data science literature, see *e.g.* D. Gorinevsky (2012); T. Fawcett (1997); Viswanathan et al. (2012). In frequently encountered practical situations and from the viewpoint embraced in this paper, anomalies coincide with rare measurements that are extremes, *i.e.* located far from central statistics such as the sample mean. In the 1-d setting, numerous statistical techniques for anomaly detection are based on a parametric representation of the tail of the observed univariate probability distribution, relying on *extreme value theory*

(EVT) (see *e.g.* Clifton et al. (2011b); Lee and Roberts (2008a); Roberts (2000); Tressou (2008) among others). In (even moderately) large dimensional situations, the modelling task becomes much harder and many nonparametric heuristics for supervised classification have been thus adapted, substituting rarity for labeling, see *e.g.* Schölkopf et al. (2001), Steinwart et al. (2005) or Liu et al. (2008). In the unsupervised setting, whereas many dimensionality reduction and visualization techniques, extending the basic linear PCA methodology, accounting for non linearities or increasing robustness for instance (*cf* Gorban et al. (2008) and Kriegel et al. (2008)), have been proposed in the statistics and data-mining literature to describe parsimoniously the 'center' of a massive data distribution (see *e.g.* Naik (2017) and the references therein), the issue of clustering extremes or outliers is only recently receiving attention, at the instigation of industrial applications such as those mentioned above and because of the increasing availability of extreme observations in databases: generally out-of-sample in the past, extreme values are becoming observable in the Big Data era.

It is the goal of the present article to propose a novel mixture model-based approach for clustering extremes in the multivariate setup, *i.e.* when the observed random vector $\mathbf{X} = (X_1, \ldots, X_d)$ takes its values in the positive orthant of the space $\mathbb{R}^d$ with $d > 1$ equipped with the sum-norm $\|(x_1, \ldots, x_d)\| = \sum_{1 \leq j \leq d} |x_j|$: 'extremes' coinciding then with values x such that $\mathbb{P}[\|\mathbf{X}\| > \|\mathbf{x}\|]$ is 'extremely small'. Precisely, it relies on a dimensionality reduction technique of the tail distribution recently introduced in Goix et al. (2017) (see also Goix et al. (2016a)) and referred to as the DAMEX algorithm. Based on multivariate extreme value theory (MEV theory), the latter method may provide a hopefully sparse representation of the support of the angular measure related to the supposedly heavy-tailed distribution of the r.v. $\mathbf{X}$. As the angular measure asymptotically describes the dependence structure of the variables X_j in the extremal domain (and, roughly speaking, permit to assign limit probabilities to directions $\mathbf{x}/\|\mathbf{x}\|$ in the unit sphere along which extreme observations may occur), this statistical procedure identifies the groups $\alpha \subset \{1, \ldots, d\}$ of feature indices such that the collection of variables $\{X_j : j \in \alpha\}$ may be simultaneously very large, while the others, the X_j's for $j \notin \alpha$, remain small. Groups of this type being in 1-to-1 correspondence with the faces $\Omega_\alpha = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1, x_j = 0 \text{ if } j \notin \alpha \text{ and } x_j > 0 \text{ if } j \in \alpha\}$ of the unit sphere composing the support of the angular measure. In practice, a sparse representation of the extremal dependence structure is obtained when only a few such groups of variables can be exhibited (compared to $2^d - 1$) and/or when these groups involve a small number of variables (with respect to d). Here we develop this framework further, in order to propose a (soft) clustering technique in the region of extremes and derive effective 2-d visual displays, sheding light on the structure of anomalies/extremes in sparse situations. By modelling the distribution of extremes as a specific *mixture model*, where each component generates a different type α of extremes, the Expectation-Maximization algorithm (EM in abbreviated form) permits to partition/cluster the set of extremal data through the statistical recovery of *latent*

observations, as well posterior probability distributions (inducing a soft clustering of the data in a straighforward manner) and, as a by-product, a similarity measure on the set of extremes: the higher the probability that their latent variables are equal, the more similar two extreme observations X and X' are considered. The similarity matrix thus obtained naturally defines a *weighted graph*, whose vertices are the anomalies/extremes observed, paving the way for the use of powerful graph-mining techniques for community detection and visualization, see *e.g.* Schaeffer (2007) and Hu and Shi (2015) as well as the references therein. Beyond its detailed description, the methodology proposed is applied to a real fleet monitoring dataset in the aeronautics domain and shown to provide useful tools for analyzing and interpreting abnormal data.

The paper is structured as follows. Basic concepts of MEV theory are briefly recalled in Section 5.2, together with the DAMEX technique proposed in Goix et al. (2016a, 2017) for estimating the (hopefully sparse) support of a heavy-tailed distribution. Section 5.3 introduces the concept of angular measure and details the mixture model we propose to describe the distribution of extreme data, based on DAMEX output, together with the EM algorithm variant we introduce in order to estimate its parameters. It is next explained in Section 5.4 how to exploit the results of this inference method to define a similarity matrix of the extremal data, reflecting a weighted graph structure of the observed anomalies, and apply dedicated community detection and visualization techniques so as to extract meaningful information from the set of extreme observations. The relevance of the approach we promote is finally illustrated by numerical experiments, on synthetic and real data in Section 5.5.

5.2 Background and preliminaries

As a first go, we start with recalling key notions of MEVT, as well a the inference method investigated in Goix et al. (2016a, 2017) to estimate its support. Here and throughout, the Dirac mass at any point x is denoed by δ_x, the indicator function of any event A by $\mathbf{1}\{A\}$. Capital letters generally refer to random quantities whereas lower case ones denote deterministic values. Finally, boldface letters denote vectors as opposed to Roman letters denoting real numbers.

5.2.1 Multivariate extreme value theory

It is the goal of Extreme Value Theory (EVT) to describe phenomena that are not governed by an 'averaging effect' but can be instead significantly impacted by very large movements. By focusing on large quantiles rather than central statistics such as the median or the sample mean, EVT provides models for the unusual rather than the usual and permits to assess the probability of occurence of rare (extreme) events. Application domains are numerous and diverse, including any field related to risk management as finance, insurance, environmental sciences or aeronautics. Risk

monitoring is a typical use case of EVT. In the univariate setting, typical quantities of interest are high quantiles of a random variable X, *i.e.* $1-p$ quantiles for $p \to 0$. When p is of the order of magnitude or smaller than $1/N$, empirical estimates become meaningless. Another issue is the estimation of the probability of an excess over a high threshold u, $p_u = \mathbb{P}(X > u)$ when few (or none) observations are available above u. In such contexts, EVT essentially consists in using a parametric model (the generalized Pareto distributions) for the tail distribution, which is theoretically justified asymptotically, *i.e.* when $p \to 0$ or $u \to \infty$. The required assumption is the existence of two sequences $\{a_n, n \geq 1\}$ and $\{b_n, n \geq 1\}$, with $a_n > 0$ and a non-degenerate cumulative distribution function (*c.d.f.*) G such that

$$n\,\mathbb{P}\left(\frac{X-b_n}{a_n} \;\geq\; x\right) \xrightarrow[n\to\infty]{} -\log G(x) \tag{5.1}$$

For all x in the continuity set of G. Notice that this assumption is satified by most textbook distributions, *e.g.* the normal, exponential, Cauchy, beta, gamma distributions. The reader is referred to Coles (2001) and the references therein for an introduction to EVT and its applications.

In the multivariate setting, EVT is concerned about the tail behaviour of a d-dimensional random variable $\mathbf{X} = (X_1, \ldots, X_d)$. The goal is to infer quantities of the kind $\mathbb{P}[X_1 > x_1, \ldots, X_d > x_d]$ for large $x_1, \ldots, x_d$. A natural first step is to standardize each component so as to work with identically distributed component and focus on the dependence structure. One convenient choice is to use the probability integral transform: For $\mathbf{x} = (x_1, \ldots, x_d)$, let $F_j(x_j) = \mathbb{P}[X_j \leq x_j]$. Assuming that F_j is continuous, the transformed variable $V_j = (1 - F_j(X_j))^{-1}$ follows a Pareto distribution, $\mathbb{P}[V_j > v] = v^{-1}$, $v \geq 1$. Consider the the Pareto-tranformed variable $\mathbf{V} = (V_1, \ldots, V_d)$. A multivariate analogue of Assumption (5.1) is

$$n\mathbb{P}[\frac{V_1}{n} > v_1, \text{ or} \ldots, \text{or } \frac{V_d}{n} > v_d] \xrightarrow[n\to\infty]{} -\log G(\mathbf{v}) \tag{5.2}$$

where $\mathbf{v} = (v_1, \ldots, v_d)$, $v_j > 0$ and G is a multivariate *c.d.f.*. Notice that the choice of Pareto margins implies normalizing sequences $a_n = n, b_n = 0$ for each component V_j and that G has unit Fréchet margins, $G_j(v) = e^{-1/v}$, $v > 0$. Other standardizations are possible which lead to alternative normalizing sequences and limits.

Exponent measure : To understand the right-hand-side of (5.2), the following result (see *e.g.*Resnick (1987, 2007b)) is key: there exists a measure μ on $E = \mathbb{R}_+^d \setminus \{0\}$ which is finite on any set A such that $\mathbf{0}$ does not belong to the closure of A, such that $-\log G(\mathbf{v}) = \mu[\mathbf{0}, \mathbf{v}]^c$. μ is called the *exponent measure*. It is homogeneous of order -1, that is $\mu(tA) = t^{-1}\mu(A)$, where $tA = \{t\mathbf{v}, \mathbf{v} \in A\}, A \subset \mathbb{R}_d^+$. Another consequence of (5.2) is that for all $A \subset \mathbb{R}_+^d$ such that $0 \notin \partial A$,

$$t\mathbb{P}[\mathbf{V} \in tA] \xrightarrow[t\to\infty]{} \mu(A). \tag{5.3}$$

This convergence property applies immediatly to the problem of estimating the probability of reaching a set tA which is far form $\mathbf{0}$ (*i.e.* t is large): one may write

$\mathbb{P}(\mathbf{V} \in tA) \approx \frac{1}{t}\mu(A)$, so that estimates of μ automatically provide estimates for such quantities.

In a word, μ may be used to characterize the distributional tail of $\mathbf{V}$.

5.2.2 Support estimation

The goal of this section is to expose the connection between the support of μ and the subsets of components which may assume large values simultaneously. **Sparse support :** Consider $\alpha \subset \{1, \ldots, d\}$ a nonempty subset of features and the associated truncated cone

$$\mathcal{C}_\alpha = \Big\{\mathbf{x} \geq 0 : \ \|\mathbf{x}\|_\infty \geq 1, \ x_i > 0 \text{ for } i \in \alpha, \\ x_i = 0 \ \text{ for } i \notin \alpha \ \Big\}. \tag{5.4}$$

as illustrated in Fig. 5.1. The family $\{\mathcal{C}_\alpha, \alpha \subset \{1, \ldots, d\}, \alpha \neq \emptyset\}$ defines a partition of $\mathbb{R}_+^d \setminus [0,1]^d$ which is of particular interest for ou purposes: indeed, $\mu(\mathcal{C}_\alpha) > 0$ if and only if the limiting rescaled probability that all feature in α are large while the others are small is non zero, see Remark 5.1.

Remark 5.1. Consider the ϵ-thickened rectangles

$$\mathcal{R}_\alpha^\epsilon \ = \Big\{\mathbf{v} \geq 0, \ \|\mathbf{v}\|_\infty \geq 1, \ v_i > \epsilon \ \text{ for } i \in \alpha, \\ v_i \leq \epsilon \ \text{ for } i \notin \alpha \ \Big\}, \tag{5.5}$$

which defines again a partition of $\mathbb{R}_+^d \setminus [0,1]^d$ for each fixed $\epsilon \geq 0$. Also $\mathcal{C}_\alpha = \cap_{\epsilon>0, \epsilon\in\mathbb{Q}} \mathcal{R}_\alpha^\epsilon$. Thus by upper continuity of μ,

$$\mu(\mathcal{C}_\alpha) = \lim_{\epsilon \to 0} \mu(\mathcal{R}_\alpha^\epsilon)$$

with

$$\mu(\mathcal{R}_\alpha^\epsilon) = \lim_{t\to\infty} t\mathbb{P}(\|\mathbf{V}\|_\infty > t \ \forall j \in \alpha : V_j > t\epsilon, \ \ \forall j \notin \alpha : V_j < t\epsilon).$$

Now the right-hand side of the latter display corresponds to the event that all features in α are large while the other are small.

In the sequel we denote

$$\mu_\alpha = \mu(\mathcal{C}_\alpha), \qquad \mathbb{M} = \Big\{\alpha \subset \{1, \ldots, d\}, \alpha \neq \varnothing, \mu_\alpha > 0\Big\}.$$

In theory, every μ_α may be positive . However a reasonable assumption in a many high dimensional contexts is that $\mu_\alpha = 0$ for the vast majority of the $2^d - 1$ cones $\mathcal{C}_\alpha$. In other words, not all combinations of coordinates of $\mathbf{V}$ can be large together, so that the support of μ is sparse.

DAMEX algorithm : Earlier works (Goix et al. (2016a)) have proposed an algorithm named DAMEX which produces the list of α's such that the empirical

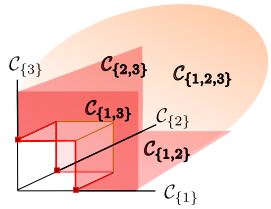


Figure 5.1: Truncated cones $\mathcal{C}_\alpha$ in 3D

counterpart of μ_α (denoted $\hat{\mu}_\alpha$ in the sequel) is non zero. Defining a threshold $m_{\min} > 0$ below which $\hat{\mu}_\alpha$ is deemed negligible, one thus obtains a list ofsubsets $\widehat{\mathbb{M}} = \{\alpha \subset \{1, \ldots, d\} : \hat{\mu}_\alpha > \mu_{\min}\}$. A uniform boud on the error $|\hat{\mu}_\alpha - \mu_\alpha|$ is derived in Goix et al. (2017) which scale roughly as $k^{-1/2}$, where k is the order of magnitude of the number of largest observations used to learn $\mathbb{M}$ and the μ_α's. Given a dataset $(\mathbf{X}_i)_{i \leq n}$ of independent data identically distributed as $\mathbf{X}$, estimation proceeds as follow: first, replace the unkown marginal distributions F_j with their empirical counterpart $\hat{F}_j(x) = \frac{1}{n}\sum \mathbf{1}\{X_{i,j} < x\}$. Define then $\hat{V}_{i,j} = (1 - \hat{F}_j(X_{i,j}))^{-1}$ and $\hat{\mathbf{V}}_i = (\hat{V}_{i,1}, \ldots, \hat{V}_{i,d})$. Then choose some $k \ll n$ large enough (typically $k = O(\sqrt{n})$) and define $\hat{\mu}_\alpha$ as the empirical counterpart of $\mu(R^\epsilon_\alpha)$ with t replaced with n/k in (5.3), that is

$$\hat{\mu}_\alpha = \frac{1}{k}\sum_{i=1}^{n} \mathbf{1}\{\hat{\mathbf{V}}_i \in \frac{n}{k}\mathcal{R}^\epsilon_\alpha\}.$$

Notice that the above definition is a variant of the orignal algorithm in Goix et al. (2016a) which uses thickened cones $\mathcal{C}^\epsilon_\alpha$ instead of $\mathcal{R}^\epsilon_\alpha$. However finite sample guarantees in Goix et al. (2017) are obtained using the latter rather than the original $\mathcal{C}^\epsilon_\alpha$'s, which is why we prefer using the $\mathcal{R}^\epsilon_\alpha$'s.

5.3 A MIXTURE MODEL FOR MULTIVARIATE EXTREME VALUES

The idea behind this section is to consider a mixture model for μ (the distribution of the largest instances of the dataset) indexed by $\alpha \in \widehat{\mathbb{M}}$, where $\widehat{\mathbb{M}}$ is the output of the DAMEX algorithm. Thus each component $\alpha \in \widehat{\mathbb{M}}$ of the mixture generated instances $\mathbf{V}$ such that V_j is likely to be large for $j \in \alpha$

In this paper we adopt a plug-in approach and identify $\mathbb{M}$ (the true support of μ) with $\widehat{\mathbb{M}}$ (the output of DAMEX).

For modeling purposes, the homogeneity property $\mu(t\cdot) = t^{-1}\mu(\cdot)$ suggests a preliminary decomposition of μ within a (pseudo)-polar coordinates system, as detailed next.

5.3.1 Angular measure

Let us consider the sum-norm $\|\mathbf{v}\| := v_1 + \ldots + v_d$ and $\mathcal{S}_d := \{\mathbf{w} \in \mathbb{R}^d_+ : \|\mathbf{w}\| = 1\}$ the d-dimensional simplex. Introduce the polar transformation $T : \mathbf{v} \mapsto T(\mathbf{v}) = (r, \mathbf{w})$

defined on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$, where where $r = \|\mathbf{v}\|$ is the radial component and $\mathbf{w} = r^{-1}\mathbf{v}$ is the angular one. Now define an *angular measure* Φ on $\mathcal{S}_d$ (see *e.g.* Resnick (2007b) or Beirlant et al. (2004) and the references therein):

$$\Phi(A) := \mu\left\{\mathbf{v} : \|\mathbf{v}\| > 1, \|\mathbf{v}\|^{-1}\mathbf{v} \in A\}\right\},$$

with $A \subset \mathcal{S}_d$. Notice that $\Phi(\mathcal{S}_d) < \infty$ and that by homogeneity, it may be shown that

$$\mu \circ T^{-1}(\mathrm{d}r, \mathrm{d}\mathbf{w}) = r^{-2}\mathrm{d}r\Phi(\mathrm{d}\mathbf{w}). \tag{5.6}$$

In other words the exponent measure μ factorizes into a radial component and an angular component. Setting $R = \|\mathbf{V}\|$ and $\mathbf{W} = R^{-1}\mathbf{V}$, a consequence of (5.3) is that

$$\mathbb{P}[\mathbf{W} \in A, R > tr \| R > t] \xrightarrow[t\to\infty]{} r^{-1}\Phi(\mathcal{S}_d)^{-1}\Phi(A) \tag{5.7}$$

for all measurable set $A \subset \mathcal{S}_d$ and $r > 1$. In other words, given that the radius R is large, the radius R and the angle $\mathbf{W}$ are approximately independent, the distribution of $\mathbf{W}$ is approximately the angular measure – up to a normalizing constant $\Phi(\mathcal{S}_d)$ – and R follows approximately a Pareto distribution.

The transformation to unit Pareto margins and the choice of the sum-norm yield the following moment constraint on Φ:

$$\int_{\mathcal{S}_d} w_i \, \Phi(\mathrm{d}\mathbf{w}) = 1, \text{ for } i = 1, \ldots, d. \tag{5.8}$$

In addition, the normalizing constant is explicit:

$$\Phi(\mathcal{S}_d) = \int_{\mathcal{S}_d} \Phi(\mathrm{d}\mathbf{w}) = \int_{\mathcal{S}_d} (w_1 + \ldots + w_d)\Phi(\mathrm{d}\mathbf{w}) = d. \tag{5.9}$$

Remark 5.2. The choice of the sum-norm here is somewhat arbitrary. Any other norm on $\mathbb{R}^d$ for the pseudo-polar transformation is equally possible, leading to alternative moment constraints and normalizing constants. The advantage of the sum-norm is that it allows convenient probabilistic modeling of the angular component $\mathbf{w}$ on the unit simplex.

5.3.2 A mixture model

The partition of $\mathbb{R}_+^d$ into cones $\mathcal{C}_\alpha$ introduced in Section 5.2.2 induces a partition of $\mathcal{S}_d$ into $2^d - 1$ sub-simplices $\mathcal{S}_\alpha$, $\varnothing \neq \alpha \subset \{1, \ldots, d\}$,

$$\mathcal{S}_\alpha = \left\{\mathbf{v} \in \mathbb{R}_+^d : \ \|v\| = 1, v_i > 0 \text{ for } i \in \alpha, v_i = 0 \text{ for } i \notin \alpha, \right\}.$$

Also by homogeneity, the following equivalence holds:

$$\Phi(\mathcal{S}_\alpha) > 0 \Leftrightarrow \mu(\mathcal{C}_\alpha) > 0. \tag{5.10}$$

Recall that our key assumption in this work is that the support of μ is sparse, namely we assume that $|\mathbb{M}| \ll 2^d$, where $\mathbb{M} = \{\alpha : \mu(\mathcal{C}_\alpha) > 0\} = \{\alpha : \Phi(\mathcal{S}_\alpha) > 0\}$.

In view of (5.10) and (5.9), the angular measure admits the decomposition

$$\Phi(\,\cdot\,) = d \sum_{\alpha \in \mathbb{M}} \pi_\alpha \Phi_\alpha(\,\cdot\,) \tag{5.11}$$

where Φ_α is a probability measure on $\mathcal{S}_\alpha$ and $\sum_{\alpha\in\mathbb{M}} \pi_\alpha = 1$. We make the simplifying assumption that the sets $\alpha \in \mathbb{M}$ are not nested, *i.e.* there does not exist two subsets $\alpha, \beta \in \mathbb{M}$ such that $\alpha \subset \beta$. Notice that this assumption could be omitted at the price of additional notational complexity.

Introduce the set of coordinates which are singletons in $\mathbb{M}$, $\mathbb{E} = \{j \in \{1, \ldots, d\} : \{j\} \in \mathbb{M}\}$, as opposed to $\mathbb{M}_2 = \{\alpha \in \mathbb{M} : |\alpha| \geq 2\}$. Up to relabeling we may assume that $\mathbb{E} = \{1, \ldots, d_1\}$ for some $1 \leq d_1 \leq d\}$ or that $\mathbb{E} = \varnothing$, in which case $d_1 = 0$. Then $\bigcup_{\alpha\in\mathbb{M}_2} \alpha = \{d_1 + 1, \ldots, d\}$. For convenience let us write $\mathbb{M}_2 = \{\alpha_1, \ldots, \alpha_K\}$ with $K = |\mathbb{M}_2|$ and let us relabel the weights as $\pi_k = \pi_{\alpha_k}$ for $k \leq K$, $\pi_{K+j} = \pi_{\{j\}}$ for $j \leq d_1$. Equipped with these notations, (5.11) becomes

$$d^{-1}\Phi(\,\cdot\,) = \sum_{k=1}^{K} \pi_k \Phi_{\alpha_k}(\,\cdot\,) + \sum_{j \leq d_1} \pi_{K+j} \delta_{\mathbf{e}_j}(\,\cdot\,) \tag{5.12}$$

where δ_a is the Dirac mass at point a and $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)$ the jth canonical basis vector of $\mathbb{R}^d$.

The singletons weights derive immediately from the moment constraint (5.8): for $i \leq d_1$,

$$d^{-1} = \underbrace{\sum_{k=1}^{K} \int_{\mathcal{S}_{\{i\}}} w_i \pi_k \Phi_{\alpha_k}(\mathrm{d}\mathbf{w})}_{=0} + \sum_{j \leq d_1} \int_{\mathcal{S}_{\{i\}}} w_i \, \pi_{K+j} \delta_{\mathbf{e}_j}(\mathrm{d}\mathbf{w}) = \pi_{K+i}$$

We obtain

$$\Phi(\,\cdot\,) = d \sum_{k=1}^{K} \pi_k \Phi_{\alpha_k}(\,\cdot\,) + \sum_{j \leq d_1} \delta_{\mathbf{e}_j}(\,\cdot\,)$$

where the vector $\pi \in [0,1]^{K+d_1}$ must satisfy

$$\sum_{k=1}^{K} \pi_k = 1 - d_1/d. \tag{5.13}$$

As is usual for mixture modeling purposes, we introduce a latent variable $\mathbf{Z} = (Z_1, \ldots, Z_K, Z_{K+1}, \ldots, Z_{K+d_1})$ such that for $k \leq K$ (*resp.* $k > K$) , $Z_k = 1$ if $\mathbf{W}$ has been generated by the mixture component Φ_{α_k} (*resp.* $\delta_{\mathbf{e}_{k-K}}$) and $Z_k = 0$ otherwise. Then for $k \leq K$, $\mathbb{P}[Z_k = 1] = \pi_k$ while for $k > K$, $\mathbb{P}[Z_k = 1] = d^{-1}$.

Dirichlet model : One natural model for probability distributions on a simplex $\mathcal{S}_\alpha$ is the Dirichlet family. Such distributions admit a density φ_α with respect to the $(|\alpha|-1)$ Lebesgue measure which we denote $\mathrm{d}\mathbf{w}$ for simplicity. It can be parameterized by a mean vector $\mathbf{m}_\alpha \in \mathcal{S}_\alpha$ and a concentration parameter $\nu_\alpha > 0$, so that for $\mathbf{w} \in \mathcal{S}_\alpha$,

$$\varphi_\alpha(\mathbf{w}|\mathbf{m}_\alpha, \nu_\alpha) = \frac{\Gamma(\nu_\alpha)}{\prod_{i\in\alpha}\Gamma(\nu_\alpha m_{\alpha,i})} \prod_{i\in\alpha} w_i^{\nu_\alpha m_{\alpha,i}-1}. \tag{5.14}$$

In this paper we model Φ_α by a Dirichlet distribution with unknown parameters m_α, ν_α. Using the standard fact that for such a distribution, $\int_{\mathbb{S}_\alpha} \mathbf{w}\varphi_\alpha(\mathbf{w}|m_\alpha, \nu_\alpha)\mathrm{d}\mathbf{w} = \mathbf{m}_\alpha$, the moment constraint (5.8) becomes:

$$\frac{1}{d} = \sum_{k=1}^{K} \pi_k \mathbf{m}_{k,j}, \quad j \in \{d_1+1, \ldots, d\}. \tag{5.15}$$

where we have set $\mathbf{m}_k = \mathbf{m}_{\alpha_k}$, $k \leq K$.

The Dirichlet mixture model may be summarized as follows:

Model 1 (Dirichlet mixture model).

1. Consider a standardized random vector $\mathbf{V}$ such that $V_j$has standard Pareto distribution (see Section 5.2.1)
2. Set $R = \|\mathbf{V}\|, \mathbf{W} = R^{-1}\mathbf{V}$.
3. Fix some high radial threshold r_0, typically a large quantile of the observed radii.
4. Let $\mathbf{Z}$ be a hidden variable indicating the mixture component responsible for $\mathbf{W}$, and k such that $Z_k = 1$. Then let $\Phi_k = \varphi_k(\cdot|m_k, \nu_k)$ if $k \leq K$, otherwise let $\Phi_k = \delta_{k-K}$. Conditionally to $\{R > r_0\}$, for $r > r_0$,

$$\mathbb{P}[R > r] = r_0 r^{-1}\ ; \quad \mathbf{W} \sim \Phi_k\ ; \quad \mathbf{W} \perp\!\!\!\perp R \tag{5.16}$$

The unknown parameters of the model are ($\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$, $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$), where $\nu_k > 0$ and where $\boldsymbol{\pi}, \mathbf{m}$ must satisfy the constraints (5.13) and (5.15)

Figure 5.2 illustrates Model 1 in the bivariate case.

Sub-asymptotic model: incorporating a noise : Recall from (5.7) that Φ is the *limiting* distribution of $\mathbf{W}$ for large R's. In practice the observed angles corresponding to radii $R > r_0$ follow a sub-asymptotic version of Φ. In particular, the marginal variables V_j have continuous Pareto distributions, so that $\mathbb{P}(V_j > 1) = 1$, $j \in \{1, \ldots, d\}$. As a consequence with probability 1, all the $\mathbf{V}_i = (V_{i,1}, \ldots, V_{i,d})$, $1 \leq i \leq n$, lie in the central cone $\mathcal{C}_{\{1,\ldots,d\}}$ and the angles $\mathbf{W}_i = (W_{i,1}, \ldots W_{i,d})$ lie in

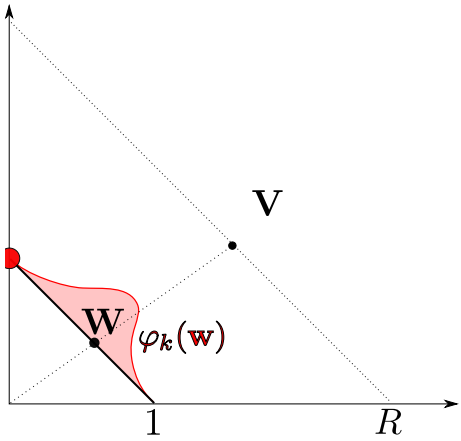


Figure 5.2: Bivariate illustration of Model 1.
Here, $\mathbf{V}$ is generated from component $\alpha_k = \{1, 2\}$ and the model has two components: $\mathbb{M} = \{\{2, \}, \{1, 2\}\}$. The red line is the Dirichlet density φ_k on the unit simplex $\mathcal{S}_{\{1,2\}}$. The red point represent the other component $\delta_{\mathbf{e}_2}$.

$\mathcal{S}_{\{1,\ldots,d\}}$, the interior of $\mathcal{S}_d$ (This is also true using the empirical versions $\hat{\mathbf{V}}_i$ defined in Section 5.2.2).

To account for the non-asymptotic nature of the data, we model the deviation from the asymptotic support of $\mathbf{V}$, which is $\bigcup_{\alpha\in\mathbb{M}} \mathcal{C}_\alpha$, as a noise $\boldsymbol{\varepsilon}$ with light tailed distribution, namely an exponential distribution. We denote by $\tilde{\mathbf{V}} = \mathbf{V} + \boldsymbol{\varepsilon}$ the resulting noisy vector and we assume that only $\tilde{\mathbf{V}}$ is observed (not $\mathbf{V}$). This sub-asymptotic model may be described as follows

Model 2 (Sub-asymptotic mixture model).

1. Let $\tilde{\mathbf{V}}$ be an observed random vector which marginal distributions are approximately Pareto (typically $\tilde{V}_j = (1 - \hat{F}_j(X_j))$ for $\hat{F}_j$ an estimate of the marginal distribution F_j of X_j)

2. Let $\mathbf{Z} \in \{0, 1\}^{K+d_1}$ be a hidden variable as in Model 1 and let $\tilde{R} = \|\tilde{\mathbf{V}}\|$. Then for $1 \leq k \leq K + d_1$, conditionally to $\{\tilde{R} > r_0, Z_k = 1\}$, the observed vector $\tilde{\mathbf{V}}$ decomposes as

$$\tilde{\mathbf{V}} \;=\; \mathbf{V}_k + \boldsymbol{\varepsilon}_k \;=\; R_k \mathbf{W}_k + \boldsymbol{\varepsilon}_k, \tag{5.17}$$

   where $\mathbf{V}_k \in \mathcal{C}_{\alpha_k}$, $\boldsymbol{\varepsilon}_k \in \mathcal{C}_{\alpha_k}^{\perp}$ are independent from each other and where $R_k = \|\mathbf{V}_k\|$, $\mathbf{W}_k = R_k^{-1}\mathbf{V}_k \in \mathcal{S}_{\alpha_k}$ are as in Model 1, *i.e.* R_k is Pareto distributed, $\mathbf{W}_k \sim \Phi_k$ and $R_k, \mathbf{W}_k$ are independent.

3. The noise's components are independent and identically distributed as a translated exponential distribution with rate λ_k: for $j \in \{1, \ldots, d\} \setminus \alpha_k$,

$$\varepsilon_j \sim 1 + \mathcal{E}xp(\lambda_k)$$

The unknown parameters are those inherited from Model 1, with the addition of the exponential rates $\boldsymbol{\lambda} = (\lambda_k,\ 1 \leq k \leq K + d_1)$ where $\lambda_k > 0$.

Figure 5.3 illustrates Model 2 in dimension $d = 3$.

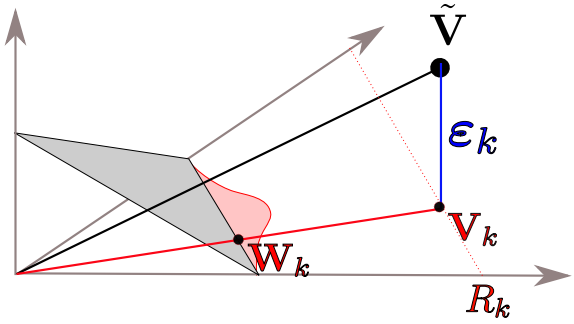


Figure 5.3: Trivariate illustation of the sub-asymptotic model 2
Here the observed point $\tilde{\mathbf{V}}$ has been generated by component $\alpha_k = \{1, 2\}$. The grey triangle is the unit simplex, the shaded red area represents the Dirichlet density φ_k.

The next paragraph describes an EM-algorithm for maximum-likelihood estimation of Model 2.

5.3.3 An EM algorithm for model inference

The likelihood for Model 2, $p(\tilde{\mathbf{v}}|\mathbf{m}, \boldsymbol{\nu}, \pi, \boldsymbol{\lambda})$, for one observation $\tilde{\mathbf{v}} \in (1, \infty)^d$, $\|\tilde{\mathbf{v}}\| \geq r_0$, follows directly from the model specification,

$$\begin{aligned} p(\tilde{\mathbf{v}}|\mathbf{m}, \boldsymbol{\nu}, \pi, \boldsymbol{\lambda}) =& r_0 \sum_{k=1}^{K} \pi_k\, r_k^{-|\alpha_k|-1}\, \varphi_k(\mathbf{w}_k|\mathbf{m}_k, \nu_k) \prod_{j \in \alpha_k^c} f_\varepsilon(\tilde{v}_j|\lambda_k) \\ &+ \frac{r_0}{d} \sum_{k=K+1}^{K+d_1} r_k^{-2} \prod_{j \in \{1,\ldots,d\} \setminus k} f_\varepsilon(\tilde{v}_j|\lambda_k) \end{aligned} \tag{5.18}$$

where $f_\varepsilon(\cdot\ |\lambda_k)$ denotes the marginal density for the noise $\boldsymbol{\varepsilon}_k$ given the noise parameter λ_k. As specified in Model 2, in this paper we set $f_\varepsilon(x|\lambda_k) = \lambda_k e^{-\lambda_k(x-1)}$, $x > 1$ (a translated exponential density), but any other light tailed distribution could be used instead. Notice that the term $r_k^{-|\alpha_k|-1} = r_k^{-2} r_k^{-|\alpha_k|+1}$ is the product of the radial Pareto density and the Jacobian term for the change of variables $T_k : \mathbf{V}_k \mapsto (R_k, \mathbf{W}_k)$.

Recall that the constraints are

$$\nu_k > 0\ (1 \leq k \leq K)\ , \qquad \lambda_k > 0\ (1 \leq k \leq K + d_1), \tag{5.19}$$

and that $\pi = (\pi_1, \ldots, \pi_K)$ and $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_K)$ satisfy (5.13) and (5.15). The latter linear constraint on $(\boldsymbol{\pi}, \mathbf{m})$ implies that $\mathbf{m}$ and $\boldsymbol{\pi}$ cannot be optimized independently, which would complicate the M-step of an EM-algorithm. Thus we begin with a re-parametrization of the model ensuring that the moment constraint (5.8) is automatically satisfied.

5.3.3.1 Re-parametrization of the moment constraint

The main idea behind the re-parametrization is to work with the parameter $\rho_{k,j} = \pi_k m_{k,j}$ instead of $(\pi_k, m_{k,j})$.

Namely, define a $K \times (d - d_1)$ matrix $\boldsymbol{\rho} = ({\boldsymbol{\rho}_1}^\top, \ldots, {\boldsymbol{\rho}_K}^\top)$ where $\rho_{k,j} > 0$ for $j \in \alpha_k$ and $\rho_{k,j} = 0$ otherwise. Then for all $k \in \{1, \ldots, K\}$, set

$$\begin{aligned} \pi_k &:= \sum_{j \in \alpha_k} \rho_{k,j} \\ m_{k,j} &:= \frac{\rho_{k,j}}{\pi_k}, \forall j \in \alpha_k. \end{aligned} \tag{5.20}$$

Then (5.13) and (5.15) together are equivalent to

$$\sum_{\{k : j \in \alpha_k\}} \rho_{k,j} = \frac{1}{d}, \quad \forall j \in \{d_1 + 1, \ldots, d\}. \tag{5.21}$$

In the sequel we denote respectively by $p(\tilde{\mathbf{v}}|\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda}) := p(\tilde{\mathbf{v}}|\boldsymbol{\pi}, \mathbf{m}, \boldsymbol{\nu}, \boldsymbol{\lambda})$ and $\varphi_k(\mathbf{w}|\boldsymbol{\rho}_k, \nu_k) := \varphi_k(\mathbf{w}|\mathbf{m}_k, \nu_k)$ the likelihood and the Dirichlet densities in the re-parameterized model, where $(\mathbf{m}, \boldsymbol{\pi})$ are obtained from $\boldsymbol{\rho}$ *via* (5.20).

5.3.3.2 EM algorithm

We summarize below the EM algorithm in our framework. Let $n_0 \leq n$ be the number of observations $\tilde{\mathbf{V}}_i$ such that $\|\tilde{\mathbf{V}}_i\| > r_0$. To alleviate notations, we may relabel the indices i so that these observations are $\tilde{\mathbf{V}}_{1:n_0} = (\tilde{\mathbf{V}}_1, \ldots, \tilde{\mathbf{V}}_{n_0})$. Let $\mathbf{Z}_i = (Z_{i,1}, \ldots, Z_{i,K+d_1}), i \leq n_0$ be the hidden variables associated with $\tilde{\mathbf{V}}_{1:n_0}$. In the sequel, θ denotes the set of parameters of the re-parameterized version of Model 2, that is $\theta = (\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda})$, and let Θ be the parameter space, that is the set of θ's such that constraints (5.19) and (5.21) hold. Also let $p(\tilde{\mathbf{v}}|\theta, z_k = 1)$ denote the conditional density of $\tilde{\mathbf{V}}$ given $(Z_k = 1, \theta)$. In view of the likelihood (5.18), it is given for for $k \leq K$ by

$$p(\tilde{\mathbf{v}}|z_k = 1, \theta) = r_k^{-|\alpha_k|-1} \, \varphi_k(\mathbf{w_k}|\rho_\mathbf{k}, \nu_k) \prod_{j \in \alpha_k^c} f_\varepsilon(\tilde{v}_j|\lambda_k) \tag{5.22}$$

and for $K < k \leq K + d_1$,

$$p(\tilde{\mathbf{v}}|z_k = 1, \theta) = \tilde{v}_k^{-2} \prod_{j \in \{1, \ldots, d\} \setminus k} f_\varepsilon(\tilde{v}_j|\lambda_k) \tag{5.23}$$

Algorithm 1 (EM algorithm for Model 2).

- **Input** Extreme standardized data $\tilde{\mathbf{V}}_{1:n_0}$,
- **Initialization** Choose a starting value for θ (See Remark 5.3)
- **Repeat until convergence**:
  - **E-step**: compute for $1 \leq i \leq n_0$ and $k \leq K + d_1$,

    $$\gamma_{i,k} = \mathbb{P}[Z_{i,k} = 1 \| \tilde{\mathbf{V}}_i, \theta]$$

    according to (5.25). Set $\boldsymbol{\gamma} = (\gamma_{i,k})_{i \leq n_0, k \leq K+d_1}$.
  - **M-step**: Solve the optimization problem $\max_{\theta \in \Theta} Q(\theta, \boldsymbol{\gamma})$where Q is a lower bound for the likelihood, namely

    $$Q(\theta, \boldsymbol{\gamma}) = \sum_{i=1}^{n_0} \sum_{k=1}^{K+d_1} \gamma_{i,k} \Big(\log \pi_k + \log p(\tilde{\mathbf{V}}_i | \theta, z_{i,k} = 1) \Big),$$

    with $\pi_k = \mathbb{P}(Z_{i,k} = 1 | \theta)$, *i.e.*

    $$\pi_k = \begin{cases} \sum_{\ell \in \alpha_k} \rho_{k,l} & \text{for } 1 \leq k \leq K \ , \\ d^{-1} & \text{for } K < k \leq K + d_1 \ , \end{cases} \tag{5.24}$$

    and where $p(\tilde{\mathbf{V}}_i | \theta, z_{i,k} = 1)$ is given by (5.22) and (5.23).
    Denote by $\theta^\star$ the solution and set $\theta = \theta^\star$.

Remark 5.3. In this work, the output of DAMEX is used for choosing the initial value for ρ. Namely, given $\widehat{\mathbb{M}}_2$ we compute the empirical means $\widehat{m}_{k,j} := \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{\mathbf{V}}_{i,j}$ for all j in α_k and k in $\{1, \ldots, K\}$ and we set $\widehat{\pi}_0 = \ldots = \widehat{\pi}_K = \frac{1}{K}$ so that we get the corresponding $\widehat{\rho}$ by $\widehat{\rho}_{k,j} = \pi_k m_{k,j}$. Although it is not likely to verify (5.21), we can easily project $\widehat{\rho}$ on Θ : $\widehat{\rho}_{k,j}^{init} = \frac{\widehat{\rho}_{k,j}}{d \sum_{h=1}^{K} \widehat{\rho}_{h,j}}$.

We now describe at length the E-step and the M-step.

E-step. : The $\gamma_{i,k}$'s are obtained using the Bayes formula, for $1 \leq k \leq K + d_1$,

$$\begin{aligned} \gamma_{i,k} &= p(Z_{i,k} = 1 | \tilde{\mathbf{V}}_i \theta) \\ &= \frac{\pi_k \, p(\tilde{\mathbf{V}}_i | z_{i,k} = 1, \theta)}{\sum_{\substack{1 \leq \ell \leq K+d_1 \\ \ell \neq k}} \pi_\ell \, p(\tilde{\mathbf{V}}_i | z_{i,\ell} = 1, \theta)}, \end{aligned} \tag{5.25}$$

where π_k is defined in (5.24) and $p(\tilde{\mathbf{V}}_i | Z_{i,k} = 1, \theta)$ is given by (5.22) and (5.23).

M-step. : Here optimization of $Q(\theta, \gamma)$ with respect ot $\theta = (\boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\lambda})$ is performed under constraints (5.19), (5.21). Since Q decomposes into a function of $(\boldsymbol{\rho}, \boldsymbol{\nu})$ and a

function of $\boldsymbol{\lambda}$, and since the constraints on $\boldsymbol{\rho}, \boldsymbol{\nu}$ and $\boldsymbol{\lambda}$ are independent, maximization can be performed separtely over the two blocks. Indeed, gathering terms not depending on θ into a constant C,

$$\begin{aligned} Q(\theta, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \Big[\sum_{k=1}^{K} \gamma_{i,k} \Big[& \log \pi_k \\ & \cdots + \log \varphi_k(\mathbf{W}_{i,k} | \boldsymbol{\rho}_k, \nu_k) + \sum_{l \in \alpha_k^c} \log f_\varepsilon(\tilde{V}_{i,l} | \lambda_k) \Big] \\ & \cdots + \sum_{k=K+1}^{K+d_1} \gamma_{i,k} \Big[\sum_{\ell \neq k} \log f_\varepsilon(\tilde{V}_{i,l} | \lambda_k) \Big] \Big] + C \\ & = Q_1(\boldsymbol{\rho}, \boldsymbol{\nu}) + Q_2(\boldsymbol{\lambda}) + C, \end{aligned}$$

where

$$\begin{aligned} Q_1(\boldsymbol{\rho}, \boldsymbol{\nu}) &= \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{i,k} \Big[\log \sum_{l \in \alpha_k} \rho_{\mathbf{k}l} + \log \varphi_k(\mathbf{W}_{i,k} | \rho_{\mathbf{k}}, \nu_k) \Big] \\ Q_2(\boldsymbol{\lambda}) &= \sum_{i=1}^{n} \sum_{k=1}^{K+d_1} \gamma_{i,k} \sum_{l \in \alpha_k^c} \log f_\varepsilon(\tilde{V}_{i,l} | \lambda_k) \,. \end{aligned}$$

Here we have denoted $\alpha_k = \{k - K\}$ for $K < k \leq K + d_1$, in accordance with the notations from Section 5.3.2. Notice that the dependence of Q_1 and Q_2 on $\boldsymbol{\gamma}$ is omitted for the sake of concision.

With these notations

$$\max_{\substack{\theta \text{ s.t.} \\ (5.19),\ (5.21)}} Q(\theta, \boldsymbol{\gamma}) = \max_{\substack{\boldsymbol{\rho}, \boldsymbol{\nu} \text{ s.t.} \\ (5.21), \nu_k > 0,\ k \leq K}} Q_1(\boldsymbol{\rho}, \boldsymbol{\nu}) + \max_{\substack{\boldsymbol{\lambda} \text{ s.t.} \\ \lambda_k > 0,\ 1 \leq k \leq K + d_1}} Q_2(\boldsymbol{\lambda})$$

The function Q_1 being non-concave we use the python package **mystic** (McKerns et al. (2012)) to maximize it. For our choice of translated exponential noise, $f_\varepsilon(v|\lambda_k) = \lambda_k e^{-\lambda_k(v-1)}$, $v \geq 1$, the maximizer of Q_2 has an explicit expression,

$$\lambda_k^* = \frac{|\alpha_k^c| \sum_{i=1}^{n} \gamma_{i,k}}{\sum_{i=1}^{n} \gamma_{i,k} \sum_{l \in \alpha_k^c} (\tilde{V}_{i,\ell} - 1)} \,, \qquad k \leq K + d_1.$$

Remark 5.4. Let $\boldsymbol{\gamma}^t$ and θ^t be the results of the t-th iteration of the algorithm then we conclude the iterative process if $Q(\theta^t, \boldsymbol{\gamma}^t) < Q(\theta^{t-1}, \boldsymbol{\gamma}^{t-1}) + \epsilon$, with ϵ a small threshold.

5.4 Graph-based visualization tools

In this section, we explain that, beyond the hard clustering that may be straightforwardly deduced from the computation of the likeliest values $z_1, \ldots, z_{n_0}$ for the

hidden variables given the $\tilde{\mathbf{v}}_i$'s and the parameter estimates produced by the algorithm described in subsection 5.3.3, the statistical model previously introduced defines a natural structure of undirected weighted graph on the set of observed extremes, which interpretable layouts (graph drawing) can be directly derived from, using classical solutions. Indeed, a partition (hard clustering) of the set of (standardized) anomalies/extremes $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{n_0}$ is obtained by assigning membership of each $\tilde{\mathbf{v}}_i$ in a cluster (or cone/sub-simplex) determined by the component of the estimated mixture model from which it arises with highest probability: precisely, one then considers that the abnormal observation $\tilde{\mathbf{v}}_i$ is in the cluster indexed by

$$k_i = \underset{k \in \{1, \ldots, K\}}{\arg\max}\ \gamma_{i,k}$$

and is of type α_{k_i}. However, our model-based approach brings much more information and the vector of posterior probabilities $(\gamma_{i,1}, \ldots, \gamma_{i,k})$ output by the algorithm actually defines soft membership and represent the uncertainty in whether anomaly $\tilde{\mathbf{v}}_i$ is in a certain cluster. It additionally induces a similarity measure between the anomalies: the higher the probability that two extreme values arise from the same component of the mixture model, the more similar they are considered. Hence, consider the undirected graph whose vertices, indexed by $i = 1, \ldots, n_0$, correspond to the extremal observations $\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_{n_0}$ and whose edgeweights are $w_\theta(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)$, $1 \leq i \neq j \leq n_0$, where

$$w_\theta(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) = \mathbb{P}\left(\mathbf{Z}_i = \mathbf{Z}_j \mid \tilde{\mathbf{V}}_i = \tilde{\mathbf{v}}_i,\ \tilde{\mathbf{V}}_j = \tilde{\mathbf{v}}_j,\ \theta\right) = \sum_{k=1}^{K} \gamma_{i,k}\gamma_{j,k}.$$

Graph visualization techniques (see *e.g.* Hu and Shi (2015)), possibly combined with (spectral) graph clustering methods (see *e.g.* Schaeffer (2007)) when the number n_0 of anomalies to be analyzed is large, can then be used to produce informative layouts. Discussing the merits and limitations of the wide variety of approaches documented in the literature in this purpose is beyond the scope of this paper. It is the goal of the next section to simply illustrate the usefulness of the weighted graph representation of the set of anomalies proposed above, when applying to it state-of-the-art graph-mining tools.

5.5 Illustrative experiments

5.5.1 Experiments on simulated data

To assess the performance of the proposed estimator of the dependence structure and of the EM algorithm, we generate synthetic data according to Model 2. The dimensionality is fixed to $d = 100$ and the mixture components, that is the elements of $\mathbb{M} = \{\alpha_1, \ldots, \alpha_K\} \cup \mathbb{E}$ are randomly chosen in the power Set of $\{1, \ldots, d\}$, with $K = 50$. The coefficients of the matrix ρ which determines the weights and centers through (5.20) is also randomly chosen, then its columns are normalized so that the

moment constraint (5.21) is satisfied. Finally we fix $\nu_k = 20, 1 \leq k \leq K$ and $\lambda_k, 1 \leq k \leq K + d_1$ are respectively set to $(1., 0.75, 0.5, 0.25, 0.1)$ in the different experiments to illustrate different levels of noise. Then each point $\tilde{\mathbf{V}}_i = R_i \mathbf{W}_i + \boldsymbol{\varepsilon}_i, i \leq n$ is generated with probability $\pi_k, k \in \{1, \ldots, K\}$ according to the mixture component k $(k \leq K)$, that is

$$\begin{aligned} R_i &\sim Pareto(1) | \{R_i > r_0\} \\ \mathbf{W}_i &\sim \Phi_k \\ \varepsilon_{i,j} &\sim 1 + \mathcal{E}xp(\lambda_k), j \in \{1, \ldots, d\} \setminus \alpha_k, \end{aligned}$$

and with probability $\frac{1}{d}$ according to component $k \in \{K, \ldots, K + d_1\}$ according to:

$$\begin{aligned} R_i &\sim Pareto(1) | \{R_i > r_0\} \\ \mathbf{W}_i &= 1 \\ \varepsilon_{i,j} &\sim 1 + \mathcal{E}xp(\lambda_k), j \in \{1, \ldots, d\} \setminus \{k\}. \end{aligned}$$

The threshold r_0 above which points are considered as extreme is fixed to 100.

On this toy example, the pre-processing step consisting in applying DAMEX for recovering $\mathbb{M}$ produces an exact estimate, so that $\hat{\mathbb{M}} = \mathbb{M}$. Then the procedure described in Algorithm 1 is applied.

Tables 5.1 and 5.2 show the average absolute errors for the estimates $\widehat{\rho}$, $\widehat{\nu}$ and $\widehat{\lambda}$ on 50 datasets of the n_0 generated extreme points, for $n_0 = 1e + 3, 2e + 3$, namely

$$\begin{aligned} \mathrm{err}(\widehat{\rho}) &= \frac{1}{50 \cdot K \cdot d} \sum_{l=1}^{50} \sum_{k=1}^{K} \sum_{j=1}^{d} |\widehat{\rho}_{k,j} - \rho_{k,j}| \\ \mathrm{err}(\widehat{\nu}) &= \frac{1}{50 \cdot K} \sum_{l=1}^{50} \sum_{k=1}^{K} |\widehat{\nu}_k - \nu_k| \\ \mathrm{err}(\widehat{\lambda}) &= \frac{1}{50 \cdot (K + d_1)} \sum_{l=1}^{50} \sum_{k=1}^{K+d_1} |\widehat{\lambda}_k - \lambda_k| \end{aligned}$$

On this toy example, estimation of the means and weights, as well as the noises parameters are almost exact. The estimator for the ν_k's is not so precise, but as shown next, this drawback does not jeopardize clusters identification.

Table 5.1: Average error on the model parameters, $n_0 = 1e3$

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$err(\widehat{\rho})$	1.39e-5	1.37e-5	1.57e-5	1.22e-5	2.11e-5
$err(\widehat{\nu})$	5.53	5.81	6.28	6.41	9.06
$err(\widehat{\lambda})$	2.65e-2	2.04e-2	1.19e-2	5.97e-3	3.66e-3

Table 5.2: Average error on the model parameters, $n_0 = 2e3$

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$err(\widehat{\rho})$	9.98e-6	1.12e-5	1.06e-5	1.62e-5	1.64e-5
$err(\widehat{\nu})$	3.23	4.13	4.08	4.29	5.05
$err(\widehat{\lambda})$	1.62e-2	1.2e-2	8.11e-3	4.28e-3	3.11e-3

The performance in terms of cluster identification is measured as follows: for each point $\tilde{\boldsymbol{v}}_i$, we compare the the true label $y_i \in \{1, \ldots, K + d_1\}$ with the label obtained *via* assignment to the highest probable component, that is $\hat{y}_i = \arg\max_{k \in \{1, \ldots, K+d_1\}} \gamma_{i,k}$. Table 5.3 shows the average number of labeling errors for different values of n_0 and λ_k.

Table 5.3: Average number of labeling errors

	$\lambda_k = 1.$	$\lambda_k = 0.75$	$\lambda_k = 0.5$	$\lambda_k = 0.25$	$\lambda_k = 0.1$
$n_0 = 1e3$	0.	0.	0.	0.6	264.4
$n_0 = 2e3$	0.	0.	0.4	1.8	537.8

Figure 5.4 illustrates the potential of the proposed approach in terms of visual display of anomalies. A testing set of size 100 is simulated as above, and the corresponding matrix $\widehat{\boldsymbol{\gamma}}$ is computed according to (5.25) with θ taken as the output of the training step (*i.e.* Algorithm 1 run with the training dataset of $n_0 = 2e3$ points). Finally an adjacency matrix $w_{\widehat{\theta}}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j)$ is obtained as detailed in Section 5.4, on which we apply the spectral clustering in order to group the points according to the similarities given by w. Graph visualization of w is performed using the python package 'Networkx' Hagber et al. (2008), that provides a spring layout of the graph according to the Fruchtermen-Reingold algorithm Fruchterman and Reingold (1991). A hard thresholding on the edges in w is applied in order to improve readability: edges (i, j) such that $w_{\widehat{\theta}}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) < \epsilon$ with ϵ a small threshold are removed. Each cluster obtained by spectral clustering is idendified with a specific color for the corresponding nodes.

5.5.2 Flights clustering and visualization

At Airbus, this algorithm is currently being tested to assist the building of health indicators in the context of condition based maintenance. Health indicators are then used for assessing the current state of some system and also for forecasting the future states and future degradation (ex : bleed, power systems, engine, APU, ...). Airlines are then informed that some systems should be maintained so as to avoid any operational event in a given time horizon (ex : such as delays, operational interruptions *etc* ...).

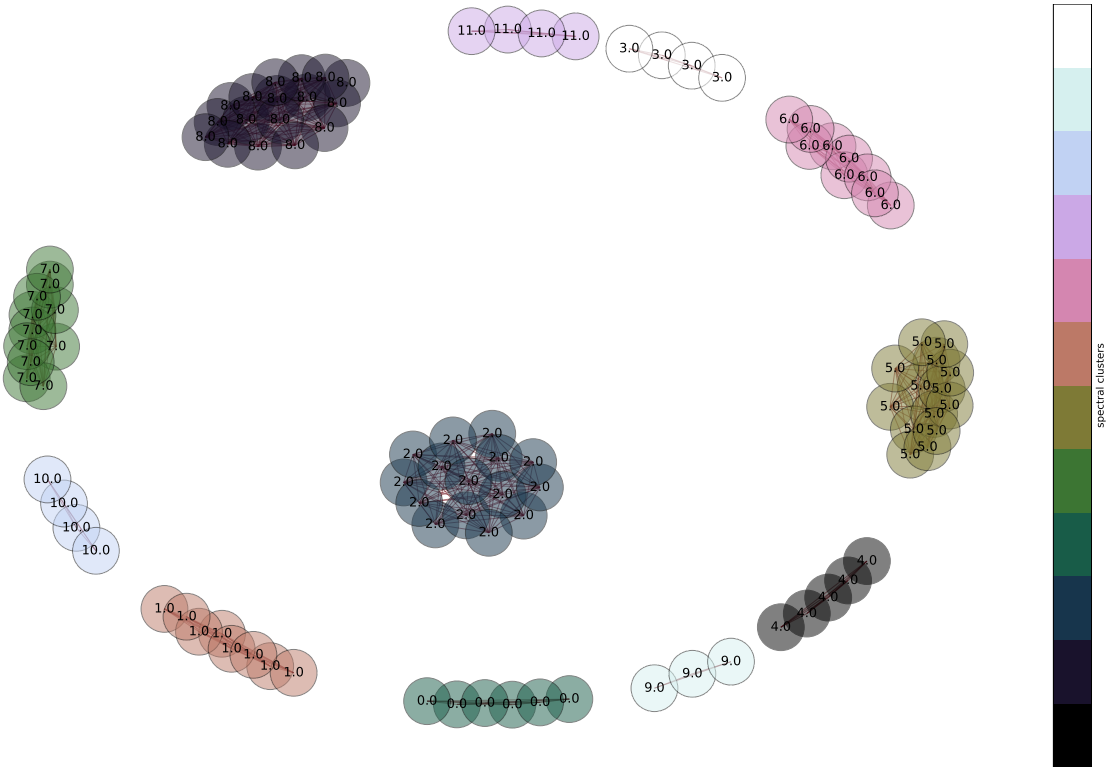


Figure 5.4: Spectral clustering visualization of a synthetic anomaly test data of size 100 with $d = 20$ and $|\mathbb{M}| = 12$
Each point is represented as a numbered node. The number is the true label, while the colors indicate the clusters obtained by spectral clustering. The spatial arrangement of the nodes is obtained by the Fruchtermen-Reingold algorithm.

The building of an health indicator can be basically summarized as follow :

1. Collect health and usage data from various aircrafts (generally one has to consider some similar ones).

2. Collect some operational events happening on these aircrafts due to some aircraft systems errors (ex : operational interruption, delays, etc ...).

3. Identify anomalies in the the health and usage data.

4. Identify some dependencies between health and usage data anomalies and operational events (thanks to statistical tests but also thanks to human expertise).

5. If some dependencies are well identified, then one can quite easily build a health indicator.

One of the tricky part is the identification and the understanding of the anomalies. Indeed different operational events are often recorded corresponding to the

degradation of different systems. Usually, a first step of anomalies is identified followed by a clustering of these anomalies for helping in the interpretation. One benefit of the approach proposed in this paper is that it directly provides some similarity measures associated to the anomalies. This strategy is illustrated by Figure 5.5. The proposed method was applied on a dataset of 18553 flights, each of which is characterized by 82 parameters. In order to differentiate between anomalies corresponding to unusual large and small values, each feature is duplicated and each copy of a given feature is defined as the positive (*resp.* negative) value of the parameter above (*resp.* below) its mean value.

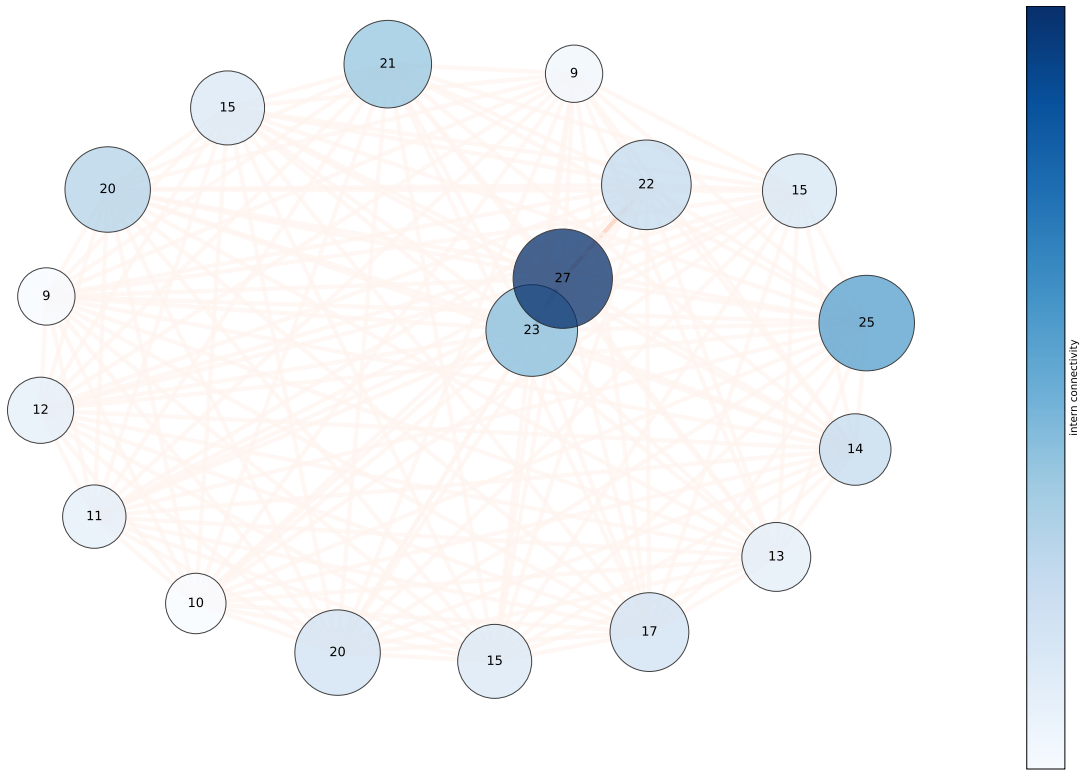


Figure 5.5: Spectral clustering visualization of flights anomalies with agglomerated Nodes
The agglomerated visualization is obtained *via* spectral clustering: each node represents a cluster. Levels of blue show the intern connectivity between the original nodes so that darker clusters have strongly connected elements. The size of each node is proportional to the number of points forming the cluster.

Figure 5.5 and figure 5.6 display the clustering of 300 'extremal' flights into 18 groups, showing on the one hand the output of the spectral clustering applied to the similarity graph $w_{\hat{\theta}}$ and on the other hand the underlying graph obtained with the same procedure as in Figure 5.4.

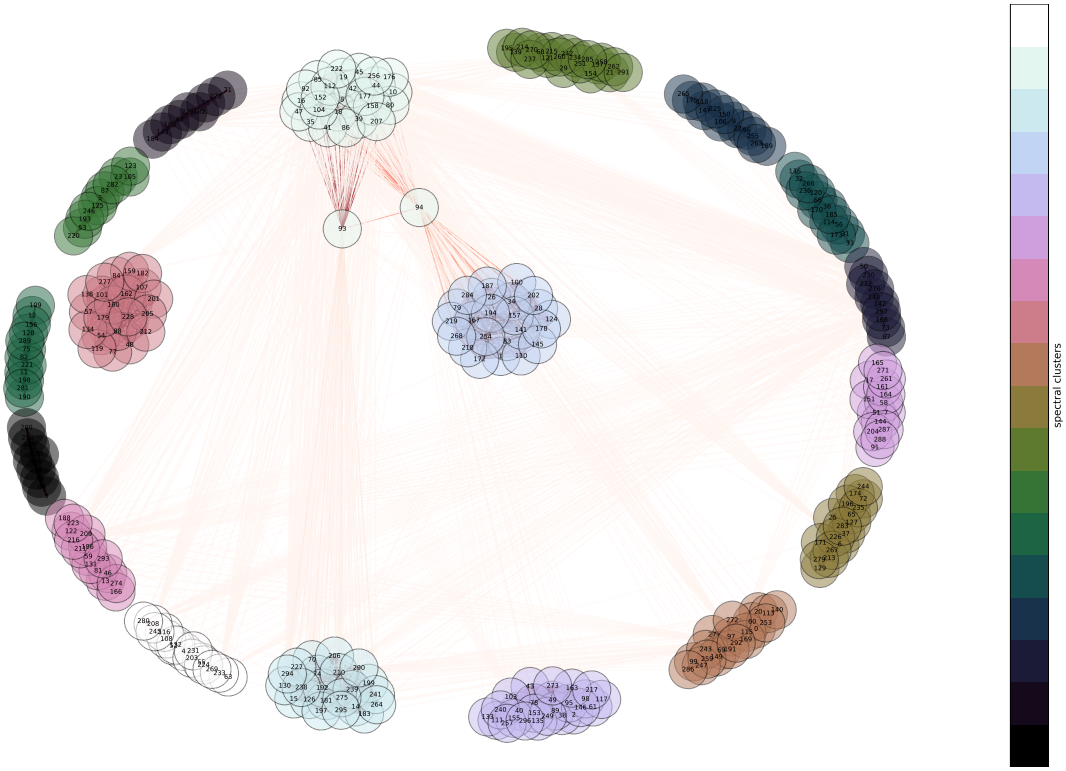


Figure 5.6: Spectral clustering visualization of flights anomalies
The number of each node is the (anonymized) flight identification number. The nodes colors and the spatial arrangement are obtained similarly to Figure 5.4.

5.6 Conclusion

Because extreme values (viewed as anomalies here) cannot be summarized by simple meaningful summary statistics such as local means or modes/centroids, clustering and dimensionality reduction techniques for such abnormal observations must be of very different nature than those developed for analyzing data lying in high probability regions. This paper is a first attempt to design a methodology fully dedicated to the clustering and visualization of anomalies, by means of a statistical mixture model for multivariate extremes that can be interpreted as a noisy version of the angular measure, which distribution on the unit sphere exhaustively describes the limit dependence structure of the extremes (up to a standardization). Localization of the mixture components is understood here as closeness of the data arising from them to a specific sub-simplex forming the support of the angular measure. Considering synthetic and real datasets, we also provided empirical evidence of the usefulness of (graph-based) techniques that can be straightforwardly implemented from the framework we developed.

Bibliography

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (2005). Automatic subspace clustering of high dimensional data. *DMKD*, 11(1):5–33. 59

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. 18, 45, 59, 63, 74

Bacro, J.-N. and Toulemonde, G. (2013). Measuring and modelling multivariate and spatial dependence of extremes. *Journal de la Société Française de Statistique*, 154(2):139–155. 77

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons. 61

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley. 103

Boldi, M.-O. and Davison, A. (2007a). A mixture model for multivariate extremes. *JRSS-B*, 69(2):217–229. 62

Boldi, M.-O. and Davison, A. C. (2007b). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229. 9, 37

Bron, C. and Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577. 67

Bücher, A. and Dette, H. (2013). Multiplier bootstrap of tail copulas with applications. *Bernoulli*, 19(5A):1655–1687. 89

Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383–418. 10, 38, 58, 59, 62

Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *ECML-PKDD 2016, workshop NFmcp2016*. 17, 74, 75, 76, 77, 85, 89, 96

Clifton, D. A., Hugueny, S., and Tarassenko, L. (2011a). Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389. 57

Clifton, D. A., Hugueny, S., and Tarassenko, L. (2011b). Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389. 98

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag, London. 65, 100

Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365. 75

Coles, S. and Tawn, J. (1991). Modeling extreme multivariate events. *JRSS-B*, 53:377–392. 9, 11, 25, 37, 39, 51, 62

Cooley, D., Davis, R., and Naveau, P. (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *JMVA*, 101(9):2103–2117. 62

D. Gorinevsky, B. Matthews, R. M. (2012). Aircraft anomaly detection using performance models trained on fleet data. In *Proceedings of the 2012 Conference on Intelligent Data Understanding*. 97

de Haan, L. and Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 40(4):317–337. 8, 35

De Haan, L. and Zhou, C. (2011). Extreme residual dependence for random vectors and processes. *Advances in Applied Probability*, 43(01):217–242. 22, 49, 76

De Haan, L. d. (1970). *On regular variation and its application to the weak convergence of sample extremes*. Amsterdam : Mathematisch Centrum. Bibliography: p. 123-124. 3, 31

Draisma, G., Drees, H., Ferreira, A., and de Haan, L. (2001). Tail dependence in independence. *Eurandom preprint*. 23, 49, 74, 83, 84

Draisma, G., Dress, H., Ferreira, A., and De Haan, L. (2004). Bivariate tail estimation: dependence in asymptotic independence. *Bernoulli*, pages 251–280. 23, 49, 74, 76, 77, 81, 83, 84, 89, 94

Drees, H. (1998a). A general class of estimators of the extreme value index. *Journal of Statistical Planning and Inference*, 66(1):95–112. 84, 94, 95

Drees, H. (1998b). On smooth statistical tail functionals. *Scandinavian Journal of Statistics*, 25(1):187–210. 84

Eastoe, E. F. and Tawn, J. A. (2012). Modelling the distribution of the cluster maxima of exceedances of subasymptotic thresholds. *Biometrika*, 99(1). 22, 49, 76

Einmahl, J. H. (1997). Poisson and gaussian approximation of weighted local empirical processes. *Stochastic processes and their applications*, 70(1):31–58. 78

Einmahl, J. H., Krajina, A., Segers, J., et al. (2012). An m-estimator for tail dependence in arbitrary dimensions. *The Annals of Stat.*, 40(3):1764–1793. 78, 79, 89

Einmahl, J. H. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Stat.*, pages 2953–2989. 61

Fougeres, A.-L., De Haan, L., Mercadier, C., et al. (2015). Bias correction in multivariate extremes. *The Annals of Stat.*, 43(2):903–934. 61

Fougeres, A.-L., Mercadier, C., and Nolan, J. P. (2013). Dense classes of multivariate extreme value distributions. *Journal of Multivariate Analysis*, 116:109–129. 10, 37, 62

Fruchterman, T. and Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164. 113

Giuntoli, I., Renard, B., Vidal, J.-P., and Bard, A. (2013). Low flows in france and their relationship to large-scale climate indices. *J. of Hydro.*, 482:105–118. 58

Goix, N., Sabourin, A., and Clémençon, S. (2016a). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS'16*. 10, 14, 38, 42, 97, 98, 99, 101, 102

Goix, N., Sabourin, A., and Clémençon, S. (2015a). Sparsity in multivariate extremes with applications to anomaly detection. *arXiv preprint arXiv:1507.05899*. 10, 38, 57, 58, 59, 62

Goix, N., Sabourin, A., and Clémençon, S. (2016b). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Proceedings of the 19th AISTAT conference*, pages 287–295. vii, 57, 58, 59, 62, 63, 67, 69, 73, 85

Goix, N., Sabourin, A., and Clémençon, S. (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31. 73, 77, 78, 98, 99, 102

Goix, N., Sabourin, A., and Clémençon, S. (2015b). Learning the dependence structure of rare events: a non-asymptotic study. In *Proceedings of the 28th Conference on Learning Theory*. 61

Gorban, A., Kégl, B., C. Wunsch, D., and Zinovyev, A. (2008). *Principal Manifolds for Data Visualisation and Dimension Reduction*. LNCSE 58. Springer. 98

Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric bayesian inference on bivariate extremes. *JRSS-B*, 73(3):377–406. 10, 37, 62

Gunopulos, D., Khardon, R., Mannila, H., Saluja, S., Toivonen, H., and Sharma, R. S. (2003). Discovering all most specific sentences. *ACM Trans. Database Syst.*, 28(2):140–174. 59

Hagber, A., Schult, D., and Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA. 113

Hu, Y. and Shi, L. (2015). Visualizing large graphs. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2):115–136. 99, 111

Huser, R., Davison, A. C., and Genton, M. G. (2016). Likelihood estimators for multivariate extremes. *Extremes*, 19(1):79–103. 9, 37

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8):1287–1304. 57

Kriegel, H.-P., Kröger, P., Schubert, E., and Zimek, A. (2008). A general framework for increasing the robustness of pca-based correlation clustering algorithms. In Ludäscher, B. and Mamoulis, N., editors, *Scientific and Statistical Database Management*, pages 418–435, Berlin, Heidelberg. Springer Berlin Heidelberg. 98

Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187. 22, 49, 74, 76

Lee, H. and Roberts, S. (2008a). On-line novelty detection using the kalman filter and extreme value theory. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. 98

Lee, H.-j. and Roberts, S. J. (2008b). On-line novelty detection using the kalman filter and extreme value theory. In *Pattern Recognition. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE. 57

Liu, F., Ting, K., and Zhou, Z. (2008). Isolation Forest. In *ICDM*. 98

McKerns, M., Strand, L., Sullivan, T., Fang, A., and Aivazis, M. (2012). Building a framework for predictive science. *arXiv preprint arXiv:1202.1056*. 110

Naik, E. G., editor (2017). *Advances in Principal Component Analysis*. Research and Development. Springer. 98

Peng, L. (1999). Estimation of the coefficient of tail dependence in bivariate extremes. *Statistics & Probability Letters*, 43(4):399–409. 23, 49, 74, 76, 80, 81, 82

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, pages 119–131. 82

Qi, Y. (1997). Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics. *Acta Mathematicae Applicatae Sinica (English series)*, 13(2):167–175. 61, 78

Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B*, 71(1):219–241. 74, 76

Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering. 5, 6, 7, 9, 33, 34, 36, 100

Resnick, S. I. (2007a). *Heavy-Tail Phenomena*. Springer Series in Operations Research and Financial Engineering. Springer, New York. 75

Resnick, S. I. (2007b). *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media. 100, 103

Resnick, S. I. (2013). *Extreme values, regular variation and point processes*. Springer. 6, 34, 61, 75

Roberts, S. (2000). Extreme value statistics for novelty detection in biomedical signal processing. In *Advances in Medical Signal and Information Processing, 2000. First International Conference on (IEE Conf. Publ. No. 476)*, pages 166–172. 98

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J. 78

Sabourin, A. and Naveau, P. (2014). Bayesian dirichlet mixture model for multivariate extremes: A re-parametrization. *CSDA*, 71:542–567. 62

Sabourin, A., Naveau, P., and Fougeres, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, 16(3):325. 9, 37, 62

Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, 1(1):27 – 64. 99, 111

Schlather, M. and Tawn, J. A. (2003). A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika*, 90(1):139–156. 75

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471. 98

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM. 95

Steinwart, I., Hush, D., and Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6:211–232. 98

Stephenson, A. (2003). Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59. 68, 86

Stephenson, A. and Tawn, J. (2005). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, 92(1):213–227. 9, 37

T. Fawcett, F. P. (1997). Adaptive fraud detection. *Data-Mining and Knowledge Discovery*, 1:291–316. 97

Tawn, J. (1990a). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253. 14, 42

Tawn, J. A. (1990b). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253. 68, 86, 89

Tomita, E., Tanaka, A., and Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, 363(1):28–42. 67

Tressou, J. (2008). Bayesian nonparametrics for heavy tailed distribution. application to food risk assessment. *Bayesian Anal.*, 3(2):367–391. 98

van der Vaart, A. W. (1998). *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge. 79, 91, 92, 95

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. 79

Viswanathan, K., Choudur, L., Talwar, V., Wang, C., Macdonald, G., and Satterfield, W. (2012). Ranking anomalies in data centers. In R.D.James, editor, *Network Operations and System Management*, pages 79–87. IEEE. 97

Wadsworth, J. (2015). On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions. *Biometrika*, 102(3):705–711. This is a pre-copy-editing, author-produced PDF of an article accepted for publication in Biometrika following peer review. The definitive publisher-authenticated version Jennifer L. Wadsworth On the occurrence times of componentwise maxima and bias in likelihood inference for multivariate max-stable distributions Biometrika (2015) 102 (3): 705-711 first published online June 25, 2015 doi:10.1093/biomet/asv029 is available online at: http://biomet.oxfordjournals.org/content/102/3/705. 9, 37

Xie, Y. and Philip, S. Y. (2010). Max-clique: a top-down graph-based approach to frequent pattern mining. In *2010 IEEE Int. Conf. Data Mining*, pages 1139–1144. IEEE. 67