# High-dimensional Learning for Extremes

Nicolas Meyer

**École doctorale de sciences mathématiques de Paris centre**
Laboratoire de Probabilités, Statistique et Modélisation - LPSM
Sorbonne Université

# Thèse de doctorat

Discipline : Mathématiques

Spécialité : Statistique

présentée par

**Nicolas Meyer**

---

# High-dimensional Learning for Extremes

---

dirigée par Olivier Wintenberger (Sorbonne Université, LPSM)

Laboratoire de Probabilités, Statistique et Modélisation
UMR 8001
Boîte courrier 158
4 place Jussieu
75 252 Paris Cedex 05

Sorbonne Université
École doctorale de Sciences mathématiques de Paris Centre
Boîte courrier 290
4 place Jussieu
75 252 Paris Cedex 05

# Contents

# List of Figures

# List of Tables

# Résumé détaillé

Ce texte rassemble les travaux effectués au cours de mes trois années de thèse. Le problème auquel je me suis intéressé concerne l'étude et la modélisation des données extrêmes en grande dimension. Je résume ici de façon succinte les résultats obtenus.

## Variation régulière

Le cadre général de cette thèse est le suivant : on se donne un échantillon de vecteurs aléatoires indépendants et identiquement distribués (i.i.d.) $\mathbf{X}_1, \ldots, \mathbf{X}_n$ et on souhaite étudier la structure de dépendence des valeurs extrêmes de cet échantillon. L'idée est alors de considérer un vecteur aléatoire $\mathbf{X}$ de même loi que l'échantillon et de s'intéresser au comportement joint des queues de dsitributions des marginales de $\mathbf{X}$. Dans ce contexte, de nombreuses questions apparaissent. Quelle est la probabilité que deux marginales sont simultanément grandes ? Est-il possible que toutes les coordonnées soient extrêmes en même temps ? À l'inverse, est-il possible que le comportement extrême de $\mathbf{X}$ ne soit dû qu'à une seule des ses marginales ? Plus simplement, parmi ces $n$ vecteurs de l'échantillon quels seront ceux que l'on considèrera comme extrêmes ?

Une hypothèse classique en Théorie des Valeurs Extrêmes est de supposer que les vecteurs considérés sont à *variation régulière*, c'est-à-dire qu'il existe une suite strictement positive $(a_n)$ vérifiant $a_n \to \infty$ quand $n \to \infty$ et une mesure de Radon non-nulle $\mu$ définie sur la tribu des boréliens de $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ telles que

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,, \tag{0.0.1}$$

voir Resnick (1987) et Resnick (2007). La mesure limite $\mu$ est appelée *mesure exposant*. Elle rassemble une grande partie de l'information concernant le comportement de la queue de $\mathbf{X}$. En effet, les directions sur lesquelles cette mesure met de la masse correspondent aux directions dans lesquelles des événements extrêmes sont susceptibles d'apparaître.

On peut montrer que la mesure exposant est homogène au sens suivant: il existe $\alpha > 0$ tel que

$$\mu(tA) = t^{-\alpha}\mu(A)\,.$$

pour tout $t > 0$ et tout borélien $A \subset \mathbb{R}_+^d \setminus \{\mathbf{0}\}$. L'indice $\alpha > 0$ est appelé l'*indice de queue*. L'homogénéité de la mesure exposant a des nombreuses conséquences. En particulier, elle permet de

décomposer cette mesure en une partie radiale et une partie angulaire. En effet, un vecteur aléatoire $\mathbf{X} \in \mathbb{R}_+^d$ est à variation régulière si et seulement s'il existe un vecteur aléatoire $\mathbf{\Theta}$ sur la sphère unité positive $\mathbb{S}_+^{d-1}$ et une variable aléatoire $Y$ de loi de Pareto de paramètre $\alpha$ tels que

$$\mathbb{P}((t^{-1}|\mathbf{X}|, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > t) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,. \tag{0.0.2}$$

Dans ce cas, le vecteur $\mathbf{\Theta}$ est indépendant de la variable alátoire $Y$ et est appelé *vecteur spectral*. Sa loi $S(\cdot) = \mathbb{P}(\mathbf{\Theta} \in \cdot)$ est appelée *mesure spectrale*. L'Equation (0.0.2) permet de décomposer l'étude des extrêmes multivariés en deux étapes. La première consiste à inférer l'intensité des extrêmes en estimant l'indice $\alpha > 0$ : plus $\alpha$ est petit, plus cette intensité est grande. D'un point de vue théorique, on est ramené à l'étude de la norme du vecteur $\mathbf{X}$ et donc à la théorie des extrêmes univariés. Ce cadre a déjà été longuement étudié dans les ouvrages de Beirlant et al. (2006), de Haan and Ferreira (2006), Embrechts et al. (2013) ou Coles (2001). De nombreuses méthodes pour estimer $\alpha$ ont été proposées, par exemple par Hill (1975) ou Pickands (1975).

Le coeur du problème réside donc dans l'étude de la mesure spectrale $S$. Cette dernière rassemble l'information sur la dépendance et la localisation des événements extrêmes : les parties de la sphère unité sur lesquelles $S$ met de la masse correspondent aux directions dans lesquelles de tels événements se produisent. Ainsi, la connaissance du support de cette mesure s'avère être un point central de l'étude des extrêmes multivariés. À cet égard, l'Equation (0.0.2) implique en particulier que

$$\mathbb{P}(\mathbf{X}/|\mathbf{X}| \in \cdot \mid |\mathbf{X}| > t) \xrightarrow{w} \mathbb{P}(\mathbf{\Theta} \in \cdot)\,, \quad t \to \infty\,. \tag{0.0.3}$$

À première vue, cette convergence permet d'estimer le vecteur spectral à partir des données $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Néanmoins, il est fréquent que $\mathbf{\Theta}$ se concentre sur des parties de la sphère de dimension $d' \ll d$. On dit alors que cette mesure est parcimonieuse. Cela signifie qu'avec grande probabilité le vecteur spectral a plusieurs coordonnées nulles. À l'inverse, le vecteur $\mathbf{X}$ modélise des données réelles donc sa loi ne met pas de masse sur de tels sous-espaces. L'Equation (0.0.3) tombe alors en défaut. Le phénomène de parcimonie a d'autant plus lieu en grande dimension. Un des objectifs de la théorie des extrêmes multivariés est alors de modéliser au mieux ce phénomène.

## Parcimonie dans les extrêmes

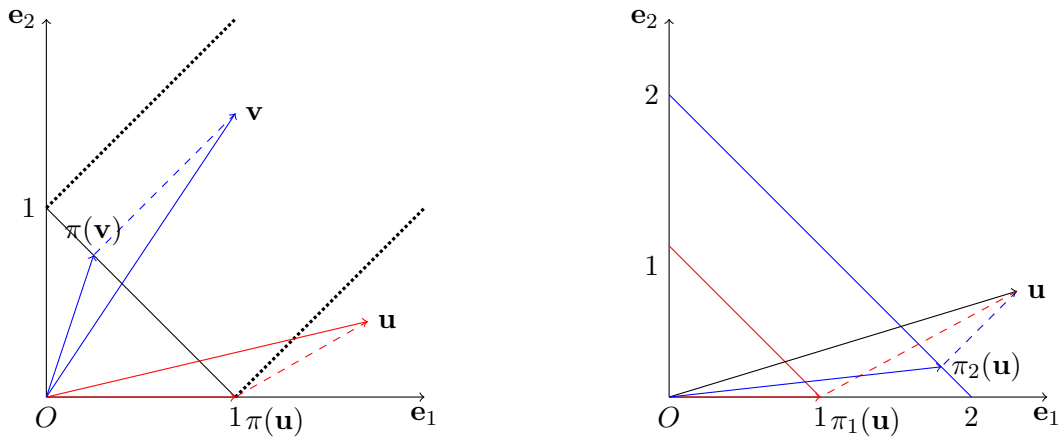
Un des obstacles majeurs à l'utilisation de la convergence (0.0.3) est la différence de support entre $\mathbf{X}$ et $\mathbf{\Theta}$. L'idée est alors d'adapter l'approche classique basée sur le concept de variation régulière en modifiant la composante angulaire $\mathbf{X}/|\mathbf{X}|$ dans l'Equation (0.0.2). Dans le but d'introduire de la parcimonie dans le vecteur initial $\mathbf{X}$, on utilise la projection euclidienne sur une sphère $\ell^1$ (Duchi et al. (2008), Gafni and Bertsekas (1984), Bertsekas (1999)). Pour $z > 0$, on consière la sphère positive $\mathbb{S}_+^{d-1} = \{\mathbf{x} \in \mathbb{R}_+^d,\ |\mathbf{x}|_1 = z\}$. La projection euclidienne sur cette sphère est alors définie par

la fonction

$$\begin{array}{rccc} \pi_z & : & \mathbb{R}_+^d & \to & \mathbb{S}_+^{d-1}(z) \\ & & \mathbf{v} & \mapsto & \mathbf{w} = (\mathbf{v} - \lambda_{\mathbf{v},z})_+ \,. \end{array}$$

où $\lambda_{\mathbf{v},z} \in \mathbb{R}$ est l'unique constante vérifiant la relation $\sum_{1 \le i \le d} (v_i - \lambda_{\mathbf{v},z})_+ = z$.

Une illustration de la projection en dimension 2 est donnée en Figure 1. Des nombreux algorithmes calculant cette projection ont été développés par Condat (2016) and Duchi et al. (2008). Dans ce dernier article, les auteurs proposent un algorithme dont la complexité est linéaire en moyenne. Ceci justifie l'utilisation de cette projection pour l'étude des extrêmes en grande dimension. Notons que le choix du seuil $z$ a des conséquences sur la parcimonie du vecteur projeté (voir Figure 1b). Ce choix fait l'objet d'une étude particulière dans le Chapitre 3.




(a) Deux vecteurs et leur image par $\pi$. Les lignes pointillées séparent l'espace en fonction de l'ensemble image de la projection : $\mathbf{e}_1$, $\mathbf{e}_2$, ou l'intérieur du simplexe.

(b) Influence du choix de $z$. L'image du vecteur $\mathbf{u}$ est $\pi_1(\mathbf{u}) = \mathbf{e}_1$ avec $z = 1$ alors qu'elle devient $\pi_2(\mathbf{u}) > 0$ avec $z = 2$. La parcimonie augmente quand le seuil diminue.

Figure 1: La projection euclidienne sur le simplexe, exemple et influence du choix du seuil.

La susbtitution du vecteur normalisé $\mathbf{X}/|\mathbf{X}|$ par le vecteur $\pi(\mathbf{X}/t)$ donne lieu à la définition suivante. On dit qu'un vecteur aléatoire $\mathbf{X} \in \mathbb{R}_+^d$ est à variation régulière parcimonieuse s'il vérifie la convergence :

$$\mathbb{P}\left( \left( \frac{|\mathbf{X}|_1}{t}, \pi\left(\frac{\mathbf{X}}{t}\right) \right) \in \cdot \;\middle|\; |\mathbf{X}|_1 > t \right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{0.0.4}$$

où $Y$ est une variable aléatoire de loi de Pareto de paramètre $\alpha > 0$ et $\mathbf{Z}$ un vecteur aléatoire sur le simplexe positif $\{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = 1\}$ (voir Section 2.4 pour plus de détails).

Contrairement à la variation régulière standard, les composantes radiale et angulaire de la limite dans (0.0.4) ne sont pas indépendantes. Néanmoins, leur structure de dépendance est déterminée dans la Proposition 2.4.1. Le résultat principal du Chapitre 2 est alors le Théorème 2.4.1 qui, sous des hypothèses assez faibles, établit l'équivalence entre variation régulière et variation régulière

parcimonieuse. Ce théorème, ainsi que la Proposition 2.3.2, met également en évidence les relations entre la loi de $\mathbf{Z}$ et celle de $\mathbf{\Theta}$.

La notion de variation régulière parcimonieuse permet, via la projection euclidienne sur le simplexe, de travailler avec un vecteur $\pi(\mathbf{X}/t)$ qui comporte un grand nombre de coordonnées nulles. Cette réduction de la dimension rend alors plus réalisable l'étude de la dépendance des composantes extrêmes de $\mathbf{X}$ via l'estimation du support de $\mathbf{Z}$.

## Estimation de la dépendance

Partant de l'Equation (0.0.4), on se propose d'estimer la dépendance des extrêmes multivariés de $\mathbf{X}$ en s'appuyant sur le vecteur angulaire $\mathbf{Z}$. Pour cela, on partitionne la simplexe en fonction de la nullité des coordonnées d'un vecteur de cet ensemble. Pour $\beta \subset \{1, \ldots, d\}$, $\beta \neq \emptyset$, on définit le sous-ensemble $C_\beta$ par

$$C_\beta = \left\{ \mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = 1,\, x_i > 0 \text{ pour } i \in \beta,\, x_i = 0 \text{ pour } i \notin \beta \right\}.$$

Ces sous-ensembles forment une partition du simplexe. Cette approche rejoint celle développée par Chautru (2015) qui utilise une partition de la sphère dans un contexte d'analyse en composante principale. De leurs côtés,Simpson et al. (2019) s'appuient sur la même partition pour étudier la dépendance sous des hypothèses de variation regulière cachée. Enfin, Goix et al. (2017) utilisent une partition similaire pour estimer le support de la mesure exposant.

L'interprétation des $C_\beta$ en terme de valeurs extrêmes est la suivante : la mesure spectrale met de la masse sur $C_\beta$ si et seulement si des événements extrêmes apparaissent dans la direction $\beta$, autrement dit, si et seulement s'il est probable que les coordonnées dans $\beta$ soient simultanément grandes alors que celles dans $\beta^c$ ne le soient pas. Néanmoins, la convergence faible de l'Equation (0.0.2) tombe en défaut dès lors qu'un $C_\beta$ vérifie $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ pour $\beta \neq \{1, \ldots, d\}$. Le vecteur spectral $\mathbf{\Theta}$ n'apparaît donc pas comme le bon modèle dans ce cadre.

À l'inverse, on établit dans le Chapitre 2 plusieurs résultats concernant le comportement du vecteur angulaire $\mathbf{Z}$ sur les sous-espaces $C_\beta$, notamment la convergence de $\pi(\mathbf{X}/t) \mid |\mathbf{X}|_1 > t$ vers $\mathbf{Z}$ sur ces ensembles. L'interprétation des probabilités $\mathbb{P}(\mathbf{Z} \in C_\beta)$ en terme de valeurs extrêmes est assez similaire à celle concernant $\mathbf{\Theta}$ (voir Proposition 2.3.3 et Théorème 2.3.1). Par ailleurs, les propriétés de la projection euclidienne sur le simplexe permettent d'étudier l'importance relative d'une coordonnée (ou d'un groupe de coordonnées) par rapport aux autres via l'étude du support de $\mathbf{Z}$ (voir Section 2.3.1). Le but des Chapitres 2 et 3 est alors de s'intéresser à l'estimation de la masse mise par la loi de $\mathbf{Z}$ sur les sous-espaces $C_\beta$.

Un cas fréquemment étudié en théorie des valeurs extrêmes est celui où la masse de la mesure spectrale se concentre sur les axes $\mathbf{e}_k = C_{\{k\}}$, pour $k = 1, \ldots, d$. On parle alors d'asymptotique indépendance (de Haan and Ferreira (2006), Section 6.2). La modélisation et l'étude de vecteurs asymptotiquement indépendants a fait l'objet de nombreux articles (voir e.g. Ledford and Tawn

(1996), Heffernan and Tawn (2004), Fougères and Soulier (2010)). Les résultats numériques proposés en fin de Chapitre 2 illustrent l'avantage de l'utilisation du vecteur $\mathbf{Z}$ pour identifier de tels vecteurs.

Dans un contexte statistique, le Chapitre 3 rassemble différents résultats concernant l'étude des vecteurs aléatoires à variation régulière parcimonieuse. On y établit des résultats de convergence à la fois pour l'étude d'un seul sous-ensemble $C_\beta$ (Proposition 3.3.1 et Théorème 3.3.1), mais également pour un vecteur qui regroupe plusieurs sous-ensembles $C_\beta$ (Théorème 3.4.2). Cette approche motive l'utilisation d'un modèle multinomial à $2^d - 1$ paramètres pour estimer les probabilités $\mathbb{P}(\mathbf{Z} \in C_\beta)$.

## Sélection de modèle

L'estimation des quantités $\mathbb{P}(\mathbf{Z} \in C_\beta)$ se fait via l'approximation

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}|_1 > t) \approx \mathbb{P}(\mathbf{Z} \in C_\beta)\,,$$

pour $t$ "assez grand". Plutôt que de fixer un seuil $t$, une méthode standard utilisée en théorie des valeurs extrêmes est de se donner un niveau $k = n\mathbb{P}(|\mathbf{X}| > t)$ correspondant au nombre de données parmi $\mathbf{X}_1, \ldots, \mathbf{X}_n$ que l'on considère comme extrêmes et de travailler uniquement sur ces $k$ vecteurs. La sélection du modèle multinomial développée dans la Section 3.5 se fait alors en reprenant la méthode de minimisation de la log-vraisemblance pénalisée introduite par Akaike (1973). Pour $k$ fixé suffisamment grand, les résultats asymptotiques obtenus permettent alors de mettre en évidence le modèle qui correspond le mieux aux données.

À ce stade, le choix du seuil $t$, ou de manière équivalente celui du niveau $k$ reste encore à déterminer. De manière générale, il s'agit de réaliser un compromis entre un choix de $k$ assez grand dans le but d'utiliser au mieux les données disponibles et un choix plus modéré qui permet de rester dans le régime extrême. Une revue de la littérature sur ce sujet à été effectuée par Scarrott and MacDonald (2012). Comme expliqué par Embrechts et al. (2013), on "ne doit pas s'attendre de voir apparaître un unique choix de $t$". Une approche standard consiste alors à considérer plusieurs niveaux $k$, de calculer les estimateurs considérés pour ces valeurs et de représenter graphiquement leur évolution en fonction de $k$.

Dans le but de faire tout de même ressortir une valeur optimale de $k$, on prolonge la sélection de modèle effectuée avec les $C_\beta$ en ajoutant le choix du niveau $k$. L'idée est donc de considérer différents modèles avec différents choix de $k$ et de déterminer le plus approprié. Mais cette approche implique de comparer des modèles de taille différente, la taille étant en l'occurrence le paramètre $k$. Dès lors, il semble naturel de tenir compte des valeurs non-extrêmes et ainsi de traiter l'ensemble des données. On réalise donc une partition entre données extrêmes et non-extrêmes et le but de la sélection de modèle est d'identifier celle qui correspond le mieux aux données. Cette étude fait l'objet de la Section 3.5.

Dans ce contexte, la pénalisation linéaire standard qui minimise le Critère d'Information d'Akaike (AIC) ne s'applique pas. Notre approche consiste alors à reprendre les calculs effectués dans le cadre

classique et à les adapter à la sélection jointe des sous-ensembles $C_\beta$ et du niveau $k$. En particulier, on fait apparaître une pénalisation multiplicative via l'étude de l'optimisation de la divergence de Kullback-Leibler entre le vrai modèle (inconnu) régissant nos données et le modèle multinomial théorique.

Notons également que l'influence du seuil dans la parcimonie des vecteurs projetés (voir Figure 1b) rend notre méthode peu coûteuse d'un point de vue algorithmique. En effet, les propriétés théoriques de la projection développées dans le Lemme 2.2.2 permettent de traiter simultanément la parcimonie des vecteurs $\pi(\mathbf{X}/t)$ pour tout $t > 0$. Le temps de calcul est ainsi assez faible et on peut traiter de manière assez efficace des dimensions de l'ordre de $d \sim 10^2$.

On applique cette méthode sur divers données simulées qui modélisent à la fois des cas de dépendance extrême et d'asymptotique indépendance. Les résultats numériques obtenus illustrent la pertinence de l'approche proposée, notamment quand la dimension devient assez grande (de l'ordre de $10^2$). L'algorithme parvient à identifier les différentes directions de l'espace sur lesquels les extrêmes se concentrent. La Section 2.5 propose une comparaison de cette procédure avec celle proposée par Goix et al. (2017). Outre l'absence d'hyper-paramètre, l'approche basée sur la la notion de variation régulière parcimonieuse semble assez robuste, notamment quand la taille des données varie.

## Indépendance conditionnelle

La variation régulière parcimonieuse est efficace pour étudier des événements extrêmes apparaissant sur des sous-espaces de dimension bien inférieure à celle de l'espace de départ. Elle permet donc principalement d'étudier les extrêmes multivariés en grande dimension. En effet, à mesure que la dimension $d$ augmente, il devient très peu probable d'avoir une structure de dépendance complète pour la queue de distribution de $\mathbf{X}$, c'est-à-dire d'observer un comportement extrême simultané de toutes les marginales de $\mathbf{X}$.

En dimension modérée, il se peut que des données vérifient une telle dépendence complète. Dans ce cas, les modèles parcimonieux proposés jusqu'à présent ne sont plus adaptés. Concernant le seuil $t$ et la condition $|\mathbf{X}| > t$ de l'Equation (0.0.2), il s'agit de trouver un conditionnement adapté qui tienne compte d'une éventuelle dépendance forte entre les composantes de $\mathbf{X}$. Un choix naturel se porte sur le minimum des marginales $X_k$ de $\mathbf{X}$. La condition $\min_{1\leq k\leq d} X_k > t$ implique en effet que tous les $X_k$ sont grands simultanément. Le Chapitre 4 développe cette approche en s'appuyant sur les travaux de Segers et al. (2017). La convergence de l'Equation (0.0.2) est alors remplacée par une hypothèse de variation régulière sur la variable $\min_{1\leq k\leq d} X_k$, ainsi que par l'hypothèse de convergence

$$\mathbb{P}\Big(\mathbf{X}/t \in \cdot \mid \min_{1\leq k\leq d} X_k > t\Big) \xrightarrow{w} \mathbb{P}(\mathbf{Y}' \in \cdot)\,, \quad t \to \infty\,,$$

sur l'espace restreint $(0,\infty)^d$. Le vecteur aléatoire $\mathbf{Y}'$ est défini comme la limite des excès de $\mathbf{X}$ au-dessus d'un certain seuil, sa loi se rapproche donc de la loi de Pareto multivariée introduite par

Rootzén and Tajvidi (2006) et étudiée ensuite par Rootzén et al. (2018a), Rootzén et al. (2018b) et Kiriliouk et al. (2019).

Dans ce contexte, l'idée est alors de définir la notion d'indépendance conditionnelle pour $\mathbf{Y}'$ et de la mettre en rapport avec celle de $\mathbf{X}$ (Proposition 4.5.2). Le concept d'indépendance conditionnelle permet de développer des modèles graphiques pour les extrêmes. Cette étude est effectuée par Engelke and Hitz (2020) qui étendent les différentes notions impliquant les modèles graphiques à la loi de Pareto multivariée. Leur analyse repose sur l'utilisation de lois conditionnelles et s'effectue sous l'hypothèse que la loi de Pareto multivariée ne met pas de masse sur les sous-espaces de $\mathbb{R}_+^d$ de dimension inférieure à $d$. Le Chapitre 4 propose alors une comparaison de cette approche avec celle reposant sur les minimum des marginales d'un vecteur aléatoire. On prouve en particulier que l'utilisation du minimum forme un cadre plus général que celui développé par Engelke and Hitz (2020).

# Chapter 1

# Introduction

## Contents

## 1.1 Context and objectives

### 1.1.1 Extreme Value Theory

The study of risk management has been booming in recent years, for instance in the environmental, industrial, or financial fields. The key issue in this framework is to assess the probability of occurrence of an exceptional event which may never have been observed. In meteorology, we can look at the flood risk due to exceptional levels of precipitation. One may also want to estimate the intensity and duration of a heat wave. Similarly in seismology, it is natural to study the maximum intensity that a potential earthquake could reach in a given region. In the industrial field, companies want to estimate as best as possible the probability of incurring heavy financial losses. Similarly, insurance companies would like to evaluate the amount of reinsurance premiums in order to limit their risk of bankruptcy.

These different examples illustrate the general idea of Extreme Value Theory (EVT): An event that occurs regularly (rain, for example) has serious consequences if its intensity is abnormally high. The aim of EVT is to quantify the frequency of occurrence of these events as well as their intensity. The main issue lies in studying an event that has occurred only rarely, if ever. There, the first lines of the famous book of de Haan and Ferreira (2006) are enlightening:

> "Approximately 40% of the Netherlands is below sea level. Much of it has to be protected against the sea by dikes. [...] The government, balancing considerations of cost and safety, has determined that the dikes should be so high that the probability of a flood (i.e., the seawater level exceeding the top of the dike) in a given year is $10^{-4}$. The question is then how high the dikes should be built to meet this requirement. Storm data have been collected for more than 100 years. In this period, at the town of Delfzijl, in the northeast of the Netherlands, 1877 severe storm surges have been identified. The collection of high-tide water levels during those storms forms approximately a set of independent observations, taken under similar conditions (i.e., we may assume that they are independent and identically distributed). No flood has occurred during these 100 years."

How can we determine the probability of occurrence of a flood and its intensity given that none has been observed in the last 100 years? How can the available data, which often contain few (if any) large events, be used to assess the probability of occurrence of an extreme event? These central questions in EVT are opposed to classical statistic which consists in studying the average trend of a sample using the usual tools: mean, variance, median, etc. At a multivariate level, a large event is often the consequence of extreme values jointly in several components. In an environmental context for instance, the air quality can be explained by several air pollutants like ozone, nitrogen dioxide, nitrogen oxide, etc. (see Heffernan and Tawn (2004) and Janßen and Wan (2020) for more details). There, a simultaneous high level of these pollutants leads to deleterious effects on human health. Therefore, the joint structure of extreme values has to be studied. This means that we have to focus

on the dependence between the different components which lead to severe events.

From a statistical point of view, three main issues arise in the study of multivariate extremes. First, the multivariate models proposed so far in the literature are non-parametric. Estimating the tail dependence boils down to studying a probability measure on the unit sphere. This leads to the second issue of multivariate EVT: The non-parametric setting is all the more challenging in a high-dimensional setting. The *curse of dimensonality* (Bellman (1957)) even arises for moderate dimensions such as $d = 10$. Using learning techniques in this context is thus computationally expensive and do not provide good results. A last key issue in EVT is the number of data points used to model the extremal behavior of the sample. While it seems natural to consider only the data which are above a threshold, the choice of this latter is however an unsolved question. This point is all the more crucial since choosing a few number of data points indirectly increases the high-dimensional setting.

From a theoretical point of view, given a random variable $X$, extreme events are characterized by the tails of the distribution of $X$. They can thus be studied through two different approaches. The first one consists in focusing on the highest values of a sample, i.e. the maximum, and to study its convergence when the sample size increases. The second approach is based on threshold exceedances: The idea is to study a conditional distribution given that $X$ is above a threshold $t$ and to investigate its limit when $t$ goes to infinity. The following paragraphs detail these two main approaches, looking separately at the univariate and the multivariate frameworks.

#### 1.1.1.1 Univariate framework

Our aim is to model the extreme behavior of a random variable $X$ taking values in $\mathbb{R}$. Historically, EVT started with the study of the maximum of a sample of independent random variables $X_1, \ldots, X_n$ with the same distribution $X$. This approach was developed in the first half of the twentieth century with the seminal works of Fréchet (1927), Fisher and Tippett (1928) and Gnedenko (1943). Many current works rely on this approach in order to develop a general theory for extremes, see Resnick (1987), Kotz and Nadarajah (2000), Beirlant et al. (2006), Resnick (2007), de Haan and Ferreira (2006), or Embrechts et al. (2013) for a textbook treatment. Nevertheless the main disadvantage of this approach lies in the use of a single observation (the largest one).

In order to better use the available data, another method has then been developed by Balkema and de Haan (1974) and Pickands (1975). It consists in studying the tail of $X$ by focusing on the distribution of $X \mid X > t$ for $t$ large enough or even $t \to \infty$. Similarly one can also look at the distribution of the excess $X - t \mid X > t$. For the study of a river's floods for instance, the idea consists in setting a threshold $t$ beyond which the water level becomes critical and then studying the behavior of the water level beyond this threshold, see Pericchi and Rodríguez-Iturbe (1985). A inventory of the different techniques used in univariate EVT can be found in Gomes and Guillou (2015).

For case studies based on these two methods, many examples are developed in the book of Beirlant et al. (2006). In the founding article on peak-over threshold, Balkema and de Haan (1974)

model residual lifetime at great age. Similarly, Lawless (2011) provides inference procedures in the context of lifetime data. EVT is also widely used to model life or non-life insurance, for instance in Beirlant and Teugels (1992), Resnick (1997), or Teugels (1984). Several applications to finance or insurance have been developed in the monograph of Embrechts et al. (2013). Finally statistic of extreme events has also been used to model natural phenomena such as maximum wind speed in Thom (1954), hydrology in Katz et al. (2002), or general meteorological phenomena in Jenkinson (1955). We refer to Kotz and Nadarajah (2000) for a huge list of applications.

At this point a remark is in order. The study of extremes has been defined as the study of the tails of the distribution. One can thus also be interested in the behavior of the minimum $m_n = \min(X_1, \dots, X_n)$ of a sample of $n$ random variables $X_1, \dots, X_n$. But this comes back to the study of the maximum via the relation $\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n)$. In the same way, if one wishes to study the distribution of $X \mid X < -t$ for $t > 0$, then one can apply the previous case by setting $Y = -X$ and by studying the opposite of the distribution of $Y \mid Y > t$. However, note that in many fields of application it is the study of the maximum or of $X \mid X > t$ with $t > 0$ that is of interest. One of the only examples where the study of extremes concerns large negative values is the case of financial losses of one (or more) company(ies). But in this case, it is sufficient to consider these losses in absolute value to come back to the standard positive framework.

#### 1.1.1.2 Multivariate framework

As pointed out by Coles and Tawn (1991), "problems concerning environmental extremes are often multivariate in character", for instance "wind speed data where maximum hourly gusts, maximum hourly mean speeds and the dependence between them are relevant to building safety". The first works addressing multivariate extremes are the articles by Tiago De Oliveira (1958), Sibuya (1960) and de Haan and Resnick (1977). More recently, Chapter 3 of Kotz and Nadarajah (2000) and Chapters 8 and 9 of Beirlant et al. (2006) expose the different theoretical results in multivariate EVT. A main reference on the multivariate framework is also the review of Fougères (2004) and the references therein.

Assume for instance that we want to study rainfall data in different points of the globe. This is modeled by a random vector $\mathbf{X} = (X_1, \dots, X_d)^\top$ with values in $\mathbb{R}^d$, where $d$ denotes the number of stations and $X_j$ denotes the rainfall level at station $j$. As for the univariate framework, two approaches coexist. The first one consists in studying the componentwise maximum $\mathbf{M}_n$ of a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ with same distribution as $\mathbf{X}$, defined as

$$\mathbf{M}_n := (M_{n,1}, \dots M_{n,d})^\top , \tag{1.1.1}$$

where

$$M_{n,k} := \max_{1 \le i \le n} X_{i,k} , \quad k = 1, \dots, d .$$

The study of the behavior of $M_{n,k}$ for $k = 1, \dots, d$ naturally relies on the univariate framework. But

we also need to study the asymptotic behavior of the dependence structure of $M_n^k$. Going back to the example of rainfall data, the idea is to identify a possible correlation between the maxima of two or several stations. For a survey on this approach, see Kotz and Nadarajah (2000), Chapter 3 or Galambos (1978), Chapter 5.

As in the univariate framework, a second approach consists in setting a threshold $t > 0$ and studying the behavior of the vector $\mathbf{X} \mid |\mathbf{X}| > t$ when $t \to \infty$, where $|\cdot|$ denotes an arbitrary norm on $\mathbb{R}^d$. Under some assumptions on $\mathbf{X}$, multivariate Pareto distributions arise as limits of the threshold exceedances of $\mathbf{X}$. This family of distributions has been introduced by Rootzén and Tajvidi (2006) and stability properties have been established by Rootzén et al. (2018a). For methods based on multivariate threshold exceedances, see Smith (1994) and Rootzén et al. (2018b).

A wide variety of applications of multivariate EVT is given by Tawn (1994). A majority of them addresses environmental issues, such as coastal flooding (Bruun and Tawn (1998)), acid rain (Joe et al. (1992)), sea-level (De Haan and De Ronde (1998) and Tawn (1992)), or air pollution (Heffernan and Tawn (2004) and Janßen and Wan (2020)). For an application on financial risk management, see Longin (2000) who deals with both univariate and multivariate frameworks. The aforementioned examples mainly concern the bivariate case. More recently, higher dimensional studies have been led to study rainfall data (De Fondeville and Davison (2016) and Cooley and Thibaud (2019)), financial return data (Cooley and Thibaud (2019) and Janßen and Wan (2020)) or anomaly detection (Goix et al. (2017)). As for the univariate case, we refer to Kotz and Nadarajah (2000) for a list of applications in the multivariate case.

For the univariate framework as well as for the multivariate one both approaches (maximum of a sample and threshold exceedances) address the behavior of the tail of a random variable or vector. The key notion which tackles this issue and manages to combine these approaches is regular variation (see Bingham et al. (1989) for a general survey on this concept) which provides an elegant and useful description of the asymptotic tail distribution. While the univariate setting can be modeled in a parametric framework, the multivariate one is based on a probability measure on the unit sphere. This non-parametric approach does not easily provide efficient estimators.

### 1.1.2 Goal of this thesis

The central problem of multivariate EVT lies in modeling and estimating the tail dependence. In this context, three general (and thus vague) questions arise. Given a data set of $n$ points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$:

**(Q1)** How can we study the extreme behavior of this data set if $d$ is large?

**(Q2)** What can we say about the dependence structure of the extremes of this data set?

**(Q3)** How many data points should we use to have enough information while being in an extreme setting?

The question **(Q1)** raises the point of the curse of dimensionality. As already explained, the study of multivariate extremes has until recently only been addressed for low-dimensional data, i.e. $d \leq 5$. In order to handle vectors in high dimension, it is necessary to reduce the dimension of the study. Hence, the question **(Q1)** falls within the realm of statistical learning which provides methods to identify some structure in a data set or to classify the data even when $d$ is large. To this end, several learning techniques have been proposed. Their common idea is to introduce sparsity into the data (for instance after an appropriate projection) in order to focus on lower-dimensional subspaces.

This is all the more relevant since extreme values are often located in subspaces of smaller dimension. It is indeed very unlikely that all directions simultaneously contribute to the tail behavior of the data when $d$ is large. Therefore, addressing question **(Q2)** requires to identify these subsets. However, it is not so straightforward to transpose standard learning methods to an extreme context since these methods often focus on the mean behavior of the data while the extreme one is sometimes hard to capture. This is why learning for extremes has only been developed very recently, for instance by Goix et al. (2016) where the major question was to distinguish anomalies from extremal -but not abnormal- data.

Finally, the question **(Q3)** is more naturally associated to model selection. If $k$ denotes the number of extreme points among the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$, then a large $k$ increases the number of data points used and hence the accuracy of the estimation, while it moves away from the extreme setting. A balanced choice should therefore be done between having enough data while staying in an extreme framework. Therefore, a standard idea is to define a family of models with different $k$ and to compute which one best fits the data.

From a theoretical point of view the goal of this thesis is to learn the tail dependence of a random vector $\mathbf{X} \in \mathbb{R}^d$. Since we essentially focus on the right tail of $\mathbf{X}$, it is not so restrictive to assume that $\mathbf{X}$ is non-negative. Most current statistical models for extremes suppose regular variation for $\mathbf{X}$. In this case, the tail dependence of $\mathbf{X}$ is summarized by a probability measure on the unit sphere, called the spectral measure, see Section 1.2 for a review on some existing results on regularly varying random vectors. Roughly speaking, this measure places mass in a direction if extreme events are likely to appear in this direction. The main information is then gathered in the support of this measure. Hence, addressing the question **(Q2)** boils down to studying the spectral measure's support. But estimating this support becomes more and more challenging when the dimension increases. Therefore, many authors first carried out the bivariate case, for instance Tawn (1988), Einmahl et al. (1993), Schmidt and Stadtmüller (2006) and Einmahl et al. (1997). For a review on parametric models in the bivariate case, see Kotz and Nadarajah (2000), Section 3.4. In moderate dimension, Heffernan and Tawn (2004) introduced a semi-parametric approach to analyze a five-dimensional air pollution data. The same data set has been used by Sabourin et al. (2013) and by Sabourin and Naveau (2014) in a Bayesian context. Smith et al. (1990) and subsequently Tawn (1990) analyze the sea level annual maxima of three sites on the south-east coast of England.

In moderate dimension, it is likely that a large event is due to an extreme behavior of all

marginals. In this dependent case, the relations between the components can be done through conditional independence. This has been established by Papastathopoulos and Strokorb (2016) in a general setting and by Gissibl et al. (2018) for discrete spectral measure. The notion of conditional independence is closely linked to the one of graphical models which provide useful representation of the dependence structure between the marginals of a random vector. Engelke and Hitz (2020) extend these concepts for multivariate extremes in order to study extreme dependent data. In Chapter 4, we discuss this approach and develop another one based on the minimum of the marginals. This is a way to address the question **(Q2)** for dependent extremes.

Nevertheless, when the dimension $d$ is large, it is very unlikely that all directions are simultaneously extreme and thus some other dependence structures arise. Until recently, no specific method has been introduced to tackle this framework and so to address the question **(Q1)**. Our aim is to use learning techniques to tackle high-dimensional data. The idea is to project the data onto lower-dimensional subspaces so that some coordinates are put to zero. In this context, only the most significant directions appear and provide some patterns on which extreme events occur. The highlighting of these significant directions is done with a specific projection widely studied in learning theory: the Euclidean projection onto the simplex (Gafni and Bertsekas (1984), Duchi et al. (2008), Bertsekas (1999)). The main advantage of this approach is the linear-time complexity of the projection which allows to handle vectors in large dimension (Condat (2016)). This particular approach to study the angular components of extreme values enjoys many properties and is almost equivalent to the standard approach of regular variation introduced in Section 1.2. Therefore, it allows to deal with simultaneously with both questions **(Q1)** and **(Q2)**: Using the Euclidean projection onto the simplex leads both to a better understanding of the tail dependence and to dimension reduction. In this context, we introduce the notion of *sparse regular variation* (see Chapter 2).

The Euclidean projection onto the simplex provides different results depending on the choice of the threshold above which the data are extreme or equivalently the choice of a level which corresponds to the number of extreme values among the data. This encourages to study more deeply the choice of a potential optimal threshold. As pointed out by Embrechts et al. (2013), one "should never expect a unique choice of [the threshold] $u$ to appear". However, obtaining an idea of which threshold or which level should be used can be done with model selection (Massart (1989), Hastie et al. (2009)). Of course, a model will all the more fits the data if it has a large number of explanatory parameters. However, it becomes hard to understand when the number of parameters becomes too large. Therefore, the goal of model selection is to provide a balanced choice of relevant parameters that explain the data. A widely used technique in this context is the Akaike Information Criterion (Akaike (1973)).

Regarding the level, that is, the number of threshold exceedances, comparing models with different levels implies a comparison between models with different data size. To circumvent this issue, it is necessary to also take the non-extreme values into account. The model selection relies then on the appropriate partition between extreme and non-extreme values. Note that the choice of the level has an impact on the dependence structure of extreme values. Indeed, the Euclidean projection

provides different results depending on the sphere on which the vectors are projected. Hence, the identification of an accurate level has to be done simultaneously with the study of the tail dependence. Therefore, the model selection procedure needs to deal with the whole data set in order to simultaneously address the questions **(Q2)** and **(Q3)**. This is the main purpose of Chapter 3 which develop a statistical framework to answer these questions.

**Outline of the Introduction** This first chapter gathers all theoretical concepts which are useful in this thesis. The main one is regular variation which is introduced in Section 1.2. We develop all tools regarding regularly varying random variables and random vectors and link them to the study of extreme events. We particularly insist on the multivariate framework for which we define the spectral measure and the spectral vector. In order to work with high-dimensional data, it is convenient to reduce the dimension and so to use techniques coming from learning theory. Therefore, Section 1.3 deals with different standard methods to tackle this issue. In particular, we highlight the Euclidean projection onto the simplex, a particular projection which provides sparse vectors and hence reduce the dimension of the study. In Section 1.4 we focus on learning techniques in the context of EVT. A particular attention is paid on the existing methods which are for most of them quite recent. We introduce the different approaches proposed in the literature and insist on the aspects we will rely on. Finally, Section 1.5 deals with the main ideas in model selection. We particularly insist on the estimation of density which will be one of the main point of Chapter 3 and discuss how to deal with threshold selection for multivariate extremes.

**Notations** Throughout all the manuscript, we will use the following notations.

Denote in bold-face elements $\mathbf{x} = (x_1, \ldots, x_d)$ of $\mathbb{R}^d$. We write $\mathbf{x} \leq \mathbf{y}$, $\mathbf{x} < \mathbf{y}$, $\mathbf{x} \geq \mathbf{y}$, etc. where $\leq$, $<$, $\geq$ refer to the componentwise partial ordering in $\mathbb{R}^d$. More generally, for $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, we write $\mathbf{x} \leq y$ if all components $x_i$ of $\mathbf{x}$ satisfy $x_i \leq y$. In the same way, $\mathbf{x} + y$ is defined as the vector $(x_1 + y, \ldots x_d + y)$. We also define $\mathbb{R}^d_+ = \{\mathbf{x} \in \mathbb{R}^d, x_1 \geq 0, \ldots, x_d \geq 0\}$ and $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^d$. For $j = 1, \ldots, d$, $\mathbf{e}_j$ denotes the $j$-th vector of the canonical basis of $\mathbb{R}^d$, which means that $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 is in position $j$. For $a \in \mathbb{R}$, $a_+$ denotes the positive part of $a$, that is $a_+ = a$ if $a \geq 0$ and $a_+ = 0$ otherwise. If $\mathbf{x} \in \mathbb{R}^d$ and $I = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$, then $\mathbf{x}_I$ denotes the vector $(x_{i_1}, ..., x_{i_r})$ of $\mathbb{R}^r$. For $p \in [1, \infty]$, we denote by $|\cdot|_p$ the $\ell^p$-norm in $\mathbb{R}^d$. We write $\xrightarrow{w}$ for the weak convergence, $\xrightarrow{v}$ for the vague convergence, and $\xrightarrow{d}$ for the convergence in distribution of random variables.

For a set $E$, we denote by $\mathcal{P}(E)$ its power set: $\mathcal{P}(E) = \{A, A \subset E\}$. We also use the notation $\mathcal{P}^*(E) = \mathcal{P}(E) \backslash \{\emptyset\}$. If $E = \{1, \ldots, r\}$, we simply write $\mathcal{P}_r = \mathcal{P}(\{1, \ldots, r\})$ and $\mathcal{P}^*_r = \mathcal{P}(\{1, \ldots, r\}) \backslash \{\emptyset\}$. For a finite set $E$, we denote by $\#E$ its cardinality. If $\#E = r \geq 1$, then $\#\mathcal{P}(E) = 2^r$. In particular, $\#\mathcal{P}_r = 2^r$ and $\#\mathcal{P}^*_r = 2^r - 1$. Finally, if $F$ is a subset of a set $E$, we denote by $F^c$ the complementary of $F$ (in $E$).

## 1.2 Regular variation

The notion of regular variation has been introduced by Karamata (1933) and then used in different mathematical contexts and particularly in applied probability theory. A main reference on this subject in the book of Bingham et al. (1989) which gathers several theoretical results. We also refer to the report of Mikosch (1999) and the thesis of Basrak (2000) for a review of the multivariate setting. This section develops all useful tools regarding regular variation in the context of EVT. We start with the univariate setting which emphasizes the key role played by regularly varying random variables when studying the maximum of a sample or the threshold exceedances. Then, we extend this framework to random vectors.

### 1.2.1 The univariate framework

We begin this section with general notions of regularly varying functions and regularly varying random variables before using them to unify different approaches in univariate EVT.

#### 1.2.1.1 General results

**Definition 1.2.1** (Regularly varying function)**.** A positive measurable function $f$ is *regularly varying* (at infinity) with index $\alpha \in \mathbb{R}$ if it is defined on some neighborhood of infinity and if

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)} = t^\alpha\,, \quad t > 0\,. \tag{1.2.1}$$

If $\alpha = 0$, then $f$ is said to be *slowly varying*.

**Remark 1.2.1.**

1. A regularly varying function $f$ with index $\alpha$ can be written as $f(x) = x^\alpha L(x)$ where $L$ is a slowly varying function.

2. It is sufficient to require that the limit

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)}$$

exists, is finite, and positive for all $t > 0$. Then, it satisfies the so-called Cauchy's functional equation and is thus a power function. It is even sufficient to suppose that for all $t$ in a set of positive Lebesgue measure the limit exists, is finite, and positive, see Theorem B.1.3 in de Haan and Ferreira (2006).

3. It is possible to define regular variation at a point $x_0 \in \mathbb{R}$. A positive measurable function $f$ is regularly varying at $x_0$ if the function $x \mapsto f(x_0 - x^{-1})$ is regularly varying at infinity.

4. If Equation (1.2.1) holds, then the convergence is actually uniform on all compact sets of $(0, \infty)$, see Bingham et al. (1989), Theorem 1.2.1.

**Example 1.2.1.** Functions converging to positive constants (and thus constant functions) are slowly varying, as well as the family of functions $f(x) = \log^{\beta}(x)$, $\beta \in \mathbb{R}$. If $\alpha \in \mathbb{R}$, the functions $x \mapsto x^{\alpha}\log(x)^{\beta}$ are regularly varying with index $\alpha$.

We now introduce the concept of regularly varying random variable. If $X$ is a random variable with distribution function $F$, i.e. $F(x) = \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$, then we denote its *survey function* by

$$\bar{F}(x) := 1 - F(x) = \mathbb{P}(X > x)\,, \quad x \in \mathbb{R}\,.$$

We also write $x_F := \sup\{x \in \mathbb{R},\, F(x) < 1\}$ and call this point the *right endpoint* of $F$.

**Definition 1.2.2** (Non-negative regularly varying random variable)**.** A non-negative random variable $X$ and its distribution $F$ are said to be regularly varying with index $\alpha \geq 0$ if $\bar{F}$ is regularly varying with index $-\alpha$.

Unless it is explicitly stated, we always consider regularly varying random variables with infinite right endpoint $x_F$. In this case, we obtain the following convergence:

$$\lim_{x\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = \lim_{x\to\infty} \frac{1 - F(tx)}{1 - F(x)} = \lim_{x\to\infty} \frac{\mathbb{P}(X > tx)}{\mathbb{P}(X > x)} = t^{-\alpha}\,, \quad t > 0\,. \tag{1.2.2}$$

One should pay attention to the difference of notation between the index of a regularly varying function and of a regularly varying random variable. In particular, it is clear from (1.2.2) that the case $\alpha < 0$ is impossible.

**Example 1.2.2** (Pareto distribution)**.** A standard example of a regularly varying distribution is the Pareto distribution whose distribution function is given by $x \mapsto 1 - (x_m/x)^{\rho}$ for $x \geq x_m$, where $x_m > 0$ is the location parameter and $\rho > 0$ is the shape parameter. We can easily check that this distribution is regularly varying with index $\rho$.

**Example 1.2.3** (Burr distribution)**.** The Burr distribution is characterized by the distribution function $F(x) = 1 - (1 + x^c)^{-k}$ for $x > 0$ and $F(x) = 0$ elsewhere, where $c > 0$ and $k > 0$ are given parameters. Then, a short computation gives

$$\frac{\bar{F}(tx)}{\bar{F}(x)} = \frac{(1 + (tx)^c)^{-k}}{(1 + x^c)^{-k}} \to t^{-ck}\,, \quad x \to \infty\,,$$

for all $t > 0$. Hence, the Burr distribution is regularly varying with tail index $ck$.

Extending the notion of regular variation to real random variables requires to take into account the two tails of the distribution of $X$. This is the purpose of the following definition.

**Definition 1.2.3** (Regularly varying random variable)**.** A real-valued random variable $X$ and its distribution are regularly varying with index $\alpha \geq 0$ if there exist constants $p_+, p_- \geq 0$ satisfying

$p_+ + p_- = 1$ such that

$$\lim_{x\to\infty} \frac{\mathbb{P}(X > tx)}{\mathbb{P}(|X| > x)} = p_+ t^{-\alpha} \quad \text{and} \quad \lim_{x\to\infty} \frac{\mathbb{P}(X < -tx)}{\mathbb{P}(|X| > x)} = p_- t^{-\alpha}\,, \tag{1.2.3}$$

for all $t > 0$.

In particular this definition implies that the random variable $|X|$ is regularly varying with index $\alpha$. Besides we notice that if $X$ is a non-negative random variable, then the notion is consistent with Definition 1.2.2. It corresponds to the case where $p_+ = 1 - p_- = 1$.

**Example 1.2.4** (Symmetric distributions)**.** If a random variable $X$ has a symmetric distribution, i.e. $X \overset{d}{=} -X$, then $\mathbb{P}(|X| > x) = 2\mathbb{P}(X > x)$ and then $X$ is regularly varying if and only if $\bar{F}$ is regularly varying. There we have necessary $p_+ = p_- = 1/2$.

For instance, consider a random variable $X$ with a Cauchy distribution given by $F(x) = 1/2 + \arctan(x)/\pi$. The density of $X$ is given by $f(x) = 1/(\pi(1 + x^2))$ which is a symmetric function. Hence, the random variable $X$ has a symmetric distribution. A Taylor expansion of arctan is given by

$$\arctan(x) = \frac{\pi}{2} - \frac{1}{x} + o\Big(\frac{1}{x}\Big)\,, \quad x \to \infty\,,$$

which leads to the equivalent $\bar{F}(x) = 1/2 - \arctan(x)/\pi \sim 1/(\pi x)$ when $x \to \infty$. This implies that

$$\frac{F(tx)}{F(x)} \to t^{-1}\,, \quad x \to \infty\,,$$

for all $t > 0$. Thus, $X$ is regularly varying with tail index 1 and constants $p_+ = p_- = 1/2$.

#### 1.2.1.2 Application to univariate EVT

As explained in Section 1.1, studying the extreme behavior of a random variable $X$ boils down to focusing on the tails of $X$. Since the study can be done equivalently on the left or on the right tail we focus here on the right one. To this end, two similar but different approaches coexist. The first one deals with max-stable distributions which arise as limits of normalized maxima of an i.i.d. sample $X_1, \ldots, X_n$ with generic distribution $X$. The second one consists in studying the behavior of $X$ conditioned on the event that $X$ exceeds a high threshold $t$. Intuitively, these two approaches are quite natural. Studying an extreme event can be done either by focusing on the behavior of the highest value, that is the maximum, or by studying only large values, that is values above a high threshold. We explain in this section that these two approaches are theoretically equivalent and that they are closely related to the notion of regular variation.

The results discussed here have been developed in the seminal articles of Fréchet (1927), Fisher and Tippett (1928) and Gnedenko (1943). More recently, one can cite the textbooks of Resnick (1987), Beirlant et al. (2006), de Haan and Ferreira (2006), Resnick (1987), or Embrechts et al. (2013) for a survey on the main theoretical results of EVT. Regarding statistical aspects we refer

to the textbooks of Gumbel (1958), Pickands (1975), Beirlant et al. (1996a), or Coles (2001). The "Bibliographical Notes" at the end of each chapter of Reiss (1989) provides relevant historical aspects.

We start with a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with generic distribution $X$. We denote by $F$ the distribution function of $X$ and by $(M_n)_{n\geq 1}$ the sequence of the partial maxima $M_n = \max_{1\leq j\leq n} X_j$ for $n \geq 1$. Recall that $x_F = \sup\{x \in \mathbb{R},\, F(x) < 1\}$ denotes the *right endpoint* of $F$. The distribution function of $M_n$ corresponds to $F^n$ so that $M_n$ converges in distribution to a Dirac mass in $x_F$. Thus a normalization is necessary to obtain a non-degenerate limit: We assume that there exist two real-valued sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, with $a_n > 0$, such that $a_n^{-1}(M_n - b_n)$ converges in distribution to a non-degenerate random variable $Y$, which is equivalent to the convergence

$$F^n(a_n x + b_n) \to H(x)\,, \quad n \to \infty\,, \tag{1.2.4}$$

for any continuity point $x$ of $H$, where $H$ is the distribution function of $Y$. If such a convergence holds, then the distribution of $Y$ is unique up to an affine transformation according to the convergence to types theorem (Theorem A.1.1). In this case, we say that $X$ belongs to the maximum domain of attraction of $Y$ and we denote this indifferently $X \in \mathrm{MDA}(H)$ or $F \in \mathrm{MDA}(H)$. The goal is then to tackle the three following points: identify the possible limit distributions $H$, characterize the convergence in Equation (1.2.4), provide suitable sequences $(a_n)$ and $(b_n)$.

**Definition 1.2.4** (Max-stable distribution)**.** A non-degenerate random variable $X$ and its distribution are said to be max-stable if there exist two real sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$, with $a_n > 0$, such that for any sample $X_1, \ldots, X_n$ of i.i.d. random variables with the same distribution as $X$, the following equality in distribution is satisfied for all $n \geq 1$:

$$a_n^{-1}(M_n - b_n) \stackrel{d}{=} X\,.$$

It is clear that any max-stable distribution is in its own domain of attraction. The following theorem ensures that these are the only possible limits.

**Theorem 1.2.1.** *The class of max–stable distributions coincides with the class of all possible (non-degenerate) limit distribution for normalized maxima of i.i.d. random variables.*

The proof can be found in Embrechts et al. (2013), Theorem 3.2.2, and is based on the convergence to types theorem (Theorem A.1.1). The last step is then to identify the max-stable distributions. It is the purpose of the following theorem which is the basis of univariate EVT.

**Theorem 1.2.2** (Fisher)**.** *The following distribution are, up to an affine transformation, the only max-stable distributions.*

1. *The Fréchet distribution with parameter $\alpha > 0$:*

$$\Phi_\alpha(x) = \begin{cases} 0 & \text{if } x \leq 0\,, \\ \exp(-x^{-\alpha}) & \text{if } x > 0\,. \end{cases} \tag{1.2.5}$$

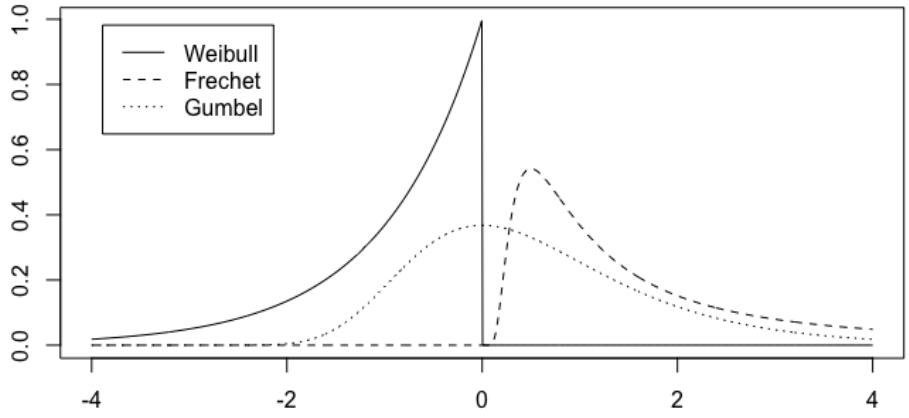



Figure 1.1: Densities of the standard max-stable distributions. We choose $\alpha = 1$ for the Fréchet and the Weibull distributions.

2. *The Weibull distribution with parameter* $\alpha > 0$*:*

$$\Psi_\alpha(x) = \begin{cases} \exp(-|x|^\alpha) & \text{if } x < 0\,, \\ 0 & \text{if } x \geq 0\,. \end{cases} \tag{1.2.6}$$

3. *The Gumbel distribution:*

$$\Lambda(x) = \exp(-\exp(-x))\,, \quad x \in \mathbb{R}\,. \tag{1.2.7}$$

The max-stable distributions are also called *Extreme Value Distributions* since they appear as the limits of normalized maxima. The densities of these three distributions are represented in Figure 1.1. We can easily check that the three distributions in Theorem 1.2.2 are max-stable.

The three distributions of Theorem 1.2.2 can be summarized into a *Generalized Extreme Value distribution* (GEV) whose distribution function is given by

$$H_{\xi,\mu,\sigma}(x) = \exp\left(-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right), \tag{1.2.8}$$

for $\xi, \mu \in \mathbb{R}$, and $\sigma > 0$. We shortly denote $H_\xi = H_{\xi,0,1}$ the *standard* GEV.

**Remark 1.2.2.** If $\xi > 0$, then $H_\xi$ is a Fréchet distribution with parameter $\alpha = 1/\xi$. Likewise, if $\xi < 0$, then $H_\xi$ is a Weibull distribution with parameter $\alpha = -1/\xi$. Finally, if $\xi = 0$, then Equation (1.2.8) is interpreted as the limit when $\xi \to 0$ which gives $H_0 = \exp(-\exp(-x))$ and corresponds to

the Gumbel distribution.

**Remark 1.2.3.** Note that since the distribution function $H_\xi$ is continuous on $\mathbb{R}$, the convergence (1.2.4) holds for all $x \in \mathbb{R}$.

According to the convergence to types theorem (Theorem A.1.1), it is always possible to choose $a_n$ and $b_n$ such that $a_n^{-1}(M_n - b_n) \to H_\xi$ as $n \to \infty$. Thus, the study of univariate EVT boils down to a parametric family of distributions $(H_\xi)_{\xi \in \mathbb{R}}$. Therefore, the shape parameter $\xi$ concentrates a lot of information regarding the behavior of the extreme values of $X$. The sign of $\xi$ implies different behavior for the right endpoint of $X$. In particular, $\xi > 0$ means that the underlying distribution is *heavy-tailed*. This case is hence naturally used in EVT.

From a statistical point of view, the estimation of $\xi$ is one of the major issues. Several estimators of $\xi$ have been proposed. If $\xi > 0$, the most common is the Hill estimator introduced by Hill (1975) and defined by

$$\hat{\xi} = \left( \frac{1}{k} \sum_{j=1}^{k} \log(X_{(j)}) - \log(X_{(k)}) \right)^{-1},$$

where $X_{(1)} \geq \ldots \geq X_{(n)}$ denotes the order statistics of the i.i.d. sample $X_1, \ldots, X_n$, and where $k = k_n$ is an intermediate sequence satisfying $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. Other estimators addressing all cases have also been proposed, e.g. the Pickands estimator (Pickands (1975)) or the Dekkers–Einmahl–de Haan estimator (Dekkers et al. (1989)). See Embrechts et al. (2013), Section 6.4.2, for some properties of these estimators.

Going back to the Fréchet case, that is, $\xi > 0$, we detail here some conditions on the distribution of $X$ which ensure the convergence to $H_\xi$.

**Theorem 1.2.3** (Domain of attraction of a Fréchet distribution)**.** *A distribution function $F$ belongs to the domain of attraction of $\Phi_\alpha$, $\alpha > 0$, if and only if the $\bar{F} = 1 - F$ is regularly varying with tail index $-\alpha$.*

*In this case, one can choose $a_n = F^{\leftarrow}(1 - n^{-1})$ or more generally a sequence $(a_n)$ such that $n\bar{F}(a_n) \to 1$, and $b_n = 0$. Then, the following convergence holds:*

$$a_n^{-1} M_n \to Y \sim \Phi_\alpha \,, \quad n \to \infty \,. \tag{1.2.9}$$

This result implies in particular that every $F \in \mathrm{MDA}(\phi_\alpha)$ has an infinite right endpoint $x_F = \infty$: The random variable $X$ is not bounded on the right. Therefore, a particular attention should be paid to this case, since the $\mathrm{MDA}(\phi_\alpha)$ contains heavy-tailed distributions. The following proposition summarizes the results we obtain for the Fréchet maximum domain of attraction.

**Proposition 1.2.1.** *For $\alpha > 0$, the following assertions are equivalent:*

1. $F \in \mathrm{MDA}(\phi_\alpha)$,
2. $\bar{F} \in \mathrm{RV}_{-\alpha}$,

3. *there exists* $a_n > 0$ *such that* $a_n^{-1} M_n \to Y \sim \Phi_\alpha$ *when* $n \to \infty$,

4. *there exists* $a_n > 0$ *such that* $n\mathbb{P}(a_n^{-1} X > x) = n\bar{F}(a_n x) \to x^{-\alpha} \mathbb{1}_{\{x>0\}}$ *when* $n \to \infty$,

5. *there exists* $a_n > 0$ *such that* $n\mathbb{P}(a_n^{-1} X \in \cdot) \xrightarrow{v} \nu_\alpha(\cdot)$, $n \to \infty$, *in* $\mathcal{M}_+((0,\infty))$, *where* $\nu_\alpha((x,\infty)) = x^{-\alpha}$, *and where* $\mathcal{M}_+((0,\infty))$ *denotes the space of all Radon measures on* $(0,\infty)$.

*In the last three assertions, a natural choice is* $a_n = F^{\leftarrow}(1 - n^{-1})$. *More generally the convergence* (1.2.9) *holds if and only if* $n\mathbb{P}(X > a_n) \to 1$. *The other choices of* $a_n$ *and* $b_n$ *are given by the convergence to types theorem (and then the limit has the same type as* $\phi_\alpha$*).*

Proposition 1.2.1 is an extension of Proposition 3.6 in Resnick (2007). Resnick explains that it is useful to consider the space $(0,\infty]$ rather than $(0,\infty)$ so that the neighborhoods of $\infty$ are relatively compact, but this compactification is not necessary.

**Remark 1.2.4.** If $X$ is a non-negative random variable, then the assumptions of Proposition 1.2.1 are equivalent to the fact that $X$ is regularly varying with tail index $\alpha$. The equivalence between regular variation of a non-negative random variable $X$ and vague convergence of the measure $n\mathbb{P}(a_n^{-1} X \in \cdot)$ will be the starting point of the study of multivariate regular variation.

**Example 1.2.5.** Every distribution function $F$ which satisfies $\bar{F}(x) \sim Cx^{-\alpha}$, $x \to \infty$, for some $C, \alpha > 0$ is in the domain of attraction of $\Phi_\alpha$. This is the case of the Pareto, Cauchy, and Burr distributions.

#### 1.2.1.3 Threshold exceedances

We still focus on the Fréchet case. An asymptotic expansion gives the equivalent

$$1 - \Phi_\alpha(x) \sim x^{-\alpha}, \quad x \to \infty.$$

Intuitively, $X$ belongs to $\mathrm{MDA}(\Phi_\alpha)$ if the behavior of the right tail of $X$ is "close" to the behavior of $1 - \phi_\alpha(x) \sim x^{-\alpha}$. The vague notion of "close" is then quantified by the fact that $\bar{F}$ is regularly varying according to Theorem 1.2.3:

$$\lim_{t\to\infty} \frac{\bar{F}(tx)}{\bar{F}(t)} = x^{-\alpha}, \quad x > 0.$$

In particular, if $x \geq 1$, then the previous convergence can be rephrased as

$$\mathbb{P}(X > tx \mid X > t) \to x^{-\alpha}, \quad t \to \infty.$$

This brings us to the other approach to EVT which consists in studying the distribution of $X$ conditioned on the event that $\{X > t\}$ when $t \to \infty$. The following theorem due to Pickands, Balkema and de Haan ensures that this approach is equivalent to the previous one and gives the limit distribution.

**Theorem 1.2.4.** *For $\xi \in \mathbb{R}$, the following assertions are equivalent:*

1. $F \in \mathrm{MDA}(H_\xi)$.
2. *There exists a positive measurable function $a(\cdot)$ such that for $1+\xi x > 0$,*

$$\lim_{t \to x_F} \frac{\bar{F}(t + a(t)x)}{\bar{F}(t)} = (1+\xi x)^{-1/\xi}\,, \tag{1.2.10}$$

*where the right-hand side in interpreted as $e^{-x}$ if $\xi = 0$.*

We define the *standard Generalized Pareto Distribution* (*standard GPD* in abbreviated form) $G_\xi$ as the limit of (1.2.10):

$$G_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi} & \text{if } \xi > 0,\ x > 0\,, \\ 1 - (1+\xi x)^{-1/\xi} & \text{if } \xi < 0,\ 0 \le x \le -1/\xi\,, \\ 1 - e^{-x} & \text{if } \xi = 0,\ x > 0\,. \end{cases}$$

The *Generalized Pareto Distribution* $G_{\xi,\mu,\sigma}$ is defined by the relation $G_{\xi,\mu,\sigma} = G_\xi(\sigma^{-1}(x-\mu))$, for $\mu \in R$ and $\sigma > 0$. This family of distributions satisfies the relation

$$1 - G_{\xi,\mu,\sigma}(x) = -\log(H_{\xi,\mu,\sigma})\,, \tag{1.2.11}$$

where $H_{\xi,\mu,\sigma}$ is the GEV distribution (see Equation (1.2.8)). Equation (1.2.10) can be rephrased in terms of conditional probability as follows:

$$\lim_{t \to x_F} \mathbb{P}\left( \frac{X-t}{a(t)} > x \middle| X > t \right) = 1 - G_\xi(x)\,, \tag{1.2.12}$$

for the $x$ defined in (1.2.10).

The Peaks over Threshold (PoT) method (Leadbetter (1991)) uses this convergence as an approximation for a threshold $t > 0$ "high enough". Then, as soon as an estimator $\hat{a}$ of $a$ is obtained, it is possible to estimate the limit $1 - G_\xi(x)$ with the available data. Regarding the question **(Q3)**, the main problem relies here in the notion of "high enough": What is the "best" choice for $t$? As pointed out by Embrechts et al. (2013), "the reader should never expect a unique choice of $t$ to appear". The authors "recommend using plots, to reinforce judgement and common sense and compare resulting estimates across a variety of $t$–values".

### 1.2.2 Multivariate regular variation

The goal is now to extend the concepts introduced in Section 1.2.1 to a multivariate framework. We thus consider a random vector $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$. According to Theorem 1.2.3 there is a close connection between regular variation and study of the maximum of a sample. We restrict the

study to the non-negative case, i.e. for $\mathbf{X} \in \mathbb{R}_+^d$, which already provides a wide theory of regularly varying random vectors.

#### 1.2.2.1 From univariate to multivariate extremes

We start with the univariate case developed previously. This means that $d = 1$ and $X \in \mathbb{R}_+$. Following Proposition 1.2.1 and Remark 1.2.4, we have equivalence of

1. $X$ is regularly varying (at infinity) with tail index $\alpha > 0$,
2. there exists $a_n \to \infty$ such that $n\mathbb{P}(a_n^{-1}X \in \cdot) \xrightarrow{v} \nu_\alpha(\cdot)$, $n \to \infty$, in $\mathcal{M}_+((0,\infty))$, where $\nu_\alpha((x,\infty)) = x^{-\alpha}$.

Our aim is to use this characterization to define multivariate regular variation in a similar way, i.e. with vague convergence. We consider a sequence of i.i.d. random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots$ on $\mathbb{R}_+^d$ with generic distribution $\mathbf{X}$ and denote by $\mathbf{M}_n$ the componentwise maximum:

$$\mathbf{M}_n := (\max_{1\le j\le n} X_{j,1}, \ldots, \max_{1\le j\le n} X_{j,d})^T, \quad n \ge 1\,. \tag{1.2.13}$$

Following the ideas of the univariate case, we focus on the convergence of the vector $a_n^{-1}\mathbf{M}_n$ when $n \to \infty$ for a real-valued sequence $(a_n)$ satisfying $a_n > 0$. To this end, we fix $\mathbf{x} \in \mathbb{R}_+^d \setminus \{\mathbf{0}\}$ and write

$$\begin{aligned}\mathbb{P}(a_n^{-1}\mathbf{M}_n \le \mathbf{x}) &= \mathbb{P}(\forall k = 1,\ldots,n,\ a_n^{-1}\mathbf{X}_k \le \mathbf{x})\\ &= \mathbb{P}(a_n^{-1}\mathbf{X} \le \mathbf{x})^n\\ &= \Big(1 - \frac{n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{x}]^c)}{n}\Big)^n,\end{aligned} \tag{1.2.14}$$

where $[\mathbf{0},\mathbf{x}]^c = \mathbb{R}_+^d \setminus [\mathbf{0},\mathbf{x}]$, and where all inequalities are meant componentwise. Then, the following assertions are equivalent.

1. The normalized componentwise maximum $a_n^{-1}\mathbf{M}_n$ converges to a limit distribution $\mathbf{Y} \sim H$ when $n \to \infty$.
2. For every continuity point $\mathbf{x} \in \mathbb{R}_+^d$ of $H$,
$$\mathbb{P}(a_n^{-1}\mathbf{M}_n \le \mathbf{x}) \to H(\mathbf{x})\,, \quad n \to \infty\,.$$
3. For every continuity point $\mathbf{x} \in \mathbb{R}_+^d$ of $H$,
$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{x}]^c) \to -\log(H(\mathbf{x}))\,, \quad n \to \infty\,, \tag{1.2.15}$$
If $H(\mathbf{x}) = 0$, then the right-hand side is interpreted as $\infty$.

Equation (1.2.15) can be rewritten in the following way: for any continuity point $\mathbf{x}$ of $H$,

$$\mu_n([\mathbf{0},\mathbf{x}]^c) := n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{x}]^c) \to -\log(H(\mathbf{x})) =: \mu([\mathbf{0},\mathbf{x}]^c)\,, \quad n \to \infty\,. \tag{1.2.16}$$

This convergence must be seen as the convergence of two measures, $\mu_n$ and $\mu$, on the sets $[\mathbf{0},\mathbf{x}]^c$. By Dynkin's theorem (Theorem A.2.1), $\mu_n$ and $\mu$ can be uniquely extended to measures on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$. Then, the convergence (1.2.16) suffices to prove that $\mu_n$ converges vaguely to $\mu$, see Resnick (2007), Lemma 6.1.

#### 1.2.2.2 Regularly varying random vectors

The considerations of the previous subsection lead to the following definition.

**Definition 1.2.5** (Regularly varying random vector)**.** Let $\mathbf{X} \in \mathbb{R}_+^d$ be a non-negative random vector. Assume that there exists a positive sequence $(a_n)$ such that $a_n \to \infty$ when $n \to \infty$. The vector $\mathbf{X}$ and its distribution are said *regularly varying* if there exists a non-zero Radon measure $\mu$ on the Borel $\sigma$-field of $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ such that

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot) \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,. \tag{1.2.17}$$

The limit measure $\mu$ is called the *tail measure* of the regularly varying vector $\mathbf{X}$.

We already explained in Section 1.2.2.1 that the convergence in Equation (1.2.17) is closely related to the convergence of the componentwise maximum $\mathbf{M}_n$. In particularly if $\mathbf{X} \in \mathbb{R}_+^d$ is a regularly varying random vector, then

$$\lim_{n\to\infty} \mathbb{P}(a_n^{-1}\mathbf{M}_n \le \mathbf{x}) \to \exp(-\mu([\mathbf{0},\mathbf{x}]^c)) =: H(\mathbf{x})\,, \tag{1.2.18}$$

for all continuity point $\mathbf{x}$ of the limit $H$ (see Equation (1.2.14)). The distribution of $H$ is called *multivariate Fréchet distribution.*

**Example 1.2.6** (Independent marginals)**.** Assume that the marginals $X_i$ of the vector $\mathbf{X}$ are independent, identically distributed, and regularly varying with the same tail index $\alpha > 0$. Choose $a_n$ such that $n\mathbb{P}(X_1 > a_n) \to 1$. Following Proposition 1.2.1, we obtain that

$$\mathbb{P}(a_n^{-1}\mathbf{M}_n \le \mathbf{x}) = \prod_{i=1}^{d} \mathbb{P}(a_n^{-1}M_{n,i} \le x_i) \to \exp(-(x_1^{-\alpha} + \ldots + x_d^{-\alpha}))\,, \quad n \to \infty\,,$$

for all $x_k > 0$. Thus, Equation (1.2.18) implies that the tail measure $\mu$ satisfies $\mu([\mathbf{0},\mathbf{x}]^c) = x_1^{-\alpha} + \ldots + x_d^{-\alpha}$, for all $x_k > 0$. This example will be developed further in Section 1.2.3.

Recall that the univariate case highlights a tail index $\alpha$ which characterizes the tail behavior of the random variable $X \in \mathbb{R}$. This index appears in particular in the limit measure $\nu_\alpha$ in Proposition 1.2.1. It is then natural that such an index also appears in the multivariate setting.

The tail measure $\mu$ satisfies the following key property: for all Borel sets $A \subset \mathbb{R}_+^d \setminus \{\mathbf{0}\}$ and for all $t > 0$,

$$\mu(tA) = t^{-\alpha}\mu(A)\,. \tag{1.2.19}$$

Intuitively, it means that the mass the measure $\mu$ places on the set $A$ decreases like a power function as this set is translated toward infinity. In this case, the index $\alpha$ is called the *tail index* and we say that the random vector $\mathbf{X}$ is *regularly varying with limit measure $\mu$ and tail index $\alpha$*.

**Consequences of the homogeneity property** The homogeneity property has plenty of consequences.

First, recall from Equation (1.2.18) that $\mu([\mathbf{0},\mathbf{x}]^c) = -\log H(\mathbf{x})$ for all continuity point $\mathbf{x}$ of $H$. In particular since for $\mathbf{x} > \mathbf{0}$ large enough $H(\mathbf{x})$ is positive it implies that the quantity $\mu([\mathbf{0},\mathbf{x}]^c)$ is finite. Hence, by the homogeneity property, for all $t > 0$ the quantity $\mu([\mathbf{0},t\mathbf{x}]^c) = t^{-\alpha}\mu([\mathbf{0},\mathbf{x}]^c)$ is finite. We say that the measure $\mu$ is finite for sets *bounded away from zero*.

Now consider the infinity norm in $\mathbb{R}^d$. Then, the set $\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > t\} = ([0,t]^d)^c$ is a $\mu$-continuity set. Indeed, its boundary corresponds to the set $\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty = t\}$ which satisfies

$$\begin{aligned}
\mu(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty = t\}) &= \lim_{\epsilon\to 0}\Big[\mu\big(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > (1-\epsilon)t\}\big) - \mu\big(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > (1+\epsilon)t\}\big)\Big] \\
&= \lim_{\epsilon\to 0}\Big[(1-\epsilon)^{-\alpha} - (1+\epsilon)^{-\alpha}\Big]\mu\big(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > t\}\big) \\
&= 0\,,
\end{aligned}$$

where we used that $\mu(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > t\}) = \mu(([0,t]^d)^c)$ is finite thanks to the previous point. Therefore, for all $t > 0$, the homogeneity property implies that

$$n\mathbb{P}(|\mathbf{X}|_\infty > a_n t) = n\mathbb{P}(a_n^{-1}|\mathbf{X}|_\infty \in ([0,t]^d)^c) \to \mu(([0,t]^d)^c) = t^{-\alpha}\mu([\mathbf{0},\mathbf{1}]^c)\,, \quad t \to \infty\,,$$

Following Proposition 1.2.1, we obtain that $|\mathbf{X}|_\infty$ is regularly varying with tail index $\alpha$. In particular the choice $t = 1$ leads to the convergence

$$n\mathbb{P}(|\mathbf{X}|_\infty > a_n) = n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{1}]^c) \to \mu([\mathbf{0},\mathbf{1}]^c)\,.$$

It is sometimes convenient to choose a sequence $(a_n)$ such that $\mu([\mathbf{0},\mathbf{1}]^c) = 1$ for reasons explained later.

Subsequently since all norms are equivalent in $\mathbb{R}^d$ a short computation provides that $|\mathbf{X}|$ is regularly varying for every norm $|\cdot|$. Similarly as for the infinity norm, as soon as a norm $|\cdot|$ is fixed it seems natural to choose $a_n$ such that $n\mathbb{P}(|\mathbf{X}| > a_n) \to 1$.

Finally the homogeneity property implies the max-stability property of the multivariate Fréchet distribution. Recall that in the univariate case a Fréchet-distributed random variable $Y$ with tail index $\alpha > 0$ satisfies the equality $n^{-1/\alpha}M_n \overset{d}{=} Y$ for all $n \geq 1$. This can be easily extended to the multivariate setting. Indeed, for $\mathbf{x} \in \mathbb{R}_+^d \setminus \{\mathbf{0}\}$, the homogeneity property with $A = [\mathbf{0},\mathbf{x}]^c$ and

$t = n^{-1/\alpha}$ implies that

$$\mu([\mathbf{0}, \mathbf{x}]^c) = n\mu([\mathbf{0}, n^{1/\alpha}\mathbf{x}]^c)\,.$$

Taking the exponential of the opposite of both members leads then to

$$H(\mathbf{x}) = \exp(-\mu([\mathbf{0}, n^{1/\alpha}\mathbf{x}]^c))^n\,.$$

Therefore, if $\mathbf{Y}$ follows a multivariate Fréchet distribution $H$, we obtain that

$$\mathbb{P}(n^{-1/\alpha}\mathbf{M}_n \leq \mathbf{x}) = \mathbb{P}(n^{-1/\alpha}\mathbf{Y} \leq \mathbf{x})^n = \exp(-\mu([\mathbf{0}, n^{1/\alpha}\mathbf{x}]^c))^n = H(\mathbf{x})\,.$$

Hence, we obtain the max-stability property

$$n^{-1/\alpha}\mathbf{M}_n \stackrel{d}{=} \mathbf{Y}\,, \quad n \geq 1\,.$$

#### 1.2.2.3 Equivalent formulation of regular variation

As in the univariate case, several characterizations of regular variation can be given. In particular, we would like to have one which brings out threshold exceedances. With this in mind, we would like to decompose the convergence in (1.2.17) into a radial convergence and and angular one. Hence, fix a norm $|\cdot|$ in $\mathbb{R}^d$ and denote by $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d,\, |\mathbf{x}| = 1\}$ its unit sphere. Denote also by $\mathbb{S}_+^{d-1}$ the intersection of this unit sphere with the positive orthant: $\mathbb{S}_+^{d-1} = \mathbb{S}^{d-1} \cap \mathbb{R}_+^d$. Following Basrak (2000), Theorem 2.1.8, we define the set $V_{r,\,A}$ by

$$V_{r,\,A} = \{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}| > r, \mathbf{x}/|\mathbf{x}| \in A\}\,.$$

for $r > 0$ and $A \subset \mathbb{S}_+^{d-1}$ (see Figure 1.2). Then, the regular variation property $\mu_n \to \mu$ applied on the sets $V_{r,A}$ gives

$$\lim_{n\to\infty} n\mathbb{P}(|\mathbf{X}| > a_n r,\, \mathbf{X}/|\mathbf{X}| \in A) = \mu(V_{r,A})\,, \tag{1.2.20}$$

as soon as the set $V_{r,A}$ is a $\mu$-continuity set. The sets $V_{r,A}$ provide a useful characterization of regular variation in terms of polar coordinates. Indeed, Equation (1.2.20) may be seen as the combination of two convergences, a radial one via the term $|\mathbf{X}| > a_n r$, and an angular one via the term $\mathbf{X}/|\mathbf{X}| \in A$.

Note that the homogeneity property implies that $\mu(V_{r,A}) = r^{-\alpha}\mu(V_{1,A})$. It encourages to consider $\mu(V_{1,A})$ as a measure on the positive unit sphere $\mathbb{S}_+^{d-1}$ evaluated in $A$. In particular, with $r = 1$ and $A = \mathbb{S}_+^{d-1}$ we obtain that

$$n\mathbb{P}(|\mathbf{X}| > a_n) \to \mu(V_{1,\,\mathbb{S}_+^{d-1}})\,, \quad n \to \infty\,.$$

As already explained, it is sometimes convenient to choose an sequence $(a_n)$ such that $\mu(V_{1,\,\mathbb{S}_+^{d-1}}) = 1$, or equivalently $n\mathbb{P}(|\mathbf{X}| > a_n) \to 1$. In this case, the measure $S(\cdot)$ defined on the Borel $\sigma$-sets of $\mathbb{S}_+^{d-1}$ by $S(A) = \mu(V_{1,\,A})$ is a probability measure. Otherwise, we obtain a finite measure on $\mathbb{S}_+^{d-1}$ with total mass $\mu(\{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}| > 1\})$.

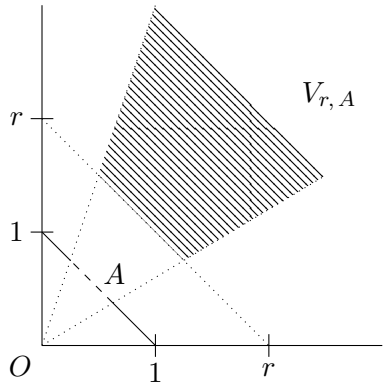



Figure 1.2: Illustration of the sets $V_{r,A}$.

To avoid dividing each terms involving the measure $S$ by $\mu(\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1\})$, we assume once and for all that $n\mathbb{P}(|\mathbf{X}| > a_n) \to 1$ when $n \to \infty$ where $|\cdot|$ is a fixed norm. This allows one to deal with a probability measure on the positive unit sphere $\mathbb{S}_+^{d-1}$. Under this assumption, it is possible to divide the left-hand side of Equation (1.2.20) by $n\mathbb{P}(|\mathbf{X}| > a_n)$ without changing the limit in the right-hand side. For $r \geq 1$ this leads to the convergence

$$\mathbb{P}(|\mathbf{X}| > ra_n,\, \mathbf{X}/|\mathbf{X}| \in A \mid |\mathbf{X}| > a_n) = \frac{n\mathbb{P}(|\mathbf{X}| > ra_n,\, \mathbf{X}/|\mathbf{X}| \in A)}{n\mathbb{P}(|\mathbf{X}| > a_n)} \to \mu(V_{r,A}) = r^{-\alpha} S(A)\,,$$

when $n \to \infty$.

**Remark 1.2.5.** Another way to define the set $V_{r,A}$ is to consider the transformation

$$\begin{array}{rccc} T & : & \mathbb{R}_+^d \setminus \{\mathbf{0}\} & \to & (0,\infty) \times \mathbb{S}_+^{d-1} \\ & & \mathbf{v} & \mapsto & (r, \boldsymbol{\theta}) = (|\mathbf{v}|, \mathbf{v}/|\mathbf{v}|)\,, \end{array}$$

and to set $V_{r,A} = T^{-1}((r,\infty) \times A)$. Then, for $A \subset \mathbb{R}_+^d$ and $r > 0$ we have the relation

$$r^{-\alpha} S(A) = \mu(V_{r,A}) = \mu\left(T^{-1}\left[(r,\infty) \times B\right]\right)\,. \tag{1.2.21}$$

This is the device developed in Beirlant et al. (2006) who even consider two different norms, one for the radial part and another for the angular part.

It can be shown that the convergence of $\mu_n(V_{r,A}) \to \mu(V_{r,A})$ is sufficient to characterize the vague convergence of measures. This leads to the following proposition which is an adapted version of Theorem 6.1 in Resnick (2007).

**Proposition 1.2.2.** *Let $\mathbf{X} \in \mathbb{R}_+^d$ be a non-negative random vector. Consider a positive sequence $(a_n)$ such that $n\mathbb{P}(|\mathbf{X}| > a_n) \to 1$ as $n \to \infty$. Then, the following assumptions are equivalent.*

1. *$\mathbf{X}$ is regularly varying with tail measure $\mu$ and tail index $\alpha$.*

2. *The following vague convergence holds:*

$$\frac{\mathbb{P}(a_n^{-1}\mathbf{X} \in \cdot)}{\mathbb{P}(|\mathbf{X}| > a_n)} \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,.$$

3. *There exist $\alpha > 0$ and a random vector $\mathbf{\Theta} \in \mathbb{S}_+^{d-1}$ such that for all $r > 0$,*

$$n\mathbb{P}(|\mathbf{X}| > a_n r,\, \mathbf{X}/|\mathbf{X}| \in \cdot) \xrightarrow{w} r^{-\alpha}\mathbb{P}(\mathbf{\Theta} \in \cdot)\,, \quad n \to \infty\,.$$

4. *There exists a random vector $\mathbf{\Theta} \in \mathbb{S}_+^{d-1}$ independent of a Pareto($\alpha$) random variable $Y$ such that*

$$\mathbb{P}((|\mathbf{X}|/a_n, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > a_n) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad n \to \infty\,. \tag{1.2.22}$$

5. *$|\mathbf{X}|$ is regularly varying with tail index $\alpha$ (in the sense of real-valued random variables) and there exist $\alpha > 0$ and a random vector $\mathbf{\Theta} \in \mathbb{S}_+^{d-1}$ such that*

$$\mathbb{P}(\mathbf{X}/|\mathbf{X}| \in \cdot \mid |\mathbf{X}| > a_n) \xrightarrow{w} \mathbb{P}(\mathbf{\Theta} \in \cdot)\,, \quad n \to \infty\,.$$

In this case, the random vector $\mathbf{\Theta}$ is called the *spectral vector* and its distribution $S(\cdot) = \mathbb{P}(\mathbf{\Theta} \in \cdot)$ is called the *spectral measure*. Equation (1.2.21) highlights the tenuous link between the spectral measure and the tail measure. Both measures have the same interpretation in terms of extreme values: They both places mass in directions where large events occur. The main difference is that the spectral measure is a probability measure whereas the tail measure is not. Therefore, it is sometimes more convenient to work with the first one.

Proposition (1.2.2) provides useful characterizations of regular variation in terms of radial and angular convergences. The point 3 allows one to look separately to the radial part $|\mathbf{X}|$ and the angular one $\mathbf{X}/|\mathbf{X}|$. Moreover, the last two convergences deal with the conditional distribution of $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > a_n$. In terms of extremes, this allows one to study the angular behavior of $\mathbf{X}$ conditioned on the event that $\mathbf{X}$ is extreme.

A final step to characterize multivariate regularly varying random vectors in $\mathbb{R}_+^d$ consists in replacing the sequential versions of convergence in 1.2.2 in a continuous form. This is the purpose of the following proposition.

**Proposition 1.2.3.** *Let $\mathbf{X} \in \mathbb{R}_+^d$ be a non-negative random vector. The following assumptions are equivalent.*

1. *$\mathbf{X}$ is regularly varying with tail measure $\mu$ and tail index $\alpha > 0$.*

2. *The following vague convergence holds*

$$\frac{\mathbb{P}(x^{-1}\mathbf{X} \in \cdot)}{\mathbb{P}(|\mathbf{X}| > x)} \xrightarrow{v} \mu(\cdot)\,, \quad x \to \infty\ ,. \tag{1.2.23}$$

3. *There exists a random vector* $\mathbf{\Theta} \in \mathbb{S}_+^{d-1}$ *independent of a Pareto*$(\alpha)$ *random variable* $Y$ *such that*

$$\mathbb{P}((x^{-1}|\mathbf{X}|, \mathbf{X}/|\mathbf{X}|) \in \cdot \mid |\mathbf{X}| > x) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad x \to \infty\,. \tag{1.2.24}$$

Equation (1.2.24) is the key convergence on which this thesis relies. Indeed, it provides a understandable and useful theoretical framework to study the dependence structure of extreme events, and hence to address the question **(Q2)**. Equation (1.2.24) highlights the behavior of extreme events of $\mathbf{X}$ via the conditioning $|\mathbf{X}| > t$. Moreover, it gives a decomposition of the tail measure in a radial part and an angular part. The radial component can be modeled through a random variable with a Pareto$(\alpha)$ distribution whereas the angular one is characterized by the spectral measure $S$. Moreover, both radial and angular parts are independent. From a statistical point of view, the study of the radial part boils down to the estimation of the tail index $\alpha$. Therefore, it comes down to the univariate case. Our major efforts must thus be focused on the study of the spectral measure. It is a challenging issue since this measure is non-parametric and also since the dimension of the study can be very large. The central point of this thesis is then to provide tools to estimate this measure in high dimension. Moreover, the estimation based on Equation (1.2.24) or on Equation (1.2.22) both rely on the fact that the convergences become approximation as soon as the threshold ($t$ or $a_n$) are chosen "large enough". A particular attention should be paid on this choice which is closely related to the question **(Q3)**.

In this context, a first major step us to study the support of the spectral measure. Indeed, starting for instance from Equation (1.2.24), we can notice that the spectral vector $\mathbf{\Theta}$ is the limit vector of the angular component $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > x$ for $x$ large enough. Therefore, the subspaces of the unit sphere where the spectral measure places mass correspond to the ones where extreme events generated by $\mathbf{X}$ appear. This is why most of the actual works in multivariate extremes concerns the estimation of the support of $S$.

In a recent work Lehtomaa and Resnick (2019) address the estimation of this support by using a bijective application to map the simplex $\mathbb{S}_+^{d-1}$ to the $d-1$ dimensional space $[0,1]^{d-1}$ in order to partition this image space into a grid of equally sized rectangles. Consistency results using the Haussdorf distance are established, and a particular attention is paid to the notion of asymptotic independence. In this approach, the authors assume that the marginals are tail equivalent, which means that they consider rank transformed data, but they explain that "it is not clear what effect such a transform applied to finite samples has on support estimation".

Other approaches based on statistical learning are developed in Section 1.4. Most of them partition the unit sphere into subsets on which the behavior of $\mathbf{X}$ can be easily interpreted.

#### 1.2.2.4 Some results on the spectral measure

Going back to Remark 1.2.5, Equation (1.2.21) can be rephrased as

$$\alpha r^{-(\alpha+1)} \mathrm{d}r S(\mathrm{d}\boldsymbol{\theta}) = \mu \circ T^{-1}(\mathrm{d}r, \mathrm{d}\boldsymbol{\theta})\,.$$

Hence, we obtain that

$$\int_{\mathbb{R}_+^d\setminus\{\mathbf{0}\}} f(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = \int_0^\infty \int_{\mathbb{S}_+^{d-1}} f(r\boldsymbol{\theta})\,\alpha r^{-(\alpha+1)}\mathrm{d}r\,\mathrm{d}S(\mathrm{d}\boldsymbol{\theta})\,, \tag{1.2.25}$$

for all non-negative function $f : \mathbb{R}_+^d \setminus \{\mathbf{0}\} \to \mathbb{R}$. On the other hand, the definition of the spectral measure $S$ via the relation $S(A) = \mu(V_{1,A})$ implies that

$$\int_{\mathbb{S}_+^{d-1}} f(\boldsymbol{\theta})\,\mathrm{d}S(\boldsymbol{\theta}) = \int_{|\mathbf{x}|>1} f(\mathbf{x}/|\mathbf{x}|)\,\mathrm{d}\mu(\mathbf{x})\,,$$

for all non-negative function $f : \mathbb{S}_+^d \to \mathbb{R}$.

In particular, for $\mathbf{x} > \mathbf{0}$, Equation (1.2.25) leads to

$$\begin{aligned}
\mu([\mathbf{0},\mathbf{x}]^c) &= \mu\Big(\Big\{\mathbf{u}\in\mathbb{R}_+^d,\ \max_{1\le i\le d}\frac{u_i}{x_i} > 1\Big\}\Big)\\
&= \int_{\mathbb{S}_+^{d-1}}\int_0^\infty \mathbb{1}\Big\{r\boldsymbol{\theta}\in\mathbb{R}_+^d,\ \max_{1\le i\le d}\frac{\theta_i}{x_i} > \frac{1}{r}\Big\}\,\mathrm{d}r\,\mathrm{d}S(\boldsymbol{\theta})\\
&= \int_{\mathbb{S}_+^{d-1}} \max_{1\le i\le d}\Big(\frac{\theta_i}{x_i}\Big)^\alpha \mathrm{d}S(\boldsymbol{\theta})\\
&= \mathbb{E}\Big[\max_{1\le i\le d}\Big(\frac{\theta_i}{x_i}\Big)^\alpha\Big]\,.
\end{aligned}$$

Therefore, we can express the multivariate Fréchet distribution in terms of the spectral measure or of the spectral vector:

$$H(\mathbf{x}) = \exp\Big(-\mathbb{E}\Big[\max_{1\le i\le d}\Big(\frac{\Theta_i}{x_i}\Big)^\alpha\Big]\Big) = \exp\Big(-\int_{\mathbb{S}_+^{d-1}}\max_{1\le i\le d}\Big(\frac{\theta_i}{x_i}\Big)^\alpha \mathrm{d}S(\boldsymbol{\theta})\Big)\,, \tag{1.2.26}$$

for all $\mathbf{x} > \mathbf{0}$.

All these results make the connections between the spectral measure $S$, the tail measure $\mu$, and the multivariate Fréchet distribution $H$. The two first ones are more related to an approach based on regular variation and threshold exceedances while the third one rather concerns a max-stable approach. In what follows, more emphasis will be placed on the spectral measure and on its support.

### 1.2.3 Examples of limit distributions, asymptotic independence

We develop in this section several examples for which we compute the associated quantities $\mu$, $\boldsymbol{\Theta}$, and $H$. The general idea is to start with a regularly varying random vector $\mathbf{X} \in \mathbb{R}_+^d$ whose marginals satisfy some dependence properties and to study the consequences on the limit measures $\mu$ and $S$ as well as on the multivariate Fréchet distribution $H$.

**Example 1.2.7.** For the first example, we consider a random vector $\mathbf{X} = (X, 0, \ldots, 0)^\top \in \mathbb{R}_+^d$ where $X$ is a regularly varying random variable. We fix a sequence $(a_n)$ such that $n\mathbb{P}(|\mathbf{X}|_\infty > a_n) =$

$n\mathbb{P}(X > a_n) \to 1$ when $n \to \infty$, and we consider $A_2, \dots, A_d \subset \mathbb{R}_+$. Then, for all $x > 0$, we have the convergence

$$\begin{aligned} n\mathbb{P}(a_n^{-1}\mathbf{X} \in (x,\infty) \times A_2 \times \ldots \times A_d) &= n\mathbb{P}(X > a_n x)\mathbb{1}\{0 \in A_2 \cap \ldots \cap A_d\} \\ &\to x^{-\alpha}\mathbb{1}\{0 \in A_2 \cap \ldots \cap A_d\}\,, \quad n \to \infty\,. \end{aligned}$$

Hence, the tail measure is given by the relation

$$\mu\big((x,\infty) \times A_2 \times \ldots \times A_d\big) = x^{-\alpha}\mathbb{1}\{0 \in A_2 \cap \ldots \cap A_d\}\,.$$

Then, we obtain that $H(\mathbf{x}) = \exp(-x_1^{-\alpha})$ for all $x_1 > 0$ and that the spectral measure is a Dirac mass on the first unit vector $\mathbf{e}_1 = (1, 0, \dots, 0)^\top$.

Example 1.2.7 provides models where only one coordinate contribute to the extreme behavior of the data. It also shows that the limit measure can concentrate on very small subspaces of $\mathbb{R}_+^d$. Actually, such situations are both interpretable in terms of extremes and easily calculable, so that they are widely studied.

**Example 1.2.8** (Continuation of Example 1.2.6)**.** We develop here an explicit example for which all previous quantities are calculable. We consider a random vector $\mathbf{X} \in \mathbb{R}_+^d$ with independent marginals $X_1, \dots, X_d$ which are all regularly varying with the same tail index $\alpha > 0$. Recall that in Example 1.2.6 we chose a sequence $(\tilde{a}_n)$ such that

$$n\mathbb{P}(X_1 > \tilde{a}_n x) \to x^{-\alpha}\,, \quad n \to \infty\,.$$

Now, in order to be consistent with the previous section, we consider the infinity norm $|\cdot|_\infty$, and choose a sequence $(a_n)$ such that

$$n\mathbb{P}(|\mathbf{X}|_\infty > a_n) \to 1\,, \quad n \to \infty\,.$$

Then, some short calculations lead to

$$\begin{aligned} n\mathbb{P}(|\mathbf{X}|_\infty > a_n) &= n[1 - \mathbb{P}(|\mathbf{X}|_\infty \le a_n)] \\ &= n[1 - \mathbb{P}(X_1 \le a_n)^d] \\ &= n[1 - (1 - \mathbb{P}(X_1 > a_n))^d] \\ &\sim nd\mathbb{P}(X_1 > a_n)\,, \quad n \to \infty\,. \end{aligned}$$

Hence, we obtain that $n\mathbb{P}(X_1 > a_n) \to 1/d$ when $n \to \infty$. With this normalization sequence, we obtain that

$$\mathbb{P}(a_n^{-1}M_{n,1} \le x) \to \exp\Big(-\frac{x^{-\alpha}}{d}\Big)\,, \quad n \to \infty\,,$$

for $x > 0$, see Proposition 1.2.1. Then, the independence of the marginals leads to the following

convergence:

$$\mathbb{P}(a_n^{-1}\mathbf{M}_n \leq \mathbf{x}) = \prod_{i=1}^{d} \mathbb{P}(a_n^{-1}M_{n,i} \leq x_i) \to \exp(-(x_1^{-\alpha} + \ldots + x_d^{-\alpha})/d) = H(\mathbf{x})\,, \quad n \to \infty\,, \quad (1.2.27)$$

for all $\mathbf{x} > \mathbf{0}$. Thus, $\mathbf{X}$ is regularly varying with tail index $\alpha$ and a tail measure $\mu$ satisfying

$$\mu([\mathbf{0}, \mathbf{x}]^c) = -\log H(\mathbf{x}) = (x_1^{-\alpha} + \ldots + x_d^{-\alpha})/d\,, \quad \mathbf{x} > \mathbf{0}\,.$$

Finally, following Equation (1.2.26), we obtain that

$$\mathrm{d}S = \frac{1}{d}(\delta_{\mathbf{e}_1} + \ldots + \delta_{\mathbf{e}_d})\,,$$

where $\mathbf{e}_k$ denotes the vector of $\mathbb{R}^d$ with a one in position $k$ and zeros elsewhere.

Regarding the max-stable distribution $H$, we observe that it factorizes as

$$H(\mathbf{x}) = \exp(-(x_1^{-\alpha} + \ldots + x_d^{-\alpha})) = \prod_{i=1}^{d} \exp(-x_i^{-\alpha})\,.$$

This means that the multivariate Fréchet distribution is equal to the product of its marginals. In this case, we say that the distribution of $H$ has the property of *asymptotic independence* (see de Haan and Ferreira (2006), Section 6.2). As mentioned in Example 1.2.8, it is equivalent to say that the spectral measure only places mass on the axes, i.e. on the vectors $\mathbf{e}_k$.

This situation models data for which extreme events are likely to appear only because one coordinate is large. It has been widely studied since it provides models which are interpretable regarding extreme values and for which all calculations can be done. There is an abundant literature on the subject, see e.g. Ledford and Tawn (1996), Heffernan and Tawn (2004), or Fougères and Soulier (2010).

**Example 1.2.9** (Equal marginals)**.** Assume that $\mathbf{X}$ has equal components, i.e. $\mathbf{X} = X(1, \ldots, 1)^T$ with $X \geq 0$. In this case, we obtain the following equality

$$\mathbb{P}(a_n^{-1}\mathbf{M}_n \leq \mathbf{x}) = \mathbb{P}(a_n^{-1}M_n^1(1, \ldots, 1)^T \leq \mathbf{x}) = \mathbb{P}(a_n^{-1}M_n^1 \leq \min_{i=1,\ldots,d} x_i)\,, \quad \mathbf{x} \in \mathbb{R}_+^d\,.$$

Assume now that $X$ is regularly varying, or equivalently that there exists $\alpha > 0$ such that $X \in \mathrm{MDA}(\Phi_\alpha)$. Then, if we choose $a_n$ such that $n\mathbb{P}(|\mathbf{X}|_\infty > a_n) = n\mathbb{P}(X > a_n) \to 1$ when $n \to \infty$, we obtain the following convergence:

$$\mathbb{P}(a_n^{-1}\mathbf{M}_n \leq \mathbf{x}) = \mathbb{P}(a_n^{-1}M_n^1 \leq \min_{i=1,\ldots,d} x_i) \to \Phi\Big(\min_{i=1,\ldots,d} x_i\Big) = \exp\Big(-\min_{i=1,\ldots,d} x_i^{-\alpha}\Big)\,, \quad n \to \infty\,,$$

for $\mathbf{x} > \mathbf{0}$. Hence, the tail measure $\mu$ satisfies $\mu([\mathbf{0}, \mathbf{x}]^c) = \min_{i=1,\ldots,d} x_i^{-\alpha}$. Finally, Equation (1.2.26)

implies that the spectral measure is a Dirac mass in $\mathbf{1}/|\mathbf{1}|$. Indeed, in this case we easily verify that

$$\mu([\mathbf{0},\mathbf{x}]^c) = \min_{i=1,\dots,d} x_i^{-\alpha} = \max_{i=1,\dots,d} \left(\frac{1}{x_i}\right)^{\alpha} = \int_{\mathbb{S}_+^{d-1}} \max_{1\le i\le d} \left(\frac{\theta_i}{x_i}\right) \mathrm{d}S(\boldsymbol{\theta}),$$

when $\mathrm{d}S = \delta_{\mathbf{1}/|\mathbf{1}|}$.

These different examples provide understandable models for multivariate extremes. Choosing which model is the most accurate for the given data is then a major point to tackle. Before choosing an optimal model, we need to learn the structure of the data and to highlight to trends, especially for the dependence structure of multivariate extremes. This is all the most necessary when the dimension $d$ is large. To this end, we use techniques developed in the learning community to deal with high-dimensional data.

## 1.3 High-dimensional learning: some techniques

In multivariate statistics, the difference between the finite sample size and the large dimensional space on which the probability measures are defined is referred to as the curse of dimensonality. This expression was coined by Bellman (1957) in the context of dynamic programming. In a statistical framework, this issue rises large variance for standard estimators (see Massart (1989), Donoho (2000) or Verleysen (2003)). In extreme value analysis, the estimation becomes all the more difficult since it is based on the largest observations of the sample which reduces the number of data points used. This is why estimating the spectral measure has mostly only been studied in the bivariate case (Einmahl et al. (1993), Einmahl et al. (1997), Einmahl et al. (2001), Einmahl and Segers (2009)).

In order to cope with high-dimensional data and thus to address the question **(Q1)**, the goal is to provide methods with lead to dimension reduction. Three main techniques are introduced here. The first one consists in grouping together data points that have similar behavior. The second one is Principal Component Analysis which classify the data in terms of their variance. Finally, we expose procedures which introduce sparsity in the parameters one wants to estimate.

### 1.3.1 Clustering methods

A natural approach when dealing with data is to aggregate them into groups (also called *clusters*) of variables with the same behavior. This does not naturally leads to dimension reduction but it often happens that variables which belong to the same cluster are likely to take values in a lower dimensional space than $\mathbb{R}^d$. In this case, the clustering approach boils down to the study of each group of variables separately. Each of these groups is included in $\mathbb{R}^{d'}$ with hopefully $d' \ll d$.

One of the most standard clustering techniques is the $k$-means procedure. The term *$k$-means* was first introduced by MacQueen (1967) but some ideas were already developed before, see Hans-Hermann (2008) for an historical review. The goal of the procedure is to identify clusters such that the distance of an observation to the closest cluster center is minimized. More precisely, given $n$

data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$ and a number $k$ of clusters, the $k$-means algorithm partitions the data points into $k$ sets $C_1, \ldots, C_k$ such that the quantity

$$\sum_{k=1}^{n} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \mathbf{m}_i\|_2 \tag{1.3.1}$$

is minimal, where $\mathbf{m}_i = \frac{1}{\#C_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$ is the center of the cluster $C_i$. MacQueen (1967) proposed an algorithm sometimes called "naive" since some faster versions of the algorithm have then been proposed, see for instance Pelleg and Moore (1999). Some alternative approaches close to the $k$-means procedures have then be developed where the Euclidean distance is replaced by a distance function $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, or more generally a dissimilarity function (Gan et al. (2007), Chapter 6).

When the observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ take values on the unit sphere $\mathbb{S}^{d-1}$, a natural way to evaluate the distance between two points is in terms of angular dissimilarity. To this end, the distance function $d$ is defined as

$$d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y}) = 1 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{x}||\mathbf{y}|},$$

where $\langle \cdot, \cdot \rangle$ denotes a scalar product on $\mathbb{R}^d$ associated to the unit sphere $\mathbb{S}^{d-1}$. This procedure developed by Dhillon and Modha (2001) seems appropriate for the study of the spectral measure which has mass on the unit sphere $\mathbb{S}_+^{d-1}$. It is for instance used by Chautru (2015) and Janßen and Wan (2020), see Section 1.4 for more details on these approaches.

### 1.3.2 Principal Components Analysis

In order to study dependence between components, a natural way to proceed is to consider the covariance matrix of the data, or its estimate. Several methods relying on linear transformations of the sample covariance matrix have been proposed. The most standard on is Principal Components Analysis (PCA), a statistical procedure which transforms a data matrix $\mathbf{X}$ with possibly correlated variables to another matrix with linearly uncorrelated components (Anderson (1963), Jolliffe (1986)). It relies on orthogonal transformations and highlights the components with high variances. It is widely used in explanatory data analysis and model prediction. Hastie et al. (2009) develop the ideas of PCA in a statistical learning framework and introduce the notion of sparse PCA (see Section 14.5 there).

More generally, PCA allows to identify linear subspaces on which a measure is concentrated. In this context, several theoretical results have been established regarding the reconstruction error (Blanchard et al. (2007), Koltchinskii and Giné (2000)) or the approximation error for the eigenspaces of the covariance matrix (Zwald and Blanchard (2006)). Some assumptions are required on the moments of the underlying distribution of the data.

Since PCA projects the data onto a subspace with equal of fewer dimension that the original one, it is a useful tool when dealing with high-dimensional data. Therefore, the transposition of standard

PCA to an extreme setting should reduce the dimension and highlight patterns into the extreme values of a data set. However, some precautions must be taken in such a context for mainly two reasons. The first one is that PCA focuses on the covariance matrix which summarizes the general dependence of the data. There exists no natural transposition of the covariance matrix in EVT since it is difficult to summarize the behavior of extreme events through linear transformations. The other main issue is that the assumptions required to obtain theoretical guarantees of PCA often fail in a context of EVT. In particular, since EVT highlights heavy tails, the distribution studied often do not have finite moments. We refer to Section 1.4.1 for a review of the existing methods which adapt PCA to an extreme setting.

Whereas clustering methods and PCA highlight some trends of the data, there is unfortunately no guarantees that a variables of the same group belong to a low-dimensional subset of $\mathbb{R}^d$. Hence, it seems natural to force some coordinates of the data points to be equal to zero. This is the approach developed in the following sections.

### 1.3.3 The LASSO procedure

In order to reduce the dimension it is necessary to introduce *sparsity* in the considered models. A vector $\mathbf{v} \in \mathbb{R}^d$ is all the more *sparse* if it contains a few number of nonzero coordinates, that is, if $|\mathbf{v}|_0$ is low[1]. In a statistical framework, the notion of sparsity is crucial since it brings out the most relevant parameters in the model. From a computational point of view, it makes the algorithms faster. The main approach in this context is to fit the best sparse model to the data we want to explain. This is mainly done by introducing a penalization on the number of parameters in the estimation procedure. For a review on statistical methods with sparsity see the monographs of Hastie et al. (2015) and Giraud (2014).

We take the example of linear regression in which we observe $n$ outcome variables $y_1, \dots, y_n$ and $d$ explanatory variables $\mathbf{x}_i = (x_{i,1}, \dots x_{i,d})^\top$. The goal is then to predict the outcome from the predictors. The linear regression model is given by

$$y_i = \beta_0 + \sum_{j=1}^{d} x_{i,j}\beta_j + \epsilon_i\,, \quad 1 \le i \le n\,, \tag{1.3.2}$$

where $\beta_0$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^\top$ are unknown parameters and $\epsilon_i$ is an error term. Setting $\mathbf{y} = (y_1, \dots, y_n)^\top$, $\mathbf{X} = (x_{i,j})_{1\le i\le n,\, 1\le j\le d}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$, we can rewrite Equation (1.3.2) in a matrix form:

$$\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,,$$

where $\mathbf{1} = (1, \dots, 1)^\top \in \mathbb{R}^n$ .

Then, the method of least squares provides estimates of the parameters by minimizing the func-

[1]Recall that $|\cdot|_0$ denotes the "$\ell^0$-norm" given by $|\mathbf{v}|_0 = \#\{k = 1, \dots, d,\, v_k > 0\}$.

tion

$$\sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_{j=1}^{d}x_{i,j}\beta_j\Big)^2=|\mathbf{y}-\beta_0\mathbf{1}+\mathbf{X}\boldsymbol{\beta}|_2^2\,. \tag{1.3.3}$$

The ordinary least squares estimates $\hat{\beta}_j$ from Equation (1.3.3) are almost always different from zero. If the dimension $p$ is large, it leads to a high-dimensional vector of estimates which is not easy to interpret. If $p>n$, minimizing (1.3.3) even leads to an infinite number of solutions. The idea introduced by Tibshirani (1996) is to penalized the number of nonzero parameters with a $\ell^1$-regularized regression. There, the minimization problem becomes

$$\underset{\beta_0,\beta_1,\ldots,\beta_d}{\text{minimize}}\ \sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_{j=1}^{d}x_{i,j}\beta_j\Big)^2\quad\text{subject to }|\boldsymbol{\beta}|_1\leq t\,, \tag{1.3.4}$$

or equivalently,

$$\underset{\beta_0,\beta_1\ldots,\beta_d}{\text{minimize}}\ |\mathbf{y}-\beta_0\mathbf{1}+\mathbf{X}\boldsymbol{\beta}|_2^2\quad\text{subject to }|\boldsymbol{\beta}|_1\leq t\,, \tag{1.3.5}$$

where $t$ is a tuning user-specified parameter. This technique is called *Least Absolute Shrinkage and Selection Operator* or *LASSO* in an abbreviated form.

**Remark 1.3.1.** As explained in Hastie et al. (2015), this approach relies on the specificity of the $\ell^1$-norm. For $q<1$, the optimization problem (1.3.4) with $|\boldsymbol{\beta}|_q$ instead of $|\boldsymbol{\beta}|_1$ is not convex while for $q>1$ it does not provide sparse estimates. Therefore, the choice of the $\ell^1$-norm combines both computational convenience (the problem is convex) and statistical interpretations (the number of non-zero parameters is low). When $q=2$, the optimization problem becomes

$$\underset{\beta_0,\ldots,\beta_d}{\text{minimize}}\ \sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_{j=1}^{d}x_{i,j}\beta_j\Big)^2\quad\text{subject to }|\boldsymbol{\beta}|_2\leq t\,,$$

which corresponds to Ridge regression introduced by Hoerl and Kennard (1970). In this case, the procedure shrinks the coefficients but does not set any of them to 0.

In terms of Lagrangian, Equation (1.3.4) can be rephrased as follows:

$$\underset{\beta_0,\ldots,\beta_d}{\text{minimize}}\ \sum_{i=1}^{n}\Big(y_i-\beta_0-\sum_{j=1}^{d}x_{i,j}\beta_j\Big)^2+\eta\big(|\boldsymbol{\beta}|_1-t\big)\,, \tag{1.3.6}$$

where $\eta\in\mathbb{R}$ is a Lagrange multiplier. The solutions to Equation (1.3.4) are shown to be

$$\beta_j^*=\text{sign}(\hat{\beta}_j)(\hat{\beta}_j-\eta)_+\,,$$

where $\eta$ is the parameter of the Lagrangian problem (1.3.6) and $\hat{\beta}_j$ corresponds to the estimator of $\beta_j$ in the ordinary least squares problem (1.3.3).

### 1.3.4 Euclidean projection onto the $\ell^1$-ball

An extension of Equation (1.3.5) is given by the following optimization problem. For a vector $\mathbf{v} \in \mathbb{R}^d$ consider the following minimization:

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}|\mathbf{w}-\mathbf{v}|_2^2 \quad \text{subject to } |\mathbf{w}|_1 \leq t\,. \tag{1.3.7}$$

This problem has been widely studied in learning theory, see e.g. Duchi et al. (2008), Kyrillidis et al. (2013), or Liu and Ye (2009). It corresponds to the Euclidean projection of the vector $\mathbf{v}$ onto the $\ell^1$-ball of radius $z$. As explained in Duchi et al. (2008), Section 4, if $\|\mathbf{v}\|_1 \leq z$, then the solution of Equation (1.3.7) is $\mathbf{w} = \mathbf{v}$. It is therefore sufficient to assume that the vector $\mathbf{v}$ satisfies $\|\mathbf{v}\|_1 > z$ and in this case the solution is on the $\ell^1$-sphere $\{\mathbf{x} \in \mathbb{R}^d,\ |\mathbf{x}|_1 = t\}$. Hence, the constraint $|\mathbf{w}|_1 \leq t$ in Equation (1.3.7) can be replaced by the equality constraint $|\mathbf{w}|_1 = t$. Moreover, the projected vector $\mathbf{w}$ satisfies $v_i w_i \geq 0$ for all $i = 1, \dots, d$, see Duchi et al. (2008), Lemma 3. Due to symmetry considerations, we can thus restrict our study to non-negative vectors $\mathbf{v} \in \mathbb{R}^d_+$. There, the minimization problem (1.3.7) becomes

$$\underset{\mathbf{w}\in\mathbb{R}^d}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}-\mathbf{v}\|_2^2 \quad \text{subject to } \sum_{i=1}^{d} w_i = t, \quad w_i \geq 0\,. \tag{1.3.8}$$

Given a vector $\mathbf{v}$ and a threshold $t$, many algorithms which compute the optimal solution $\mathbf{w}$ have been proposed. The main one was the one introduced by Gafni and Bertsekas (1984) and Bertsekas (1999), and then rediscovered by Crammer and Singer (2002) in the context of support-vector machines and by Hazan (2006) in the context of online convex programming. This algorithm starts with the sorting of the coordinates of the vector $\mathbf{v}$ which requires $O(n \log(n))$ time. Then, based on the sorted vector, the projection can be calculated in exactly linear time. Hence, the complexity of this algorithm is $O(n \log(n))$. Several refinements of this procedure have then been proposed. Duchi et al. (2008) use the randomized median finding algorithm introduced by Cormen et al. (2001) to build an efficient algorithm which computes the projected vector $\mathbf{w}$ in expected linear time. Other algorithms have been proposed, see Condat (2016) for a review of the existing methods.

The number of articles dealing with the Euclidean projection onto the simplex has increased significantly in the recent years. In learning theory, this approach is for instance used in portfolio optimization (see Brodie et al. (2009)) and in imaging problems, such as segmentation (see Lellmann et al. (2009)) and multispectral unmixing (see Bioucas-Dias et al. (2012)). The main goal in this fields is to introduce sparsity in the data in order to reduce the dimension. The purpose of Chapter 2 is to develop a similar idea for extreme values. The sparsity is crucial since it allows to work on lower-dimensional subspaces. In this context, the linear complexity of the algorithm developed by Duchi et al. (2008) in order to handle high-dimensional data will be a crucial point. Besides, the number of null-coordinates for a projected vector is deeply linked to the choice of the sphere considered, that is, to $t$. Therefore, sparsity and choice of a threshold $t$ are two issues that can be

tackled simultaneously. This is the main idea of Section 3.5 in which we address both questions **(Q1)** and **(Q3)**.

In terms of extreme values, the first two methods (clustering and PCA) have been recently used to study multivariate extremes. The goal of the following section is to make a literature review regarding both approaches.

## 1.4 Learning for extremes

The study of extreme events in high dimension is a challenging issue particularly because we focus only on the largest values and thus reduce the size of the data used. Therefore, it is necessary to provide accurate methods which highlight the main trend in the tail dependence structure. To this end, the knowledge of the spectral measure is crucial since it gathers almost all information regarding dependence between extreme events. The inference of the spectral measure is an issue that has been widely studied in a low-dimensional framework but the high-dimensional setting has been addressed only recently. To the best of our knowledge, the first work which deals with the estimation of a high-dimensional spectral measure is the one by Chautru (2015). This article tackles the problem of exhibiting groups (hopefully of lower size) of asymptotically dependent variables. The approach is divided into two steps, the first one to circumvent the issue of high dimension, the second one to cluster the data. More precisely, the first step relies on the Principal Nested Spheres (PNS) technique developed by Jung et al. (2012) and consists of an iterative projection of the data on smaller sub-spheres. This approach is close to PCA. In the second step, the projected data are classified through a spherical $k$-means procedure Dhillon et al. (2002). Subsequently, both ideas, i.e. matrix analysis close to PCA and a clustering techniques, have then been studied by other authors.

### 1.4.1 Principal Component Analysis

As already explained in Section 1.3.2, adapting PCA to the study of extreme events requires some precautions since EVT does not provide a linear framework to apply matrix transformation. This explains why this approach has only been tackled very recently.

Cooley and Thibaud (2019) summarize the tail dependence structure via a matrix of pairwise tail dependence metrics which has similarity with the covariance matrix. Unfortunately, PCA relies on linear algebra techniques, while non-negative regularly varying random vectors are part of $\mathbb{R}_+^d$. To tackle this issue, the authors define a bijective transformation which maps $\mathbb{R}_+$ to $\mathbb{R}$ as well as a vector space structure on $\mathbb{R}_+^d$ based on this transformation, which preserves regular variation. The eigendecomposition of the matrix of pairwise tail dependence metrics allows then interpretation of the dependence via the eigenbasis. However, this work is restricted to pairwise summaries and does not provide information on larger groups of variables.

Subsequently, Sabourin and Drees (2019) focus on the study of the support of the tail measure $\mu$ (see Section 1.2.2.2). They assume that the vector space spanned by this support has dimension

$p < d$ and the main goal is then to find this linear subspace, which encourages the use of PCA. However, since the data may not satisfy some finite moment assumptions the authors work with re-scaled data on which they establish consistency results. A common choice for the scaling function is to divide the vectors by their norms, which boils down to a study of unit vectors. In this context the authors obtain finite sample bounds on the reconstruction error.

### 1.4.2 Clustering approaches for extremes

One of the challenging issues of multivariate EVT being the high-dimensional setting, some clustering approaches have been recently used to reduce the dimension. The general idea of the proposed method is to determine which subsets of variables can take their largest values simultaneously while the others are of smaller order. A naive use of the data $\mathbf{X}$ to study the spectral vector $\boldsymbol{\Theta}$ does not lead to efficient results. This is why some techniques have been introduced so that a clustering of the data corresponds to groups of components with tail dependence.

One of the first approach proposed in this regard is the one by Goix et al. (2016) who focus on the subsets $R_{\beta,\infty}$ defined by

$$R_{\beta,\infty} = \left\{\mathbf{x} \geq \mathbf{0},\ |\mathbf{x}|_\infty \geq 1,\ x_i > 0 \text{ for } i \in \beta,\ x_i = 0 \text{ for } i \notin \beta\right\}, \tag{1.4.1}$$

for $\beta \subset \{1, \ldots, d\}$ and $\epsilon > 0$. The quantity $\mu(R_\beta)$ is then approximated by $\mu(R^\epsilon_\beta(1))$, where the "truncated cones" $R^\epsilon_\beta$ are defined as

$$R^\epsilon_{\beta,\infty} = \left\{\mathbf{v} \geq 0, |\mathbf{v}|_\infty \geq 1,\ v_i > \epsilon|\mathbf{v}|_\infty \text{ for } i \in \beta,\ v_i \leq \epsilon|\mathbf{v}|_\infty \text{ for } i \notin \beta\right\}, \tag{1.4.2}$$

The authors choose the infinity norm and base their work on the approximation $\mu(R^\epsilon_{\beta,\infty}) \approx \mu(R_\beta)$ for $\epsilon$ small enough without giving theoretical guarantees. An algorithm called DAMEX (for *Detecting Anomaly with Multivariate EXtremes*) with complexity $O(dn\log(n))$ is provided, where $n$ corresponds to the number of data points.

Subsequently, Goix et al. (2017) slightly modify the previous approach by defining "$\epsilon$-thickened rectangles" in the following way:

$$R'^\epsilon_{\beta,\infty} = \left\{\mathbf{v} \geq 0,\ |\mathbf{v}|_\infty \geq 1,\ v_i > \epsilon \text{ for } i \in \beta,\ v_i \leq \epsilon \text{ for } i \notin \beta\right\}. \tag{1.4.3}$$

The authors propose to estimate the quantity $\mu(R_\beta)$ by

$$\mu_n(R'^\epsilon_{\beta,\infty}) = \frac{1}{k}\sum_{j=1}^{n} \mathbb{1}\{\hat{\mathbf{V}}_j \in (n/k)R^\epsilon_\beta(2)\},$$

where $k = k_n$ satisfies $k \to \infty$, $k/n \to 0$ when $n \to \infty$, and $\mathbf{V}_j$ is the empirical standardized variable of $\mathbf{X}_j$. Theoretical guarantees of this method are given via finite-sample error bounds. An application of the DAMEX algorithm (with the rectangles instead of the cones) to anomaly detection

provides good results. However, no precision is given on the choice of the hyperparameter $\epsilon$ and of the level $k$.

The sparse representation provided by Goix et al. (2017) may lead to a very large number of subsets $R_{\beta,\infty}$. The idea proposed by Chiapino and Sabourin (2016) is then to relax the condition "$v_i \leq \epsilon$ for $i \notin \beta$" in Equation (1.4.3). There, the authors provide an incremental-type algorithm (CLustering Extreme Features, CLEF) to group together components which may be large together. This algorithm is based on the DAMEX algorithm and also requires an hyperparameter $\kappa_{\min}$ which must be "chosen in 'stability regions' of relevant summaries of the output" the authors specify. Several variants of the CLEF algorithm have then been proposed by Chiapino et al. (2019). These approaches differ in the stopping criteria which are based on asymptotic results of the coefficient of tail dependence.

Always in the idea of classifying extreme events, Janßen and Wan (2020) propose an approach based on $k$-means clustering. The authors adapt the spherical $k$-means algorithm to the extremal setting and construct a non-parametric estimator for the theoretical cluster centers. A consistency result is provided, as well as numerical experiments on data examples. A major point not addressed in this article is the choice of $k$ which is a priori unknown.

In order to group together components that are likely to be simultaneously extreme, Simpson et al. (2019) focus on hidden regular variation, a concept introduced by Resnick (2002). The authors provide a set of parameters $(\tau_\beta(\delta))_{\beta\subset\{1,\dots,d\}}$, depending on a parameter $\delta \in [0,1]$, and which characterizes hidden regular variation on the family of cones

$$\mathbb{E}_\beta = \{\mathbf{x} \in [0,\infty]^d \setminus \{\mathbf{0}\},\, x_j \in (0,\infty] \text{ for } j \in \beta,\, x_j = 0 \text{ for } j \notin \beta\}\,.$$

The family $(\tau_\beta(\delta))_{\beta\subset\{1,\dots,d\}}$ reveals whether the measure $\mu$ places mass on the cone $\mathbb{E}_\beta$. The authors assume that there exists a $\delta^* < 1$ such that $\tau_\beta(\delta^*) = 1$ for all $\beta$ such that $\mu(\mathbb{E}_\beta) > 0$, and $\tau_\beta(\delta^*) < 1$ for all $\beta$ such that $\mu(\mathbb{E}_\beta) = 0$. The goal is then to estimate the parameters $\tau_\beta(\delta)$ in order to classify the cones $\mathbb{E}_\beta$.

These different approaches show a general method for the study of multivariate extremes. The idea is to identify some interpretable subsets of $\mathbb{S}_+^{d-1}$ (respectively $\mathbb{R}^d + \setminus\{\mathbf{0}\}$) on which the spectral measure (respectively the tail measure) places mass. In this context, the idea developed in this thesis in to consider the $\ell^1$-norm and to focus on the study of the angular components of multivariate extremes through a Euclidean projection onto the simplex. This method leads to sparse projected vectors which is a key point in order to work in high dimension. With this new approach called *sparse regular variation*, we lose the independence between the angular and the radial components (see Equation(1.2.24)) and there is no notion of homogeneity anymore. However, we highlight the dependence which arises between the radial and the angular components (Proposition 2.4.1) and we prove that under mild assumptions sparse regular variation and standard regular variation are equivalent notions (see Theorem 2.4.1).

Regarding the positive unit sphere $\mathbb{S}_+^{d-1}$ which contains the support of the spectral measure, an idea similar to the one of Simpson et al. (2019) and Goix et al. (2017) is to consider the following subsets

$$C_\beta = \left\{\mathbf{x} \in \mathbb{S}_+^d,\, x_i > 0 \text{ for } i \in \beta,\, x_i = 0 \text{ for } i \notin \beta\right\}. \tag{1.4.4}$$

for $\beta \in \mathcal{P}_d^*$. By construction, the subsets $C_\beta$ form a partition of the positive unit sphere $\mathbb{S}_+^{d-1}$. Regarding the question **(Q1)**, this partition allows to deal with high-dimensional data. Indeed, for $\beta \in \mathcal{P}_d^*$ with cardinality $b$, the subset $C_\beta$ can be seen as part of the sphere $\mathbb{S}_+^{b-1}$. Therefore, as soon as $b$ is moderate compared to $d$, the use of $C_\beta$ reduces the dimension of the study. The idea is then to provide methods to learn on which of these subsets the spectral measure puts mass. This is the idea developed in Chapter 2 and Chapter 3.

All the approaches developed in this section rely on asymptotic results of multivariate random vectors (see Proposition 1.2.2 and Proposition 1.2.3). However, in a statistical context we only have a finite data set at our disposal. Therefore, the convergences that appear in the aforementioned propositions become approximation. In particular, Equation (1.2.24) can be used to study the behavior of the spectral vector $\mathbf{\Theta}$ as soon as the threshold $t$ is "large enough". This is way a particular attention should be paid on the choice of this threshold $t$, or equivalently on the number of data considered to be extreme (see the question **(Q3)**). One way to deal with this issue is to use model selection to identify for which threshold $t$ the approximation is the more accurate.

## 1.5 Model selection

We develop in this section some classical aspects of model selection. As pointed out by Birgé and Massart (2001), "choosing a proper parameter set is a difficult task in many estimation problems. [...] Both excessively complicated or oversimplified models should be avoided. The dilemma of the choice between many possible models, of one which is adequate for the situation at hand, depending on both the unknown complexity of the true parameter to be estimated and the known amount of noise or number of observations, is often a nightmare for the statistician."

The first part of this section is devoted to the general setting of model selection. We particularly insist on the context of density estimation, which will be used in Chapter 3. Then, we develop the notion of penalization which enables not to choose a too large model. In this context, we discuss the existing works regarding model selection for extremes and expose how the general theory can be used for selecting an optimal threshold.

### 1.5.1 General framework

We start with the general framework of model selection introduced for instance by Massart (2007). Consider i.i.d. random variables $\xi_1, \ldots, \xi_n$ (which can be random variables, random vectors, or random processes) with unknown distribution and assume that this distribution depends on some

quantity $s \in \mathcal{S}$. The general goal in statistics is then to infer on the unknown parameter $s$ based on the sample $\xi_1, \dots, \xi_n$. Regarding model selection, the idea is to consider a subset $S$ of $\mathcal{S}$ (called a model) and to provide an estimator of $s$ in $S$ which is the closest among all parameters in $S$. This first step is already not so easy since it often appears that no natural choice of $S$ arises. While this choice should be done so that the true parameter $s$ is close to $S$, taking a too large subset $S$ does not provide good results (Bahadur (1958), Birgé and Massart (1993)).

**Example 1.5.1** (Density estimation)**.** For instance, consider $\xi_1, \dots, \xi_n$ with a common unknown density $s$ with respect to a given measure $\mu$. In this case, the set $\mathcal{S}$ corresponds to the set of all probability densities with respect to $\mu$. If $\mu$ is the counting measure on $\mathbb{N}$, then $\mathcal{S}$ denotes all discrete probability distributions.

Consider now an empirical criterion $\gamma_n$ based on the observations $\xi_1, \dots, \xi_n$ such that

$$s = \arg\min_{t \in \mathcal{S}} \mathbb{E}[\gamma_n(t)] \,.$$

In this case, we say that $\gamma_n$ is an empirical contrast. Then, a *minimum contrast estimator* $\hat{s}$ of $s$ is a minimizer of $\gamma_n$ over $S$. In order to prove that an empirical criterion is an empirical contrast, it is often convenient to show that the associated loss function

$$l(s,t) = \mathbb{E}[\gamma_n(t)] - \mathbb{E}[\gamma_n(s)]$$

is non-negative for all $t \in \mathcal{S}$.

**Example 1.5.2** (Continuation of Example 1.5.1)**.** Coming back to the example of density estimation, a standard choice is

$$\gamma_n(t) = -\log(t) \,,$$

for $t \in \mathcal{S}$ ($t$ is a function!). We obtain the maximum likelihood criterion. The associated loss function is then given by the Kullback-Leibler divergence $K(s,t)$:

$$l(s,t) = K(s,t) \,. \tag{1.5.1}$$

The Kullback-Leibler divergence computes the distance between two distributions $P$ and $Q$ (Kullback and Leibler (1951)). If $P$ has density $p$ and $Q$ density $q$ with respect to a measure $\mu$, then the divergence is given by

$$K(p,q) = \int p \log\left(\frac{p}{q}\right) \mathrm{d}\mu \,.$$

Regarding our context, we can rephrase the loss function in the following way:

$$l(s,t) = K(s,t) = \mathbb{E}_{\xi \sim s}\left[\log\left(\frac{s(\xi)}{t(\xi)}\right)\right] .$$

Since the distribution $s$ is fixed (but unknown), the goal is to minimize the quantity

$$-\mathbb{E}_{\xi\sim s}[\log t(\xi)]\,,$$

among all $t \in S$.

### 1.5.2 Penalization

We consider now a finite collection of models $(S_m)_{m\in\mathcal{M}}$ and an empirical contrast $\gamma_n$. For each of this model, we can compute a minimum contrast estimator $\hat{s}_m$ and the goal is to select the "best" of theses estimators. We would like to choose this estimator by minimizing the risk $\mathbb{E}[l(s,\hat{s}_m)]$ but this quantity depends on the parameter $s$ and is thus unknown. The main idea to circumvent this issue is to consider a penalization $\text{pen} : \mathcal{M} \to \mathbb{R}_+$ and to minimize over $\mathcal{M}$ the penalized quantity

$$\gamma_n(\hat{s}_m) + \text{pen}(m)\,. \tag{1.5.2}$$

This leads to an optimal parameter $\hat{m}$ and then a selected model $S_{\hat{m}}$ and a selected estimator $\hat{s}_{\hat{m}}$.

In the context of penalized log-likelihood with density estimation (see Example 1.5.1), this approach has been widely studied. The main one consists in choosing a penalization equal to $D_m/n$, where $D_m$ is the number of parameters of the model $S_m$ (Akaike (1973)). This method leads to good empirical results whereas no theoretical guarantees have been provided so far.

Roughly speaking, the penalization in Akaike (1973), called *Akaike Information Criterion* (AIC) allows not to consider models with too many parameters. Indeed, models with a high number of parameters necessarily lead to good estimations but can not be well interpreted. On the other hand, choosing to few parameter provides models that are too far from the true one. This implies that a balanced choice should therefore be done. A good compromise is to penalize the minimization criterion with the number of parameters. Hurvich and Tsai (1989) propose to slightly modify the AIC criterion by taking adding a penalization based on the sample size $n$ and the number of parameters $D_m$. This leads to the $\text{AIC}_\text{c}$ criterion which is used for small sample. Other penalization criterion in density estimation have been provided, for instance by Beran (1977) where the minimization relies on the Hellinger distance. We refer to Massart (2007), Chapter 7, for an overview of the different techniques density estimation via model selection.

Regarding density estimation, several results related to the Kullback-Leibler divergence have been established. Since this divergence is not symmetric, a main issue is to study what happens where we choose $l(s,t)$ or $l(t,s)$ is Equation (1.5.1). This has been for instance analyzed by Seghouane and Bekara (2004) which shows that the standard approach that is proposed here may better reflect the error due to overfitting, while the opposite one may better reflect the error due to underfitting. Subsequently, Seghouane and Amari (2007) focus on a way to symmetrize the divergence between the true model and the approximating candidate model. The symmetrization procedure is done with operations like average, geometric, and harmonic means.

### 1.5.3 Model selection in EVT

In the multivariate setting, Kiriliouk et al. (2019) use Akaike information to compare the different multivariate Pareto models. Subsequently, Engelke and Hitz (2020) use this approach to propose model selection for extremal graph structures. In spatial extremes, Ribatet (2013) focus on model selection for max-stable process. Alternatives to AIC is used, such as Takeuchi's information criterion or Bayesian information criterion.

Regarding the question **(Q3)**, that is, the choice of the threshold $u$ in (1.2.24), several works have been done for marginal threshold selection, see for instance Caeiro and Gomes (2015) or the review of Scarrott and MacDonald (2012). On the other hand, rather few articles focus on dependence model. We can cite the work of Lee et al. (2015) who deal with Bayesian threshold selection based on measure of surprise.

One way to deal with the choice of the threshold is to reverse the problem and to rather focus on $k = k_n = n\mathbb{P}(|\mathbf{X}| > u)$ which defines a *level*, i.e. the number of exceedances. Selecting an appropriate $k$ means obtaining a balanced choice between having enough data and remaining in the extremal framework. To this end, it seems natural to consider models with different levels $k$ and to choose the most accurate one. But such an approach leads to a comparison of models with different sizes. Indeed, each model concerns only the $k$ largest values of the original data set of size $n$. Hence, it seems natural to take the non-extreme values into account and to deal with the whole data set. In this case, a partition between extreme and non-extreme values is done and the model selection should provide which partition is the most appropriate to the data. This approach is tackled in Chapter 3.

In this context, the natural additive penalization given in Equation (1.5.2) does not apply. Indeed, the model selection is done on a parameter $k$ which counts the number of data in a given subgroup (actually the extreme values). Since this parameter is expressed as a proportion of the data size, the standard procedure are not helpful. We need to adapt Akaike's method to our context, which leads to a multiplicative penalization (see Chapter 3).

This approach is all the more justified by some particular properties of the projection $\pi$ which highlight the impact of the threshold $t$ regarding the sparsity of the projected vectors (Lemma 2.2.2). These results motivates the joint study of the threshold with the sparse structure of extreme values. Moreover, they allow to compute the projected vectors for any given threshold $t$, which reduces the computational cost of our approach. We apply for instance our method on large sample of dimension $d = 10^2$ and obtain good results with a low computing time. Regarding our question **(Q1)**, the numerical examples proposed are promising in order to tackle the large dimension.

## 1.6 Outline of the thesis

This thesis is devoted to modeling and estimating dependence for high-dimensional extremes. Its goal is to tackle the estimation of the tail dependence of a regularly varying random vector $\mathbf{X}$ in $\mathbb{R}_+^d$ by addressing the two issues already mentioned in the previous paragraphs: the curse of

dimensionality and the choice of the threshold. To deal with this point, the idea is to divide the study of extremes in $\mathbb{R}_+^d$ into the lower dimensional subspaces $C_\beta$ defined in Equation (1.4.1). Since standard regular variation fails to capture the behavior of the spectral measure on these susbsets, we propose a method based on the Euclidean projection onto the simplex for which several algorithms are known (see the review of Condat (2016)). These procedures have the key property to have an expected linear complexity which is a crucial point in order to deal with high-dimensional data. Moreover, this projection manages to reduce the dimension by introducing sparsity in the considered vectors. These considerations lead to the definition of sparse regular variation and are the purpose of Chapter 2. In the end of this chapter, we provide numerical evidence of our theoretical findings and compare our method with a recent one developed by Goix et al. (2017). This latter requires a hyperparameter $\epsilon$ in order to introduce sparsity while this is done directly by the projection in our setting.

The theoretical context of sparsely regularly varying random vectors being defined, the second step is to develop a statistical framework to study the tail dependence. The idea is to provide a learning approach to identify on which subspaces $C_\beta$ extreme events appear. Since the number of $C_\beta$ is relatively high, we use model selection to highlight only the most relevant ones. We use the ideas developed in Section 1.5 by focusing on the Kullback-Leibler divergence minimization of multinomial models. This optimization takes into account the choice of an optimal threshold above which the data are considered as extreme. Chapter 3 ends with some examples on simulated data that model both asymptotic independence and extreme dependence. These numerical results illustrate the relevance of our proposed approach, especially when the dimension is large (of the order of $10^2$). The algorithm succeeds in identifying the different directions of the space on which extreme events are concentrated. . The results we obtain show that the approach based on the notion of sparse regular variation seems quite robust, especially when the size of the data varies.

The dependence structure of multivariate extremes being analyzed, some other aspects as conditional independence can be investigated. This is done is Chapter 4 which consists in a discussion on the article of Engelke and Hitz (2020). In this paper, the authors introduce the notion of conditional independence for threshold exceedances. We discuss the assumptions made in this article and provide another approach to define conditional independence for a multivariate Pareto distribution. This approach relies on the minimum of the marginals of a regularly varying random vector for which some results are established. It provides accurate models for strong extremal dependence. In this context, no sparsity can be introduced so that sparse regular variation is not helpful. Therefore, the approach developed in this chapter is a good complement to the one introduced in Chapter 2 and 3.

# Chapter 2

# Sparse regular variation

**Abstract**

Regular variation provides a convenient theoretical framework to study large events. In the multivariate setting, the dependence structure of the positive extremes is characterized by a measure - the spectral measure - defined on the positive orthant of the unit sphere. This measure gathers information on the localization of extreme events and is often sparse since severe events do not simultaneously occur in all directions. Unfortunately, it is defined through weak convergence which does not provide a natural way to capture this sparsity structure. In this chapter, we introduce the notion of sparse regular variation which allows to better learn the dependence structure of extreme events. This concept is based on the Euclidean projection onto the simplex for which efficient algorithms are known. We show several results for sparsely regularly varying random vectors and prove that under mild assumptions sparse regular variation and regular variation are two equivalent notions. Finally, we provide numerical evidence of our theoretical findings and compare our method with a recent one developed by Goix et al. (2017).

***Keywords***— Euclidean projection onto the simplex, high dimension, multivariate extremes, regular variation, sparse regular variation, spectral measure

**Regarding our questions**

**(Q1)** This chapter introduces the concept of sparse regular variation which allows to study the dependence structure of extreme events on lower-dimensional subspaces. One of the main results concerns the behavior of a regularly varying random vector $\mathbf{X}$ on the subsets $C_\beta$ defined in (1.4.1).

**(Q2)** Theorem 2.4.1 states that under mild assumption sparse regular variation is equivalent to standard regular variation. Therefore, the dependence structure of extreme values can be studied with this new concept, for which several theoretical results are given.

**(Q3)** Regarding the threshold, the study is conducted with the condition $|\mathbf{X}|_1 > t$. The $\ell^1$-norm is used in order to apply the Euclidean projection onto the simplex. We do not discuss the

selection of the threshold in this chapter.

**Contents**



## 2.1 Introduction

Estimating the dependence structure of extreme events has proven to be a major issue in many applications. The standard framework in multivariate Extreme Value Theory (EVT) is based on the concept of regularly varying random vectors (see Section 1.2.2.2). Regular variation has first been defined in terms of vague convergence on the compactified space $[-\infty,\infty]^d$ and several characterizations have subsequently been established, see e.g. Resnick (1987), Resnick (2007), Beirlant et al. (2006), or Embrechts et al. (2013). Hult and Lindskog (2006) extend the notion of regular variation on a general (possibly infinite dimensional) metric space. They introduce the concept of $M_0$-convergence of Borel measures which is based on bounded continuous test functions with support bounded away from the origin.

In this chapter, we use Resnick's setting and define multivariate regular variation through the convergence of the radial and polar coordinates of a random vector (see Resnick (1987), Proposition 5.17 and Corollary 5.18, or Resnick (2007), Theorem 6.1). A random vector $\mathbf{X} \in \mathbb{R}_+^d$ is said to be regularly varying with tail index $\alpha > 0$ and spectral measure $S$ on the positive orthant $\mathbb{S}_+^{d-1}$ of the unit sphere if

$$\mathbb{P}\left(|\mathbf{X}| > tx, \mathbf{X}/|\mathbf{X}| \in B \mid |\mathbf{X}| > t\right) \to x^{-\alpha} S(B)\,, \quad t \to \infty\,, \tag{2.1.1}$$

for all $x > 0$ and for all continuity set $B \in \mathbb{S}_+^{d-1}$ of $S$. This convergence (2.1.1) can be interpreted as follows: The limit of the radial component $|\mathbf{X}|/t$ has a Pareto distribution with parameter $\alpha > 0$, while the angular component $\mathbf{X}/|\mathbf{X}|$ has limit measure $S$. Moreover, both components of the limit are independent. The measure $S$, called the *spectral measure*, summarizes the tail dependence of the regularly varying random vector $\mathbf{X}$. Note that the choice of the norm in (2.1.1) is arbitrary. Actually, it is even possible to choose two different norms for the radial and angular parts (see Beirlant et al. (2006), Section 8.2.3).

Based on convergence (2.1.1), several non-parametric estimation techniques have been proposed to estimate $S$. In the bivariate case, some useful representations of the spectral measure has been introduced by Einmahl et al. (1993), Einmahl et al. (1997), Einmahl et al. (2001) and Einmahl and Segers (2009). In Einmahl et al. (1997), the authors replace the tails of the marginals by fitted Pareto tails in order to estimate $S$ by an empirical measure. The latter is consistent and asymptotically normal under suitable assumptions. Einmahl and Segers (2009) focus on the choice of the $\ell^p$-norm, for $p \in [1, \infty]$, in order to construct an estimator of the spectral measure which satisfies moment constraints. Inference on the spectral measure has also been studied in a Bayesian framework, for instance by Guillotte et al. (2011). In this chapter, the authors use censored likelihood methods in the context of infinite dimensional spectral measures. Parametric approaches have also been introduced to tackle the study of extremes in moderate ($d \leq 10$) dimensions, for instance by Coles and Tawn (1991) and Sabourin et al. (2013).

In higher dimensions, mixtures of Dirichlet distributions are often used to model the spectral densities. Boldi and Davison (2007) show that under some conditions these distributions are weakly dense in the set of spectral measures. They propose both frequentist and Bayesian inferences based on EM algorithms and MCMC simulations. Subsequently, Sabourin and Naveau (2014) introduce a re-parametrization of the Bayesian Dirichlet mixture model.

More recently, the study of the spectral measure's support has become an active topic of research. Indeed, this support gathers information on the dependence structure of extreme values: The subspaces on which the spectral measure puts mass correspond to these where extreme events occur. Thus, estimating the spectral measure is a major issue in multivariate EVT but it is a challenging problem, especially in high dimensions. Unfortunately, the complete support's estimation is often difficult to capture, so that a main goal in the tail dependence's study is rather to identify clusters of components which are likely to be extreme together. This approach has firstly been introduced by Chautru (2015) who uses a clustering technique to exhibit groups of variables with asymptotic dependence. In the same way, Janßen and Wan (2020) use spherical $k$-means in order to find clusters with the same extremal behavior.

Recently, two approaches based on Principal Component Analysis (PCA) for high-dimensional extremes have been developed. Cooley and Thibaud (2019) define a vector space on the positive orthant $\mathbb{R}_+^d$ in order to conciliate both PCA and regular variation. They summarize the tail dependence through a matrix of pairwise tail dependence metrics and apply some usual decomposition on this matrix. They illustrate their approach with simulations on Swiss rainfall data and financial

return data. In a recent work, Sabourin and Drees (2019) assume that the spectral measure concentrates on $\mathbb{S}_+^{p-1}$ with $p < d$, and assume that the parameter $p$ is known. The aim of their chapter is to identify this support with an empirical risk minimization's technique.

Some more algorithmic approaches have also been recently introduced. Goix et al. (2017) consider $\epsilon$-thickened rectangles to estimate the directions on which the spectral measure concentrates. This estimation is based on a tolerance parameter $\epsilon > 0$ and brings out a sparse representation of the dependence structure. It leads to an algorithm called DAMEX (for Detecting Anomalies among Multivariate EXtremes) of complexity $O(dn \log n)$, where $n$ corresponds to the number of data points. Subsequently, Chiapino and Sabourin (2016) propose another algorithm (CLEF for CLustering Extremal Features) to group together subsets that are likely to be simultaneously extreme. A $O(dn \log n)$ complexity has also been reached by Simpson et al. (2019) who base their method on hidden regular variation. They introduce a set of parameters $(\tau_C)_{C \subset \{1,\dots,d\}}$ which describe to what extent the feature $C$ gathers extreme values. Most of these approaches are based on the rank transform and try to identify groups of asymptotically dependent extremes.

In a recent work, Lehtomaa and Resnick (2019) analyze tail dependence with application to risk management. They study the support of the spectral measure by using a grid estimator. The simplex is firstly mapped to the space $[0,1]^{d-1}$ before being partitioned in equally sized rectangles. The estimation of the support is based on a standard estimator of the spectral measure, see Resnick (2007), Section 9.2.2. The second step is then to build an asymptotically normal test statistic to validate the support estimate.

The main issue in the study of the spectral measure is that the self-normalized extreme $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > t$ that appears in (2.1.1) is inefficient to estimate $S$ in subspaces of dimension smaller than $d-1$, while these types of subsets often concentrate large events. Indeed, in many situations it is very unlikely that a lot of coordinates are simultaneously extreme. In other words, extreme events occur in few directions $i_1, \dots, i_r \in \{1, \dots, d\}$, with $r \ll d$. In this case, the spectral measure puts mass on $\mathrm{Vect}(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_r}) \cap \mathbb{S}_+^{d-1}$, where $\mathbf{e}_1, \dots, \mathbf{e}_d$ denote the vectors of the canonical basis of $\mathbb{R}^d$. We say then that the spectral measure is *sparse*. Unfortunately, as soon as $r < d$, the weak convergence (2.1.1) does not hold for subspaces like $\mathrm{Vect}(\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_r}) \cap \mathbb{S}_+^{d-1}$, since they are not continuity sets for $S$. This is why the difficulty to identify the possible sparsity of $S$ is at the core of the multivariate extremes' study.

Since the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ fails to identify the regions on which the spectral measure puts mass, our aim is to introduce another way of projecting onto the unit sphere. This new projection should take the sparsity of the spectral measure into account by introducing some sparsity in the vector $\mathbf{X}$. In other words, as the limit measure $S$ in (2.1.1) is likely to be sparse, we need to replace $\mathbf{X}/|\mathbf{X}|$ by a unit vector based on $\mathbf{X}$ which is also likely to be sparse. To this end, we use the Euclidean projection of $\mathbf{X}/t$ onto the simplex $\{\mathbf{x} \in \mathbb{R}_+^d, x_1 + \ldots + x_d = 1\}$. This projection has been widely studied in learning theory (see e.g. Duchi et al. (2008), Kyrillidis et al. (2013), or Liu and Ye (2009)). Many different efficient algorithms have been proposed, for instance by Duchi et al. (2008) and Condat (2016).

Based on this projection, we define the concept of sparse regular variation for which the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ is replaced by $\pi(\mathbf{X}/t)$, where $\pi$ denotes the Euclidean projection onto the simplex. The limit measure obtained after this substitution is slightly different from the spectral measure $S$. We study this new angular limit and show that it better captures the possible sparsity structure of the extremes. Besides, we prove that under mild conditions both concepts of regular variation are equivalent and we give the relation between both limit measures.

**Outline** The structure of this chapter is as follows. Section 2.2 gathers all theoretical results useful in this chapter. Firstly, we introduce the multivariate EVT framework. We detail why the knowledge of the subspaces on which the spectral measure puts mass is a main issue for the study of extreme events and explain which difficulties appear in this context. Secondly, we introduce the Euclidean projection onto the simplex and list several results which are of constant use for our study. Section 2.3 is dedicated to the study of this projection in a regular variation context. We focus on the angular part of the limit after substituting the usual projected vector $\mathbf{X}/|\mathbf{X}|$ in (2.1.1) by a vector based on the Euclidean projection onto the simplex. We also provide some interpretations of this new angular vector and discuss to what extent this way of projecting allows us to better capture the sparsity structure of the extremes. The concept of sparsely regularly varying random vector is then introduced in Section 2.4. We establish the equivalence, under mild conditions, between this notion and the standard regular variation's concept. Finally, we illustrate in Section 2.5 the performance of our method on simulated data and compare it with the approach of Goix et al. (2017).

## 2.2 Theoretical background

### 2.2.1 Regular variation and spectral measure

We consider a non-negative random vector $\mathbf{X} = (X_1, \ldots, X_d)$ and our goal is to assess its tail structure. It is customary in EVT to assume that the random vector $\mathbf{X}$ is regularly varying: There exist a random vector $\mathbf{\Theta}$ on $\mathbb{S}_+^{d-1}$ and a non-degenerate random variable $Y$ such that the following limit holds:

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \frac{\mathbf{X}}{|\mathbf{X}|}\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,. \tag{2.2.1}$$

In this case, there exists $\alpha > 0$ such that $Y$ follows a Pareto distribution with parameter $\alpha$. Moreover, the radial limit $Y$ is independent of the angular limit $\mathbf{\Theta}$. The random vector $\mathbf{\Theta}$ is called the *spectral vector* and its distribution $S(\cdot) := \mathbb{P}(\mathbf{\Theta} \in \cdot)$ is called the *spectral measure*.

Equation (2.2.1) brings out the two quantities which characterize the regular variation property of $\mathbf{X}$. On the one hand, the tail index $\alpha$ highlights the intensity of the extremes: The smaller this index is, the larger the extremes are. On the other hand, the spectral vector $\mathbf{\Theta}$ informs on their localization and their dependence structure: The spectral measure puts mass in a direction of $\mathbb{S}_+^{d-1}$ if and only if extreme events appear in this direction. Hence, estimating the spectral measure is a crucial (but challenging) problem in multivariate EVT.

In arbitrary dimensions, several authors recently focus on the estimation of the spectral measure's support. The main purpose is to detect features that are likely to be extreme together. In other words, we would like to identify some specific subsets of $\mathbb{S}_+^{d-1}$ on which the spectral measure puts mass. To this end, it is convenient to partition the positive unit sphere $\mathbb{S}_+^{d-1}$ in the following way. For $\beta \in \mathcal{P}_d^*$, we define the subsets

$$C_\beta = \left\{ \mathbf{x} \in \mathbb{S}_+^d,\, x_i > 0 \text{ for } i \in \beta,\, x_i = 0 \text{ for } i \notin \beta \right\}, \tag{2.2.2}$$

This approach can be related to the one developed by Goix et al. (2017) (see Remark 2.3.1 and Section 2.5). Note that, by construction, the subsets $C_\beta$ are pairwise disjoint and form a partition of $\mathbb{S}_+^{d-1}$:

$$\mathbb{S}_+^{d-1} = \bigsqcup_{\beta \in \mathcal{P}_d^*} C_\beta \,,$$

where $\bigsqcup$ denotes a disjoint union. An illustration of these subsets in dimension 3 are given in Figure 2.1.

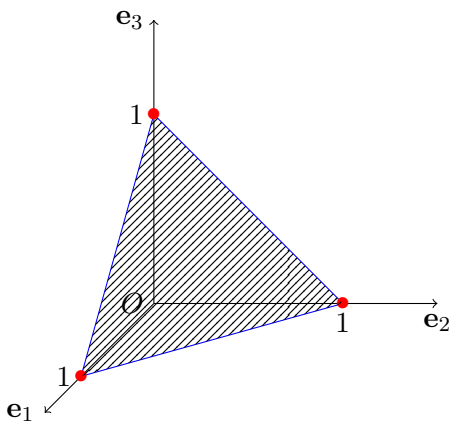



Figure 2.1: The subsets $C_\beta$ in dimension 3 for to the $\ell^1$-norm. In red, the subsets $C_{\{1\}}$, $C_{\{2\}}$, and $C_{\{3\}}$. In blue, the subsets $C_{\{1,2\}}$, $C_{\{1,3\}}$, and $C_{\{2,3\}}$. The shaded part corresponds to the interior of the simplex, that is, the subset $C_{\{1,2,3\}}$.

This partition is helpful to study the extremal structure of $\mathbf{X}$. Indeed, for $\beta \in \mathcal{P}_d^*$, the inequality $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ means that it is likely to observe simultaneously large values in the directions $i \in \beta$ and small values in the directions $i \in \beta^c$. Then, identifying the subsets $C_\beta$ which concentrate the mass of the spectral measure allows us to bring out clusters of coordinates which can be simultaneously large. Hence, the main first step of the spectral measure's estimation consists in classifying the $2^d - 1$ probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ depending on their nullity or not. Note that if $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, for $\beta \neq \{1, \ldots, d\}$, it means that some coordinates of $\mathbf{\Theta}$ are equal to zero with positive probability. In this case, we say that the spectral vector (and hence the spectral measure) is sparse.

**Remark 2.2.1.** In EVT, the notion of sparsity can be defined in two different ways. The first

one concerns the number of subsets $C_\beta$ which gather the mass of the spectral measure. "Sparse" means then that this number is much smaller than $2^d - 1$. This is for instance the device of Goix et al. (2017). The second notion deals with the number of 0 in the spectral vector $\mathbf{\Theta}$. In this case, "sparse" means that with high probability $|\mathbf{\Theta}|_0 \ll d$, where $|\cdot|_0$ denotes the $\ell^0$-norm of $\mathbf{\Theta}$, that is, $|\mathbf{\Theta}|_0 = \#\{i = 1, \ldots, d,\, \theta_i \neq 0\}$. In all this thesis we refer to this second notion. Our aim is to provide a suitable model for extremes which takes this possible sparsity into account.

A standard example of sparsity is the one where the spectral measure only puts mass on the axis: $\mathbb{P}(\mathbf{\Theta} \in \sqcup_{1\leq j\leq d} \{\mathbf{e}_j\}) = \mathbb{P}(\mathbf{\Theta} \in \sqcup_{1\leq j\leq d} C_{\{j\}}) = 1$. This means that there is never more than one direction which contributes to the extremal behavior of the data. In this case, we say that the extremes are asymptotically independent (see Section 1.2.3). This concept has been studied by many authors, for instance Ledford and Tawn (1996) or Ramos and Ledford (2009).

Even in cases of asymptotic dependence the mass of the spectral measure often only spreads on low-dimensional subsets $C_\beta$, that is, for $\beta$ such that $\#\beta \ll d$. This is all the more true in high dimension. Indeed, when $d$ is large, it is very unlikely that all coordinates are extreme together. Regarding the spectral vector, this means that $\mathbb{P}(\mathbf{\Theta} \in C_{\{1,\ldots,d\}}) = 0$. In such cases, it is then interesting to identify the larger groups of variables $\beta \in \mathcal{P}_d^*$ such that $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$. This motivates the notion of maximal subset.

**Definition 2.2.1** (Maximal subset for $\mathbf{\Theta}$)**.** Let $\beta \in \mathcal{P}_d^*$. We say that a subset $C_\beta$ is *maximal* for $\mathbf{\Theta}$ if

$$\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0 \quad \text{and} \quad \mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) = 0\,, \quad \text{for all } \beta' \supsetneq \beta\,. \tag{2.2.3}$$

In terms of extreme values, the notion of maximality can be rephrased in the following way. Firstly, $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ means that the coordinates of $\beta$ may be extreme together. Secondly, the condition $\mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) = 0$, for all $\beta' \supsetneq \beta$, means that $\beta$ is not included in a larger group of coordinates $\beta'$ such that the coordinates of $\beta'$ may be simultaneously extreme.

**Remark 2.2.2.** A straightforward but useful consequence of Definition 2.2.1 is that each subset $C_\beta$ such that $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ is included in a maximal subset of $\mathbf{\Theta}$. Indeed, if there exists no $\beta' \supsetneq \beta$, such that $\mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) = 0$, then $C_\beta$ is a maximal subset itself. If not, it means that there exists $\beta' \supsetneq \beta$, such that $\mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) > 0$. If $C_{\beta'}$ is not maximal, then we repeat this procedure with $\beta'$. Since the length of the $\beta$'s is finite, the procedure stops and provides $\gamma \in \mathcal{P}_d^*$ such that $\beta \subset \gamma$, $\mathbb{P}(\mathbf{\Theta} \in C_\gamma) > 0$ and $\mathbb{P}(\mathbf{\Theta} \in C_{\gamma'}) = 0$, for all $\gamma' \supsetneq \gamma$.

**Why the support's estimation is difficult** While the interpretation of the subspaces $C_\beta$ is rather intuitive, it is quite difficult to estimate the probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$. A natural estimator of the spectral vector $\mathbf{\Theta}$ is based the second component of convergence (2.2.1). Indeed, the polar component of $\mathbf{X}$ satisfies

$$\mathbb{P}\left(\mathbf{X}/|\mathbf{X}| \in \cdot \mid |\mathbf{X}| > t\right) \xrightarrow{d} \mathbb{P}(\mathbf{\Theta} \in \cdot)\,, \quad t \to \infty\,. \tag{2.2.4}$$

This means that the spectral vector $\mathbf{\Theta}$ can be approximated by the self-normalized extreme $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > t$, for $t$ large enough. Unfortunately, the supports of $\mathbf{\Theta}$ and $\mathbf{X}/|\mathbf{X}|$ often drastically differ. Indeed, since $\mathbf{X}$ could model real-world data, the components of $\mathbf{X}$ are almost surely positive. In other words, except for degenerate cases, the random vector $\mathbf{X}/|\mathbf{X}|$ concentrates on the central subspace $C_{\{1,\dots,d\}}$. Equivalently, if $\beta \neq \{1,\dots,d\}$, then $\mathbb{P}(\mathbf{X}/|\mathbf{X}| \in C_\beta) = 0$. This arises while the probability $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ is often positive for some $\beta \neq \{1,\dots,d\}$.

This means that Equation (2.2.4) is not helpful to study the support of the spectral vector $\mathbf{\Theta}$. The self-normalized extreme $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > t$ does not inform on the behavior of $\mathbf{\Theta}$ on the $C_\beta$'s. This kind of problems arises since the spectral measure may put mass on subspaces included in the boundary of the unit sphere $\mathbb{S}_+^{d-1}$ (in our case the $C_\beta$'s, for $\beta \neq \{1,\dots,d\}$), whereas the data generally do not concentrate on such subspaces. Our goal is thus to circumvent this problem by using another projection. This projection has to capture the dependence structure of extremes by taking into account the potential sparsity of the spectral measure. The solution we propose in this chapter is to replace the quantity $\mathbf{X}/|\mathbf{X}|$ by the Euclidean projection onto the simplex of $\mathbf{X}/t$. To this end, we have to adapt Equation (2.2.1).

From now until the end of this chapter, $|\cdot|$ denotes the $\ell^1$-norm and $\mathbb{S}_+^{d-1}$ denotes the simplex in dimension $d$:

$$\mathbb{S}_+^{d-1} := \{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = 1\}\,.$$

In particular, the subsets $C_\beta$ defined in (2.2.2) are now associated to the $\ell^1$-norm. More generally $\mathbb{S}_+^{d-1}(z) := \{\mathbf{x} \in \mathbb{R}_+^d,\, x_1 + \ldots + x_d = z\}$ for $z > 0$.

### 2.2.2 The Euclidean projection onto the simplex

In this subsection, we introduce the Euclidean projection onto the simplex. For more details, see Duchi et al. (2008) and the references therein.

Let $z > 0$ and $\mathbf{v} \in \mathbb{R}_+^d$. We consider the following optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}}\ \frac{1}{2}|\mathbf{w} - \mathbf{v}|_2^2 \quad \text{s.t.} \quad |\mathbf{w}|_1 = z\,. \tag{2.2.5}$$

Since $\mathbf{v} \geq 0$, the minimization problem (2.2.5) is equivalent to

$$\underset{\mathbf{w}}{\text{minimize}}\ \frac{1}{2}|\mathbf{w} - \mathbf{v}|_2^2 \quad \text{s.t.} \quad \sum_{i=1}^d w_i = z\,, \quad w_i \geq 0\,.$$

(see Duchi et al. (2008), Lemma 3). The Lagrangian of this problem and the complementary slackness KKT condition imply that this problem has a unique solution $\mathbf{w} \in \mathbb{R}_+^d$ which satisfies $w_i = (v_i - \lambda_{\mathbf{v},z})_+$ for $\lambda_{\mathbf{v},z} \in \mathbb{R}$. The constant $\lambda_{\mathbf{v},z}$ is defined by the relation $\sum_{1\leq i\leq d}(v_i - \lambda_{\mathbf{v},z})_+ = z$.

Based on these considerations, we define the application $\pi_z$ which maps $\mathbf{v}$ to $\mathbf{w}$:

$$\begin{array}{rccc} \pi_z & : & \mathbb{R}_+^d & \to & \mathbb{S}_+^{d-1}(z) \\ & & \mathbf{v} & \mapsto & \mathbf{w} = (\mathbf{v} - \lambda_{\mathbf{v},z})_+ . \end{array}$$

This application is called the *projection onto the positive sphere* $\mathbb{S}_+^{d-1}(z)$. An algorithm which computes $\pi_z(\mathbf{v})$ for $\mathbf{v} \in \mathbb{R}_+^d$ and $z > 0$ is given in Duchi et al. (2008). It is based on a median-search procedure whose expected time complexity is $O(d)$. Unfortunately, this approach is not very intuitive and introduces many variables. Hence, we include it in Section 2.8 and detail here a more understandable version of this algorithm with complexity $O(d \log(d))$. Algorithm 1 emphasizes the number of positive coordinates $\rho$ of the projected vector $\pi_z(\mathbf{v})$:

$$\rho = \max \left\{ j \in \{1, \dots, d\},\ \mu_j - \frac{1}{j} \left( \sum_{r=1}^{j} \mu_j - z \right) > 0 \right\}, \tag{2.2.6}$$

where $\mu_1 \geq \dots, \mu_d$ denote the order coordinates of $\mathbf{v}$, see Duchi et al. (2008), Lemma 2. The integer $\rho$ corresponds to the $\ell^0$-norm of $\pi_z(\mathbf{v})$ and thus informs on the sparsity of this projected vector. It will therefore be crucial in what follows.

**Data:** A vector $\mathbf{v} \in \mathbb{R}_+^d$ and a scalar $z > 0$
**Result:** The projected vector $\mathbf{w} = \pi(\mathbf{v})$
Sort $\mathbf{v}$ in $\boldsymbol{\mu}$ : $\mu_1 \geq \dots \geq \mu_d$;
Find $\rho = \max \left\{ j \in \{1, \dots, n\},\ \mu_j - \frac{1}{j} \left( \sum_{r=1}^{j} \mu_j - z \right) > 0 \right\}$;
Define $\eta = \frac{1}{\rho} \left( \sum_{r=1}^{\rho} \mu_j - z \right)$;
Define $\mathbf{w}$ s.t. $w_i = \max(v_i - \eta, 0)$.

**Algorithm 1:** Euclidean projection onto the simplex.

**Remark 2.2.3.** The expected linear complexity with respect to the dimension $d$ is essential. Indeed, multivariate extremes have already been studied in low dimensions, especially in two dimensions (for instance in Einmahl et al. (2001) or Einmahl and Segers (2009)). But when the dimension increases, the study of large events becomes a difficult issue. The recent algorithmic approaches developed by Simpson et al. (2019) or Goix et al. (2017) reach a complexity $O(dn \log(n))$, where $n$ denotes the number of data points. Based on Algorithm 3, we manage to reach a complexity $O(dn)$.

**Remark 2.2.4.** Algorithm 1 highlights the quantity $\eta = \eta(z)$ which is decreasing when $z$ increases. Therefore, when $z$ is large, $\eta$ is small and $w_i = \max(v_i - \eta, 0)$ is more likely to be positive. This means that the projected vector $\mathbf{w}$ is sparser when $z$ is large (see also Lemma 2.2.2). This relation between the parameter $z$ and the sparsity structure of $\mathbf{w} = \pi_z(\mathbf{v})$ will be a crucial point regarding the choice of a threshold for the extremes.

Note that the projection satisfies the relation $\pi_z(\mathbf{v}) = z\pi_1(\mathbf{v}/z)$ for all $\mathbf{v} \in \mathbb{R}_+^d$ and $z > 0$. This is why we mainly focus on the projection $\pi_1$ onto the simplex $\mathbb{S}_+^{d-1}$. In this case, we shortly denote

$\pi$ for $\pi_1$ and $\lambda_{\mathbf{v}}$ for $\lambda_{\mathbf{v},1}$:

$$\begin{array}{rccc} \pi & : & \mathbb{R}_+^d & \to & \mathbb{S}_+^{d-1} \\ & & \mathbf{v} & \mapsto & (\mathbf{v} - \lambda_{\mathbf{v}})_+ \,. \end{array}$$

An illustration of $\pi$ for $d = 2$ is given in Figure 2.2.

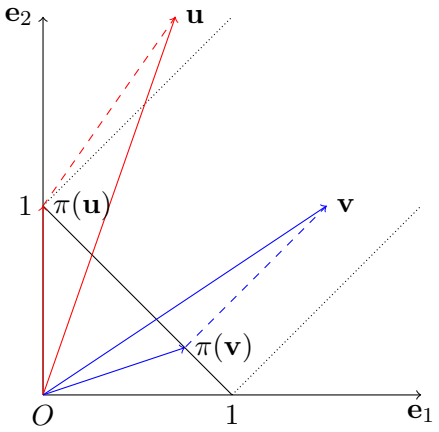



Figure 2.2: The Euclidean projection onto the simplex $\mathbb{S}_+^1$.

We list below some straightforward results on the projection.

P1. The projection preserves the order of the coordinates: If $v_{\sigma(1)} \geq \ldots \geq v_{\sigma(d)}$ for a permutation $\sigma$, then $\pi(\mathbf{v})_{\sigma(1)} \geq \ldots \geq \pi(\mathbf{v})_{\sigma(d)}$ for the same permutation.

P2. If $\pi(\mathbf{v})_j > 0$, then $v_j > 0$. Equivalently, $v_j = 0$ implies $\pi(\mathbf{v})_j = 0$.

P3. The projection $\pi$ is continuous, as every projection on a convex closed set in a Hilbert space.

The last property will be useful in what follows since $\pi$ is used to tackle the weak convergence's issue in the spectral measure's definition (2.2.1). The idea is indeed to substitute the quantity $\mathbf{X}/|\mathbf{X}|$ in (2.2.1) for $|\cdot| = |\cdot|_1$ by $\pi(\mathbf{X}/t)$ and to manage to get same convergence results. A natural way to do this relies on the continuous mapping theorem.

We end this section with two important properties satisfied by the projection.

**Lemma 2.2.1.** *If $0 < z \leq z'$, then $\pi_z \circ \pi_{z'} = \pi_z$.*

This means that projecting onto a sphere and then onto a smaller one is the same as directly projecting onto the smaller sphere. This lemma will be useful to prove some technical results gathering the projection $\pi$ and regular variation.

Finally, in order to study the sparsity structure of extreme events, we are interested in computing probabilities like $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ and $\mathbb{P}(\mathbf{\Theta}_{\beta^c} = 0)$, for $\beta \in \mathcal{P}_d^*$. To this end, next lemma will be helpful.

**Lemma 2.2.2.** *Let* $\mathbf{v} \in \mathbb{R}_+^d$ *and* $\beta \in \mathcal{P}_d^*$. *The following equivalences hold:*

$$\pi(\mathbf{v})_{\beta^c} = 0 \quad \textit{if and only if} \quad 1 \leq \min_{i \in \beta^c} \sum_{k=1}^{d} (v_k - v_i)_+ \,, \tag{2.2.7}$$

*and*

$$\pi(\mathbf{v}) \in C_\beta \quad \textit{if and only if} \quad \begin{cases} \max_{i \in \beta} \sum_{j \in \beta} (v_j - v_i) < 1 \,, \\ \min_{i \in \beta^c} \sum_{j \in \beta} (v_j - v_i) \geq 1 \,. \end{cases} \tag{2.2.8}$$

If $\pi(\mathbf{v}) > 0$ (that is, if $\beta = \{1, \ldots, d\}$), then $\pi(\mathbf{v})$ has necessary the following form (see Algorithm 1):

$$\pi(\mathbf{v}) = \mathbf{v} - \frac{1}{d}\left(\sum_{k=1}^{d} v_k - 1\right) = \mathbf{v} - \frac{|\mathbf{v}| - 1}{d} \,.$$

Thus, for $\mathbf{x} \geq 0$, we have the following characterization:

$$\pi(\mathbf{v}) > \mathbf{x} \quad \text{if and only if} \quad \mathbf{v} > \mathbf{x} + \frac{|\mathbf{v}| - 1}{d} \,. \tag{2.2.9}$$

This equivalence will be of constant use in the proofs.

**Remark 2.2.5.** Note that the projection $\pi$ is not homogeneous. Recall that a function $f$ is said to be homogeneous if there exists $q > 0$ such that for all $t > 0$, $f(t\mathbf{x}) = t^q f(\mathbf{x})$. If $f$ is a continous and homogeneous function and $\mathbf{X}$ is a regularly random vector in $\mathbb{R}_+^d$ with tail index $\alpha > 0$, then the random vector $f(\mathbf{X})$ is regularly varying with tail index $\alpha/q$ (see Jessen and Mikosch (2006)). Such a result cannot be used for the Euclidean projection onto the simplex.

The theoretical framework being defined, we now want to use the projection $\pi$ in a regular variation context. This is the purpose of next section.

## 2.3 Spectral measure and projection

The aim of this section is twofold. In the first part, we use the Euclidean projection onto the simplex to introduce a new convergence based on (2.2.1). This new convergence brings out an angular limit vector which differs from the spectral vector. Some results on this limit and its relation with the spectral vector are introduced. Secondly, we establish sparsity results for this new vector. Finally, we develop a model with a discrete spectral vector $\mathbf{\Theta}$ and study how it affects the vector $\mathbf{Z}$.

### 2.3.1 Regular variation and projection

From now on, and until the end of Section 2.3, we consider a regularly varying random vector $\mathbf{X}$ on $\mathbb{R}_+^d$:

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \frac{\mathbf{X}}{|\mathbf{X}|}\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot) \,, \quad t \to \infty \,. \tag{2.3.1}$$

In this case, we know that there exists $\alpha > 0$ such that $Y$ follows a Pareto($\alpha$) distribution and also that the limits $Y$ and $\mathbf{\Theta}$ are independent. We emphasized in Subsection 2.2.1 that convergence (2.3.1) is not helpful to capture the possible sparsity structure of the spectral vector $\mathbf{\Theta}$. Our idea is to substitute the self-normalized extremes $\mathbf{X}/|\mathbf{X}|$ by another vector on the simplex which better highlights this sparsity.

Here is an intuitive idea to see how the Euclidean projection can solve this kind of issue. As explained in Section 2.2.1, for $\beta \in \mathcal{P}_d^*$, the quantity $\mathbb{P}(\mathbf{X}/|\mathbf{X}| \in C_\beta \mid |\mathbf{X}| > t)$ is always equal to 0 (except for degenerate cases), whereas $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ could be positive. This arises since for $t > 0$, the sets $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1, \mathbf{x}/|\mathbf{x}| \in C_\beta\}$ have zero Lebesgue measure for $\beta \neq \{1, \dots, d\}$, and real-world data do not concentrate on such subspaces. Our idea is to replace these subsets by closer ones, but with positive Lebesgue measure. Based on the projection $\pi$, we use the subsets $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1, \pi(\mathbf{x}/t) \in C_\beta\}$.

**Example 2.3.1.** Let us take the example of the two-dimensional case illustrated in Figure 2.2. Here, estimating for instance the probability $\mathbb{P}(\mathbf{\Theta} \in C_{\{2\}}) = \mathbb{P}(\Theta_1 = 0)$ with the set of zero Lebesgue measure $\{\mathbf{x}, \mathbf{x}/|\mathbf{x}| \in C_{\{2\}}\}$ seems unachievable. Our idea here is to rather use the set $\{\mathbf{x}, \pi(\mathbf{x}) \in C_{\{2\}}\} = \{\mathbf{x}, x_2 \geq x_1 + 1\}$ which has positive Lebesgue measure. In a sense, the projection allows us to give more weight to the subsets $C_\beta$, for $\beta \neq \{1, \dots, d\}$.

**Remark 2.3.1.** The idea of substituting the subspaces $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1, \mathbf{x}/|\mathbf{x}| \in C_\beta\}$ which have zero Lebesgue measure by closer subspaces with positive Lebesgue measure has already been used in the literature. For instance, Goix et al. (2017) define $\epsilon$-thickened rectangles $R_\beta^\epsilon$, defined by

$$R_\beta^\epsilon = \left\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}|_\infty > 1, x_i > \epsilon \text{ for } i \in \beta, x_i \leq \epsilon \text{ for } i \notin \beta\right\},$$

for $\beta \in \mathcal{P}_d^*$ and $\epsilon > 0$. Unfortunately, these considerations are based on a hyperparameter $\epsilon > 0$ which has to be tuned in practice. One of the advantages of the projection $\pi$ is that it is does not need any hyperparameter. A more detailed comparison of these two methods will be discussed in Section 2.5.

With this in mind, we substitute the usual projection $\mathbf{X}/|\mathbf{X}|$ by $\pi(\mathbf{X}/t)$. The first step is to see how this affects the spectral vector. The continuity of the projection $\pi$ implies that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \pi\left(\frac{\mathbf{X}}{t}\right)\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \pi(Y\mathbf{\Theta})) \in \cdot), \quad t \to \infty. \tag{2.3.2}$$

The limit of the angular component is now $\pi(Y\mathbf{\Theta})$. In particular, we lose independence between the radial component $Y$ and the angular component $\pi(Y\mathbf{\Theta})$ of the limit. The dependence relation between both components will be detailed in Proposition 2.4.1.

Following Equation (2.3.2), we set $\mathbf{Z} = \pi(Y\mathbf{\Theta}) \in \mathbb{S}_+^{d-1}$. The aim of this section is to study to what extent the new angular limit $\mathbf{Z}$ differs from the spectral vector $\mathbf{\Theta}$ and how it helps to study the tail dependence of $\mathbf{X}$. A first crucial point is that convergence (2.3.2) holds for Borel sets

$A \in \mathbb{S}_+^{d-1}$ which satisfy $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$. Next proposition states that the subsets $C_\beta$ and $\mathrm{Vect}(\mathbf{e}_j,\, j \in \beta)$, $\beta \in \mathcal{P}_d^*$, satisfy this condition.

**Proposition 2.3.1.** *Let $\mathbf{X}$ be a regularly varying random vector in $\mathbb{R}_+^d$ with spectral vector $\boldsymbol{\Theta}$ and tail index $\alpha > 0$. Set $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$, where $Y$ is a Pareto($\alpha$)-distributed random variable independent of $\boldsymbol{\Theta}$. For any $\beta \in \mathcal{P}_d^*$, the following convergences hold:*

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \quad t \to \infty\,, \tag{2.3.3}$$

$$\mathbb{P}(\pi(\mathbf{X}/t)_{\beta^c} = 0 \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z}_{\beta^c} = 0)\,, \quad t \to \infty\,. \tag{2.3.4}$$

Both convergences imply that the sparsity structure of $\mathbf{Z}$ can be studied through the projected vector $\pi(\mathbf{X}/t)$. We insist on the fact that for $\beta \neq \{1,\dots,d\}$, the convergences (2.3.3) and (2.3.4) do not hold if we replace $\mathbf{Z}$ by $\boldsymbol{\Theta}$ and $\pi(\mathbf{X}/t)$ by $\mathbf{X}/|\mathbf{X}|$. From a statistical point of view, Proposition 2.3.1 is helpful since it allows us to estimate the sparse behavior of $\mathbf{Z}$ based on the one of $\mathbf{X}$. This will be used on numerical results in Section 2.5 and developed in a statistical framework in Chapter 3.

**A first interpretation of Z** At first glance, using the vector $\mathbf{Z}$ instead of $\boldsymbol{\Theta}$ in order to capture the tail dependence of $\mathbf{X}$ seems less interpretable. Nevertheless, some properties of the projection make this new vector more understandable, in particular its interpretation regarding $\mathbf{X}$.

The first property deals with the extreme behavior of a component with respect to the others. For $j = 1,\dots,d$, we apply Equation (2.2.8) of Lemma 2.2.2 with $\beta = \{j\}$ and obtain the following equivalences:

$$\pi(\mathbf{X}/t) \in C_{\{j\}} \iff \min_{i\neq j}(X_j/t - X_i/t) \geq 1 \iff X_j \geq \max_{i\neq j} X_i + t\,, \quad t > 0.$$

Then, applying Proposition 2.3.1 to the subset $C_{\{j\}}$ leads to the convergence

$$\mathbb{P}\big(X_j \geq \max_{i\neq j} X_i + t \mid |\mathbf{X}| > t\big) = \mathbb{P}\big(\pi(\mathbf{X}/t) \in C_{\{j\}} \mid |\mathbf{X}| > t\big) \to \mathbb{P}(\mathbf{Z} \in C_{\{j\}}) = \mathbb{P}(Z_j = 1)\,,$$

when $t \to \infty$. At a non-asymptotic level, this means that for $t$ "high enough", we have the approximation

$$\mathbb{P}\big(X_j \geq \max_{i\neq j} X_i + t \mid |\mathbf{X}| > t\big) \approx \mathbb{P}(Z_j = 1)\,.$$

This means that $\mathbf{Z}$ concentrates on the $j$-th axis if the $j$-th coordinate of $\mathbf{X}$ is much larger than the others, that is, if extreme values appear in this direction.

More generally, if we fix $\beta \in \mathcal{P}_d^*$ with length $r = \#\beta$, then Equation (2.2.7) of Lemma 2.2.2 leads

to the following equivalences:

$$\pi(\mathbf{X}/t)_{\beta^c} = 0 \iff 1 \leq \min_{i\in\beta^c} \sum_{k=1}^{d} (X_k/t - X_i/t)_+ \iff t \leq \sum_{k=1}^{d} (X_k - \max_{i\in\beta^c} X_i)_+$$
$$\iff t \leq \sum_{k\in\beta} (X_k - \max_{i\in\beta^c} X_i)_+ \iff t \leq \sum_{k\in\beta}^{d} X_k - r \max_{i\in\beta^c} X_i \,,$$

where the positive part can be withdrawn since the projection keeps the order of the coordinates (see the property P1, Subsection 2.2.2). All in all, we obtain the equivalence

$$\pi(\mathbf{X}/t)_{\beta^c} = 0 \iff r^{-1} \sum_{k\in\beta} X_k \geq \max_{i\in\beta^c} X_i + t \,.$$

Then, following Proposition 2.3.1, we obtain

$$\mathbb{P}\Big(r^{-1} \sum_{k\in\beta} X_k \geq \max_{i\in\beta^c} X_i + t \mid |\mathbf{X}| > t\Big) = \mathbb{P}\big(\pi(\mathbf{X}/t)_{\beta^c} = 0 \mid |\mathbf{X}| > t\big) \to \mathbb{P}(\mathbf{Z}_{\beta^c} = 0) \,, \quad t \to \infty \,,$$

and it leads to the following approximation:

$$\mathbb{P}\Big(r^{-1} \sum_{k\in\beta} X_k \geq \max_{i\in\beta^c} X_i + t \mid |\mathbf{X}| > t\Big) \approx \mathbb{P}(\mathbf{Z}_{\beta^c} = 0) \,,$$

for $t$ "high enough". This can be interpreted in the following way: The vector $\mathbf{Z}$ does not concentrate on the directions $j \in \beta^c$ if the average value of $X_j$ for $j \in \beta$ is much larger than all the components of $\mathbf{X}$ on $\beta^c$. In other words, there is an important gap between the average value of the marginals on $\beta$ and the value of the marginals on $\beta^c$.

### 2.3.2 The distribution of Z

The new angular vector $\mathbf{Z}$ being defined, the aim is now to explicit some links between $\mathbf{\Theta}$ and $\mathbf{Z}$. To this end, we define the function $G_{\mathbf{Z}}$ by

$$G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{P}(\mathbf{Z} > \mathbf{x}) = \mathbb{P}(Z_1 > x_1, \ldots, Z_d > x_d) \,, \quad \mathbf{x} \in \mathbb{R}^d \,. \tag{2.3.5}$$

The function $G_{\mathbf{Z}}$ characterizes the distribution of $\mathbf{Z}$. However, note that there is no simple relation between $G_{\mathbf{Z}}$ and the cumulative distribution function of $\mathbf{Z}$ as soon as $d \geq 2$. Since $\mathbf{Z} \in \mathbb{S}_+^{d-1}$, we only focus on $G_{\mathbf{Z}}(\mathbf{x})$ for $\mathbf{x}$ in $\mathbb{R}_+^d$ such that $\sum_j x_j < 1$, this means for $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d$, where $\mathcal{B}(0,1)$ denotes the (open) unit ball for the $\ell^1$-norm. Thus, we write

$$G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) \,,$$

where the sets $A_{\mathbf{x}}$ are defined by

$$A_{\mathbf{x}} = \{\mathbf{u} \in \mathbb{S}_+^{d-1},\, x_1 < u_1, \ldots, x_d < u_d\}\,. \tag{2.3.6}$$

Since the family $\mathcal{A} = \{A_{\mathbf{x}},\, \mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d\}$ generates the Borel $\sigma$-algebra of the simplex $\mathbb{S}_+^{d-1}$, the distribution of $\mathbf{Z}$ is completely characterized by $G_{\mathbf{Z}}(\mathbf{x})$ for $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d$.

Following Equation (2.2.9), we can express the condition $\mathbf{Z} > \mathbf{x}$ in terms of $\mathbf{\Theta}$.

**Proposition 2.3.2.** *Let $\mathbf{X}$ be a regularly varying random vector of $\mathbb{R}_+^d$ with tail index $\alpha > 0$ and spectral vector $\mathbf{\Theta}$. For $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d$, such that for all $j = 1, \ldots, d$, $x_j \neq 1/d$, define $J_+ = \{j,\, x_j > 1/d\}$ and $J_- = \{j,\, x_j < 1/d\}$. Then, we have*

$$G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{E}\left[\left(1 \wedge \min_{j \in J_+}\left(\frac{\Theta_j - 1/d}{x_j - 1/d}\right)_+^{\alpha} - \max_{j \in J_-}\left(\frac{\Theta_j - 1/d}{x_j - 1/d}\right)_+^{\alpha}\right)_+\right], \tag{2.3.7}$$

*with $G_{\mathbf{Z}}$ defined in (2.3.5).*

Proposition 2.3.2 gives an interesting relation between the distribution of $\mathbf{Z}$ and the one of $\mathbf{\Theta}$. Unfortunately, its complexity makes it difficult to use. But specific choices for $\mathbf{x}$ will give some useful results.

A convenient particular case is the one where $\mathbf{x}$ satisfies $\mathbf{x} < 1/d$. There, we obtain

$$G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{E}\left[1 - \max_{1 \leq j \leq d}\left(\frac{1/d - \Theta_j}{1/d - x_j}\right)^{\alpha}\right].$$

In particular, for $\mathbf{x} = \mathbf{0}$, we get

$$G_{\mathbf{Z}}(\mathbf{0}) = 1 - \mathbb{E}\left[\max_{1 \leq j \leq d}(1 - d\Theta_j)^{\alpha}\right]. \tag{2.3.8}$$

Thus, the probability for $\mathbf{Z}$ to have a null component is

$$\mathbb{P}(\exists j = 1, \ldots, d\,, Z_j = 0) = \mathbb{E}\left[\max_{1 \leq j \leq d}(1 - d\Theta_j)^{\alpha}\right]. \tag{2.3.9}$$

This quantity is null if and only if for all $j = 1, \ldots, d$, $\Theta_j = 1/d$ a.s. and is equal to 1 if and only if $\min_{1 \leq j \leq d} \Theta_j = 0$ a.s. As expected, the new angular vector $\mathbf{Z}$ is more likely to be sparse. In particular, all usual spectral models on $\mathbf{\Theta}$ that are not supported on the axis are not suitable for $\mathbf{Z}$. The goal of the next subsection is to study more into details the sparsity structure of $\mathbf{Z}$.

### 2.3.3 Sparsity structure of Z

Since the projection is introduced in order to better capture the sparsity structure of the extremes, we give here different results of sparsity for the angular component $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. The general aim

is thus to compute probabilities like $\mathbb{P}(\mathbf{Z} \in C_\beta)$ or $\mathbb{P}(\mathbf{Z}_{\beta^c} = 0)$, for $\beta \in \mathcal{P}_d^*$, in order to generalize Equation (2.3.9).

**Proposition 2.3.3.** *Let $\mathbf{X}$ be a regularly varying random vector of $\mathbb{R}_+^d$ with spectral vector $\mathbf{\Theta}$ and tail index $\alpha > 0$. Set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$, where $Y$ is a Pareto($\alpha$)-distributed random variable independent of $\mathbf{\Theta}$. For any $\beta \in \mathcal{P}_d^*$, we have*

$$\mathbb{P}(\mathbf{Z}_{\beta^c} = 0) = \mathbb{E}\left[\min_{j\in\beta^c}\left(\sum_{k=1}^{d}(\Theta_k - \Theta_j)_+\right)^\alpha\right], \tag{2.3.10}$$

*and*

$$\mathbb{P}(\mathbf{Z} \in C_\beta) = \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+\right]. \tag{2.3.11}$$

If we consider the case where $\beta = \{1, \dots, d\}$, then we obtain the probability that all coordinates are positive. This has already been computed in (2.3.8). It is equal to $G_{\mathbf{Z}}(\mathbf{0}) = 1 - \mathbb{E}\left[\max_{1\le j\le d}(1 - d\Theta_j)^\alpha\right]$.

Another particular case of Proposition 2.3.3 is the one where $\beta$ corresponds to a single coordinate $j_0$. In this case, since $\mathbf{Z}$ belongs to the simplex, both probabilities $\mathbb{P}(\mathbf{Z}_{\beta^c} = 0)$ and $\mathbb{P}(\mathbf{Z} \in C_\beta)$ are equal. Their common value corresponds to the probability that $\mathbf{Z}$ concentrates on the $j_0$-th axis, which is equal to

$$\mathbb{P}(Z_{j_0} = 1) = \mathbb{E}\left[\min_{j\neq j_0}(\Theta_{j_0} - \Theta_j)_+^\alpha\right]. \tag{2.3.12}$$

Then, Equation (2.3.12) can be developed in the following way:

$$\begin{aligned}\mathbb{P}(Z_{j_0} = 1) &= \mathbb{E}\left[\min_{j\neq j_0}(\Theta_{j_0} - \Theta_j)_+^\alpha \mathbb{1}_{\{\Theta_{j_0}=1\}}\right] + \mathbb{E}\left[\min_{j\neq j_0}(\Theta_{j_0} - \Theta_j)_+^\alpha \mathbb{1}_{\{\Theta_{j_0}<1\}}\right] \\ &= \mathbb{P}(\Theta_{j_0} = 1) + \mathbb{E}\left[\min_{j\neq j_0}(\Theta_{j_0} - \Theta_j)_+^\alpha \mathbb{1}_{\Theta_{j_0}<1}\right] \ge \mathbb{P}(\Theta_{j_0} = 1).\end{aligned}$$

This shows again that the vector $\mathbf{Z}$ is more likely to be sparse than the spectral vector $\mathbf{\Theta}$.

**Remark 2.3.2.** Following Equation (2.3.10), we write

$$\mathbb{P}(\mathbf{Z}_{\beta^c} = 0) \ge \mathbb{E}\left[\min_{j\in\beta^c}\left(\sum_{k=1}^{d}(\Theta_k - \Theta_j)_+\right)^\alpha \mathbb{1}_{\{\mathbf{\Theta}_{\beta^c}=0\}}\right] = \mathbb{E}\left[\left(\sum_{k=1}^{d}\Theta_k\right)^\alpha \mathbb{1}_{\{\mathbf{\Theta}_{\beta^c}=0\}}\right] = \mathbb{P}(\mathbf{\Theta}_{\beta^c} = 0). \tag{2.3.13}$$

This can also be seen as a direct consequence of Property P2, see Subsection 2.2.2. This property also gives

$$\mathbb{P}(\mathbf{Z}_\beta > 0) \le \mathbb{P}(\mathbf{\Theta}_\beta > 0). \tag{2.3.14}$$

This inequality will be useful in some proofs.

Our goal is now to compare the probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ and $\mathbb{P}(\mathbf{Z} \in C_\beta)$, for $\beta \in \mathcal{P}_d^*$. Based on Proposition 2.3.3, we state a first inequality between these two quantities.

**Corollary 2.3.1.** *We use the same notations as in Proposition 2.3.3. For $\beta \in \mathcal{P}_d^*$, if $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, then $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$.*

Corollary 2.3.1 implies that we do not lose any information on the support of the spectral measure by studying $\mathbf{Z}$ instead of $\mathbf{\Theta}$. But it is possible that the distribution of $\mathbf{Z}$ puts some mass on a subset $C_\beta$ while the one of $\mathbf{\Theta}$ does not. Nevertheless, if the overestimation is not too large, $\mathbf{Z}$ gives a reduce numbers of directions (regarding the total number $2^d - 1$) in which extreme events could appear. So the use of $\mathbf{Z}$ provides some trends in the dependence structure of $\mathbf{X}$.

**Example 2.3.2.** We detail here an example which shows that the converse implication of Corollary 2.3.1 does not hold. We consider a spectral vector $\mathbf{\Theta}$ in $\mathbb{S}_+^1$ with a first component $\Theta_1$ uniformly distributed (and hence $\Theta_2 = 1 - \Theta_1$ is also uniformly distributed). On the one hand, the probability that $\mathbf{\Theta}$ belongs to an axis is equal to

$$\mathbb{P}(\mathbf{\Theta} \in C_{\{1\}} = 0) = \mathbb{P}(\Theta_2 = 0) = 0\,.$$

On the other hand, following Lemma 2.2.2, the probability that $\mathbf{Z}$ belongs to an axis is equal to

$$\mathbb{P}(\mathbf{Z} \in C_{\{1\}}) = \mathbb{P}(Y\Theta_1 - Y\Theta_2 \geq 1) = \mathbb{P}(2\Theta_1 - 1 \geq 1/Y)\,.$$

If we assume that $\alpha = 1$ in order to simplify the calculations, then $1/Y$ is uniformly distributed, and thus, by independence of $\mathbf{\Theta}$ and $Y$, we obtain

$$\mathbb{P}(\mathbf{Z} \in C_{\{1\}}) = \int_0^1 \mathbb{P}(2\Theta_1 - 1 \geq u)\mathrm{d}u = \int_0^1 \mathbb{P}\Big(\Theta_1 \geq \frac{u+1}{2}\Big)\mathrm{d}u = \int_0^1 \frac{1-u}{2}\mathrm{d}u = \frac{1}{4}\,.$$

Example 2.3.2 shows that it is possible to find some $\beta$ such that $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$ and $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = 0$. In order to have a partial converse result, and similarly to Definition 2.3.1, we introduce the notion of maximal subset for $\mathbf{Z}$.

**Definition 2.3.1** (Maximal subset for $\mathbf{Z}$)**.** Let $\beta \in \mathcal{P}_d^*$. We say that a subset $C_\beta$ is *maximal* for $\mathbf{Z}$ if

$$\mathbb{P}(\mathbf{Z} \in C_\beta) > 0 \quad \text{and} \quad \mathbb{P}(\mathbf{Z} \in C_{\beta'}) = 0\,, \text{ for all } \beta' \supsetneq \beta\,. \tag{2.3.15}$$

Next Theorem states that maximal subsets for $\mathbf{\Theta}$ and $\mathbf{Z}$ are equivalent notions.

**Theorem 2.3.1.** *We use the same notations as in Proposition 2.3.3 and fix $\beta \in \mathcal{P}_d^*$. Then, $C_\beta$ is a maximal subset for $\mathbf{\Theta}$ if and only if $C_\beta$ is a maximal subset for $\mathbf{Z}$.*

Example 2.3.2 shows it may exists $\beta \in \mathcal{P}_d^*$ such that $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$ and $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = 0$. In this case, Theorem 2.3.1 states that the subset $C_\beta$ is not maximal for $\mathbf{Z}$ since it is not maximal for $\mathbf{\Theta}$. Following Remark 2.2.2, we consider a maximal subset $\gamma$ for $\mathbf{Z}$ such that $\beta \subset \gamma$. Then, Theorem

2.3.1 states that $\mathbb{P}(\mathbf{\Theta} \in C_\gamma) > 0$. This means that even $\beta$ does not gather itself coordinates on which extreme values simultaneously occur, there exists a superset of $\beta$ which actually contains extremes. Thus, $\beta$ still gives information on the study of large events.

### 2.3.4 A discrete model for the spectral measure

We introduce here a known discrete model on $\mathbf{\Theta}$ and compute the corresponding distribution of $\mathbf{Z}$.

**Asymptotic independence and complete dependence** We first study two particular cases in multivariate EVT which have already been studied in Section 1.2.3. The first one is the complete dependence's case, which is defined by the relation $\mathbb{P}(\forall i = 1, \dots, d,\ \Theta_i = 1/d) = 1$. Equivalently, the spectral measure is a Dirac mass at $(1/d, \dots, 1/d)$. In terms of extremes, it means that all coordinates simultaneously contribute to large events. Note that if $\mathbf{u} = r(1/d, \dots, 1/d) \in \mathbb{R}_+^d$, $r \geq 1$, the projected vector $\pi(\mathbf{u})$ corresponds to the self-normalization: $\pi(\mathbf{u}) = (1/d, \dots, 1/d) = \mathbf{u}/|\mathbf{u}|$. This implies that in case of complete dependence, $\mathbf{Z} = \mathbf{\Theta} = (1/d, \dots, 1/d)$ a.s.

Another standard case is the asymptotic independence's one, which appears when $\mathbf{\Theta}$ only concentrates on the axis. It means that $\mathbb{P}(\mathbf{\Theta} \in \sqcup_{1 \leq k \leq d} \mathbf{e}_k) = 1$. Note that this case has already been partially discussed in Section 2.2. As for the complete dependence's case, we want to express asymptotic independence in terms of $\mathbf{Z}$. To this end, we write

$$\begin{aligned}
\mathbb{P}(\exists 1 \leq i \leq d,\ Z_i = 1) &= \mathbb{P}(\exists 1 \leq i \leq d,\ \forall j \neq i,\ Z_j = 0) \\
&= \mathbb{P}\left(\exists 1 \leq i \leq d,\ \forall j \neq i,\ 1 \leq Y(\Theta_i - \Theta_j)_+\right) \\
&= \mathbb{P}\Big(\exists 1 \leq i \leq d,\ Y^{-\alpha} \leq \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha\Big) \\
&= \mathbb{P}\Big(Y^{-\alpha} \leq \max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha\Big) \\
&= \int_0^1 \mathbb{P}\Big(u \leq \max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha\Big)\, \mathrm{d}u \\
&= \mathbb{E}\Big[\max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha\Big]\,.
\end{aligned}$$

Thus, since $\max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha \leq 1$, we have the equivalence

$$\mathbb{P}(\exists 1 \leq i \leq d,\ Z_i = 1) = 1 \quad \text{if and only if} \quad \mathbb{P}\Big(\max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha = 1\Big) = 1\,.$$

This last probability can be rewritten as follows:

$$\mathbb{P}\Big(\max_{1 \leq i \leq d} \min_{j \neq i}(\Theta_i - \Theta_j)_+^\alpha = 1\Big) = \mathbb{P}\Big(\exists 1 \leq i \leq d,\ \min_{j \neq i}(\Theta_i - \Theta_j)_+ = 1\Big) = \mathbb{P}\big(\exists 1 \leq i \leq d,\ \Theta_i = 1\big)\,.$$

This proves the equivalence between $\mathbb{P}(\exists 1 \leq i \leq d,\ Z_i = 1) = 1$ and $\mathbb{P}\left(\exists 1 \leq i \leq d,\ \Theta_i = 1\right) = 1$. Based on this result and Proposition 2.3.1, it is thus possible to test asymptotic independence by studying $\pi(\mathbf{X}/t)$. This justifies afterwards the choice of the projection $\pi$ to study the extremal

dependence structure.

All in all, these two standard cases of multivariate EVT can be studied through the distribution of $\mathbf{Z}$. We do not lose any information by studying $\mathbf{Z}$ instead of $\mathbf{\Theta}$ in the asymptotically independent and completely dependent settings.

**A discrete model** We now extend the previous examples to a general discrete model. If $\beta \in \mathcal{P}_d^*$, we denote by $\mathbf{e}(\beta)$ the vector with 1 in position $i$ if $i \in \beta$ and 0 otherwise. Note that for all $\beta \in \mathcal{P}_d^*$, the vector $\mathbf{e}(\beta)/\#\beta$ belongs to the simplex $\mathbb{S}_+^{d-1}$.

We consider the following class of discrete distributions on the simplex:

$$\sum_{\beta \in \mathcal{P}_d^*} p(\beta)\, \delta_{\mathbf{e}(\beta)/\#\beta}\,, \tag{2.3.16}$$

where $(p(\beta))_\beta$ is a $2^d - 1$ vector with non-negative components summing to 1. This is the device developed in Segers (2012). Note that this class of distributions includes the previous cases, with respectively $p(\{1,\dots,d\}) = 1$ for complete dependence, and $p(\{j\}) = 1/d$, for all $j = 1,\dots,d$, for asymptotic independence.

The family of distributions (2.3.16) is stable after multiplying by a positive random variable and projecting onto the simplex with $\pi$. Hence, if $\mathbf{\Theta}$ has a distribution of type (2.3.16), then $\mathbf{Z} = \mathbf{\Theta}$ a.s. This shows that (2.3.16) forms an accurate model for the angular vector $\mathbf{Z}$. Indeed, it is stable for the transformation $\mathbf{\Theta} \mapsto \mathbf{Z}$. Besides, the distributions of this class have sparse supports. Finally, they put mass on some particular points of the simplex on which extremes values often concentrate in practice.

## 2.4 Sparse regular variation

We consider in this section a random vector $\mathbf{X}$ in $\mathbb{R}_+^d$. In Section 2.3, we assumed that $\mathbf{X}$ was regularly varying. In this case, convergence (2.3.2) holds and allows us to study the properties of $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. Our aim is now to establish a converse result. Thus, we do not assume anymore that $\mathbf{X}$ is regularly varying. We only start from convergence (2.3.2) which encourages to introduce the following definition.

**Definition 2.4.1** (Sparse regular variation)**.** A random vector $\mathbf{X}$ on $\mathbb{R}_+^d$ is sparsely regularly varying if there exist a random vector $\mathbf{Z}$ defined on the simplex $\mathbb{S}_+^{d-1}$ and a non-degenerate random variable $Y$ such that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \pi\left(\frac{\mathbf{X}}{t}\right)\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y,\mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{2.4.1}$$

In this case, the general theory of regular variation states that there exists $\alpha > 0$ such that $Y$ is Pareto($\alpha$)-distributed. With this definition in mind, we can rephrase the ideas of the beginning of Section 2.3, and particularly Equation (2.3.2), in the following way: Regular variation with limit

$(Y, \mathbf{\Theta})$ implies sparse regular variation with limit $(Y, \pi(Y\mathbf{\Theta}))$.

From now on, we consider a sparsely regularly varying random vector $\mathbf{X}$. Recall that we defined the function $G_{\mathbf{Z}}$ by $G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{P}(\mathbf{Z} > \mathbf{x})$ for $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d$. However, note that for the moment we can not write $G_{\mathbf{Z}}(\mathbf{x}) = \mathbb{P}(\pi(Y\mathbf{\Theta}) > \mathbf{x})$ since the existence of $\mathbf{\Theta}$ is not guaranteed. Our aim then is twofold. The first goal is to study the dependence between the radial limit $Y$ and the angular limit $\mathbf{Z}$ in (2.4.1). Secondly, we prove that under some assumptions on $G_{\mathbf{Z}}$ the vector $\mathbf{X}$ is regularly varying.

**Proposition 2.4.1.** *Let $\mathbf{X}$ be a sparsely regularly varying random vector on $\mathbb{R}_+^d$. Then, for all $r \geq 1$,*

$$\mathbf{Z} \mid Y > r \stackrel{d}{=} \pi(r\mathbf{Z})\,. \tag{2.4.2}$$

As already mentioned in Subsection 2.3.1, we do not have independence between the angular component $\mathbf{Z}$ and the radial one $Y$. However, the dependence between $\mathbf{Z}$ and $Y$ is completely determined by Equation (2.4.2) and will be helpful in the proof of Theorem 2.4.1.

Our aim is now to prove that, under some conditions, if $\mathbf{X}$ is a sparsely regularly varying vector, then $\mathbf{X}$ is regularly varying. Note that if convergence (2.4.1) holds, then $|\mathbf{X}|$ is regularly varying. So we need to focus on the convergence of the angular component, that is, of the self-normalized extreme $\mathbf{X}/|\mathbf{X}| \mid |\mathbf{X}| > t$ when $t \to \infty$. This idea is thus to provide a result which characterizes regular variation for a vector $\mathbf{X}$ when $|\mathbf{X}|$ is already regularly varying. This is the purpose of next lemma.

**Lemma 2.4.1.** *Let $\mathbf{X}$ be a random vector on $\mathbb{R}_+^d$ and $\alpha > 0$. The following assumptions are equivalent.*

1. *$\mathbf{X}$ is regularly varying with tail index $\alpha$.*
2. *$|\mathbf{X}|$ is regularly varying with tail index $\alpha$ and there exists a finite measure $l$ on $\mathbb{S}_+^{d-1}$ such that*

$$\lim_{\epsilon \to 0} \liminf_{t \to \infty} \epsilon^{-1} \mathbb{P}\left( \frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t \right) = l(A)\,, \tag{2.4.3}$$

*and*

$$\lim_{\epsilon \to 0} \limsup_{t \to \infty} \epsilon^{-1} \mathbb{P}\left( \frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t \right) = l(A)\,, \tag{2.4.4}$$

*for all continuity set $A \in \mathcal{B}(\mathbb{S}_+^{d-1})$ of $l$.*

*In this case, $l(A) = \alpha \mathbb{P}(\mathbf{\Theta} \in A)$, where $\mathbf{\Theta}$ is the spectral vector of $\mathbf{X}$.*

**Remark 2.4.1.** The assertion 2 of Lemma 2.4.1 can be weakened by taking $A$ in a family of Borel sets that generates $\mathcal{B}(\mathbb{S}_+^{d-1})$. In what follows, we will consider the family $\mathcal{A} = \{A_{\mathbf{x}}, \mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}_+^d\}$, where the $A_{\mathbf{x}}$ are defined in (2.3.6).

**Remark 2.4.2.** In Lemma 2.4.1, $|\cdot|$ denotes any norm of $\mathbb{R}^d$, but in what follows we will use this lemma for the $\ell^1$-norm.

We now prove that under mild assumptions on $G_{\mathbf{Z}}$, a random vector $\mathbf{X}$ which satisfies (2.4.1) is regularly varying. We denote by $\lambda$ the Lebesgue measure on the positive unit sphere $\mathcal{B}(0,1) \cap \mathbb{R}^d_+$. The assumptions on $G_{\mathbf{Z}}$ are the following ones:

(A1) The function $G_{\mathbf{Z}}$ is differentiable for $\lambda$-almost every $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}^d_+$.

(A2) $\mathbb{P}(\mathbf{Z} \in \partial A_{\mathbf{x}}) = 0$ for $\lambda$-almost every $\mathbf{x} \in \mathcal{B}(0,1) \cap \mathbb{R}^d_+$.

Let us denote by $\mathcal{Z}(G_{\mathbf{Z}})$ the set of vectors $\mathbf{x}$ in $\mathcal{B}(0,1) \cap \mathbb{R}^d_+$ which satisfy (A1) and (A2). Then, the family $\mathcal{A}_{\mathcal{Z}(G_{\mathbf{Z}})} := \{A_{\mathbf{x}}, \mathbf{x} \in \mathcal{Z}(G_{\mathbf{Z}})\}$ generates the Borel sets of $\mathbb{S}^{d-1}_+$. If there is no confusion, we will simply write $\mathcal{Z}$ for $\mathcal{Z}(G_{\mathbf{Z}})$ and $\mathcal{A}_{\mathcal{Z}}$ for $\mathcal{A}_{\mathcal{Z}(G_{\mathbf{Z}})}$.

**Theorem 2.4.1.** *Let $\mathbf{X}$ be a sparsely regularly varying random vector on $\mathbb{R}^d_+$. Assume that $G_{\mathbf{Z}}(\cdot) = \mathbb{P}(\mathbf{Z} > \cdot)$ satisfies (A1) and (A2). Then $\mathbf{X}$ is regularly varying with spectral vector $\mathbf{\Theta}$ which satisfies*

$$\mathbb{P}(\mathbf{\Theta} \in A_{\mathbf{x}}) = \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) + \alpha^{-1} \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)\,, \tag{2.4.5}$$

*for all $\mathbf{x} \in \mathcal{Z}$.*

This shows under mild assumptions the equivalence of regular variation and sparse regular variation. Moreover, the distribution of $\mathbf{Z}$ completely characterizes the one of $\mathbf{\Theta}$. Equation (2.4.5) completes the result (2.3.7) obtained in Proposition 2.3.2.

Let us summarize the results we obtained. Proposition 2.3.2 characterizes the distribution of $\mathbf{Z} = \pi(Y\mathbf{\Theta})$ when $\mathbf{X}$ is regularly varying with spectral vector $\mathbf{\Theta}$. Conversely, suppose that $\mathbf{X}$ is a sparsely regularly varying random vector. Then Theorem 2.4.1 states that $\mathbf{X}$ is regularly varying with a spectral vector $\mathbf{\Theta}$ which satisfies Equation (2.4.5). This ensures that $\mathbf{Z} = \pi(Y\mathbf{\Theta})$, where $Y$ is a Pareto($\alpha$)-distributed random variable independent of $\mathbf{\Theta}$. In other words, we have an almost complete equivalence between the usual regular variation's concept and sparse regular variation.

## 2.5 Numerical results

This section is devoted to a statistical illustration of sparse regular variation. The idea is to highlight how our approach manages to detect the tail dependence's sparsity. In the first subsection, we provide a method in order to approximate the probabilities $\mathbb{P}(\mathbf{Z} \in C_\beta)$, $\beta \in \mathcal{P}^*_d$, and we introduce the approach developed by Goix et al. (2017). The second subsection is dedicated to numerical results. A more general statistical framework will be developed in Chapter 3.

### 2.5.1 The framework

We consider an i.i.d. sequence of regularly varying random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with generic distribution $\mathbf{X}$, and with tail index $\alpha$ and spectral vector $\mathbf{\Theta} \in \mathbb{S}^{d-1}_+$. We set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$, where $Y$

follows a Pareto($\alpha$) distribution independent of $\mathbf{\Theta}$. Our aim is to capture the features $\beta \in \mathcal{P}_d^*$ in which the extreme values of $\mathbf{X}$ occur. Recall that these directions are characterized by the fact that $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, and therefore, by Corollary 2.3.1, $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$. Thanks to Proposition 2.3.1, the latter probability is defined through the limit

$$\mathbb{P}(\mathbf{Z} \in C_\beta) = \lim_{t\to\infty} \mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) = \lim_{t\to\infty} \frac{\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta,\, |\mathbf{X}| > t)}{\mathbb{P}(|\mathbf{X}| > t)}\,. \tag{2.5.1}$$

The goal is then to approximate this probability with the sample $\mathbf{X}_1, \dots, \mathbf{X}_n$. For $t > 0$, and $\beta \in \mathcal{P}_d^*$, we define the quantity

$$\hat{T}_n(t, \beta) = \frac{\sum_{j=1}^n \mathbb{1}\{\pi(\mathbf{X}_j/t) \in C_\beta, |\mathbf{X}_j| > t\}}{\sum_{j=1}^n \mathbb{1}\{|\mathbf{X}_j| > t\}}\,, \tag{2.5.2}$$

which corresponds to the proportion of data $\mathbf{X}_j$ whose projected vector $\pi(\mathbf{X}_j/t)$ belongs to $C_\beta$ among the data whose $\ell^1$-norm is above $t$. Intuitively, the larger the variable $\hat{T}_n(t, \beta)$, the more likely the feature $\beta$ concentrates extreme values.

The Law of Large Numbers ensures that

$$\lim_{n\to\infty} \hat{T}_n(t, \beta) = \frac{\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta,\, |\mathbf{X}| > t)}{\mathbb{P}(|\mathbf{X}| > t)}\,, \quad \text{a.s.} \tag{2.5.3}$$

Hence, Equations (2.5.3) and (2.5.1) lead to the following approximation:

$$\hat{T}_n(t, \beta) \approx \frac{\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta,\, |\mathbf{X}| > t)}{\mathbb{P}(|\mathbf{X}| > t)} \approx \mathbb{P}(\mathbf{Z} \in C_\beta)\,, \tag{2.5.4}$$

where the first approximation holds for $n$ large, and the second one for $t$ large. With this approximation, we can classify the subsets $C_\beta$ depending on the nullity or not of the associated quantity $\hat{T}_n(t, \beta)$.

**The approach proposed by Goix et al. (2017)** In order to detect anomalies among multivariate extremes, Goix et al. (2017) propose a similar approach by using the $\ell^\infty$-norm. They define the $\epsilon$-thickened rectangles by

$$R_\beta^\epsilon := \{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > 1, x_j > \epsilon \text{ for all } j \in \beta,\, x_j \le \epsilon \text{ for all } j \in \beta^c\}\,,$$

for $\beta \in \mathcal{P}_d^*$ (see Remark 2.3.1). In order to go back to the $\ell^\infty$ positive unit sphere $\mathbb{S}_{+,\infty}^{d-1}$, we define

$$C_{\beta,\infty}^\epsilon := \{\mathbf{x}/|\mathbf{x}|_\infty,\, \mathbf{x} \in R_\beta^\epsilon\}\,, \quad \text{and} \quad C_{\beta,\infty} := \left\{\mathbf{x} \in \mathbb{S}_{+,\infty}^d,\, x_i > 0 \text{ for } i \in \beta,\, x_i = 0 \text{ for } i \notin \beta\right\}.$$

Denoting by $\mathbf{\Theta}_\infty$ the spectral vector with respect to the $\ell^\infty$-norm, the convergence

$$\mathbb{P}(\mathbf{\Theta}_\infty \in C_{\beta,\infty}^\epsilon) \to \mathbb{P}(\mathbf{\Theta}_\infty \in C_{\beta,\infty})\,, \quad \epsilon \to 0\,,$$

proved in Goix et al. (2017) ensures that $\mathbb{P}(\mathbf{\Theta}_\infty \in C^\epsilon_{\beta,\infty})$ approximates the quantity $\mathbb{P}(\mathbf{\Theta}_\infty \in C_{\beta,\infty})$.

Similarly to Equation (2.5.2), we define the quantity

$$\hat{T}^\epsilon_n(t,\beta) = \frac{\sum_{j=1}^n \mathbb{1}\{\mathbf{X}_j/|\mathbf{X}_j|_\infty \in C^\epsilon_{\beta,\infty}, |\mathbf{X}_j|_\infty > t\}}{\sum_{j=1}^n \mathbb{1}\{|\mathbf{X}_j|_\infty > t\}}\,, \tag{2.5.5}$$

and the Law of Large Numbers ensures that, for $\epsilon > 0$ and $t > 0$ fixed,

$$\lim_{n\to\infty} \hat{T}^\epsilon_n(t,\beta) = \frac{\mathbb{P}(\mathbf{X}/|\mathbf{X}|_\infty \in C^\epsilon_{\beta,\infty},\ |\mathbf{X}|_\infty > t)}{\mathbb{P}(|\mathbf{X}|_\infty > t)} = \mathbb{P}(\mathbf{X}/|\mathbf{X}|_\infty \in C^\epsilon_{\beta,\infty} \mid |\mathbf{X}|_\infty > t)\,, \quad \text{a.s.} \tag{2.5.6}$$

Hence, the estimation of $\mathbb{P}(\mathbf{\Theta}_\infty \in C_\beta)$ is based on the following sequence of approximations:

$$\hat{T}^\epsilon_n(t,\beta) \approx \frac{\mathbb{P}(\mathbf{X}/|\mathbf{X}|_\infty \in C^\epsilon_{\beta,\infty},\ |\mathbf{X}|_\infty > t)}{\mathbb{P}(|\mathbf{X}|_\infty > t)} \approx \mathbb{P}(\mathbf{\Theta}_\infty \in C^\epsilon_{\beta,\infty}) \approx \mathbb{P}(\mathbf{\Theta}_\infty \in C_{\beta,\infty})\,, \tag{2.5.7}$$

where the first approximation holds for $n$, the second one for $t$ large, and the last one for $\epsilon$ close to zero. All these considerations lead to an algorithm, called DAMEX, introduced in Goix et al. (2017), Section 4.2.

**Remark 2.5.1** (On the choice of the norm)**.** We already mentioned that the spectral vector can be defined for any norm in $\mathbb{R}^d$. The choice of the $\ell^1$-norm in this chapter is deeply linked to the use of the projection $\pi$. On the other hand, Goix et al. (2017) choose the $\ell^\infty$-norm. After some calculations, we observe that both spectral vectors $\mathbf{\Theta}$ and $\mathbf{\Theta}_\infty$ satisfy the relation

$$\mathbb{P}(\mathbf{\Theta} \in B) = \frac{\mathbb{E}[|\mathbf{\Theta}_\infty|^\alpha \mathbb{1}_{\{\mathbf{\Theta}_\infty/|\mathbf{\Theta}_\infty| \in B\}}]}{\mathbb{E}[|\mathbf{\Theta}_\infty|^\alpha]}\,,$$

for all $B \in \mathbb{S}_+^{d-1}$. In particular,

$$\mathbb{P}(\mathbf{\Theta} \in C_\beta) = \frac{\mathbb{E}[|\mathbf{\Theta}_\infty|^\alpha \mathbb{1}_{\{\mathbf{\Theta}_\infty/|\mathbf{\Theta}_\infty| \in C_\beta\}}]}{\mathbb{E}[|\mathbf{\Theta}_\infty|^\alpha]}\,.$$

Since $\mathbf{\Theta}_\infty/|\mathbf{\Theta}_\infty| \in C_\beta$ if and only if $\mathbf{\Theta}_\infty \in C_{\beta,\infty}$, we obtain the equivalence

$$\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0 \quad \text{if and only if} \quad \mathbb{P}(\mathbf{\Theta}_\infty \in C_{\beta,\infty}) > 0\,.$$

In other words, the choice of the norm has no impact on the directions in which extremes gather.

**Remark 2.5.2.** At the end of our procedure, we obtain groups of directions $\beta$ such that $\hat{T}_n(t,\beta) > 0$. Since we deal with non-asymptotic data, many $\hat{T}_n(t,\beta)$ have small values while the theoretical quantities $\mathbb{P}(\mathbf{Z} \in C_\beta)$ are null. We follow the idea of Goix et al. (2017), Remark 4, to deal with this issue. We define a threshold value under which the empirical quantities $\hat{T}_n(t,\beta)$ are set to 0. We use a threshold of the form $p/\#B$, where $B = \{\beta,\ \hat{T}_n(t,\beta) > 0\}$ and where the hyperparameter $p \geq 0$ is fixed by the user. Of course, it is possible to set $p$ to 0 and then we select all subsets $C_\beta$ such that

$\hat{T}_n(t,\beta) > 0$. In this case, the number of selected $C_\beta$ is still much smaller than the total number $2^d - 1$. We do not detail more the choice of $p$ and defer this issue to future work.

Taking this hyperparameter $p$ into account, we are now able to introduce the algorithm used to study the dependence structure of sparsely regularly varying random vectors.

**Data:** $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}_+^d$, $t > 0$, $p \geq 0$
**Result:** A list $\mathcal{C}$ of subsets $C_\beta$
Compute $\pi(\mathbf{X}_j/t)$, $j = 1, \dots, n$;
Assign to each $\pi(\mathbf{X}_j/t)$ the subsets $C_\beta$ it belongs to;
Compute $\hat{T}_n(t,\beta)$;
Set to 0 the $\hat{T}_n(t,\beta)$ below the threshold discussed in Remark 2.5.2;
Define $\mathcal{C} = \{C_\beta, \hat{T}_n(t,\beta) > 0\}$.

**Algorithm 2:** Extremal dependence structure of sparsely regularly varying random vectors.

### 2.5.2 Experimental results

We consider two different cases for the numerical results. For each case, we generate datasets of size $n \in \{10^4, 5 \cdot 10^4, 10^5\}$ and we compute the quantities $\hat{T}_n(t,\beta)$ and $\hat{T}_n^\epsilon(t,\beta)$, for $\beta \in \mathcal{P}_d^*$. We repeat this procedure over $N = 100$ simulations. Note that there are two different types of errors which could arise. The first one corresponds to the occurrence of a feature $\beta$ while it should not appear theoretically. The second one corresponds to the absence of a feature $\beta$ while it should appear theoretically. The results correspond then to the average number of errors among the $N = 100$ simulations. The code can be found at https://github.com/meyernicolas/phd_thesis/blob/master/chap_2.

The purpose of these experiments is twofold. The first idea is to study Algorithm 2 and to see if it manages to detect the sparsity structure of the extremal data. The second goal of these simulations is to highlight some evidence in favor of our method compared to the DAMEX algorithm, which is based on a hyperparameter $\epsilon$. The results will show that there exists no natural choice for this hyperparameter. In other words, it may happen that for a fixed simulation study, there exists a specific hyperparameter $\epsilon_0$ for which the DAMEX algorithm leads to better results that our approach. But as soon as we use different simulated data, this specific $\epsilon_0$ is no longer appropriate.

**Remark 2.5.3** (Choice of the parameters)**.** It is common in EVT to define a level of exceedances $k = n\mathbb{P}(|\mathbf{X} > t)$ and to rather work with $k$ instead of $t$. For our simulations, we choose $k = \sqrt{n}$, following Goix et al. (2017), who also suggest choosing $\epsilon$ of order $k^{-1/4}$, that is, of order $n^{-1/8}$. This choice of $\epsilon$ is based on theoretical results (Goix et al. (2017), Theorem 1), but the authors then advise to rather choose $\epsilon = 0.01$, which gives better results on their simulations. In order to have a very large scale of comparison, we use different $\epsilon \in \{0.01, 0.1, 0.5, 1, 5, 10\}$. Finally, we consider $p = 0.3$ which is larger than the value chosen in Goix et al. (2017) but leads to better results for both methods.

**Asymptotic independence** We consider i.i.d. vectors $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^{40}$ with all marginals independent and Pareto(1)-distributed. This leads to asymptotic independence, which has already been discussed in Subsection 2.3.4. Equivalently, $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = 1/d$ for $\beta$ such that $\#\beta = 1$ (and therefore $\mathbb{P}(\mathbf{\Theta} \in C_\beta) = 0$ elsewhere). In other words, the spectral measure concentrates on the axis. Our aim is thus to recover these 40 directions among the $2^{40} - 1 \approx 10^{12}$ subsets $C_\beta$.

Table 2.1 shows the average number of errors among the 100 experiments. For the Euclidean projection, the number of errors is quite low in compared to the total number of subsets $C_\beta$, especially when $n$ increases. As expected, the angular vector $\mathbf{Z}$ is helpful to detect asymptotic independence since it is likely to concentrate on the axis.

For the DAMEX algorithm, a large $\epsilon$ leads theoretically to more mass assigned on the axis. The asymptotic independent case should therefore gives better results for large $\epsilon$. It is the case for our numerical results which become better when $\epsilon$ increases. This algorithm also gives better results than the one we propose for $\epsilon \geq 5$. However, some results seems difficult to interpret. Firstly, for $n = 10^4$, a choice of $\epsilon = 0.01$ leads to better results than $\epsilon = 0.1$ or $\epsilon = 0.5$. Secondly, for $\epsilon < 0.5$, the number of errors increases with $n$.

| | Euclidean projection | DAMEX $\epsilon = 0.01$ | DAMEX $\epsilon = 0.1$ | DAMEX $\epsilon = 0.5$ | DAMEX $\epsilon = 1$ | DAMEX $\epsilon = 5$ | DAMEX $\epsilon = 10$ |
|---|---|---|---|---|---|---|---|
| $n_1 = 10^4$ | 15.04 | 41 | 136.43 | 64.97 | 37.39 | 10.41 | 6.51 |
| $n_2 = 5 \cdot 10^4$ | 1.36 | 264 | 196.54 | 64.33 | 33.27 | 1.05 | 0.98 |
| $n_3 = 10^5$ | 0.47 | 356 | 221.30 | 64.81 | 0.99 | 0.39 | 0.53 |

Table 2.1: Average number of errors in an asymptotically independent case ($d = 40$).

**A dependent case** We now consider a dependent case where extremes do not appear on the axis. In order to include dependence we start from a regularly varying random variable $V \in \mathbb{R}_+$ with tail index $\alpha > 0$. Then, for $r \geq 2$, we consider $r - 1$ independent variables $P_1, \dots, P_{r-1} \in \mathbb{R}_+$, independent of $X$, such that $P_j$ is regularly varying with tail index $\alpha' > \alpha$. Finally, we consider the vector $\mathbf{V} \in \mathbb{R}_+^r$ whose components are defined as follows:

$$V_1 = V \quad \text{and} \quad V_{j+1} = a_j V + P_j, \text{ for } j = 1, \dots, r-1\,, \tag{2.5.8}$$

where $a_1, \dots, a_{r-1}$ are positive constants. In this case, the random vector $\mathbf{X}$ is regularly varying with tail index $\alpha$ and a spectral vector $\mathbf{\Theta}$ which concentrates on the interior of $\mathbb{S}_+^{r-1}$, that is, on the subset $C_{\{1,\dots,r\}}$.

For our simulations, we consider two vectors $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}_+^{10}$ defined as in (2.5.8), with $\alpha = 1$, $\alpha' = 2$, and $a_j \in \{0.1, 10\}$. This choice of $a_j$ implies that the vector $\mathbf{V}_i$, $i = 1, 2$, does not concentrate too much in the center of the subset $C_{\{1,\dots,10\}}$ but rather near the axis. Finally, we define the vector $\mathbf{X} \in \mathbb{R}_+^{20}$ as the concatenation of $\mathbf{V}_1$ and $\mathbf{V}_2$. In this dependent case, the mass of the spectral measure associated to $\mathbf{X}$ concentrates on both subsets $C_{\{1,\dots,10\}}$ and $C_{\{11,\dots,20\}}$. Our aim is then to

recover these two subsets among the $2^{20} \approx 10^6$ subsets $C_\beta$, based on i.i.d. vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ with the same distribution as $\mathbf{X}$.

Table 2.2 shows the average number of errors among the 100 experiments. As for the previous case, the Euclidean projection leads to a quite low number of errors compared to the total number of subsets $C_\beta$. This means that the vector $\mathbf{Z}$ mainly concentrates on the desired subsets $C_{\{1,\dots,10\}}$ and $C_{\{11,\dots,20\}}$. Besides, the number of errors slightly decreases when $n$ increases.

Regarding the DAMEX algorithm, it is more difficult to interpret the results. For $n = 10^4$, the choice of $\epsilon = 0.01$ seems very efficient but the number of errors then drastically increases with $n$. Contrary to the previous example, large values of $\epsilon$ do not appear suitable here, even if they provide more stable results than in the asymptotic independent case. It seems that the best compromise for the different values of $n$ is $\epsilon = 1$. Besides, on this type of data, the Euclidean projection provides better results than the DAMEX algorithm for all different choice of $\epsilon$ we make.

| | Euclidean projection | DAMEX $\epsilon = 0.01$ | DAMEX $\epsilon = 0.1$ | DAMEX $\epsilon = 0.5$ | DAMEX $\epsilon = 1$ | DAMEX $\epsilon = 5$ | DAMEX $\epsilon = 10$ |
|---|---|---|---|---|---|---|---|
| $n_1 = 10^4$ | 5.02 | 3.00 | 22.26 | 5.27 | 4.20 | 8.75 | 7.52 |
| $n_2 = 5 \cdot 10^4$ | 4.48 | 68.05 | 9.95 | 8.32 | 6.02 | 8.39 | 7.10 |
| $n_3 = 10^5$ | 4.18 | 47.42 | 8.21 | 7.11 | 5.81 | 7.82 | 6.88 |

Table 2.2: Average number of errors in a dependent case ($d = 20$).

These simulations show that there is no easy way to find an optimal value for $\epsilon$. Large $\epsilon$ provide good results in the asymptotic independent case, even slightly better that the ones obtained with the Euclidean projection. On the contrary, in the dependent case we propose, the DAMEX seems less efficient, and it is not obvious which value of $\epsilon$ should be chosen.

## 2.6 Conclusion

In this chapter, we introduce the notion of sparsely regularly varying random vectors in order to tackle the issues that arise with the standard notion of regular variation in a high dimensional setting. The idea to replace the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ by the projected one $\pi(\mathbf{X}/t)$ allows us to better capture the sparsity structure of the tail dependence. Our main result is the equivalence between sparse regular variation and regular variation under some mild assumptions.

The benefits of this new way of projecting are multiple. The first one is the sparser structure of the new angular vector $\mathbf{Z}$ compared to the one of $\mathbf{\Theta}$, which implies that the new vector $\mathbf{Z}$ seems more suitable to study extremes in high dimensions. Besides, contrary to the standard regular variation's framework, the sparsity of $\mathbf{Z}$ can be directly captured by studying $\pi(\mathbf{X}/t)$, as stated in Proposition 2.3.1. This means that the projection $\pi$ manages to circumvent to issue of the weak convergence in the definition of regularly varying random vectors. Finally, the results of Proposition 2.3.2 and Theorem 2.4.1 state that under some assumptions, there is a bijection between the spectral vector

$\mathbf{\Theta}$ and the new angular vector $\mathbf{Z}$.

Practically speaking, the advantages of the projection $\pi$ are twofold. Firstly, the Euclidean projection onto the simplex does not introduce any extra parameter. The introduction of $\epsilon$-thickened rectangles in Goix et al. (2017) leads to the choice of a suitable $\epsilon$. The numerical results introduced in Section 2.5 provide empirical evidence that there is no optimal $\epsilon$. Secondly, the algorithm which computes the projection $\pi$ takes linear time. Hence, the study of extreme events with $\pi$ can be done in reasonable time in high dimension. More generally, the numerical results provide some good results for our approach and encourage to further develop the statistical study of sparsely regularly varying random vectors.

Based on Theorem 2.4.1, we can now focus on the behavior of the vector $\mathbf{Z}$ rather than on the one of $\mathbf{\Theta}$. Relying on the theoretical results established in this chapter, we now propose on estimation procedure to learn the dependence structure of a regularly varying random vector $\mathbf{X}$. This is the purpose of Chapter 3 which tackles the estimation of the quantities $\mathbb{P}(\mathbf{Z} \in C_\beta)$ and the choice of the threshold $t$, or equivalently of the level $k$.

## 2.7 Proofs

*Proof of Lemma 2.2.1.* We use the relation $\pi_z(\mathbf{v}) = z\pi(\mathbf{v}/z)$ to simplify the problem:

$$\begin{aligned}
&\forall\, 0 < z \leq z',\ \forall\, \mathbf{v} \in \mathbb{R}_+^d,\ \pi_z(\pi_{z'}(\mathbf{v})) = \pi_z(\mathbf{v}) \\
\iff &\forall\, 0 < z \leq z',\ \forall\, \mathbf{v} \in \mathbb{R}_+^d,\ z\pi(z^{-1}\pi_{z'}(\mathbf{v})) = z\pi(\mathbf{v}/z) \\
\iff &\forall\, 0 < z \leq z',\ \forall\, \mathbf{v} \in \mathbb{R}_+^d,\ \pi(z'z^{-1}\pi(\mathbf{v}/z')) = \pi(\mathbf{v}/z) \\
\iff &\forall\, a \geq 1,\ \forall\, \mathbf{u} \in \mathbb{R}_+^d,\ \pi(a\pi(\mathbf{u})) = \pi(a\mathbf{u})\,.
\end{aligned}$$

So we need to prove this last equality. Let $a \geq 1$ and $\mathbf{u} \in \mathbb{R}_+^d$. We divide the proof into three steps. Recall that an expression of $\rho$ is given in (2.2.6).

STEP 1: We prove that $\rho_{a\mathbf{u}} \leq \rho_{\mathbf{u}}$.

Fix $j \in \{1, \ldots, d\}$ such that $\pi(a\mathbf{u})_j > 0$. This means that

$$au_j - \frac{1}{j}\left(\sum_{r=1}^{j} au_{(r)} - 1\right) > 0\,.$$

and thus,

$$u_j - \frac{1}{j}\sum_{r=1}^{j} u_{(r)} + \frac{1}{ja} > 0\,.$$

Since $a \geq 1$, we obtain

$$u_j - \frac{1}{j}\sum_{r=1}^{j} u_{(r)} + \frac{1}{j} > 0\,,$$

which means that $\pi(\mathbf{u})_j > 0$. This proves that $\rho_{a\mathbf{u}} \leq \rho_{\mathbf{u}}$.

STEP 2: We prove that $\rho_{a\pi(\mathbf{u})} = \rho_{a\mathbf{u}}$.

We recall that the definition of $\pi(\mathbf{u})$ is given by $\pi(\mathbf{u})_k = (u_k - \lambda_{\mathbf{u}})$ for $1 \leq k \leq \rho_{\mathbf{u}}$ and $\pi(\mathbf{u})_k = 0$ for $\rho_{\mathbf{u}} < k \leq d$.

- We first prove that $\rho_{a\mathbf{u}} \leq \rho_{a\pi(\mathbf{u})}$. Fix $j \in \{1, \ldots, d\}$ such that $\pi(a\mathbf{u})_j > 0$. Then

$$au_j - \frac{1}{j}\left(\sum_{r=1}^{j} au_{(r)} - 1\right) > 0\,.$$

Since $\pi(a\mathbf{u})_j > 0$, we have $j \leq \rho_{a\mathbf{u}}$, and with STEP 1 we obtain $j \leq \rho_{a\mathbf{u}} \leq \rho_{\mathbf{u}}$. So for all $r \leq j \leq \rho_{\mathbf{u}}$, $\pi(\mathbf{u})_r = (u_r - \lambda_{\mathbf{u}})$. Thus,

$$a(\pi(u)_j - \lambda_{\mathbf{u}}) - \frac{1}{j}\left(\sum_{r=1}^{j} a(\pi(u)_{(r)} - \lambda_{\mathbf{u}}) - 1\right) > 0\,,$$

which gives

$$a\pi(u)_j - \frac{1}{j}\left(\sum_{r=1}^{j} a\pi(u)_{(r)} - 1\right) > 0\,.$$

This means that $\pi(a\pi(\mathbf{u}))_j > 0$. Hence, $\rho_{a\mathbf{u}} \leq \rho_{a\pi(\mathbf{u})}$.

- We now prove that $\rho_{a\pi(\mathbf{u})} \leq \rho_{a\mathbf{u}}$. Fix $j \in \{1, \ldots, d\}$ such that $\pi(a\mathbf{u})_j = 0$. Then

$$au_j - \frac{1}{j}\left(\sum_{r=1}^{j} au_{(r)} - 1\right) \leq 0\,.$$

If $j \leq \rho_{\mathbf{u}}$, then for all $r \leq j$, $u_r = \pi(\mathbf{u})_r + \lambda_{\mathbf{u}}$, so that

$$a(\pi(\mathbf{u})_j + \lambda_{\mathbf{u}}) - \frac{1}{j}\left(\sum_{r=1}^{j} a(\pi(\mathbf{u})_{(r)} + \lambda_{\mathbf{u}}) - 1\right) \leq 0\,,$$

and finally

$$a\pi(\mathbf{u})_j - \frac{1}{j}\left(\sum_{r=1}^{j} a\pi(\mathbf{u})_{(r)} - 1\right) \leq 0\,,$$

which means that $\pi(a\pi(\mathbf{u}))_j = 0$.

If $j > \rho_{\mathbf{u}}$, then $\pi(\mathbf{u})_j = 0$, so $a\pi(\mathbf{u})_j = 0$, and finally $\pi(a\pi(\mathbf{u}))_j = 0$. Hence, $\rho_{a\pi(\mathbf{u})} \leq \rho_{a\mathbf{u}}$.

All in all, we proved that if $j \in \{1, \ldots, d\}$, then $\pi(a\mathbf{u})_j > 0$ if and only if $\pi(a\pi(\mathbf{u}))_j > 0$, which concludes STEP 2.

STEP 3: We prove that $\pi(a\mathbf{u}) = \pi(a\pi(\mathbf{u}))$.

With STEP 2, we know that $\rho := \rho_{a\pi(\mathbf{u})} = \rho_{a\mathbf{u}}$. This proves that for $j > \rho$, $\pi(a\mathbf{u})_j$ and $\pi(a\pi(\mathbf{u}))_j$ are both null. Moreover, by definition of the projection $\pi$, if $j \leq \rho$,

$$\pi(a\mathbf{u})_j = au_j - \frac{1}{\rho}\Big(\sum_{r=1}^{\rho} au_{(r)} - 1\Big)\,.$$

Since $\rho \leq \rho_{\mathbf{u}}$ (with STEP 1), we use that for all $r \leq \rho$, $\pi(\mathbf{u})_{(r)} = u_{(r)} - \lambda_{\mathbf{u}}$. Thus, we obtain

$$\pi(a\mathbf{u}) = a(\pi(\mathbf{u})_j - \lambda_{\mathbf{u}}) - \frac{1}{\rho}\Big(\sum_{r=1}^{\rho} a(\pi(\mathbf{u})_{(r)} - \lambda_{\mathbf{u}}) - 1\Big) = au_j - \frac{1}{\rho}\Big(\sum_{r=1}^{\rho} au_{(r)} - 1\Big) = \pi(a\pi(\mathbf{u}))_j\,,$$

which concludes the proof. □

*Proof of Lemma 2.2.2.* Let $\mathbf{v} \in \mathbb{R}^d_+$. We sort $\mathbf{v}$ in $\boldsymbol{\mu}$ such that $\mu_1 \geq \ldots \geq \mu_d$. Firstly, note that if two coordinates of $\mathbf{v}$ are equal, then the corresponding coordinates of $\pi(\mathbf{v})$ are equal too. Thus, they are both null or both positive. So the way these two coordinates are ordered in $\boldsymbol{\mu}$ does not matter.

Let us prove the equivalence (2.2.7) For $i \in \beta^c$, let $j \in \{1,\ldots,d\}$ such that $\mu_j = v_i$, and let $\gamma^c \subset \{1,\ldots,d\}$ be the subset of such $j$. By definition of $\rho_{\mathbf{v}}$, the projected vector $\pi(\mathbf{v})$ satisfies $\pi(\mathbf{v})_{\beta^c} = 0$ if and only if for all $j \in \gamma^c$, $j > \rho_{\mathbf{v}}$, which means

$$\mu_j - \frac{1}{j}\Big(\sum_{k=1}^{j} \mu_k - 1\Big) \leq 0\,. \tag{2.7.1}$$

Note that $j = \sum_{k=1}^d \mathbb{1}_{v_k \geq v_i}$ and $\sum_{k=1}^j \mu_k = \sum_{k=1}^d v_k \mathbb{1}_{v_k \geq v_i}$, so that condition (2.7.1) can be rephrased as

$$v_i - \frac{1}{\sum_{k=1}^d \mathbb{1}_{v_k \geq v_i}}\Big(\sum_{k=1}^{d} v_k \mathbb{1}_{v_k \geq v_i} - 1\Big) \leq 0\,.$$

This inequality is equivalent to

$$1 \leq \sum_{k=1}^{d} (v_k - v_i)_+\,,$$

which proves (2.2.7).

For (2.2.8), set $r = |\beta| \geq 1$ (note that $\beta = \emptyset$ is not possible). Then, the condition $\pi(\mathbf{v}) \in C_\beta$ imply that $\rho_{\mathbf{v}} = r$. Thus, we obtain

$$\forall i \in \beta,\ v_i = \pi(\mathbf{v})_i + \frac{1}{r}\Big(\sum_{j\in\beta} v_j - 1\Big) \quad \text{and} \quad \forall i \in \beta^c,\ v_i \leq \frac{1}{r}\Big(\sum_{j\in\beta} v_j - 1\Big)\,.$$

On the one hand, since $\pi(\mathbf{v})_i > 0$ for $i \in \beta$, the first equality is equivalent to

$$\max_{i\in\beta} \sum_{j\in\beta} (v_j - v_i) < 1\,.$$

On the other hand, the second equality is equivalent to

$$\min_{i\in\beta^c} \sum_{j\in\beta} (v_j - v_i) \geq 1\,.$$

□

*Proof of Proposition 2.3.1.* We only prove (2.3.3). The proof of (2.3.4) is similar.

Let $\beta \in \mathcal{P}_d^*$. Following Lemma 2.2.2, we have the equivalence

$$\pi(Y\boldsymbol{\Theta}) \in C_\beta \quad \text{if and only if} \quad \begin{cases} \max_{i\in\beta} \sum_{j\in\beta} (\Theta_j - \Theta_i) < 1/Y\,, \\ \min_{i\in\beta^c} \sum_{j\in\beta} (\Theta_j - \Theta_i) \geq 1/Y\,. \end{cases}$$

Hence, (2.3.3) is equivalent to

$$\mathbb{P}\big((|\mathbf{X}|/t, \mathbf{X}/|\mathbf{X}|) \in D_\beta \mid |\mathbf{X}| > t\big) \to \mathbb{P}((Y, \boldsymbol{\Theta}) \in D_\beta)\,, \tag{2.7.2}$$

with

$$D_\beta = \Big\{(r,\theta) \in (1,\infty) \times \mathbb{S}_+^{d-1},\, \forall i \in \beta,\, \sum_{j\in\beta} (\theta_j - \theta_i) < 1/r,\ \text{and } \forall i \in \beta^c,\, \sum_{j\in\beta} (\theta_j - \theta_i) \geq 1/r\Big\}\,.$$

This convergence holds if $\mathbb{P}((Y, \boldsymbol{\Theta}) \in \partial D_\beta) = 0$.

The boundary $\partial D_\beta$ of $D_\beta$ is included in the union of the subsets

$$\bigcup_{i\in\beta} \partial\Big\{(r,\theta) \in (1,\infty) \times \mathbb{S}_+^{d-1},\, \sum_{j\in\beta}^{d} (\theta_j - \theta_i) < 1/r\Big\}\,,$$

and

$$\bigcup_{i\in\beta^c} \partial\Big\{(r,\theta) \in (1,\infty) \times \mathbb{S}_+^{d-1},\, \sum_{j\in\beta}^{d} (\theta_j - \theta_i) \geq 1/r\Big\}\,,$$

and for all $i = 1, \ldots, d$, we have the equality

$$\partial\Big\{(r,\theta) \in (1,\infty) \times \mathbb{S}_+^{d-1},\, \sum_{j\in\beta}^{d} (\theta_j - \theta_i) < 1/r\Big\} = \Big\{(r,\theta) \in (1,\infty) \times \mathbb{S}_+^{d-1},\, \sum_{j\in\beta}^{d} (\theta_j - \theta_i) = 1/r\Big\}\,,$$

and similarly

$$\partial\Big\{(r,\theta)\in(1,\infty)\times\mathbb{S}_+^{d-1},\ \sum_{j\in\beta}^{d}(\theta_j-\theta_i)\geq 1/r\Big\}=\Big\{(r,\theta)\in(1,\infty)\times\mathbb{S}_+^{d-1},\ \sum_{j\in\beta}^{d}(\theta_j-\theta_i)=1/r\Big\}.$$

This implies that

$$\mathbb{P}((Y,\boldsymbol{\Theta})\in\partial D_\beta)\leq\sum_{i\in\beta}\mathbb{P}\Big(\sum_{j\in\beta}(\Theta_j-\Theta_i)=Y^{-1}\Big)+\sum_{i\in\beta^c}\mathbb{P}\Big(\sum_{j\in\beta}(\Theta_j-\Theta_i)=Y^{-1}\Big),$$

and all these probabilities are null since $Y$ is a continous random variable independent of $\boldsymbol{\Theta}$. Hence, we proved that $\mathbb{P}((Y,\boldsymbol{\Theta})\in\partial D_\beta)=0$ which implies that convergence (2.7.2) holds and then convergence (2.3.3) holds as well. □

*Proof of Proposition 2.3.2.* Fix $\mathbf{x}\in\mathcal{B}(0,1)\cap\mathbb{R}_+^d$, with $x_j\neq 1/d$ for all $j=1,\dots,d$. We use (2.2.9) and the independence of $\boldsymbol{\Theta}$ and $Y$ to write

$$G_{\mathbf{Z}}(\mathbf{x})=\mathbb{P}(\mathbf{Z}>x)=\mathbb{P}(\pi(Y\boldsymbol{\Theta})>x)=\int_1^\infty\mathbb{P}(y\boldsymbol{\Theta}-(y-1)/d>\mathbf{x})\,\mathrm{d}(-y^{-\alpha}).$$

Set $J_+=\{j,\ x_j>1/d\}$ and $J_-=\{j,\ x_j<1/d\}$. Then, for $j\in J_+$, the condition $y\Theta_j-(y-1)/d>x_j$ becomes $[(\Theta_j-1/d)/(x_j-1/d)]_+>1/y$. Similarly, for $j\in J_-$, the condition $y\Theta_j-(y-1)/d>x_j$ becomes $[(\Theta_j-1/d)/(x_j-1/d)]_+<1/y$. So we can rewrite the previous integral as

$$\begin{aligned}G_{\mathbf{Z}}(\mathbf{x})&=\mathbb{P}(\pi(Y\boldsymbol{\Theta})>x)=\int_1^\infty\mathbb{P}(y\boldsymbol{\Theta}-(y-1)/d>\mathbf{x})\,\mathrm{d}(-y^{-\alpha})\\&=\int_1^\infty\mathbb{P}\left(\left\{\forall j\in J_+,\ y^{-\alpha}<\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right\}\bigcap\left\{\forall j\in J_-,\ y^{-\alpha}>\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right\}\right)\mathrm{d}(-y^{-\alpha}).\end{aligned}$$

Thus, by the change of variable $u=y^{-\alpha}$, we obtain

$$\begin{aligned}G_{\mathbf{Z}}(\mathbf{x})&=\int_0^1\mathbb{P}\left(\left\{\forall j\in J_+,\ u<\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right\}\bigcap\left\{\forall j\in J_-,\ u>\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right\}\right)\mathrm{d}u\\&=\int_0^1\mathbb{P}\left(\max_{j\in J_-}\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha<u<\min_{j\in J_+}\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right)\mathrm{d}u\\&=\mathbb{E}\left[\left(1\wedge\min_{j\in J_+}\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha-\max_{j\in J_-}\left(\frac{\Theta_j-1/d}{x_j-1/d}\right)_+^\alpha\right)_+\right].\end{aligned}$$

□

*Proof of Proposition 2.3.3.* We fix $\beta\in\mathcal{P}_d^*$ and use Lemma 2.2.2. The probability that $\mathbf{Z}_{\beta^c}$ is null is

equal to

$$\begin{aligned}\mathbb{P}(\mathbf{Z}_{\beta^c}=0) &= \mathbb{P}\Big(1 \le \min_{j\in\beta^c}\sum_{k=1}^{d}(Y\Theta_k - Y\Theta_j)_+\Big)\\ &= \mathbb{P}\Big(Y^{-\alpha} \le \min_{j\in\beta^c}\Big(\sum_{k=1}^{d}(\Theta_k-\Theta_j)_+\Big)^{\alpha}\Big)\\ &= \int_0^1 \mathbb{P}\Big(u \le \min_{j\in\beta^c}\Big(\sum_{k=1}^{d}(\Theta_k-\Theta_j)_+\Big)^{\alpha}\Big)\,\mathrm{d}u\\ &= \mathbb{E}\left[\min_{j\in\beta^c}\Big(\sum_{k=1}^{d}(\Theta_k-\Theta_j)_+\Big)^{\alpha}\right],\end{aligned}$$

which proves (2.3.10).

For Equation (2.3.11), we use Lemma 2.2.2, so that the probability that $\mathbf{Z}$ is concentrated on $C_\beta$ is equal to

$$\begin{aligned}\mathbb{P}(\mathbf{Z}\in C_\beta) &= \mathbb{P}\Big(\max_{j\in\beta}\sum_{k\in\beta}(Y\Theta_k - Y\Theta_j) < 1,\ \min_{j\in\beta^c}\sum_{k\in\beta}(Y\Theta_k - Y\Theta_j) \ge 1\Big)\\ &= \mathbb{P}\Big(\Big(\max_{j\in\beta}\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha} < Y^{-\alpha},\ \min_{j\in\beta^c}\Big(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha} \ge Y^{-\alpha}\Big)\\ &= \int_0^1 \mathbb{P}\Big(\max_{j\in\beta}\Big(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha} < u \le \min_{j\in\beta^c}\Big(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha}\Big)\,\mathrm{d}u\\ &= \mathbb{E}\left[\Big(\min_{j\in\beta^c}\Big(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha} - \max_{j\in\beta}\Big(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\Big)^{\alpha}\Big)_+\right].\end{aligned}$$

This concludes the proof of the proposition. □

*Proof of Corollary 2.3.1.* The proof of Corollary 2.3.1 is based on the following lemma, whose result will also be used in other proofs.

**Lemma 2.7.1.** *Let $\beta \in \mathcal{P}_d^*$. Then we have the inequality*

$$\mathbb{P}(\mathbf{\Theta}\in C_\beta) \le \mathbb{P}\Big(\max_{j\in\beta}\sum_{k\in\beta}(\Theta_k-\Theta_j)_+ < 1\Big). \tag{2.7.3}$$

*Proof of Lemma 2.7.1.* While Lemma 2.7.1 is stated and used in this way, we rather prove the following inequality:

$$\mathbb{P}\Big(\max_{j\in\beta}\sum_{k\in\beta}(\Theta_k-\Theta_j)_+ = 1\Big) \le \mathbb{P}(\mathbf{\Theta}\notin C_\beta).$$

The first probability can be rephrased as follows:

$$\mathbb{P}\Big(\max_{j\in\beta}\sum_{k\in\beta}(\Theta_k-\Theta_j)_+=1\Big)=\mathbb{P}\Big(\sum_{k\in\beta}(\Theta_k-\min_{j\in\beta}\Theta_j)=1\Big)=\mathbb{P}\Big(\sum_{k\in\beta}\Theta_k=1+\#\beta\min_{j\in\beta}\Theta_j\Big)\,.$$

But since $\boldsymbol{\Theta}\in\mathbb{S}_+^{d-1}$, the equality $\sum_{k\in\beta}\Theta_k=1+\#\beta\min_{j\in\beta}\Theta_j$ holds only if there exists $k\in\beta$ such that $\Theta_k=0$. Thus, we obtain the inequality

$$\mathbb{P}\Big(\max_{j\in\beta}\sum_{k\in\beta}(\Theta_k-\Theta_j)_+=1\Big)\le\mathbb{P}(\exists k\in\beta\,,\Theta_k=0)\le\mathbb{P}(\boldsymbol{\Theta}\notin C_\beta)\,,$$

which concludes the proof. □

We now prove Corollary 2.3.1. We fix $\beta\in\mathcal{P}_d^*$ and assume that $\mathbb{P}(\boldsymbol{\Theta}\in C_\beta)>0$. Then, starting from Equation (2.3.11), we write

$$\begin{aligned}
\mathbb{P}(\mathbf{Z}\in C_\beta)&\ge\mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\right)^\alpha-\max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\right)^\alpha\right)_+\mathbb{1}_{\boldsymbol{\Theta}\in C_\beta}\right]\\
&=\mathbb{E}\left[\left(\left(\sum_{k\in\beta}\Theta_k\right)^\alpha-\max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\right)^\alpha\right)_+\mathbb{1}_{\boldsymbol{\Theta}\in C_\beta}\right]\\
&=\mathbb{E}\left[\left(1-\max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\right)^\alpha\right)\mathbb{1}_{\boldsymbol{\Theta}\in C_\beta}\right]\\
&=\mathbb{E}\left[\left(1-\max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k-\Theta_j)_+\right)^\alpha\right)\mid\boldsymbol{\Theta}\in C_\beta\right]\mathbb{P}(\boldsymbol{\Theta}\in C_\beta)\,.
\end{aligned}$$

The expectation is positive by Lemma 2.7.1 and the probability $\mathbb{P}(\boldsymbol{\Theta}\in C_\beta)$ is positive by assumption. This shows that $\mathbb{P}(\mathbf{Z}\in C_\beta)>0$. □

*Proof of Theorem 2.3.1.* We separatly prove both implications.

We first consider $\beta\in\mathcal{P}_d^*$ such that $C_\beta$ is a maximal subset for $\boldsymbol{\Theta}$:

$$\mathbb{P}(\boldsymbol{\Theta}\in C_\beta)>0\quad\text{and}\quad\mathbb{P}(\boldsymbol{\Theta}\in C_{\beta'})=0,\ \text{for}\ \beta'\supsetneq\beta\,.$$

By Corollary 2.3.1, we already know that $\mathbb{P}(\mathbf{Z}\in C_\beta)>0$. Besides, if $\beta'\supsetneq\beta$, then Equation (2.3.14) gives

$$\mathbb{P}(\mathbf{Z}\in C_{\beta'})\le\mathbb{P}(\mathbf{Z}_{\beta'}>0)\le\mathbb{P}(\boldsymbol{\Theta}_{\beta'}>0)\,.$$

and this last probability equals zero since $C_\beta$ is a maximal subset for $\boldsymbol{\Theta}$. This proves that $C_\beta$ is a maximal subset for $\mathbf{Z}$.

We now consider $\beta\in\mathcal{P}_d^*$ such that $C_\beta$ is a maximal subset for $\mathbf{Z}$:

$$\mathbb{P}(\mathbf{Z}\in C_\beta)>0\quad\text{and}\quad\mathbb{P}(\mathbf{Z}\in C_{\beta'})=0,\ \text{for}\ \beta'\supsetneq\beta\,.$$

First note that, for $\beta' \supsetneq \beta$, $\mathbb{P}(\mathbf{\Theta} \in C_{\beta'}) = 0$. If not, Corollary 2.3.1 implies that $\mathbb{P}(\mathbf{Z} \in C_{\beta'}) > 0$, which contradicts the maximality of $C_\beta$ for $\mathbf{Z}$.

Secondly, Equation (2.3.11) of Proposition 2.3.3 gives

$$\begin{aligned}
\mathbb{P}(\mathbf{Z} \in C_\beta) &= \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+\right] \\
&= \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+ \mathbb{1}_{\mathbf{\Theta}\in C_\beta}\right] \\
&+ \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+ \mathbb{1}_{\mathbf{\Theta}\notin C_\beta}\right] \\
&= (A) + (B)\,.
\end{aligned} \tag{2.7.4}$$

The first term $(A)$ has already been calculated in the proof of Corollary 2.3.1. It is equal to

$$(A) = \mathbb{E}\left[1 - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha \mid \mathbf{\Theta} \in C_\beta\right]\mathbb{P}\Big(\mathbf{\Theta} \in C_\beta\Big)\,.$$

For the second term $(B)$, note that the assumption $\mathbf{\Theta} \notin C_\beta$ implies that there exists $l \in \beta$ such that $\Theta_l = 0$, or that there exists $r \in \beta^c$ such that $\Theta_r > 0$. We then decompose $(B)$ into two terms:

$$\begin{aligned}
&\mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+ \mathbb{1}_{\mathbf{\Theta}\notin C_\beta}\right] \\
&\le \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+ \mathbb{1}_{\exists l\in\beta,\,\Theta_l=0}\right] \\
&+ \mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha\right)_+ \mathbb{1}_{\exists\beta'\supsetneq\beta,\,\mathbf{\Theta}\in C_{\beta'}}\right]\,.
\end{aligned}$$

The first expectation is equal to

$$\mathbb{E}\left[\left(\min_{j\in\beta^c}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha - \left(\sum_{k\in\beta}(\Theta_k)_+\right)^\alpha\right)_+ \mathbb{1}_{\exists l\in\beta,\,\Theta_l=0}\right]\,,$$

and is thus zero. The second expectation is smaller than $\mathbb{P}(\exists\beta' \supsetneq \beta,\, \mathbf{\Theta} \in C_{\beta'})$ which is zero. Indeed, if $\mathbb{P}(\exists\beta' \supsetneq \beta,\, \mathbf{\Theta} \in C_{\beta'}) > 0$, then by Corollary 2.3.1, we also have $\mathbb{P}(\exists\beta' \supsetneq \beta,\, \mathbf{Z} \in C_{\beta'}) > 0$, which contradicts the maximality of $C_\beta$ for $\mathbf{Z}$. All in all, this proves that $(B) = 0$.

Going back to Equation (2.7.4), we have proved that

$$\mathbb{P}(\mathbf{Z} \in C_\beta) = (A) = \mathbb{E}\left[1 - \max_{j\in\beta}\left(\sum_{k\in\beta}(\Theta_k - \Theta_j)_+\right)^\alpha \mid \mathbf{\Theta} \in C_\beta\right]\mathbb{P}(\mathbf{\Theta} \in C_\beta)\,.$$

By Lemma 2.7.1, we know that the expectation is positive. Hence, the assumption $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$ implies that $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$, which proves that $C_\beta$ is a maximal subset of $\mathbf{\Theta}$. □

*Proof of Proposition 2.4.1.* Fix $r \geq 1$ and $A \in \mathcal{B}(\mathbb{S}_+^{d-1})$. For $t > 0$, the following sequence of equalities holds:

$$\begin{aligned}
\mathbb{P}\Big(\pi(\mathbf{X}/t) \in A,\, |\mathbf{X}|/t > r \mid |\mathbf{X}| > t\Big) &= \mathbb{P}\Big(\pi(\mathbf{X}/t) \in A,\, |\mathbf{X}|/t > r\Big)\frac{1}{\mathbb{P}(|\mathbf{X}| > t)} \\
&= \mathbb{P}\Big(\pi(\mathbf{X}/t) \in A \mid |\mathbf{X}|/t > r\Big)\frac{\mathbb{P}(|\mathbf{X}| > tr)}{\mathbb{P}(|\mathbf{X}| > t)} \\
&= \mathbb{P}\Big(\pi\big(r\mathbf{X}/(tr)\big) \in A \mid |\mathbf{X}| > tr\Big)\mathbb{P}(|\mathbf{X}| > tr \mid |\mathbf{X}| > t) \\
&= \mathbb{P}\left(r\pi_{1/r}\left(\frac{\mathbf{X}}{tr}\right) \in A \,\middle|\, |\mathbf{X}| > tr\right)\mathbb{P}(|\mathbf{X}| > tr \mid |\mathbf{X}| > t) \\
&= \mathbb{P}\left(r\pi_{1/r}\left(\pi\left(\frac{\mathbf{X}}{tr}\right)\right) \in A \,\middle|\, |\mathbf{X}| > tr\right)\mathbb{P}(|\mathbf{X}| > tr \mid |\mathbf{X}| > t)\,,
\end{aligned}$$

where last equality results from Lemma 2.2.1. Now, when $t \to \infty$, assumption (2.4.1) and the continuity of $\pi_{1/r}$ and $\pi$ give

$$\mathbb{P}(\mathbf{Z} \in A,\, Y > r) = \mathbb{P}\left(r\pi_{1/r}\left(\mathbf{Z}\right) \in A\right)\mathbb{P}(Y > r)\,.$$

Finally, we conclude the proof with Lemma 2.2.1:

$$\mathbb{P}(\mathbf{Z} \in A \mid Y > r) = \mathbb{P}(\pi(r\mathbf{Z}) \in A)\,.$$

□

*Proof of Lemma 2.4.1.* We first prove that 1 implies 2: assume that $\mathbf{X}$ is regularly varying with index $\alpha$. Then $|\mathbf{X}|$ is regularly varying with the same index. Denote by $\mathbf{\Theta}$ the spectral vector of $\mathbf{X}$ and consider a random variable $Y$ which follows a Pareto($\alpha$) distribution and is independent of $\mathbf{\Theta}$. For $A \in \mathcal{B}(\mathbb{S}_+^{d-1})$ such that $\mathbb{P}(\mathbf{\Theta} \in \partial A) = 0$, and $\epsilon > 0$, we have

$$\begin{aligned}
\epsilon^{-1}\lim_{t\to\infty}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon],\, \frac{\mathbf{X}}{|\mathbf{X}|} \in A \,\middle|\, |\mathbf{X}| > t\right) &= \epsilon^{-1}\mathbb{P}(Y \in (1, 1+\epsilon], \mathbf{\Theta} \in A) \\
&= \epsilon^{-1}\mathbb{P}(Y \leq 1+\epsilon)\mathbb{P}(\mathbf{\Theta} \in A) \\
&= \epsilon^{-1}(1 - (1+\epsilon)^{-\alpha})\mathbb{P}(\mathbf{\Theta} \in A)\,.
\end{aligned}$$

This last quantity converges to $\alpha\mathbb{P}(\mathbf{\Theta} \in A)$ when $\epsilon \to 0$, which proves that $\mathbf{X}$ satisfies (2.4.3) and

(2.4.4) with $l(\cdot) = \alpha\mathbb{P}(\mathbf{\Theta} \in \cdot)$.

We now prove that 2 implies 1. Fix $\epsilon > 0$, $u > 1$, and $A \in \mathcal{B}(\mathbb{S}_+^{d-1})$ such that $l(\partial A) = 0$. Denote by $l_\epsilon^+(A)$ the limsup in (2.4.3) when $t \to \infty$, and by $l_\epsilon^-(A)$ the liminf in (2.4.4) when $t \to \infty$. For $u \geq 1$, we decompose the interval $(u, \infty)$ as follows:

$$(u, \infty) = \bigsqcup_{k=0}^{\infty} \left(u(1+\epsilon)^k, u(1+\epsilon)^{k+1}\right].$$

Then for $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{|\mathbf{X}|}{t} > u, \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t\right) \\
&\quad = \sum_{k=0}^{\infty} \mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t\right) \\
&\quad = \sum_{k=0}^{\infty} \frac{\mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A\right)}{\mathbb{P}(|\mathbf{X}| > t)} \\
&\quad = \epsilon \sum_{k=0}^{\infty} \epsilon^{-1} \mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; \frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right) \frac{\mathbb{P}\left(\frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right)}{\mathbb{P}(|\mathbf{X}| > t)}.
\end{aligned}$$

Since $|\mathbf{X}|$ is regularly varying with tail index $\alpha$, the limit of the right part of the sum can be computed as follows:

$$\frac{\mathbb{P}\left(\frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right)}{\mathbb{P}(|\mathbf{X}| > t)} = \mathbb{P}\left(|\mathbf{X}| > tu(1+\epsilon)^k \mid |\mathbf{X}| > t\right) \to \left(u(1+\epsilon)^k\right)^{-\alpha}, \quad t \to \infty. \tag{2.7.5}$$

Besides, we know by (2.4.3) that

$$\liminf_{t\to\infty} \epsilon^{-1} \mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; \frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right) = l_\epsilon^-(A). \tag{2.7.6}$$

We now gather (2.7.5) and (2.7.6) and use Fatou's lemma to conclude:

$$\begin{aligned}
&\liminf_{t\to\infty} \mathbb{P}\left(\frac{|\mathbf{X}|}{t} > u, \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t\right) \\
&\quad \geq \epsilon \sum_{k=0}^{\infty} \liminf_{t\to\infty} \epsilon^{-1} \mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; \frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right) \frac{\mathbb{P}\left(\frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right)}{\mathbb{P}(|\mathbf{X}| > t)} \\
&\quad = \epsilon \sum_{k=0}^{\infty} l_\epsilon^-(A) \left(u(1+\epsilon)^k\right)^{-\alpha} \\
&\quad = u^{-\alpha} l_\epsilon^-(A) \frac{\epsilon}{1 - (1+\epsilon)^{-\alpha}},
\end{aligned}$$

and this last quantity converges to $u^{-\alpha} l(A) \alpha^{-1}$ when $\epsilon \to 0$.

In the same way, we know by (2.4.4) that

$$\limsup_{t\to\infty} \epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1,1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; \frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right) = l_\epsilon^+(A)\,. \tag{2.7.7}$$

Thus, Equations (2.7.5) and (2.7.7) and Fatou's lemma allow us to write

$$\begin{aligned}
&\limsup_{t\to\infty} \mathbb{P}\left(\frac{|\mathbf{X}|}{t} > u, \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t\right) \\
&\qquad \le \epsilon \sum_{k=0}^{\infty} \limsup_{t\to\infty} \epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{tu(1+\epsilon)^k} \in (1,1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; \frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right) \frac{\mathbb{P}\left(\frac{|\mathbf{X}|}{u(1+\epsilon)^k} > t\right)}{\mathbb{P}(|\mathbf{X}| > t)} \\
&\qquad = \epsilon \sum_{k=0}^{\infty} l_\epsilon^+(A)\big(u(1+\epsilon)^k\big)^{-\alpha} \\
&\qquad = u^{-\alpha} l_\epsilon^+(A) \frac{\epsilon}{1-(1+\epsilon)^{-\alpha}}\,,
\end{aligned}$$

and this last quantity converges to $u^{-\alpha}l(A)\alpha^{-1}$ when $\epsilon \to 0$.

This proves that

$$\mathbb{P}\left(\frac{|\mathbf{X}|}{t} > u, \frac{\mathbf{X}}{|\mathbf{X}|} \in A \;\middle|\; |\mathbf{X}| > t\right) \to u^{-\alpha} l(A)\alpha^{-1}\,, \quad t \to \infty\,,$$

for all $u > 1$ and all $A \in \mathcal{B}(\mathbb{S}_+^{d-1})$ such that $l(\partial A) = 0$. Thus, the random vector $\mathbf{X}$ is regularly varying with tail index $\alpha$ and spectral vector $\mathbf{\Theta}$ defined by $\mathbb{P}(\mathbf{\Theta} \in \cdot) = \alpha^{-1} l(\cdot)$. □

*Proof of Theorem 2.4.1.* The proof is based on Lemma 2.4.1. Firstly, note that if (2.4.1) holds, then $|\mathbf{X}|$ is regularly varying with tail index $\alpha$. Hence, the main part of the proof is to show that convergences (2.4.3) and (2.4.4) hold for all $A = A_{\mathbf{x}}$, $\mathbf{x} \in \mathcal{Z}$, where the $A_{\mathbf{x}}$ are defined in (2.3.6). We divide our proof into two steps.

Before dealing with these two steps, we make a brief remark which will be of constant use. For $\epsilon > 0$ and $\mathbf{x} > 0$, we have the following equivalence:

$$\pi((1+\epsilon)\mathbf{Z}) > \mathbf{x} \iff \mathbf{Z} > \frac{\mathbf{x} + \epsilon/d}{1+\epsilon}\,. \tag{2.7.8}$$

This is a consequence of Equation (2.2.9) and the fact that $\mathbf{Z}$ belongs to the simplex.

Let us move to the proof. We fix $\mathbf{x} \in \mathcal{Z}$ and $\epsilon > 0$. The first step consists in proving that

$$\epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1,1+\epsilon], \pi\left(\frac{\mathbf{X}}{t}\right) \in A_{\mathbf{x}} \;\middle|\; |\mathbf{X}| > t\right)$$

converges when $t \to \infty$, $\epsilon \to 0$. Following Equation (2.4.1) and assumption (A2), we know that this

quantity converges to $\epsilon^{-1}\mathbb{P}(Y \in (1,1+\epsilon],\, \mathbf{Z} \in A_{\mathbf{x}})$ when $t \to \infty$. Then, Proposition 2.4.1 gives

$$\begin{aligned}\mathbb{P}(Y \in (1,1+\epsilon], \mathbf{Z} \in A_{\mathbf{x}}) &= \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) - \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}} \mid Y > 1+\epsilon)\mathbb{P}(Y > 1+\epsilon) \\ &= \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) - \mathbb{P}(\pi((1+\epsilon)\mathbf{Z}) \in A_{\mathbf{x}})(1+\epsilon)^{-\alpha} \\ &= \left[1-(1+\epsilon)^{-\alpha}\right]\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) \\ &\quad + \left[\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) - \mathbb{P}(\pi((1+\epsilon)\mathbf{Z}) \in A_{\mathbf{x}})\right](1+\epsilon)^{-\alpha}\,.\end{aligned} \tag{2.7.9}$$

The first term divided by $\epsilon$ converges to $\alpha\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}})$ when $\epsilon \to 0$. We use (2.7.8) to compute the second term:

$$\begin{aligned}\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) - \mathbb{P}(\pi((1+\epsilon)\mathbf{Z}) \in A_{\mathbf{x}}) &= \mathbb{P}(\mathbf{Z} > \mathbf{x}) - \mathbb{P}\left(\mathbf{Z} > \frac{\mathbf{x}+\epsilon/d}{1+\epsilon}\right) \\ &= G_{\mathbf{Z}}(\mathbf{x}) - G_{\mathbf{Z}}\left(\mathbf{x} + \frac{\epsilon}{1+\epsilon}(1/d - \mathbf{x})\right).\end{aligned}$$

Since $\mathbf{x}$ is a differentiability point of $G_{\mathbf{Z}}$, we obtain

$$\epsilon^{-1}\mathbb{P}(Y \in (1,1+\epsilon],\, \mathbf{Z} \in A_{\mathbf{x}}) = \alpha\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) + \frac{1}{1+\epsilon}\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d) + o(1)\,,$$

when $\epsilon \to 0$. This means that

$$\epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1,1+\epsilon], \pi\left(\frac{\mathbf{X}}{t}\right) \in A_{\mathbf{x}} \;\middle|\; |\mathbf{X}| > t\right)$$

converges to $\alpha\mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) + \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)$ when $t \to \infty$, $\epsilon \to 0$.

For the second step, we define

$$\begin{aligned}(\star) := \epsilon^{-1}\Bigg[&\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1,1+\epsilon],\;\; \frac{\mathbf{X}}{|\mathbf{X}|} \in A_{\mathbf{x}} \;\middle|\; |\mathbf{X}| > t\right) \\ &- \mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1,1+\epsilon], \pi\left(\frac{\mathbf{X}}{t}\right) \in A_{\mathbf{x}} \;\middle|\; |\mathbf{X}| > t\right)\Bigg]\,,\end{aligned}$$

and the goal is to prove that $\lim_{\epsilon\to 0} \limsup_{t\to\infty}(\star) = \lim_{\epsilon\to 0} \liminf_{t\to\infty}(\star) = 0$.

We first deal with the lim sup. Assume that $|\mathbf{X}|/t \in (1,1+\epsilon]$. Then $(|\mathbf{X}|/t - 1 - \epsilon)/d \le 0$. Thus, if $x_j < X_j/|\mathbf{X}|$, then $x_j + (|\mathbf{X}|/t - 1 - \epsilon)/d < X_j/|\mathbf{X}| < X_j/t$. This implies that $x_j - \epsilon/d < X_j/|\mathbf{X}| - (|\mathbf{X}|/t - 1)/d$. The left member is positive for $\epsilon > 0$ small enough, so we proved that if $x_j < X_j/|\mathbf{X}|$, then $x_j - \epsilon/d < \pi(\mathbf{X}/t)$.

These considerations imply that

$$(\star) \le \epsilon^{-1}\Bigg[\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1,1+\epsilon], \pi\left(\frac{\mathbf{X}}{t}\right) \in A_{\mathbf{x}-\epsilon/d} \;\middle|\; |\mathbf{X}| > t\right)$$

$$-\mathbb{P}\left(\frac{|\mathbf{X}|}{t}\in(1,1+\epsilon],\pi\left(\frac{\mathbf{X}}{t}\right)\in A_{\mathbf{x}}\;\middle|\;|\mathbf{X}|>t\right)\Bigg]\,,$$

and thus

$$\limsup_{t\to\infty}(\star)\leq\epsilon^{-1}[\mathbb{P}\left(Y\in(1,1+\epsilon],\mathbf{Z}\in A_{\mathbf{x}-\epsilon/d}\right)-\mathbb{P}\left(Y\in(1,1+\epsilon],\mathbf{Z}\in A_{\mathbf{x}}\right)]=:\epsilon^{-1}[P_1(\epsilon)-P_2(\epsilon)]\,.$$

We use Proposition 2.4.1 and Equation (2.7.8) to compute (1) and (2). For (1), we have the following equalities:

$$\begin{aligned}\mathbb{P}(Y\in(1,1+\epsilon],\mathbf{Z}\in A_{\mathbf{x}-\epsilon/d})&=\mathbb{P}(\mathbf{Z}\in A_{\mathbf{x}-\epsilon/d})-\mathbb{P}(\mathbf{Z}\in A_{\mathbf{x}-\epsilon/d}\mid Y>1+\epsilon)\mathbb{P}(Y>1+\epsilon)\\&=\mathbb{P}(\mathbf{Z}>\mathbf{x}-\epsilon/d)-\mathbb{P}(\pi((1+\epsilon)\mathbf{Z})>\mathbf{x}-\epsilon/d)(1+\epsilon)^{-\alpha}\\&=\mathbb{P}(\mathbf{Z}>\mathbf{x}-\epsilon/d)-\mathbb{P}(\mathbf{Z}>\mathbf{x}/(1+\epsilon))(1+\epsilon)^{-\alpha}\\&=G_{\mathbf{Z}}(\mathbf{x}-\epsilon/d)[1-(1+\epsilon)^{-\alpha}]+[G_{\mathbf{Z}}(\mathbf{x}-\epsilon/d)\\&\quad-G_{\mathbf{Z}}(\mathbf{x}-\epsilon\mathbf{x}/(1+\epsilon))](1+\epsilon)^{-\alpha}\end{aligned}$$

The first term is equal to $G(\mathbf{x})\alpha\epsilon+o(\epsilon)$ when $\epsilon\to0$, whereas the second one is equal to

$$G_{\mathbf{Z}}(\mathbf{x}-\epsilon/d)-G_{\mathbf{Z}}(\mathbf{x})+G_{\mathbf{Z}}(\mathbf{x})-G_{\mathbf{Z}}(\mathbf{x}-\epsilon\mathbf{x}/(1+\epsilon))=\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(-\epsilon/d)-\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(-\epsilon\mathbf{x}/(d(1+\epsilon)))+o(\epsilon)\,,$$

when $\epsilon\to0$. This proves that $\epsilon^{-1}P_1(\epsilon)$ converges to $\alpha G_{\mathbf{Z}}(\mathbf{x})+\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x}-1/d)$ when $\epsilon\to0$. For $P_2(\epsilon)$, we refer to (2.7.9) in which we proved that $\epsilon^{-1}P_2(\epsilon)$ converges to $\alpha G_{\mathbf{Z}}(\mathbf{x})+\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x}-1/d)$ when $\epsilon\to0$. All in all we proved that $\epsilon^{-1}[P_1(\epsilon)-P_2(\epsilon)]\to0$, when $\epsilon\to0$.

We similarly proceed for the lim inf. Assume that $|\mathbf{X}|/t\in(1,1+\epsilon]$. Thus, if $\pi(\mathbf{X}/t)_j>x^j(1+\epsilon)$, then $X_j/t-(|\mathbf{X}|/t-1)/d>x_j(1+\epsilon)$, and therefore $X_j/t>x_j(1+\epsilon)$. Finally we obtain that $X_j/|\mathbf{X}|>x_j$. So we proved that if $\pi(\mathbf{X}/t)_j>x^j(1+\epsilon)$, then $X_j/|\mathbf{X}|>x_j$. These considerations give the following inequality:

$$\begin{aligned}(\star)\geq\epsilon^{-1}\Bigg[&\mathbb{P}\left(\frac{|\mathbf{X}|}{t}\in(1,1+\epsilon),\pi\left(\frac{\mathbf{X}}{t}\right)\in A_{(1+\epsilon)\mathbf{x}}\;\middle|\;|\mathbf{X}|>t\right)\\&-\mathbb{P}\left(\frac{|\mathbf{X}|}{t}\in(1,1+\epsilon),\pi\left(\frac{\mathbf{X}}{t}\right)\in A_{\mathbf{x}}\;\middle|\;|\mathbf{X}|>t\right)\Bigg]\,,\end{aligned}$$

and thus

$$\liminf_{t\to\infty}(\star)\geq\epsilon^{-1}[\mathbb{P}\left(Y\in(1,1+\epsilon),\mathbf{Z}\in A_{(1+\epsilon)\mathbf{x}}\right)-\mathbb{P}\left(Y\in(1,1+\epsilon),\mathbf{Z}\in A_{\mathbf{x}}\right)]=:\epsilon^{-1}[P_3(\epsilon)-P_4(\epsilon)]\,.$$

We use again Proposition 2.4.1 and Equation (2.7.8) to compute $P_3(\epsilon)$:

$$\begin{aligned}\mathbb{P}(Y\in(1,1+\epsilon],\mathbf{Z}\in A_{\mathbf{x}-\epsilon/d})&=\mathbb{P}(\mathbf{Z}\in A_{(1+\epsilon)\mathbf{x}})-\mathbb{P}(\mathbf{Z}\in A_{(1+\epsilon)\mathbf{x}}\mid Y>1+\epsilon)\mathbb{P}(Y>1+\epsilon)\\&=\mathbb{P}(\mathbf{Z}>(1+\epsilon)\mathbf{x})-\mathbb{P}(\pi((1+\epsilon)\mathbf{Z})>(1+\epsilon)\mathbf{x})(1+\epsilon)^{-\alpha}\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}(\mathbf{Z} > (1+\epsilon)\mathbf{x}) - \mathbb{P}(\mathbf{Z} > \mathbf{x} + \epsilon/((1+\epsilon)d))(1+\epsilon)^{-\alpha} \\
&= G_{\mathbf{Z}}((1+\epsilon)\mathbf{x})[1-(1+\epsilon)^{-\alpha}] \\
&\quad + [G_{\mathbf{Z}}((1+\epsilon)\mathbf{x}) - G_{\mathbf{Z}}(\mathbf{x} + \epsilon/(d(1+\epsilon)))](1+\epsilon)^{-\alpha} \\
&= G_{\mathbf{Z}}((1+\epsilon)\mathbf{x})\alpha\epsilon + [\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\epsilon(\mathbf{x}-1/d)/(1+\epsilon))] + o(\epsilon)\,,
\end{aligned}
$$

when $\epsilon \to 0$. The first term is equal to $G_{\mathbf{Z}}(\mathbf{x})\alpha\epsilon + o(\epsilon)$, when $\epsilon \to 0$, whereas the second one is equal to

$$
G_{\mathbf{Z}}((1+\epsilon)\mathbf{x}) - G_{\mathbf{Z}}(\mathbf{x}) + G_{\mathbf{Z}}(\mathbf{x}) - G_{\mathbf{Z}}(\mathbf{x} + \epsilon/(d(1+\epsilon))) = \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\epsilon(\mathbf{x}-1/d)) + o(\epsilon), \quad \epsilon \to 0\,.
$$

This proves that $P_3(\epsilon)$ converges to $\alpha G_{\mathbf{Z}}(\mathbf{x}) + \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)$ when $\epsilon \to 0$. Note that $P_4(\epsilon) = P_2(\epsilon)$, so that $P_4(\epsilon)$ converges to $\alpha G_{\mathbf{Z}}(\mathbf{x}) + \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)$ when $\epsilon \to 0$. All in all we proved that $\epsilon^{-1}[P_3(\epsilon) - P_4(\epsilon)] \to 0$, when $\epsilon \to 0$.

Gathering all these results together, we can write

$$
\epsilon^{-1}[P_3(\epsilon) - P_4(\epsilon)] \le \liminf_{t\to\infty}(\star) \le \limsup_{t\to\infty}(\star) \le \epsilon^{-1}[P_1(\epsilon) - P_2(\epsilon)]\,.
$$

Since $\epsilon^{-1}[P_1(\epsilon) - P_2(\epsilon)]$ and $\epsilon^{-1}[P_3(\epsilon) - P_4(\epsilon)]$ converge to 0 as $\epsilon \to 0$, we proved that $\lim_{\epsilon\to 0} \liminf_{t\to\infty}(\star) = \lim_{\epsilon\to 0} \limsup_{t\to\infty}(\star) = 0$.

To conclude the proof, we write

$$
\begin{aligned}
&\epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon],\ \frac{\mathbf{X}}{|\mathbf{X}|} \in A_{\mathbf{x}} \,\middle|\, |\mathbf{X}| > t\right) \\
&\qquad = (\star) + \epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon], \pi\left(\frac{\mathbf{X}}{t}\right) \in A_{\mathbf{x}} \,\middle|\, |\mathbf{X}| > t\right)\,,
\end{aligned}
$$

and both steps lead to

$$
\lim_{\epsilon\to 0} \liminf_{t\to\infty} \epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A_{\mathbf{x}} \,\middle|\, |\mathbf{X}| > t\right) = \alpha G_{\mathbf{Z}}(\mathbf{x}) + \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)\,,
$$

and

$$
\lim_{\epsilon\to 0} \limsup_{t\to\infty} \epsilon^{-1}\mathbb{P}\left(\frac{|\mathbf{X}|}{t} \in (1, 1+\epsilon], \frac{\mathbf{X}}{|\mathbf{X}|} \in A_{\mathbf{x}} \,\middle|\, |\mathbf{X}| > t\right) = \alpha G_{\mathbf{Z}}(\mathbf{x}) + \mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)\,.
$$

Since $|\mathbf{X}|$ is regularly varying with tail index $\alpha$, we apply Lemma 2.4.1 to conclude that $\mathbf{X}$ is regularly varying with tail index $\alpha$ and with spectral vector $\boldsymbol{\Theta}$ satisfying $\mathbb{P}(\boldsymbol{\Theta} \in A_{\mathbf{x}}) = \mathbb{P}(\mathbf{Z} \in A_{\mathbf{x}}) + \alpha^{-1}\mathrm{d}G_{\mathbf{Z}}(\mathbf{x})(\mathbf{x} - 1/d)$.

□

## 2.8 Appendix

We introduce here the linear-time algorithm given in Duchi et al. (2008). It is based on a random selection of the coordinates.

**Data:** A vector $\mathbf{v} \in \mathbb{R}_+^d$ and a scalar $z > 0$
**Result:** The projected vector $\mathbf{w} = \pi(\mathbf{v})$
Initialize $U = \{1, \ldots, d\}$, $s = 0$, $\rho = 0$;
**while** $U \neq \emptyset$ **do**
  Pick $k \in U$ at random;
  Partition $U$: $G = \{j \in U,\, v_j \geq v_k\}$ and $L = \{j \in U,\, v_j < v_k\}$;
  Calculate $\Delta\rho = |G|, \quad \Delta s = \sum_{j \in G} v_j$;
  **if** $(s + \Delta s) - (\rho + \Delta\rho)v_k < z$ **then**
    $s = s + \Delta s$;
    $\rho = \rho + \Delta\rho$;
    $U \leftarrow L$;
  **else**
    $U \leftarrow G \setminus \{k\}$;
  **end**
**end**
Set $\eta = (s - z)/\rho$.;
Define $\mathbf{w}$ s.t. $w_i = v_i - \eta$.

**Algorithm 3:** Linear time projection onto the positive sphere $\mathbb{S}_+^{d-1}(z)$.

# Chapter 3

# Tail inference for high-dimensional data

**Abstract**

In this chapter, we transpose the framework developed in Chapter 2 in a statistical context to study the dependence structure of extreme events. This approach relies on the Euclidean projection onto the simplex which exhibits the sparsity of the spectral measure and reduces the dimension of the extremes' study. Given a data set, we provide an algorithmic approach to tackle two questions: On which directions do extremes appear and which threshold is the most accurate the separate the data into an extreme category and a non-extreme one. Theses issues are addressed with multinomial model selection. Finally, we apply our method on numerical experiments to illustrate the relevance of our setting.

***Keywords***— dimension reduction, Euclidean projection onto the simplex, model selection, regular variation, sparse regular variation, tail inference

**Regarding our questions**

**(Q1)** In this chapter, we use the setting developed previously to tackle high-dimensional data. We mostly focus on the subsets $C_\beta$. The model selection we propose manages to tackle these $2^d - 1$ subsets in a very convenient fashion.

**(Q2)** The issue of determining on which directions extremes appear is addressed with a multinomial model. We identify the most relevant features through a model selection approach.

**(Q3)** The same model selection allows us to identify an optimal level $k$. This selection is made with a Kullback-Leibler minimization. From an algorithmic point of view, we use the impact of the threshold on the sparsity structure of the projected vectors to obtain a efficient algorithm.

## Contents

## 3.1 Introduction

In many applications, identifying the tail structure of the data is useful to evaluate the risk and predict future large events. Severe events are indeed often a consequence of the simultaneous extreme behavior of several factors. In financial quantitative risk management, we are willing to detect the probability that several firms make together huge losses. In the climate field, it is important to identify areas which can be impacted simultaneously by a severe event (for instance a heavy rainfall, a heat wave, or a flood). In an oceanographic context, the sea-level process can be explained by several factors like the tidal level, the mean-sea level, or the surge level (see Tawn (1992)). Therefore, high sea-levels are often due to the simultaneous occurrence of extreme values among these components.

These applications fit into the context of EVT which provides models to study large events and to assess the tail structure of a given distribution (see e.g. Resnick (1987), de Haan and Ferreira (2006), Resnick (2007), or Embrechts et al. (2013)). In the multivariate setting, EVT focuses on the intensity and the dependence structure of large events (see for instance Beirlant et al. (2006), Chapter 8). From a theoretical point of view, the study of extreme events is closely related to regular variation.

### 3.1.1 Regular variation

For a random vector $\mathbf{X}$ in $\mathbb{R}_+^d$, the purpose of EVT is to assess the tail structure of $\mathbf{X}$. As explained in Chapter 1, it is customary in this case to assume that $\mathbf{X}$ is regularly varying: There exist a positive sequence $(a_n)$, $a_n \to \infty$ as $n \to \infty$, and a non-null Radon measure $\mu$ on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ such that

$$n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in \cdot\right) \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,, \tag{3.1.1}$$

where $\xrightarrow{v}$ denotes vague convergence in $\mathcal{M}_+(\mathbb{R}_+^d \setminus \{\mathbf{0}\})$, the space of non-null Radon measures on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ (see Section 1.2.2.2). The limit measure $\mu$ is called the *tail measure* and satisfies the homogeneity property $\mu(tA) = t^{-\alpha}\mu(A)$ for any set $A \subset \mathbb{R}_+^d \setminus \{\mathbf{0}\}$ and any $t > 0$. The parameter $\alpha > 0$ is called the *tail index*.

It is often more convenient to use a polar representation for $\mu$ in order to separately study the radial part and the angular part of $\mathbf{X}$. We use the device of Beirlant et al. (2006), Chapter 8, in which regular variation is characterized in terms of polar coordinates (see also Resnick (1986)). A non-negative random vector $\mathbf{X}$ is regularly varying if there exist a parameter $\alpha > 0$ and a finite measure $S$ on the positive unit sphere $\mathbb{S}_+^{d-1}$ such that

$$\mathbb{P}\big(|\mathbf{X}| > rt, \mathbf{X}/|\mathbf{X}| \in B \mid |\mathbf{X}| > t\big) \to r^{-\alpha}S(B)\,, \quad t \to \infty\,, \tag{3.1.2}$$

for any $S$-continuity set $B$ of $\mathbb{S}_+^{d-1}$ and any $r > 0$ (see also Proposition 1.2.3). Equivalently, it means that there exist a random vector $\mathbf{\Theta}$ on $\mathbb{S}_+^{d-1}$ and a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$ such that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \frac{\mathbf{X}}{|\mathbf{X}|}\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{\Theta}) \in \cdot)\,, \quad t \to \infty\,. \tag{3.1.3}$$

The random variable $Y$ is the limit of the radial component $|\mathbf{X}|/t$ and thus models the intensity of extreme events. It is characterized by the tail index $\alpha$. The smaller $\alpha$ is, the larger the extremes could be. On the other , the vector $\mathbf{\Theta}$, called the *spectral vector*, and its distribution $S$, the *spectral measure*, are associated to the angular component of $\mathbf{X}$ and thus describe the behavior in space of large events: The subspaces of the positive unit sphere on which the spectral vector concentrates correspond to the directions where large events occur. Therefore, the knowledge of the spectral measure's support is a crucial but challenging topic of multivariate EVT, especially in high dimensions.

From a statistical point of view, studying multivariate extremes consists in estimating the parameter $\alpha$ and the spectral measure. The former estimation is parametric and has been widely studied, for instance by Hill (1975), Smith (1987) or Beirlant et al. (1996b). On the contrary, providing helpful estimators of the spectral measure is a challenging problem even more in high dimensions. Until recently, useful representations of the spectral measure have only been introduced in the bivariate case, see e.g. Einmahl et al. (1993), Einmahl et al. (1997), Einmahl et al. (2001) and Einmahl and Segers (2009). Parametric approaches have also been introduced to tackle the study of extremes in moderate dimensions ($d \leq 10$), for instance by Coles and Tawn (1991) and Sabourin et al. (2013).

In higher dimensions, it is common that large events only appear on specific directions. In other words, there are many parts of the unit sphere on which the spectral measure does not place mass. In this case, we say that this measure (or equivalently, the spectral vector $\mathbf{\Theta}$) is *sparse*. Equivalently, it means that the probability $\mathbb{P}(|\mathbf{\Theta}|_0 \ll d)$ is close to 1[1]. Thus, identifying the low-dimensional subspaces on which the spectral measure concentrates leads to dimension reduction. In this context, Lehtomaa and Resnick (2019) maps the unit sphere to the $d-1$ dimensional space $[0,1]^{d-1}$ in order to partition it in equally sized rectangles. The study of the spectral measure's support is thus done with grid estimators.

### 3.1.2 Estimation of the spectral measure

Since the complete support's estimation is often difficult to capture, a main objective in the study of the tail dependence is rather to identify the directions on which the spectral measure puts mass. Different techniques have been recently proposed. A first type of approaches highlights the use of Principal Component Analysis (PCA) in an extremal setting (Cooley and Thibaud (2019), Sabourin and Drees (2019)). In high dimensions, several authors recently focus on the directions in the data that are likely to be extreme together. This clustering approach for multivariate EVT has firstly been introduced by Chautru (2015) who uses spherical data analysis to capture the dependence structure of the data. In this context, it is convenient to partition the space $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1\}$ or the positive unit sphere $\mathbb{S}_+^{d-1}$ in terms of the nullity of the coordinates (see Section 1.4.2). For $\beta \in \mathcal{P}_d^*$, the subspaces $R_\beta$ and $C_\beta$ are defined as follows:

$$R_\beta = \left\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1, x_i > 0 \text{ for } i \in \beta, x_i = 0 \text{ for } i \notin \beta\right\}. \tag{3.1.4}$$

and

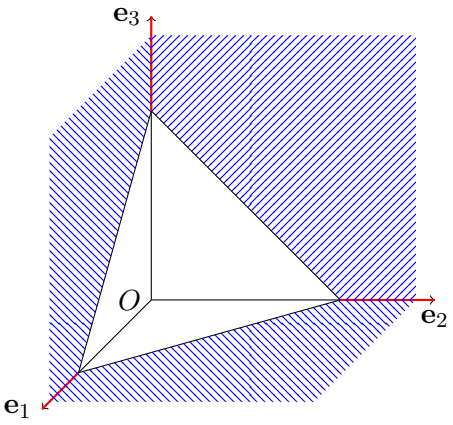
$$C_\beta = \left\{\mathbf{x} \in \mathbb{S}_+^d, x_i > 0 \text{ for } i \in \beta, x_i = 0 \text{ for } i \notin \beta\right\}. \tag{3.1.5}$$

The subsets $R_\beta$ (respectively $C_\beta$) are pairwise disjoint and form a partition of $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1\}$ (respectively $\mathbb{S}_+^{d-1}$):

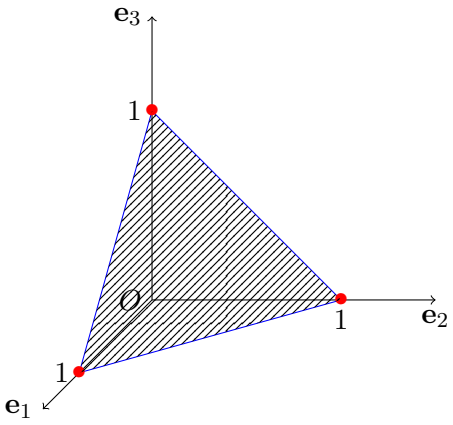
$$\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1\} = \bigsqcup_{\beta \in \mathcal{P}_d^*} R_\beta \quad \text{and} \quad \mathbb{S}_+^{d-1} = \bigsqcup_{\beta \in \mathcal{P}_d^*} C_\beta,$$

[1] $|\cdot|_0$ denotes the $\ell^0$-norm, that is, the number of non-null coordinates of a vector.

where $\bigsqcup$ denotes a disjoint union. An illustration of these subsets in dimension 3 are given in Figure 3.1.




(a) The subspaces $R_\beta$ in dimension 3 for the $\ell^1$-norm. In red the one-dimensional cones spanned by $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. In shaded blue the two-dimensional cones. For readability purposes, the full subspace $\{|\mathbf{x}| > 1, \forall i \in \{1,2,3\}, x_i > 0\}$ is not represented.




(b) The subspaces $C_\beta$ in dimension 3 for to the $\ell^1$-norm. In red the points $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. In blue the lines of the simplex's edges. The shaded part correspond to the interior of the simplex.

Figure 3.1: The subspaces $R_\beta$ and $C_\beta$ in dimension 3 for the $\ell^1$-norm.

Regarding the regularly varying random vector $\mathbf{X}$, both partitions defined by the subsets $R_\beta$ and $C_\beta$ highlight its extremal structure. Indeed, for a fixed $\beta$ in $\mathcal{P}_d^*$, the inequality $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ means that it is likely to observe simultaneously large values in the directions $\beta$ and small values in the directions $\beta^c$. Detecting these features allows one to bring out clusters of coordinates which can be large together. Hence, the main goal of the spectral measure's estimation consists in classifying the $2^d - 1$ probabilities $\mathbb{P}(\mathbf{\Theta} \in C_\beta)$ depending on their nullity or not.

With this in mind, several ideas have been developed on this topic. Goix et al. (2017) focus on the tail measure $\mu$ and estimate the mass this measure puts on the subsets $R_\beta$. This estimation is based on a hyperparameter $\epsilon > 0$ and brings out a sparse representation of the dependence structure. An algorithm called DAMEX (for Detecting Anomalies among Multivariate EXtremes) is proposed and reaches a complexity $O(dn \log n)$, where $n$ corresponds to number of data points. Based on this method, Chiapino and Sabourin (2016) provide another algorithm, (CLEF for CLustering Extremal Features) in order to gather the features that are likely to be extreme simultaneously. Simpson et al. (2019) base their method on the concept of hidden regular variation, introduced by Resnick (2002). They introduce a set of parameters $(\tau_\beta)_{\beta \in \mathcal{P}_d^*}$ which describe the extremal behavior of the data on the subsets $(C_\beta)_{\beta \in \mathcal{P}_d^*}$. An algorithm of complexity $O(dn \log n)$ is also provided.

All these approaches rely on the classical definition of regular variation which does not provide a natural estimator for the spectral measure. Indeed, assume for instance that a subset $C_\beta$ with $\beta \neq \{1, \ldots, d\}$ satisfies $\mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$. In this case, $C_\beta$ it is not a $S$-continuity set since $\mathbb{P}(\mathbf{\Theta} \in \partial C_\beta) = \mathbb{P}(\mathbf{\Theta} \in C_\beta) > 0$ and then convergence (3.1.3) fails. From an empirical point of view, this

example can be rephrased in the following way. On the one hand, it is likely that $\mu$ puts mass on some subspaces $R_\beta$, $\beta \neq \{1, \dots, d\}$, which are of zero Lebesgue measure. On the other hand, the vector $\mathbf{X}$ does not concentrate on such subspaces since it models real-world data.

In order to circumvent this issue and to take the potential sparse structure of the spectral measure into account, we use the ideas developed in Chapter 2. The self-normalization which appears in the second component of Equation (3.1.3) is replaced by the Euclidean projection onto the simplex, which has been studied by Duchi et al. (2008), Kyrillidis et al. (2013), or Liu and Ye (2009) among others. This substitution leads then to a angular vector which differs from the spectral vector $\mathbf{\Theta}$. With this new concept called *sparse regular variation*, the sparsity structure of extreme events can be more easily captured.

The purpose of this chapter is to use the theoretical results introduced by Meyer and Wintenberger (2019) in order to provide a useful inference method of the extremal dependence structure. These theoretical results lead to natural estimators for sparsely regularly varying random vectors for which consistency and asymptotic normality are proven. The identification of the subspaces $C_\beta$ on which extremes gather is done through model selection.

### 3.1.3 Choice of the threshold via model selection

From a non-asymptotic point of view, Equation (3.1.3) becomes an approximation when the threshold $t$ is "high enough". In a statistical context, if $\mathbf{X}_1, \dots, \mathbf{X}_n$ are i.i.d. regularly varying random vectors satisfying Equation (3.1.3), then choosing an optimal threshold is equivalent to choosing a number $k$ of data which will be considered to be extreme. Of course, the less data you keep the more you are in an extreme context. On the other hand, it is still needed to keep a substantial number of data to correctly learn their structure. This balanced choice is a major problem in EVT and no theoretical result has been obtained yet in a multivariate setting.

Several authors point out that the choice of an appropriate threshold above which Equation (3.1.3) is accurate is a challenging task in practice (see for instance Rootzén and Tajvidi (2006)). This choice has been tackled by Abdous and Ghoudi (2005) who propose an automatic selection technique in the bivariate case which is based on a double kernel technique (see Devroye (1989)). A more abundant literature deals with marginals threshold selection (see Caeiro and Gomes (2015) for a review). In a multivariate framework, threshold selection for dependence models has been studied by Lee et al. (2015) who apply Bayesian selection, and by Kiriliouk et al. (2019) who use stability properties of the multivariate Pareto distribution.

A general idea is to use model selection (see Section 1.5) to highlight which threshold provides the better estimation (Massart (1989)). For practical reasons, it is often more convenient to focus on the number of exceedances rather on the threshold. There, the selection consists in choosing the appropriate number of data which are considered to be extreme. In this chapter, we provide a selection procedure based on Akaike Information Criterion (AIC) (Akaike (1973)). The AIC procedure is based on a the minimization of a penalized maximum likelihood and holds for a constant sample size. To this end, this method has to be adapted in order to use it in a extreme setting. It

is indeed necessary to include the non-extreme values in the models and to separate them into an "extreme" group and a "non-extreme" one. This separation into two groups is actually based on a different choice of the level $k$. The idea is then to apply an AIC criterion which highlights the $k$ for which the separation is optimal. This method does not provided a generic choice of $k$ for all data sets but suggests an ad hoc selection based on a penalized maximum likelihood minimization.

### 3.1.4 Outline

The structure of this chapter is as follows. The theoretical background used in this chapter is introduced in Section 3.2. We deal with the notions of regular variation and sparse regular variation and detail how the new projection affects the convergence (3.1.3). We also discuss some convergence results and explain why this approach is useful to capture the sparse structure of large events. In Section 3.3, we apply our theoretical results on a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in order to introduce convenient estimators for the estimation of the proportion of extreme values in a given subspace $A$ of the positive unit sphere. Consistency and asymptotic normality are proven at a univariate level. In Section 3.4, we restrict the study to the $C_\beta$ and we extend the convergence results at a multivariate level. Section 3.5 is devoted to the selection of the more significant subspaces $C_\beta$ and to the choice of an optimal level $k$. Finally, Section 3.6 deals with some simulations which provide numerical evidence of our theoretical findings.

## 3.2 Sparse regular variation

### 3.2.1 Regular variation and spectral measure

We consider a regularly varying random vector $\mathbf{X} \in \mathbb{R}_+^d$:

$$n\mathbb{P}\left(a_n^{-1}\mathbf{X} \in \cdot\right) \xrightarrow{v} \mu(\cdot)\,, \quad n \to \infty\,, \tag{3.2.1}$$

where $\mu$ is a non-null Radon measure on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$.

**Remark 3.2.1.** As discussed in Resnick (2007), convergence (3.1.1) may be seen as standard regular variation in contrast with the nonstandard one: There exists a non-negative Radon measure $\tilde{\mu}$ on $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ such that

$$n\mathbb{P}\left((X_i/a_{n,i})_{1\le i\le d} \in \cdot\right) \xrightarrow{v} \tilde{\mu}(\cdot)\,, \quad n \to \infty\,,$$

where the sequences $(a_{n,i})_n$ are satisfying $n\mathbb{P}(X_i > a_{n,i}) \to 1$ as $n \to \infty$. Actually the standard regular variation (3.1.1) is more general since it allows the tail measure to be sparse. In our context, a measure is said to be sparse if it places mass on some lower-dimensional subspaces of $\mathbb{R}_+^d$. On the other hand, in the non-standard case the condition satisfied by the $(a_{n,i})$ implies that $\tilde{\mu}(\{\mathbf{x} \in \mathbb{R}_+^d,\, x_i > 1\}) = 1$ for all $i = 1, \ldots, d$. This means that the measure $\tilde{\mu}$ concentrates in all directions. There, the study of extreme values begins with a modification of the marginals, called *rank transform*, which gives the same distribution to all components (see Resnick (1987), Proposition 5.10). If $F_i$

denotes the distribution of the marginal $X_i$, then a natural transformation consists in considering the vector $(1/(1-F_i(X_i)))_{1\leq i\leq d}$ whose marginals are Pareto(1)-distributed. In this chapter, we do not consider any transformation of the marginals and we only focus on standard regular variation defined by (3.1.1). This means that we assume that the tail measure $\mu$ is likely to be sparse.

Following Proposition 1.2.3, the convergence in Equation (3.2.1) is equivalent to the existence of a random vector $\mathbf{\Theta}$, the spectral vector, on $\mathbb{S}_+^{d-1}$ and a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$ such that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t},\frac{\mathbf{X}}{|\mathbf{X}|}\right)\in\cdot\;\middle|\;|\mathbf{X}|>t\right)\xrightarrow{w}\mathbb{P}((Y,\mathbf{\Theta})\in\cdot)\,,\quad t\to\infty\,. \tag{3.2.2}$$

According to Remark 3.2.1, it is likely that the tail measure has a sparse structure, i.e. that it places mass on low-dimensional subspaces. Regarding the spectral vector $\mathbf{\Theta}$ defined in Equation (3.2.2), it means that it is likely that only few components of this vector are non-null. To this end, we focus on the behavior of $\mathbf{\Theta}$ on the subsets $C_\beta$ defined in (3.1.5). The advantage of these subsets is that they are interpretable in terms of extreme dependence while they reduce the dimension. The main goal is then to identify the subspaces on which the distribution of $\mathbf{\Theta}$ places mass, i.e. the ones which satisfy $\mathbb{P}(\mathbf{\Theta}\in C_\beta)>0$. To this end, we define the set

$$\mathcal{S}(\mathbf{\Theta})=\{\beta\in\mathcal{P}_d^*,\,\mathbb{P}(\mathbf{\Theta}\in C_\beta)>0\}\,. \tag{3.2.3}$$

Following the ideas of Section 2.2.1, the estimation of the probabilities $\mathbb{P}(\mathbf{\Theta}\in C_\beta)$ can not be easily addressed based on the self-normalized vector $\mathbf{X}/|\mathbf{X}|$. Indeed, as soon as the marginals of $\mathbf{X}$ are non-degenerate, the probabilities $\mathbb{P}(\mathbf{X}/|\mathbf{X}|\in C_\beta)$ are equal to zero for all $\beta\neq\{1,\dots,d\}$. Since the standard regular variation framework is too restrictive to infer the support of the spectral measure, the idea is to introduce sparsity into the vector $\mathbf{X}$. We address this issue with the ideas introduced in Chapter 2 in which the self-normalized vector $\mathbf{X}/|\mathbf{X}|$ is replaced by $\pi(\mathbf{X}/t)$, where $\pi$ denotes the Euclidean projection onto the simplex. This leads to the notion of sparse regular variation. The purpose of this section is to summarize the theoretical results obtained in Chapter 2 in order to use them in a statistical setting.

From now on and in all this chapter, $|\cdot|$ denotes $\ell^1$-norm and $\mathbb{S}_+^{d-1}$ the simplex:

$$\mathbb{S}_+^{d-1}=\{\mathbf{x}\in\mathbb{R}_+^d,\,x_1+\ldots+x_d=1\}\,.$$

More generally, we write $\mathbb{S}_+^{d-1}(z)=\{\mathbf{x}\in\mathbb{R}_+^d,\,x_1+\ldots+x_d=z\}$ for $z>0$.

### 3.2.2 The Euclidean projection onto the simplex

The useful tools regarding the Euclidean projection onto the simplex has been introduced in Section 2.2.2. Recall that for $z>0$, the Euclidean projection $\pi_z$ onto the positive sphere $\mathbb{S}_+^{d-1}(z)$ is defined

with the function

$$\begin{array}{rccc} \pi_z & : & \mathbb{R}_+^d & \to & \mathbb{S}_+^{d-1}(z) \\ & & \mathbf{v} & \mapsto & \mathbf{w} = (\mathbf{v} - \lambda_{\mathbf{v},z})_+ \,, \end{array}$$

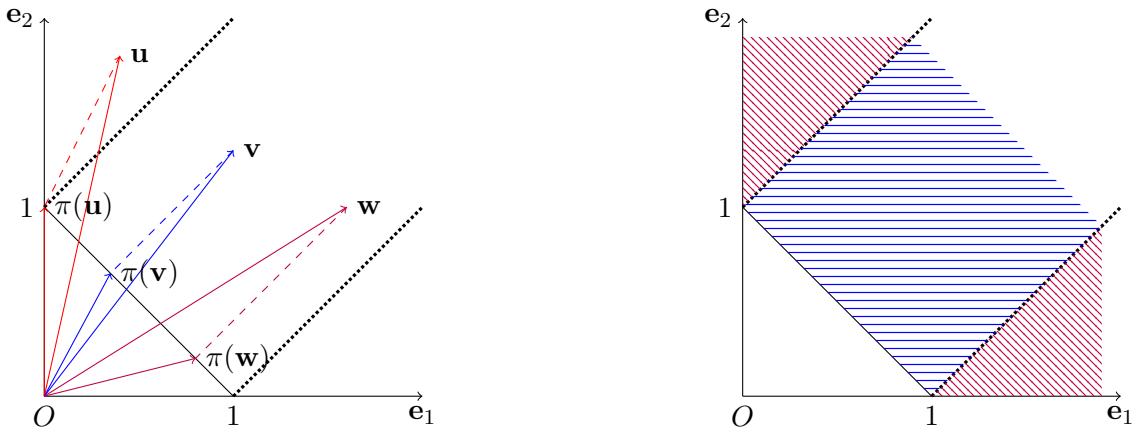
where $\lambda_{\mathbf{v},z} \in \mathbb{R}$ is the only constant satisfying the relation $\sum_{1 \le i \le d}(v_i - \lambda_{\mathbf{v},z})_+ = z$. Algorithm 3 provides a way to compute this projection in an expected time complexity of $O(d)$. This allows to deal with high-dimensional data without any complexity constraints.

If $z = 1$ we shortly denote $\pi$ for $\pi_1$. The projection satisfies the relation

$$\pi_z(\mathbf{v}) = z\pi(\mathbf{v}/z)\,, \tag{3.2.4}$$

for all $\mathbf{v} \in \mathbb{R}_+^d$ and $z > 0$. This is why we mainly focus on the projection $\pi$ onto the simplex $\mathbb{S}_+^{d-1}$. Equation (3.2.4) will be useful regarding extreme values. Indeed, the parameter $z$ will play the role of the threshold and an optimal choice of $z$ is addressed in Section 3.5 (see also Remark 3.2.2).

An illustration of $\pi$ in the bivariate case is given in Figure 3.2. It highlights the fundamental difference between $\pi$ and the self-normalization on the subspaces $C_\beta$. For the latter, the subspaces $\{|\mathbf{x}| > 1, \mathbf{x}/|\mathbf{x}| \in C_\beta\} = R_\beta$ are of zero Lebesgue measure, as soon as $\beta \neq \{1, \dots, d\}$. On the contrary, this is not true for the subspaces $\{|\mathbf{x}| > 1, \pi(\mathbf{x}) \in C_\beta\}$. For instance in dimension 2, the subspaces of $\{\mathbf{x} \in \mathbb{R}_+^d, |\mathbf{x}| > 1\}$ in which the vectors are projected on the axis $\mathbf{e}_1$ or $\mathbf{e}_2$ correspond to the purple shaded areas of Figure 3.2b. This means that the issue that arises with the weak convergence vanishes with the projection $\pi$.




(a) Three vectors and their image by $\pi$. The dotted lines partitions the subspaces depending on the localization of the projected vectors: $\mathbf{e}_1$, $\mathbf{e}_2$, or the interior of the simplex.

(b) The preimages of the subcones $C_\beta$ by $\pi$. In purple $\pi^{-1}(C_{\{1\}})$ and $\pi^{-1}(C_{\{12\}})$, and in blue $\pi^{-1}(C_{\{1,2\}})$.

Figure 3.2: Euclidean projection onto the simplex $\mathbb{S}_+^1$.

**Some results on the projection** We list here some properties of the projection $\pi_z$ that have been established in Chapter 2. All of them will be useful in the statistical setting developed in the

following sections.

P1. The projection preserves the order of the coordinates: If $v_{\sigma(1)} \geq \ldots \geq v_{\sigma(d)}$ for a permutation $\sigma$, then $\pi(\mathbf{v})_{\sigma(1)} \geq \ldots \geq \pi(\mathbf{v})_{\sigma(d)}$ for the same permutation.

P2. If $\pi(\mathbf{v})_j > 0$, then $v_j > 0$. Equivalently, $v_j = 0$ implies $\pi(\mathbf{v})_j = 0$.

P3. The projection $\pi$ is continuous, as every projection on a convex, closed set in a Hilbert space.

P4. For $\mathbf{v} \in \mathbb{R}_+^d$, we have the following equivalence:

$$\pi(\mathbf{v}) \in C_\beta \quad \text{if and only if} \quad \begin{cases} \max_{i\in\beta} \sum_{j\in\beta}(v_j - v_i) < 1\,, \\ \min_{i\in\beta^c} \sum_{j\in\beta}(v_j - v_i) \geq 1\,. \end{cases} \tag{3.2.5}$$

**Remark 3.2.2** (Choice of the threshold)**.** Another main aspect of the projection $\pi$ that will be used in Section 3.5 is the choice of the threshold $z$. This choice is indeed closely linked to the sparse structure of extreme values. For a vector $\mathbf{v} \in \mathbb{R}_+^d$ with $\ell^1$-norm $|\mathbf{v}|$, the number of null coordinates of the projected vector $\pi_z(\mathbf{v})$ strongly depends on the choice of $z$. If $z$ is close to $|\mathbf{v}|$, then $\pi_z(\mathbf{v})$ has almost only non-null coordinates (as soon as $\mathbf{v}$ itself as non-null coordinates). But if $z \ll |\mathbf{v}|$, then $\pi_z(\mathbf{v})$ tends to be sparse. The impact of the threshold $z$ on the sparsity of the projected vectors is illustrated in Figure 3.3. From a statistical point of view, we consider a sample of $n$ i.i.d. sparsely regularly varying random vectors. In order to focus on extreme values we have to select only the vectors with the largest norms, that is, vectors whose norm is above a certain threshold. For a large threshold $z$, only extreme data are selected but many vectors are close to this threshold. This implies that these vectors are projected on subsets $C_\beta$ with large $\beta$'s which means that they do not tend to be very sparse. On the other hand, if we select a low threshold, then we move away from the extreme regime. In this case, the largest vectors are projected on subsets $C_\beta$ with small $\beta$'s, i.e. they are very sparse. Hence, we have to make a balanced choice in order to keep the sparse structure of the data while staying in the extreme regime. This choice will be done based on a model selection discussed in Section 3.5.

### 3.2.3 Sparse regular variation

The substitution of the self-normalized $\mathbf{X}/|\mathbf{X}|$ by $\pi(\mathbf{X}/t)$ is at the core of Chapter 2. Recall that a random vector $\mathbf{X}$ on $\mathbb{R}_+^d$ is sparsely regularly varying if there exist a random vector $\mathbf{Z}$ defined on the simplex $\mathbb{S}_+^{d-1}$ and a non-degenerate random variable $Y$ such that

$$\mathbb{P}\left(\left(\frac{|\mathbf{X}|}{t}, \pi\left(\frac{\mathbf{X}}{t}\right)\right) \in \cdot \;\middle|\; |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}((Y, \mathbf{Z}) \in \cdot)\,, \quad t \to \infty\,. \tag{3.2.6}$$

In this case, there exists $\alpha > 0$ such that the random variable $Y$ is Pareto($\alpha$)-distributed. The limit vector $\mathbf{Z}$ must be seen as the angular limit obtained after replacing the self-normalization by $\pi$. By continuity of $\pi$, standard regular variation with tail index $\alpha$ and spectral vector $\boldsymbol{\Theta}$ implies sparse

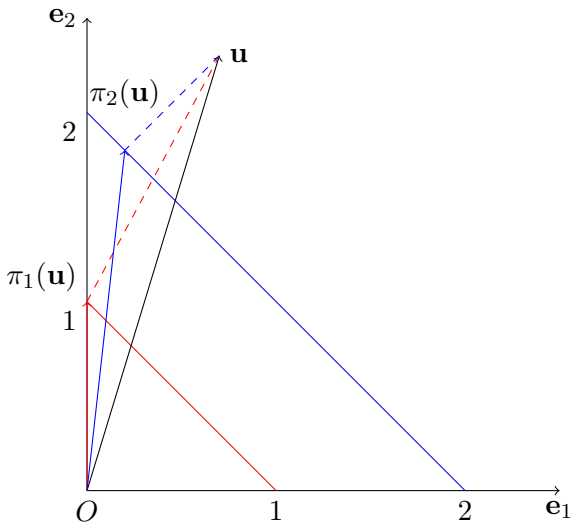



Figure 3.3: Consequence of the choice of the threshold on the sparsity. The image of the vector $\mathbf{u}$ is $\pi_1(\mathbf{u}) = (0, 1)$ with the threshold $z = 1$ while it is $\pi_2(\mathbf{u}) > 0$ with the threshold $z = 2$. The sparsity increases when the threshold decreases.

regular variation with tail index $\alpha$ and angular limit $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$, where $Y$ is a Pareto($\alpha$)-distributed random variable independent of $\boldsymbol{\Theta}$. We have proved in Chapter 2 that under mild assumptions the converse implication holds (see Theorem 2.4.1).

Regarding the subsets $C_\beta$, some useful properties for the limit vector $\mathbf{Z}$ which are not satisfied by the vector $\boldsymbol{\Theta}$ have been established. We summarize these main results are gathered in the next proposition. The proof can be found in Section 2.3.

**Proposition 3.2.1.** *Let $\mathbf{X}$ be a regularly varying random vector of $\mathbb{R}^d_+$ with spectral vector $\boldsymbol{\Theta}$ and tail index $\alpha > 0$. Set $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$, where $Y$ denotes a Pareto($\alpha$) random variable independent of $\boldsymbol{\Theta}$.*

1. *If $A$ is a Borel subset of $\mathbb{S}^{d-1}_+$ satisfying $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$, then*

$$\mathbb{P}(\pi(\mathbf{X}/t) \in A \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in A), \quad t \to \infty\,. \tag{3.2.7}$$

*In particular, for $\beta \in \mathcal{P}^*_d$, the following convergence holds:*

$$\mathbb{P}(\pi(\mathbf{X}/t) \in C_\beta \mid |\mathbf{X}| > t) \to \mathbb{P}(\mathbf{Z} \in C_\beta), \quad t \to \infty\,. \tag{3.2.8}$$

2. *For $\beta \in \mathcal{P}^*_d$, if $\mathbb{P}(\boldsymbol{\Theta} \in C_\beta) > 0$, then $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$.*

3. *Recall that a subset $C_\beta$ is maximal for $\mathbf{Z}$ if*

$$\mathbb{P}(\mathbf{Z} \in C_\beta) > 0 \quad \textit{and} \quad \mathbb{P}(\mathbf{Z} \in C_{\beta'}) = 0, \ \textit{for all } \beta' \supsetneq \beta\,,$$

*see Definition 2.3.1. For $\beta \in \mathcal{P}^*_d$, the subset $C_\beta$ is maximal for $\boldsymbol{\Theta}$ if and only if it is maximal*

*for* $\mathbf{Z}$.

**Remark 3.2.3.** Note that the first convergence of Proposition 3.2.1 is a general result than the one stated in Chapter 2 (in which we only deals with the subsets $C_\beta$ and another type of subsets) but the generalization is straightforward.

**From theoretical results...** Let us briefly discuss these results. In order to capture the sparse tail dependence structure of extreme events, we consider the projected vector $\pi(\mathbf{X}/t)$ instead of $\mathbf{X}/|\mathbf{X}|$. Equation (2.3.3) ensures that the angular component $\pi(\mathbf{X}/t)$ approximates well the angular vector $\mathbf{Z}$ on the subspaces $C_\beta$. Of course, the behavior of $\mathbf{Z}$ on $C_\beta$ has to be related to the one of $\mathbf{\Theta}$ on these same subsets. This issue is addressed by the last two results which ensures that

1. the distribution of $\mathbf{Z}$ puts mass on $C_\beta$ if the distribution of $\mathbf{\Theta}$ does,
2. the notion of maximal subsets coincide for $\mathbf{\Theta}$ and $\mathbf{Z}$.

Therefore, assessing the tail dependence of a regularly varying vector $\mathbf{X} \in \mathbb{R}_+^d$ can be done through the study of the behavior of the unit vector $\pi(\mathbf{X}/t) \mid |\mathbf{X}| > t$ on the subsets $C_\beta$ which approximates well the quantity $\mathbb{P}(\mathbf{Z} \in C_\beta)$.

**...to a statistical approach** Our goal is to infer the support of the distribution of $\mathbf{Z}$ with the estimation of the probabilities $\mathbb{P}(\mathbf{Z} \in C_\beta)$, for $\beta \in \mathcal{P}_d^*$. Let us denote by $p(\beta)$ these quantities. We are willing to decide which of these probabilities are positive. Similarly to the set $\mathcal{S}(\mathbf{\Theta})$ defined in (3.2.3), we define the set $\mathcal{S}(\mathbf{Z}) \subset \mathcal{P}_d^*$ as follows:

$$\mathcal{S}(\mathbf{Z}) = \{\beta \in \mathcal{P}_d^*,\, \mathbb{P}(\mathbf{Z} \in C_\beta) > 0\} = \{\beta \in \mathcal{P}_d^*,\, p(\beta) > 0\}\,, \tag{3.2.9}$$

and we denote by $s^*$ its cardinality. In other words, $\mathcal{S}(\mathbf{Z})$ gathers all features $\beta$ on which the angular vector $\mathbf{Z}$ places mass. The main goal of this chapter is to build a statistical approach to decide which $\beta$ belong to $\mathcal{S}(\mathbf{Z})$. This method first relies on asymptotic results obtained for estimators of $p(\beta)$. Then, the idea it to used model selection to identify not only the most relevant features $\beta$ but also an optimal threshold for which Equation (3.2.8) approximately holds.

## 3.3 Asymptotic results

### 3.3.1 Statistical framework

From now on we consider a sequence of i.i.d. regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index $\alpha$ and spectral vector $\mathbf{\Theta}$. We also consider a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. With these notations, we have the following weak convergence:

$$\mathbb{P}\left(\pi(\mathbf{X}/t) \in \cdot \mid |\mathbf{X}| > t\right) \xrightarrow{w} \mathbb{P}(\mathbf{Z} \in \cdot)\,, \quad t \to \infty\,. \tag{3.3.1}$$

Our aim is to infer the behavior of the angular vector $\mathbf{Z}$ with the sample $(\mathbf{X}_n)_{n\in\mathbb{N}}$. To this end we focus on the probabilities $p(\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta)$ for $\beta \in \mathcal{P}_d^*$ since they emphasize the extremal joint behavior of the components of $\mathbf{X}$ (see Subsection 3.1.2). Our aim is to classify which ones belong to $\mathcal{S}(\mathbf{Z})$ and which ones do not.

**Classical approach for extremes** The classical assumption to provide a statistical approach in EVT is to consider a positive sequence $(u_n)_{n\in\mathbb{N}}$, $u_n \to \infty$, which plays the role of the threshold $t$ in Equation (3.3.1) (see Beirlant et al. (2006), Section 9.4.1). This means that for $n \in \mathbb{N}$, the quantity $u_n$ must be seen as the threshold above which the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ are extreme. It is customary in EVT to define a level $k = k_n = n\mathbb{P}(|\mathbf{X}| > u_n)$ and to assume that $k_n \to \infty$ when $n \to \infty$. Note that the assumption $u_n \to \infty$ implies that $k_n/n = \mathbb{P}(|\mathbf{X}| > u_n) \to 0$. So $k_n$ tends to infinity at a slower rate than $n$. A natural unbiased estimator for $k_n$ is $\hat{k} = \hat{k}_n = \sum_{j=1}^n \mathbb{1}_{|\mathbf{X}_j|>u_n}$ which corresponds to the number of exceedances above the threshold $u_n$ i.e. the number of extreme values. Thus, the assumption $k_n \to \infty$ means that we focus on more and more extreme observations as the sample size increases. The estimator $\hat{k}_n$ satisfies the following convergence properties:

$$\hat{k}_n/k_n \to 1 \text{ a.s.} \quad \text{and} \quad \sqrt{k_n}\left(\hat{k}_n/k_n - 1\right) \xrightarrow{d} N\,, \quad n \to \infty\,, \tag{3.3.2}$$

where $N$ is a standard normal random variable. Another similar approach is to define a level $k$ which corresponds to the number of exceedances, and to assume that $k \to \infty$ and $k/n \to 0$, when $n \to \infty$. In this case, we consider $u_n = |\mathbf{X}|_{(k)}$, which corresponds to the $k$-th largest vector with respect to the $\ell^1$-norm.

In order to identify the set $\mathcal{S}(\mathbf{Z})$ defined in (3.2.9), we need to provide suitable estimators for the quantities $p(\beta) = \mathbb{P}(\mathbf{Z} \in C_\beta)$. Following Equation (3.2.8), we know that the aforementioned probabilities appear as the limits of the pre-asymptotic quantities $\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$. The key issue is then to decide based on the sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ whether $p(\beta)$ is positive or null.

We adopt here a couple of notations which will be of constant use in what follows. For a Borel subset $A$ of $\mathbb{S}_+^{d-1}$, we set

$$\begin{aligned} p(A) &= \mathbb{P}(\mathbf{Z} \in A)\,, \\ p_n(A) &= \mathbb{P}(\pi(\mathbf{X}/u_n) \in A \mid |\mathbf{X}| > u_n)\,, \\ T_n(A) &= \sum_{j=1}^n \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\}\,. \end{aligned}$$

The random variable $T_n(A)$ corresponds to the data which are projected in the subset $A$ among the extreme values.For sake of simplicity we keep our previous notations regarding $C_\beta$ and write $p(\beta) = p(C_\beta)$ and similarly $p_n(\beta) = p_n(C_\beta)$ and $T_n(\beta) = T_n(C_\beta)$. The expectation of $T_n(A)$ is equal to $\mathbb{E}\big[T_n(A)\big] = k_n p_n(A)$. Besides, the first point of Proposition 3.2.1 implies that, under a regularity assumption on $A$, the probability $p_n(A)$ converges to $p(A)$. This encourages to estimate

the probability $p(A)$ through a classical bias-variance decomposition:

$$\frac{T_n(A)}{k_n} - p(A) = \left[\frac{T_n(A)}{k_n} - p_n(A)\right] + \left[p_n(A) - p(A)\right]. \tag{3.3.3}$$

The first term is addressed in the following sections. For the second one, Equation (3.2.7) ensures that it vanishes at infinity. However, it is common to assume a stronger condition like $\sqrt{k_n}(p_n(A) - p(A)) \to 0$ as $n \to \infty$. This will be discussed more in detail in what follows.

**Estimation of $\mathcal{S}(\mathbf{Z})$** Going back to the subsets $C_\beta$, the decomposition in Equation (3.3.3) highlights the fact that the study of $p(\beta)$ will be conducted through the analysis of the pre-asymptotic probabilities $p_n(\beta)$. In particular, we would like to define a similar set as $\mathcal{S}(\mathbf{Z})$ but for $p_n(\beta)$. In order to avoid that such a set depends on $n$, we replace the natural condition $p_n(\beta) > 0$ for all $n \geq 1$ by the stronger one $k_n p_n(\beta) \to \infty$. This leads to the following subset of features:

$$\mathcal{R}_k(\mathbf{Z}) = \{\beta \in \mathcal{P}_d^*,\ k_n p_n(\beta) \to \infty \text{ when } n \to \infty\}. \tag{3.3.4}$$

We denote by $r^*$ the cardinality of $\mathcal{R}_k(\mathbf{Z})$. This definition implies two straightforward consequences. The first one is that for all feature $\beta \in \mathcal{R}_k(\mathbf{Z})$, the probability $p_n(\beta)$ is positive for $n$ large enough. Second, we have the following inclusion: $\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z})$. In particular, the cardinalities of these sets satisfy the inequality $s^* \leq r^*$.

**Outline of the statistical study** The rest of this section is devoted to asymptotic results for the estimator $T_n(A)$ under some assumptions on the Borel set $A \subset \mathbb{S}_+^{d-1}$. A law of large numbers ensures the convergence of $T_n(A)/k_n$ to $p(A)$ as $n$ increases (Proposition 3.3.1). Then a central limit theorem is established in order to exhibit a limit distribution for the previous estimators when $p(A) \notin \{0, 1\}$ (Theorem 3.3.1). This latter convergence holds under an assumption of convergence of the bias $p_n(A) - p(A)$ and brings out a rate of convergence of order $\sqrt{k_n}$. Regarding the features $\beta$, the purpose of Section 3.4 is then to extend these results at a multivariate level by considering all subsets $C_\beta$ simultaneously.

### 3.3.2 A univariate approach

We start our statistical study with a proposition which establishes the consistency of the estimator $T_n(A)$ for a Borel subset $A \subset \mathbb{S}_+^{d-1}$. This result is proved for both pre-asymptotic probability $p_n(A)$ and asymptotic one $p(A)$. For this latter, a regularity assumption on the Borel set $A$ is needed.

**Proposition 3.3.1.** *We consider a sequence of i.i.d. regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index $\alpha$ and spectral vector $\mathbf{\Theta}$, a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$.*

1. *For all Borel set $A \subset \mathbb{S}_+^{d-1}$, the following convergence in probability holds:*

$$\frac{T_n(A)}{k_n} - p_n(A) \to 0\,, \quad n \to \infty\,. \tag{3.3.5}$$

2. *For all Borel set $A \subset \mathbb{S}_+^{d-1}$ such that $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$ the following convergence in probability holds:*

$$\frac{T_n(A)}{k_n} \to p(A)\,, \quad n \to \infty\,. \tag{3.3.6}$$

In particular, the convergence in Equation (3.3.6) holds for the subsets $A = C_\beta$ and ensures that it is possible to estimate $p(\beta)$ with $T_n(\beta)/k_n$. If $p(\beta) = 0$, that is, if $\mathbf{Z}$ does not place mass on the subset $C_\beta$, then $T_n(\beta)/k_n$ becomes smaller and smaller as $n$ increases. Actually, as soon as the dimension $d$ is large, a lot of $T_n(\beta)$'s are even equal to 0 since the level $k_n$ is far below the number of subsets $2^d - 1$.

**Remark 3.3.1.** Under the assumption that $k_n p_n(A) \to \infty$, we can extend the convergence in (3.3.5) to the following convergence in probability

$$\frac{T_n(A)}{k_n p_n(A)} \to 1\,, \quad n \to \infty\,.$$

In particular, this convergence holds for all $C_\beta$ such that $\beta \in \mathcal{R}_k(\mathbf{Z})$.

We now establish asymptotic normality for $T_n(A)$.

**Theorem 3.3.1.** *We consider a sequence of i.i.d. regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index $\alpha$ and spectral vector $\boldsymbol{\Theta}$, a Pareto($\alpha$)-distributed random variable $Y$ independent of $\boldsymbol{\Theta}$ and we set $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$. Finally, we fix a Borel set $A \in \mathbb{S}_+^{d-1}$.*

1. *If $k_n p_n(A) \to \infty$, then*

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.3.7}$$

2. *If $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$ and $p(A) > 0$, then*

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.3.8}$$

3. *if $\mathbb{P}(Y\boldsymbol{\Theta} \in \partial\pi^{-1}(A)) = 0$ and $p(A) > 0$, and if we assume that*

$$\sqrt{k_n}(p_n(A) - p(A)) \to 0\,, \quad n \to \infty\,, \tag{3.3.9}$$

*then*

$$\sqrt{k_n}\frac{T_n(A)/k_n - p(A)}{\sqrt{p(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,. \tag{3.3.10}$$

Theorem 3.3.1 ensures that $T_n(A)/k_n$ is asymptotically normal as soon as $k_n p_n(A) \to \infty$. This assumption implies that $p_n(A)$ is positive for $n$ large enough. It is for instance true as soon as $p(A) > 0$. From convergence (3.3.7) to convergence (3.3.8), the denominator $\sqrt{p_n(A)}$ has been replaced by $\sqrt{p(A)}$. This requires that $p_n(A) \to p(A) > 0$ which justifies the regularity assumption. Regarding the subsets $C_\beta$, the convergences (3.3.8) and (3.3.10) only hold for the features $\beta$ such that $p(\beta) > 0$, that is, for $\beta \in \mathcal{S}(\mathbf{Z})$. On the contrary, the convergence (3.3.7) holds for the features $\beta$ such that $k_n p_n(\beta) \to \infty$, that is, for $\beta \in \mathcal{R}_k(\mathbf{Z})$.

The results obtained in Proposition 3.3.1 and in Theorem 3.3.1 highlight the asymptotic behavior of $T_n(A)$ when $A \subset \mathbb{S}_+^{d-1}$ is a fixed Borel set. Regarding the subsets $C_\beta$, these results allow the features $\beta$ to be studied individually. The next step is to establish results which address the question of the joint estimation of the probabilities $p(\beta)$.

**Remark 3.3.2.** If we consider $r$ features $\beta_1, \ldots, \beta_r$ and if we assume that $k_n p_n(C_{\beta_j}) \to \infty$ for all $j = 1, \ldots, r$, then we obtaint that $k_n p_n(\cup_j C_{\beta_j}) \to \infty$. Thus Theorem 3.3.1 yields to the following convergence

$$\sqrt{k_n}\frac{T_n(\cup_j C_{\beta_j})/k_n - p_n(\cup_j C_{\beta_j})}{\sqrt{p_n(\cup_j C_{\beta_j})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

which can be rephrased as follows

$$\sqrt{k_n}\frac{\sum_{j=1}^r T_n(C_{\beta_j})/k_n - \sum_{j=1}^r p_n(C_{\beta_j})}{\sqrt{\sum_{j=1}^r p_n(C_{\beta_j})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

This convergence will be useful in Section 3.5.

## 3.4 General results at a multivariate level

We move on to the multivariate setting. There, we only focus on the subsets $A = C_\beta$, $\beta \in \mathcal{P}_d^*$, for which we study the behavior of $\mathbf{X}$ and $\mathbf{Z}$. Our main goal is identify the set of features $\mathcal{S}(\mathbf{Z})$ defined in (3.2.9), i.e. to distinguish the positive probabilities $p(\beta)$ from the null ones. In a nutshell, the idea is to define a cutoff on the $T_n(\beta)$ to identify the features $\beta$ which belong to $\mathcal{S}(\mathbf{Z})$. To this end, we consider the set

$$\widehat{\mathcal{S}}_n(\mathbf{Z}) = \left\{\beta \in \mathcal{P}_d^*,\, T_n(\beta) > 0\right\}, \tag{3.4.1}$$

and denote by $\hat{s}_n$ its cardinality.

### 3.4.1 Estimation of the set $\mathcal{S}(\mathbf{Z})$

Some properties of the empirical set $\widehat{\mathcal{S}}_n(\mathbf{Z})$ are developed here. First, for all $\beta \in \mathcal{S}(\mathbf{Z})$, $p_n(\beta) \to p(\beta) > 0$ when $n \to \infty$, which implies that $p_n(\beta)$ is positive for $n$ large enough. The corresponding result for $T_n(\beta)$ is a consequence of the following lemma.

**Lemma 3.4.1.** *For $\beta \in \mathcal{P}_d^*$, we have the following equivalent*

$$\log(\mathbb{P}(T_n(\beta) = 0)) \sim -k_n p_n(\beta)\,, \quad n \to \infty\,.$$

If $\beta \in \mathcal{R}_k(\mathbf{Z})$, then $-k_n p_n(\beta) \to -\infty$ and thus Lemma 3.4.1 implies that $\mathbb{P}(T_n(\beta) = 0) \to 0$ when $n \to \infty$. This proves that for all $\beta \in \mathcal{R}_k(\mathbf{Z})$ the observations $T_n(\beta)$ are positive with probability converging to 1. In particular, this remark is true for all $\beta \in \mathcal{S}(\mathbf{Z})$. In this case, it means that if the vector $\mathbf{Z}$ places some mass in the direction $\beta$, then at least one extreme observation appears in this direction.

A consequence of Lemma 3.4.1 is that

$$\mathbb{P}\big(\mathcal{R}_k(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})\big) = 1 - \mathbb{P}\big(\exists \beta \in \mathcal{R}_k(\mathbf{Z}),\, \beta \notin \widehat{\mathcal{S}}_n(\mathbf{Z})\big) \geq 1 - \sum_{\beta \in \mathcal{R}_k(\mathbf{Z})} \mathbb{P}(T_n(\beta) = 0) \to 1\,,$$

when $n \to \infty$. Consequently, since $\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z})$, we also have the following convergence

$$\mathbb{P}\big(\mathcal{S}(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})\big) \to 1\,, \quad n \to \infty\,.$$

It is not so easy to obtain the converse inclusion between $\mathcal{R}_k(\mathbf{Z})$ and $\widehat{\mathcal{S}}_n(\mathbf{Z})$. However, if $\beta \notin \mathcal{S}(\mathbf{Z})$, then $p(\beta) = 0$. In this case, Lemma 3.4.1 implies that $\mathbb{P}(T_n(\beta) = 0) \to 1$ if and only if $k_n p_n(\beta) \to 0$. If this latter convergence holds for all $\beta \in \mathcal{S}(\mathbf{Z})^c$, then we obtain that

$$\mathbb{P}\big(\mathcal{S}(\mathbf{Z})^c \subset \widehat{\mathcal{S}}_n(\mathbf{Z})^c\big) = 1 - \mathbb{P}\big(\exists \beta \in \mathcal{S}(\mathbf{Z})^c,\, \beta \in \widehat{\mathcal{S}}_n(\mathbf{Z})\big) \geq 1 - \sum_{\beta \in \mathcal{S}(\mathbf{Z})^c} \mathbb{P}(T_n(\beta) > 0) \to 1\,,$$

when $n \to \infty$. We gather these results in the following proposition.

**Proposition 3.4.1.**

1. *With probability converging to 1, we have the inclusions*

$$\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z}) \subset \widehat{\mathcal{S}}_n(\mathbf{Z})\,.$$

2. *If for all $\beta \in \mathcal{S}(\mathbf{Z})^c$,*

$$k_n p_n(\beta) \to 0\,, \quad n \to \infty\,, \tag{3.4.2}$$

*then with probability converging to 1, $\widehat{\mathcal{S}}_n(\mathbf{Z}) \subset \mathcal{S}(\mathbf{Z})$.*

A consequence of the first point of Proposition 3.4.1 is that the cardinality of the sets $\mathcal{S}(\mathbf{Z})$, $\mathcal{R}_k(\mathbf{Z})$, and $\widehat{\mathcal{S}}_n(\mathbf{Z})$ are satisfying the inequality $s^* \leq r^* \leq \hat{s}_n$ with probability converging to 1.

Regarding the second point, the assumption in (3.4.2) is quite strong compared to the one given in Equation (3.3.9) and implies in particular that $\mathcal{S}(\mathbf{Z}) = \mathcal{R}_k(\mathbf{Z})$. The numerical results introduced in Section 3.6 show that this assumption is not satisfied on simulated data. This is why, unless stated otherwise, we do not assume that 3.4.2 holds.

At this point, the statistical setting is the following one. For a fixed $n$ large enough, we have a collection of features $\beta$ such that $T_n(\beta) > 0$ and the following inclusions are satisfied:

$$\mathcal{S}(\mathbf{Z}) \subset \mathcal{R}_k(\mathbf{Z}) \subset \{\beta \in \mathcal{P}_d^*,\, p_n(\beta) > 0\} \overset{a.s.}{=} \widehat{\mathcal{S}}_n(\mathbf{Z})\,, \tag{3.4.3}$$

where the last equality results from Equation (3.3.5) in Proposition 3.3.1. These inclusions highlight the fact that the observations tend to overestimate the number of relevant directions $\beta$. Therefore, the goal is to decide which ones are indeed in $\mathcal{S}(\mathbf{Z})$. In other words, we need to build a statistical method which brings out a cutoff dividing the empirical set $\widehat{\mathcal{S}}_n(\mathbf{Z})$ into two subsets: a first one corresponding to the features $\beta$ which belongs to $\mathcal{S}(\mathbf{Z})$ and a second one which contains the features $\beta$ which appear because of a possible bias between the probabilities arising from the non-asymptotic sample and the ones representing the theoretical asymptotic framework. In this context, it is customary to use either a concentration inequality or model selection. We develop the first aspect in the following section. However, since it does not provide relevant results on numerical examples, we prefer to focus on model selection which is the purpose of Section 3.5.

### 3.4.2 A concentration result

We first establish a non-asymptotic result for a fixed number of data $n$. The idea is to control the difference $T_n(\beta)/k_n - p_n(\beta)$ for $\beta$ in a fixed subset $\mathcal{B} \subset \mathcal{P}_d^*$. Hence, we establish a concentration result which holds uniformly on $\mathcal{B}$. This result emphasizes the dependence of the difference $T_n(\beta)/k_n - p_n(\beta)$ on the level $k_n$ and on the number of cones $b = \#\mathcal{B}$.

**Theorem 3.4.1.** *We consider a sequence of i.i.d. regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \ldots$ with tail index $\alpha$ and spectral vector $\mathbf{\Theta}$, a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$ and we set $\mathbf{Z} = \pi(Y\mathbf{\Theta})$. We consider a threshold $u_n \to \infty$ and assume that $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$. For a non-empty subset $\mathcal{B} \subset \mathcal{P}_d^*$ of cardinal $b = \#\mathcal{B} \in \{1, \ldots, 2^d - 1\}$ we define*

$$c_n(\mathcal{B}) = \max_{\beta \in \mathcal{B}} p_n(\beta) \tag{3.4.4}$$

*Then, for all $\delta > 0$, the following inequality holds:*

$$\mathbb{P}\left( \max_{\beta \in \mathcal{B}} \left| \frac{T_n(\beta)}{k_n} - p_n(\beta) \right| > f(\delta, n, b) \right) \leq 2e^{-\delta}\,, \tag{3.4.5}$$

*with*

$$f(\delta, n, \mathcal{B}) = \sqrt{2}\sqrt{c_n(\mathcal{B})}\sqrt{\frac{\log(b) + \delta}{k_n}} + \frac{\log(b) + \delta}{3k_n}\,. \tag{3.4.6}$$

Theorem 3.4.1 highlights a rate of convergence of order $\sqrt{k_n}$ which has already been established in Theorem 3.3.1. While this result is useful in order to work at a non-asymptotic level, it turns out that it does not highlight a natural method to identify $\mathcal{S}(\mathbf{Z})$. Besides, our tries on numerical simulations do not yield to relevant results. This is why we will rather turn to multivariate convergence results in order to apply model selection.

### 3.4.3 Ordering the $\beta$'s

In order to study the common behavior of the components $T_n(\beta)$ we need to glue them together and to build a vector of $\mathbb{R}^{2^d-1}$. Note that this can not be easily addressed since there is no specific order on $\mathcal{P}_d^*$. Therefore, we need to fix an order between the $\beta$'s, i.e. to define a bijection

$$\sigma : \{1, \dots, 2^d - 1\} \to \mathcal{P}_d^* \,.$$

The idea is to choose an order $\sigma$ which takes into account the values of $p(\beta)$. However, such an order can only be introduced for $\beta \in \mathcal{S}(\mathbf{Z})$ since the other ones are all equal to zero. Therefore, the bijection $\sigma$ is defined in two steps. First, we consider the $s^* = \#\mathcal{S}(\mathbf{Z})$ first values. In order to define them without any ambiguity, we make the following assumption on $\mathbf{p}$.

**Assumption 3.4.1.** For all $\beta, \beta' \in \mathcal{S}(\mathbf{Z})$, $p(\beta) \neq p(\beta')$.

Under Assumption 3.4.1, we define $\sigma(j)$ for $j = 1, \dots, s^*$ by considering the probabilities in $\mathcal{S}(\mathbf{Z})$ in the decreasing order, that is

$$\begin{aligned}
\sigma(1) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^*} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z})} p(\beta) \,, \\
\sigma(2) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^* \setminus \sigma(1)} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z}) \setminus \sigma(1)} p(\beta) \,, \\
&\vdots \\
\sigma(s^*) &= \operatorname*{arg\,max}_{\beta \in \mathcal{P}_d^* \setminus \{\sigma(1), \dots, \sigma(s^*-1)\}} p(\beta) = \operatorname*{arg\,max}_{\beta \in \mathcal{S}(\mathbf{Z}) \setminus \{\sigma(1), \dots, \sigma(s^*-1)\}} p(\beta) \,.
\end{aligned} \tag{3.4.7}$$

Then, we need to define $\sigma(j)$ for $j = s^* + 1, \dots, 2^d - 1$. Since the remaining values of $p(\beta)$ are null, no natural order appears here. Therefore, we fix an arbitrary order once and for all, i.e. we define distinct images $\sigma(j) \in \mathcal{P}_d^* \setminus \{\sigma(1), \dots, \sigma(s^*)\}$ for all $j = s^* + 1, \dots, 2^d - 1$. This order being now fixed for the rest of the chapter, all vectors of $\mathbb{R}^{2^d-1}$ whose components are indexed by $\mathcal{P}_d^*$ will be written based on this order. Moreover, we simplify the notations by setting $\beta_j = \sigma(j)$ for all $j = 1, \dots, 2^d - 1$.

With these considerations we define the vector $\mathbf{p} \in \mathbb{R}^{2^d-1}$ whose components are associated to the order defined in (3.4.7), i.e. $p_j = p(\beta_j) = p(\sigma(j))$. By construction, the vector $\mathbf{p}$ satisfies

$$p_1 = p(\beta_1) \geq \ldots \geq p_{s^*} = p(\beta_{s^*}) > p_{s^*+1} = \ldots = p_{2^d-1} = 0 \,.$$

In particular, the aforementioned order is also taken into account for the set $\mathcal{S}(\mathbf{Z})$ so that $\mathbf{p}_{\mathcal{S}(\mathbf{Z})} = \mathbf{p}_{\{1,\dots,s^*\}}$. Finally, we use the same order to define the vectors $\mathbf{T}_n$ and $\mathbf{p}_n$, whose components are given by

$$T_{n,j} = T_n(\beta_j) \quad \text{and} \quad p_{n,j} := p_n(\beta_j), \quad j = 1, \dots, 2^d - 1\,. \tag{3.4.8}$$

Contrary to the components of the vector $\mathbf{p}$, the ones of the vectors $\mathbf{T}_n$ and $\mathbf{p}_n$ are not necessary ordered in a decreasing order. However, this is asymptotically true, as stated in the following section.

### 3.4.4 Multivariate convergence

We discuss some convergence results for the random vector $\mathbf{T}_n$. With the order defined below, the components $T_{n,j}$, $p_{n,j}$, and $p_j$ of the vectors $\mathbf{T}_n$, $\mathbf{p}_n$, and $\mathbf{p}$ are all the three associated to the same subset $C_\beta$. Therefore, the consistency of $\mathbf{T}_n$ is a straightforward extension of Proposition 3.3.1:

$$\frac{\mathbf{T}_n}{k_n} - \mathbf{p}_n \to 0\,, \quad n \to \infty\,, \quad \text{in probability}\,, \tag{3.4.9}$$

and

$$\frac{\mathbf{T}_n}{k_n} - \mathbf{p} \to 0\,, \quad n \to \infty\,, \quad \text{in probability}\,. \tag{3.4.10}$$

For $r \geq 1$, we consider the subset $\mathrm{Ord}_r = \{\mathbf{x} \in \mathbb{R}^r,\, x_1 \geq \ldots \geq x_r\}$ whose boundary is given by $\partial\mathrm{Ord}_r = \{\mathbf{x} \in \mathbb{R}^r,\, \exists j \neq k,\, x_j = x_k\}$. On the one hand, the definition of the vector $\mathbf{p}$ ensures that $\mathbf{p} \in \mathrm{Ord}_{2^d-1}$. On the other hand, Assumption 3.4.1 ensures that $\mathbf{p}_{\{1,\dots,s^*\}} \notin \partial\mathrm{Ord}_{s^*}$. Since $\mathbf{T}_n/k_n \to \mathbf{p}$ in probability, it follows from the Portmanteau theorem and Equation (3.4.10) that

$$\mathbb{P}(\mathbf{T}_{n,\,\{1,\dots,s^*\}} \in \mathrm{Ord}_{s^*}) = \mathbb{P}(k_n^{-1}\mathbf{T}_{n,\,\{1,\dots,s^*\}} \in \mathrm{Ord}_{s^*}) \to \mathbb{P}(\mathbf{p}_{\{1,\dots,s^*\}} \in \mathrm{Ord}_{s^*}) = 1\,, \quad n \to \infty\,.$$

Hence, if we fix $\delta > 0$, then there exists $n_0$ such that for all $n \geq n_0$

$$\mathbb{P}(\mathbf{T}_{n,\,\{1,\dots,s^*\}} \in \mathrm{Ord}_{s^*}) \geq 1 - \delta\,. \tag{3.4.11}$$

In other words, if $n$ is large, then the vector $\mathbf{T}_{n,\,\{1,\dots,s^*\}}$ have its components ordered in the decreasing order with high probability.

In order to apply a model selection we need to obtain an asymptotic distribution for the vector $\mathbf{T}_n$. The idea is to extend the results obtained in Theorem 3.3.1. Recall that the convergence (3.3.7) in Theorem 3.3.1 holds only for subsets $A$ such that $k_n p_n(A) \to \infty$. Therefore, in order to obtain a multivariate convergence for the subsets $C_\beta$ it is necessary to restrict ourselves to the features $\beta \in \mathcal{R}_k(\mathbf{Z})$, where $\mathcal{R}_k(\mathbf{Z})$ is defined in (3.3.4). Consequently, the restricted vectors $\mathbf{p}_{\mathcal{R}_k(\mathbf{Z})}$, $\mathbf{p}_{n,\,\mathcal{R}_k(\mathbf{Z})}$, and $\mathbf{T}_{n,\,\mathcal{R}_k(\mathbf{Z})}$ of $\mathbb{R}^{r^*}$ are considered.

**Theorem 3.4.2.** *We consider a sequence of i.i.d. regularly varying random vectors $\mathbf{X}, \mathbf{X}_1, \mathbf{X}_2, \dots$ with tail index $\alpha$ and spectral vector $\mathbf{\Theta}$, a Pareto($\alpha$)-distributed random variable $Y$ independent of $\mathbf{\Theta}$*

*and we set* $\mathbf{Z} = \pi(Y\boldsymbol{\Theta})$. *We consider a threshold* $u_n \to \infty$ *and assume that* $k_n = n\mathbb{P}(|\mathbf{X}| > u_n) \to \infty$.

1. *The following weak convergence on* $\mathcal{R}_k(\mathbf{Z})$ *holds:*

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{r^*})\,, \quad n \to \infty\,. \tag{3.4.12}$$

2. *The following weak convergence on* $\mathcal{S}(\mathbf{Z})$ *holds:*

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{S}(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{S}(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,. \tag{3.4.13}$$

3. *Moreover, if we assume that*

$$\forall \beta \in \mathcal{S}(\mathbf{Z})\,, \quad \sqrt{k_n}(p_n(\beta) - p(\beta)) \to 0\,, \quad n \to \infty\,, \tag{3.4.14}$$

*then the following weak convergence on* $\mathcal{S}(\mathbf{Z})$ *holds:*

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{S}(\mathbf{Z})}}{k_n} - \mathbf{p}_{\mathcal{S}(\mathbf{Z})}\right) \xrightarrow{d} \mathcal{N}(0, Id_{s^*})\,, \quad n \to \infty\,. \tag{3.4.15}$$

The multivariate convergence in (3.4.12) (respectively in (3.4.13) and in (3.4.15)) is the extension of the univariate convergence in (3.3.7) (respectively in (3.3.8) and in (3.3.10). Similarly, the bias assumption in (3.4.14) corresponds to the assumption in (3.3.9).

From Equation (3.4.12) we obtain that the vector

$$\mathbf{U}_n = \sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right)$$

satisfies the convergence

$$\mathbf{U}_n^\top \cdot \mathbf{U}_n = k_n\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right)^\top \mathrm{Diag}(\mathbf{p}_{\mathcal{R}_k(\mathbf{Z})})^{-1}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n} - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right) \xrightarrow{d} \chi^2(r^*)\,, \tag{3.4.16}$$

when $n \to \infty$ and where $\chi^2(r^*)$ denotes a chi-squared distribution with $r^*$ degrees of freedom. This convergence can be rephrased as follows:

$$k_n \sum_{j=1}^{r^*} \frac{(\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}/k_n - \mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^2}{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}} \xrightarrow{d} \chi^2(r^*)\,, \quad n \to \infty\,. \tag{3.4.17}$$

**Remark 3.4.1.** If we fix $s < r^*$, then the subsets $C_{\beta_{s+1}}, \ldots, C_{\beta_{r^*}}$ satisfy the property

$$\sum_{j=s+1}^{r^*} T_{n,j} = T_n(\cup_{j=s+1}^{r^*} C_{\beta_j})\,,$$

see Remark 3.3.2. Hence, the vector

$$\mathbf{U}_n(s) = \sqrt{k_n}\left(\frac{T_{n,1}/k_n - p_{n,1}}{\sqrt{p_{n,j}}}, \ldots, \frac{T_{n,s}/k_n - p_{n,s}}{\sqrt{p_{n,s}}}, \frac{\sum_{j=s+1}^{r^*}\left(T_{n,s}/k_n - p_{n,s}\right)}{\sum_{j=s+1}^{r^*}\sqrt{p_{n,s}}}\right)^\top$$

converges in distribution to a random vector of $\mathbb{R}^{s+1}$ with distribution $\mathcal{N}(0, Id_{s+1})$. Then, similarly to Equation (3.4.16), we have the convergence

$$\mathbf{U}_n(s)^\top \cdot \mathbf{U}_n(s) \xrightarrow{d} \chi^2(s+1)\,, \quad n \to \infty\,. \tag{3.4.18}$$

This convergence will be central in order to provide a suitable procedure for model selection, the key point of the method being to identify $s^*$.

## 3.5 Model selection

Based on the asymptotic results established in Section 3.4, we provide a model selection which addresses two issues. The first one concerns the identification of the set $\mathcal{S}(\mathbf{Z})$ and the second one is the choice of an optimal level $k_n$. As already mentioned in Remark 3.2.2, the choice of the threshold $u_n$ has a direct impact on the result given by the Euclidean projection onto the simplex. Therefore, the identification of $\mathcal{S}(\mathbf{Z})$ and the choice of an optimal level $k_n$ are issues deeply related and which should therefore be addressed simultaneously.

### 3.5.1 Generalities

In all what follows we assume that $n$ is large enough. Thus, it seems natural to assume that the inclusions in Proposition 3.4.1 holds, i.e. that there exists no feature $\beta$ in $\mathcal{S}(\mathbf{Z})$ which does not belong to $\widehat{\mathcal{S}}_n(\mathbf{Z})$. Reciprocally, Assumption 3.4.2 may not hold which means that some observations could appear in a direction $\beta$ on which the distribution of $\mathbf{Z}$ does not place mass. In this case, it seems reasonable to assume that the quantity $T_n(\beta)$ associated to this direction is not very large. Since all the work is now done at a non-asymptotic level, the strict inclusion mainly arises because of a possible bias which appears on the observations. All in all, the sequence of inclusions in (3.4.3) implies that we make the following assumptions on the observations:

- If a feature $\beta$ does not appear in $\widehat{\mathcal{S}}_n(\mathbf{Z})$ we conclude that the distribution of $\mathbf{Z}$ does not place mass in this direction.
- If a feature $\beta$ satisfies $T_n(\beta) \gg 0$, then we infer that $\mathbf{Z}$ concentrates on the associated subset $C_\beta$ .
- If a feature $\beta$ satisfies $T_n(\beta) \approx 0$, then it is likely that this direction appears in $\widehat{\mathcal{S}}_n(\mathbf{Z})$ only because of the bias which arises due to the non-asymptotic setting. There, we assume that the distribution of $\mathbf{Z}$ does not place mass in this direction.

The core of the study is now to provide a suitable procedure which classify the directions $\beta$ which appear in the last two cases.

### 3.5.2 A multinomial model

Our goal is to identify which probabilistic model fits the best the data. To this end, we use Akaike Information Criterion (see Section 1.5) in order to highlight the number of relevant directions $\beta$ on which extreme values appear. Subsequently, we use a similar approach to highlight an optimal level $k_n$. In order to tackle this latter issue it is necessary to consider all observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and not only the extreme ones. This is why an extra-category is created to gather the non-extreme observations. This category is defined as

$$T_{n,\,2^d} := \sum_{j=1}^{n} \mathbb{1}\{|\mathbf{X}_j| \leq u\} = n - \sum_{\beta \in \mathcal{P}_d^*} T_n(\beta)\,.$$

Finding an optimal level $k_n$ boils down to identifying the size of this extra-category. At this stage we have $2^d$ different categories, the first $2^d - 1$ corresponding to the components of the vector $\mathbf{T}_n$ in the order explained in Section 3.4.3, and the last one denoted by $T_{n,\,2^d}$. We add this non-extreme category $T_{n,\,2^d}$ at the end of the vector $\mathbf{T}_n$ to build the vector $\mathbf{T}'_n = (T_{n,1}, \ldots, T_{n,\,2^d-1}, T_{n,\,2^d})^\top \in \mathbb{R}^{2^d}$. In what follows the apostrophe will be associated to the whole data set i.e. including the non-extreme category.

For $n \geq 1$, the vector $\mathbf{T}'_n$ follows a multinomial distribution with size $n$ and probability vector $\mathbf{p}'_n \in \mathbb{R}^{2^d}$ where $\mathbf{p}'_n$ is defined as

$$\mathbf{p}'_n = (q_n p_{n,1}, \ldots, q_n p_{n,\,2^d-1}, 1 - q_n)^\top, \tag{3.5.1}$$

with $q_n = \mathbb{P}(|\mathbf{X}| > u_n)$. We use the following notation: $\mathbf{T}'_n \sim \mathcal{M}(n, \mathbf{p}'_n)$. Similarly, for a fixed level $k = k_n$ the vector $\mathbf{T}_n$ satisfies the relation $T_{n,1} + \ldots + T_{n,\,2^d-1} = k$. Therefore, its distribution is multinomial with size $k$ and probability vector $\mathbf{p}_n$: $\mathbf{T}_n \sim \mathcal{M}(k, \mathbf{p}_n)$. The two steps of our model selection are then the following ones. First, we identify the optimal proportion of extremes and choose the associated level $k$. Second, for this $k$ we work with the vector $\mathbf{T}_n$ and we would like to identify the set $\mathcal{S}(\mathbf{Z})$. From a theoretical point of view, it is easier to first study the $\mathbf{T}_n$ for a fixed (but unknown) level $k$ and then to add the choice of an optimal level. In terms of model selection we define two family of models, one for the observation $\mathbf{T}_n$ when the level $k$ is fixed and one for the whole observation $\mathbf{T}'_n$.

Recall from Equation (3.4.11) that for $n$ large enough, the probability $\mathbb{P}(T_{n,1} \geq T_{n,2} \geq \ldots \geq T_{n,\,s^*})$ is close to 1. From now on we fix a $n$ large enough and work conditionally on this event. This order encourages the use of the following families of models.

First for a fixed level $k$, we consider a multinomial model denoted by $\mathbf{M}(k; \tilde{\mathbf{p}})$, where the param-

eter $\tilde{\mathbf{p}}$ is defined as

$$\tilde{\mathbf{p}} = (\overbrace{\tilde{p}_1, \ldots, \tilde{p}_s, \underbrace{\tilde{p}, \ldots, \tilde{p}}_{r-s}, 0, \ldots, 0}^{2^d-1 \text{ components}}),$$

with $\tilde{p}_1 \geq \ldots \geq \tilde{p}_s, \tilde{p} \in (0,1)$ satisfying the constraint

$$\tilde{p}_1 + \ldots + \tilde{p}_s + (r-s)\tilde{p} = 1\,. \tag{3.5.2}$$

Such a model is entirely characterized by the parameters $\tilde{p}_1, \ldots, \tilde{p}_s, \tilde{p}$ and $r$. The model $\mathbf{M}(k; \tilde{\mathbf{p}})$ highlights the $s$ relevant features $\beta$ which gather the mass of the distribution of $\mathbf{Z}$. The parameter $\tilde{p}$ models the bias and should thus be considered as small and converging to zero when $n$ increases. It emphasizes the idea that among the directions $\beta$ which contain at least one observation some of them indeed belong to the support of $\mathbf{Z}$ while others only appear because of a bias. The first $s$ features correspond to the most relevant ones: It is likely to observe extreme on the associated subsets.

We also consider a multinomial model denoted by $\mathbf{M}'(n; \tilde{\mathbf{p}})$ where the parameter $\tilde{\mathbf{p}}'$ is defined as

$$\tilde{\mathbf{p}}' = (\overbrace{\tilde{q}'\tilde{p}'_1, \ldots, \tilde{q}'\tilde{p}'_{s'}, \underbrace{\tilde{q}'\tilde{p}', \ldots, \tilde{q}'\tilde{p}'}_{r'-s'}, 0, \ldots, 0, 1-\tilde{q}'}^{2^d \text{ terms}}),$$

with $\tilde{p}'_1 \geq \ldots \geq \tilde{p}'_{s'}, \tilde{p}', \tilde{q}' \in (0,1)$ are satisfying the same constraint (3.5.2). Such a model is entirely characterized by the parameters $\tilde{p}'_1, \ldots, \tilde{p}'_{s'}, \tilde{p}', \tilde{q}'$ and $r'$. The model $\mathbf{M}'(n; \tilde{\mathbf{p}}')$ is the extension of the first one when all the data are considered. The parameter $\tilde{q}'$ models the theoretical proportion of extreme values taken among the data.

### 3.5.3 Estimation of the parameters

The estimation of the parameters for a multinomial model has already been widely studied. Therefore, the purpose of this section is to introduce the likelihood of the aforementioned models and the associated estimators with our notations.

**For the model $\mathbf{M}'(n; \tilde{\mathbf{p}}')$** For $\mathbf{x} \in \mathbb{S}_+^{2^d-1}(n)$ the likelihood of the model $\mathbf{M}'(n; \tilde{\mathbf{p}}')$ is given by

$$L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}'; \mathbf{x}) = \frac{n!}{\prod_{i=1}^{2^d} x_i!} \prod_{i=1}^{s'} (\tilde{q}'\tilde{p}'_i)^{x_i} \prod_{i=s'+1}^{r'} (\tilde{p}'\tilde{q}')^{x_i} (1-\tilde{q}')^{x_{2^d}} \mathbb{1}_{\{\forall j = r'+1, \ldots, 2^d-1,\, x_j = 0\}}\,.$$

For $\mathbf{x}$ such that $x_j = 0$ for all $j = r'+1, \ldots, 2^d - 1$ the log-likelihood evaluated in $\mathbf{x}$ is equal to

$$\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}'; \mathbf{x}) = \log(n!) - \sum_{i=1}^{2^d} \log(x_i!) + \sum_{i=1}^{s'} x_i \log(\tilde{q}'\tilde{p}'_i) + \Big( \sum_{i=s'+1}^{2^d-1} x_i \Big) \log(\tilde{p}'\tilde{q}')$$

$$+ x_{2^d} \log(1 - \tilde{q}') \,. \tag{3.5.3}$$

The optimization of this log-likelihood under the constraint (3.5.2) leads, with our notations, to the following estimators:

$$\widehat{\tilde{q}'} = \frac{\sum_{i=1}^{2^d-1} x_i}{n} \,, \quad \widehat{\tilde{p}'\tilde{q}'} = \frac{\sum_{i=s'+1}^{2^d-1} x_i}{(r'-s')n} \,, \quad \widehat{\tilde{q}'\tilde{p}'_j} = \frac{x_j}{n}, \quad 1 \le j \le s' \,.$$

Note that the condition $x_j = 0$ for all $j = r'+1, \dots, 2^d - 1$ can be rewritten as

$$r' \ge \#\{j = 1, \dots, 2^d - 1,\, x_j > 0\} \,,$$

so that the log-likelihood $\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}'; \mathbf{x})$ is maximum for $\widehat{r'} = \#\{j = 1, \dots, 2^d - 1,\, x_j > 0\}$. All in all, these considerations lead to the following maximum likelihood estimators:

$$\widehat{r'} = \#\{j = 1, \dots, 2^d - 1,\, T_{n,j} > 0\} = \hat{s}_n \,, \tag{3.5.4}$$

$$\widehat{\tilde{q}'} = \frac{\sum_{i=1}^{n} T_{n,i}}{n} = \frac{n - T_{n,2^d}}{n} \,, \tag{3.5.5}$$

$$\widehat{\tilde{p}'} = \frac{\sum_{i=s'+1}^{2^d-1} T_{n,i}}{(\hat{s}_n - s') \sum_{i=1}^{2^d-1} T_{n,i}} = \frac{\sum_{i=s'+1}^{2^d-1} T_{n,i}}{(\hat{s}_n - s')(n - T_{n,2^d})} \,, \tag{3.5.6}$$

$$\widehat{\tilde{p}'_j} = \frac{\tilde{T}_{n,j}}{\sum_{i=1}^{2^d-1} T_{n,i}} = \frac{T_{n,j}}{n - T_{n,2^d}} \,, \quad 1 \le j \le s' \,. \tag{3.5.7}$$

**For the model $\mathbf{M}(k; \tilde{\mathbf{p}})$** We fix $k \le n$ and consider the model $\mathbf{M}(k; \tilde{\mathbf{p}})$. For $\mathbf{x} \in \mathbb{S}_+^{2^d-2}(k)$ the likelihood of this model is given by

$$L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{x}) = \frac{k!}{\prod_{i=1}^{2^d-1} x_i!} \prod_{i=1}^{s} (\tilde{p}_i)^{x_i} \prod_{i=s+1}^{r} (\tilde{p})^{x_i} \mathbb{1}_{\{\forall j = r+1, \dots, 2^d-1,\, x_j = 0\}} \,.$$

Then, for $\mathbf{x}$ such that $x_j = 0$ for all $j = r+1, \dots, 2^d - 1$ the log-likelihood evaluated in $\mathbf{x}$ is equal to

$$\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{x}) = \log(k!) - \sum_{i=1}^{2^d-1} \log(x_i!) + \sum_{i=1}^{s} x_i \log(\tilde{p}_i) + \Big( \sum_{i=s+1}^{2^d-1} x_i \Big) \log(\tilde{p}) \,, \tag{3.5.8}$$

and the maximum likelihood estimators evaluated in $\mathbf{T}_n$ are defined as follows:

$$\begin{aligned}
\widehat{r} &= \hat{s}_n \,, \\
\widehat{\tilde{p}} &= \frac{\sum_{i=s+1}^{\hat{s}_n} T_{n,i}}{(\hat{s}_n - s)k} = \frac{\sum_{i=s+1}^{2^d-1} T_{n,i}}{(\hat{s}_n - s)k} \,, \\
\widehat{\tilde{p}_j} &= \frac{T_{n,j}}{k} \,, \quad 1 \le j \le s \,.
\end{aligned}$$

### 3.5.4 An AIC approach for the model M(k)

In this section we fix $k \leq n$ and consider the associated random vector $\mathbf{T}_n$. The unknown distribution of this vector is denoted by $\mathbf{P}_k$. We consider the model $\mathbf{M}(k;\tilde{\mathbf{p}})$ and we work conditionally on the event that $r = \hat{s}_n$. There, we recall that $\widehat{p_1}, \dots, \widehat{p_s}, \widehat{p}$ denote the maximum likelihood estimators:

$$\widehat{\tilde{\mathbf{p}}} = \underset{\tilde{p}_1+\dots+\tilde{p}_s+(r-s)\tilde{p}=1}{\arg\max} L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\,. \tag{3.5.9}$$

We also define the parameter $\tilde{\mathbf{p}}^* = (\tilde{p}_1^* + \tilde{p}^*, \dots, \tilde{p}_s^* + p^*, \tilde{p}^*)^\top \in \mathbb{R}^{s+1}$ as the optimum of the expectation of the log-likelihood:

$$\tilde{\mathbf{p}}^* = \underset{\tilde{p}_1+\dots+\tilde{p}_s+(r-s)\tilde{p}=1}{\arg\max} \mathbb{E}\big[L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\big]\,. \tag{3.5.10}$$

A similar computation as for the estimators $\widehat{\tilde{\mathbf{p}}}$ gives the relations

$$\forall j = 1,\dots,s\,, \quad \tilde{p}_j^* = p_{n,j}\,, \quad \text{and} \quad \tilde{p}^* = \frac{\sum_{i=1}^r p_{n,i}}{r-s}\,.$$

For all $j = 1, \dots 2^d - 1$, we define

$$m_j = \min\big(\widehat{\tilde{p}_j}, \tilde{p}_j^*\big) = \min\Big(\frac{T_{n,j}}{k}, p_{n,j}\Big) \quad \text{and} \quad M_j = \max\big(\widehat{\tilde{p}_j}, \tilde{p}_j^*\big) = \max\Big(\frac{T_{n,j}}{k}, p_{n,j}\Big)\,.$$

**Assumption 3.5.1.** For all $j = 1, \dots, 2^d - 1$,

$$\frac{p_{n,j}}{m_j^2}\left|\frac{M_j^2}{m_j^2} - 1\right| \to 0\,, \quad \text{and} \quad \frac{1}{m_j^2}\left|\frac{T_{n,j}}{k} - p_{n,j}\right| \to 0\,, \quad n \to \infty\,.$$

Note that this assumption is automatically satisfied for $j \leq s^*$ since in this case $m_j$ and $M_j$ converge to $p_j > 0$.

Our aim is to identify which model $\mathbf{M}(k,\tilde{\mathbf{p}})$ best fits the observations $\mathbf{T}_n$. Following the criterion introduced by Akaike (1973), we choose the model which minimizes the Kullback-Leibler divergence

$$KL\Big(\mathbf{P}_k \Big\| \mathbf{M}(k;\,\tilde{\mathbf{p}})\Big) = \mathbb{E}\left[\log\left(\frac{L_{\mathbf{P}}(\mathbf{T}_n)}{L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)}\right)\right] = \mathbb{E}\Big[\log L_{\mathbf{P}}(\mathbf{T}_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\,, \tag{3.5.11}$$

which must be seen as a function of $\tilde{\mathbf{p}}$. In particular, the first term is constant with respect to the parameter $\tilde{\mathbf{p}}$. Regarding the second term, Equation (3.5.8) entails that

$$\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big] = \log(k!) - \mathbb{E}\left[\sum_{i=1}^{2^d-1}\log(T_i!)\right] + k\sum_{i=1}^{s} p_{n,i}\log(\tilde{p}_i) + k\left(\sum_{i=s+1}^{2^d-1} p_{n,i}\right)\log(\tilde{p})\,. \tag{3.5.12}$$

We will use several time the following result known as "Cauchy's Mean-Value Theorem" (Hille

(1964)).

**Lemma 3.5.1.** *Let $f$ and $g$ be two continuous functions on the closed interval $[a,b]$, $a<b$, and differentiable on the open interval $(a,b)$. Then there exists some $c\in(a,b)$ such that*

$$(f(b)-f(a))g'(c)=(g(b)-g(a))f'(c)\,.$$

Following Lemma 3.5.1, we obtain an Taylor expansion for

$$-\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\,.$$

**Lemma 3.5.2.** *There exists $c_1\in(0,1)$ such that*

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*} \tag{3.5.13}$$

$$\frac{1}{2}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,. \tag{3.5.14}$$

Note that since the quantity $\tilde{\mathbf{p}}^*$ is deterministic, the first term of the right-hand side can be written as follows

$$KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}=\mathbb{E}\Big[\log L_{\mathbf{P}}(\mathbf{T}_n)\Big]-\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n)\Big]\,.$$

The idea is then to provide an Taylor expansion of $\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n)$ around the point $\widehat{\tilde{\mathbf{p}}}$. This is the purpose of the following lemma.

**Lemma 3.5.3.** *There exists $c_2\in(0,1)$ such that*

$$\begin{aligned}-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*;\mathbf{T}_n)=&-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\\&-\frac{1}{2}(\tilde{\mathbf{p}}^*-\widehat{\tilde{\mathbf{p}}})^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(c_2\tilde{\mathbf{p}}^*+(1-c_2)\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)(\tilde{\mathbf{p}}^*-\widehat{\tilde{\mathbf{p}}})\,.\end{aligned} \tag{3.5.15}$$

Now, taking the expectation with respect to $\widehat{\tilde{\mathbf{p}}}$ in Equations (3.5.13) and (3.5.3), and combining both equations, we obtain the following expression for the expectation of the divergence with respect to $\widehat{\tilde{\mathbf{p}}}$:

$$\begin{aligned}&\mathbb{E}\Big[KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\Big]\\&=\mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big]-\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big]\\&+\mathbb{E}\Big[\underbrace{(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)}_{(\star)}\Big]\\&+\frac{1}{2}\mathbb{E}\Big[(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\Big[-\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big|_{c_2\tilde{\mathbf{p}}^*+(1-c_2)\widehat{\tilde{\mathbf{p}}}}\end{aligned} \tag{3.5.16}$$

$$+\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}\Big](\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\Big]\,.$$

The last two steps consist in dealing with the last term of the right-hand side in Equation (3.5.16) and with the term $(\star)$. For the first one, we prove that it converges to zero.

**Lemma 3.5.4.** *Under Assumption 3.5.1, the following convergences in probability holds:*

$$\sup_{(c,c')\in(0,1)^2}\left|k^{-1}\left(\frac{\partial^2\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}}{\partial\tilde{\mathbf{p}}^2}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}-\mathbb{E}\Big[\frac{\partial^2\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}}{\partial\tilde{\mathbf{p}}^2}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{c'\widehat{\tilde{\mathbf{p}}}+(1-c')\tilde{\mathbf{p}}^*}\right)\right|_\infty\to 0\,,$$

*when $n\to\infty$ and where $|\cdot|_\infty$ denotes the infinity norm.*

Since $\sqrt{k}\,\mathrm{Diag}(\mathbf{p}_{n,\{1,\dots,s\}})^{-1/2}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)$ converges to a Gaussian distribution thanks to Theorem 3.4.2, the last term in the right-hand side of Equation (3.5.16) converges to zero when $n\to\infty$.

Moving on to the term $(\star)$, we prove that it converges in distribution to a chi-square-distributed random variable with $s+1$ degrees of freedom.

**Lemma 3.5.5.** *For all $c\in(0,1)$, the following weak convergence holds:*

$$(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}\Big]\Big|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\xrightarrow{d}\chi(s+1)\,,\quad n\to\infty\,.$$

Based on Lemma 3.5.4 and Lemma 3.5.5, Equation (3.5.16) entails, for $n$ large enough, the following approximation:

$$\begin{aligned}\mathbb{E}\Big[KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\Big]&\approx\mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big]-\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big]+\mathbb{E}[\chi^2(s+1)]\\&\approx\mathbb{E}\Big[\log L_{\mathbf{P}_k}(\mathbf{T}_n)\Big]-\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big]+(s+1)\,.\end{aligned}$$

Therefore, for a given level $k$, the idea is to choose the parameter $s$ which minimizes the quantity

$$-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)+(s+1)\,.$$

The last theoretical step of our study is to include the choice of $k$ in our procedure. This is the purpose of the following section.

### 3.5.5 From the extreme values to the whole dataset

The choice of $s^*$ has been done for a fixed level $k$ and by considering the model $\mathbf{M}(k,\tilde{\mathbf{p}})$. The last step, which should practically speaking be done first, is to provide an accurate method to select an optimal level $k$. To this end, we need to work on the whole data set in order to have a fixed number of data. Therefore, we focus on the model $\mathbf{M}(n,\tilde{\mathbf{p}}')$. We denote by $\mathbf{P}_n$ the true distribution of $\mathbf{T}'_n$.

We start by writing down the Kullback-Leibler divergence between $\mathbf{P}_n$ and $\mathbf{M}'(n,\tilde{\mathbf{p}}')$ and de-

compose it as follows

$$KL\Big(\mathbf{P}_n\Big\|\mathbf{M}(n;\,\tilde{\mathbf{p}}')\Big) = \mathbb{E}\left[\log\left(\frac{L_{\mathbf{P}_n}(\mathbf{T}'_n)}{L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n)}\right)\right] = \mathbb{E}\Big[\log L_{\mathbf{P}_n}(\mathbf{T}'_n)\Big] - \mathbb{E}\Big[\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n)\Big]\,. \tag{3.5.17}$$

We focus on the second term of the right-hand side. We decompose the log-likelihood $\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}$ defined in Equation (3.5.3) as follows

$$\begin{aligned}\log L_{\mathbf{M}'(n;\tilde{\mathbf{p}}')}(\tilde{\mathbf{p}}';\mathbf{T}'_n) &= \log((n-T'_{n,\,2^d})!) - \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!) + \sum_{j=1}^{s'}T'_{n,\,j}\log(\tilde{p}'_j) + \log(\tilde{p}')\sum_{j=s'+1}^{2^d-1}T'_{n,\,j}\\ &\quad + \log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!) + (n-T'_{n,\,2^d})\log(\tilde{q}') + T'_{n,\,2^d}\log(1-\tilde{q}')\\ &= \log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n) + \phi(n,\tilde{q}',T'_{n,\,2^d})\,,\end{aligned}$$

where

$$\phi(n,\tilde{q}',T'_{n,\,2^d}) = \log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!) + (n-T'_{n,\,2^d})\log(\tilde{q}') + T'_{n,\,2^d}\log(1-\tilde{q}')\,.$$

Hence, after evaluating the expression in (3.5.17) in $\widehat{\tilde{\mathbf{p}}}'$ and taking the expectation, we obtain that

$$\begin{aligned}\mathbb{E}\Big[KL\Big(\mathbf{P}_n\Big\|\mathbf{M}'(n;\,\tilde{\mathbf{p}}')\Big)\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big] &= \mathbb{E}\Big[\log L_{\mathbf{P}_n}(\mathbf{T}'_n)\Big] + \mathbb{E}\Big[\mathbb{E}\Big[-\log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\mid T'_{n,\,2^d}\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\Big]\\ &\quad - \mathbb{E}\Big[\mathbb{E}\Big[\phi(n,\tilde{q}',T'_{n,\,2^d})\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big]\,.\end{aligned} \tag{3.5.18}$$

The first term of the right-hand side is a constant. For the second term, we remark that the log terms

$$-\mathbb{E}\Big[\log((n-T'_{n,\,2^d})!) - \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!)\Big]$$

are constant. The idea is then to condition the remainder with respect to $T'_{n,\,2^d} = n-k$ in order to apply the results of the previous section. To this end, we use the approximation $T'_{n,\,j} \approx p_{n,\,j}(n-T'_{n,\,2^d})$ which holds when $(n-T'_{n,\,2^d})$ is large. Indeed, a similar result as in Proposition 3.3.1 can be obtained for the vector $\mathbf{T}'_n$:

$$\frac{T'_{n,\,j}}{nq_n} - p_{n\,j} \to 0\,, \quad \text{and} \quad \frac{n-T'_{n,\,2^d}}{nq_n} \to 1\,, \quad n\to\infty\,, \quad \text{in probability}\,,$$

as soon as $nq_n \to \infty$. By combining these two convergences, we obtain that

$$\frac{T'_{n,\,j}}{n - T'_{n,\,2^d}} - p_{n,\,j} \to 0\,, \quad n \to \infty\,, \quad \text{in probability}\,,$$

which justifies our approximation. Thus, after removing the log terms, we consider the quantity

$$\begin{aligned}
&\mathbb{E}\Big[ - \log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) + \log((n - T'_{n,\,2^d})!) - \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d}\Big] \\
&\qquad \approx (n - T'_{n,\,2^d})\Big( \sum_{j=1}^{s'} p_{n,\,j} \log(\tilde{p}'_j) + \log(\tilde{p}') \sum_{j=s'+1}^{2^d-1} p_{n,\,j}\Big) \\
&\qquad \approx \frac{n - T'_{n,\,2^d}}{k}\Big( k \sum_{j=1}^{s'} p_{n,\,j} \log(\tilde{p}'_j) + k \log(\tilde{p}') \sum_{j=s'+1}^{2^d-1} p_{n,\,j}\Big) \\
&\qquad \approx \frac{n - T'_{n,\,2^d}}{k}\left( \mathbb{E}\Big[ - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)\Big] + \log(k!) - \mathbb{E}\Big[ \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d} = n - k\Big]\right),
\end{aligned}$$

from (3.5.12) since $\mathbf{T}_n$ is given $T'_{n,\,2^d} = n - k$ implicitly in the model $\mathbf{M}(k; \tilde{\mathbf{p}})$. This entails that

$$\begin{aligned}
&\mathbb{E}\Big[- \log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) + \log((n - T'_{n,\,2^d})!) - \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d}\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}} \\
&\qquad \approx \frac{n - T'_{n,\,2^d}}{k}\left( \mathbb{E}\Big[ - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}} + \log(k!) - \mathbb{E}\Big[ \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d} = n - k\Big]\right).
\end{aligned}$$

Using the preceding result, for $k$ large enough we obtain the following unbiased approximation of the second term in (3.5.18):

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\Big[ - \log L_{\mathbf{M}(n-T'_{n,\,2^d};\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) + \log((n - T'_{n,\,2^d})!) - \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d}\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\right] \\
&\qquad \approx \frac{n(1-q_n)}{k}\left( \mathbb{E}\left[\mathbb{E}\Big[ - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}}\right] + \log(k!) - \mathbb{E}\left[\mathbb{E}\Big[ \sum_{j=1}^{2^d-1} \log(T'_{n,\,j}!) \mid T'_{n,\,2^d} = n - k\Big]\right]\right).
\end{aligned}$$

For the third term in Equation (3.5.18), we have

$$\begin{aligned}
\mathbb{E}\Big[\phi(n, \tilde{q}', T'_{n,\,2^d})\Big]\Big|_{\widehat{\tilde{\mathbf{p}}}'} &= \mathbb{E}\Big[ \log\left(\frac{n!}{(n - T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!)\Big] + \mathbb{E}\Big[(n - T'_{n,\,2^d}) \log(\tilde{q}') + T'_{n,\,2^d} \log(1 - \tilde{q}')\Big]\Big|_{\widehat{\tilde{q}}'} \\
&= \mathbb{E}\Big[ \log\left(\frac{n!}{(n - T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!)\Big] + nq_n \log(k/n) + n(1 - q_n) \log(1 - k/n)\,,
\end{aligned}$$

as $\widehat{\tilde{q}}' = k/n$.

All in all, the Kullback-Leibler divergence in (3.5.18) satisfies the relation

$$\mathbb{E}\Big[KL\Big(\mathbf{P}_n\Big\|\mathbf{M}(n;\,\tilde{\mathbf{p}}')\Big)\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big] \approx \mathbb{E}\Big[\log L_{\mathbf{P}_n}(\mathbf{T}'_n)\Big] + \frac{n(1-q_n)}{k}\Big(\mathbb{E}\Big[-\log L_{\mathbf{M}'(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big] + (s+1)\Big) + R_{n,\,k}\,, \tag{3.5.19}$$

where $R_{n,\,k}$ is defined as

$$\begin{aligned} R_{n,\,k} = \mathbb{E}\Big[-\log((n-T'_{n,\,2^d})!) + \sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!)\Big] - \mathbb{E}\Big[\log\left(\frac{n!}{(n-T'_{n,\,2^d})!}\right) - \log(T'_{n,\,2^d}!)\Big] - nq_n\log(k/n) \\ - n(1-q_n)\log(1-k/n) + \frac{n(1-q_n)}{k}\left(\log(k!) - \mathbb{E}\Big[\mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T'_{n,\,j}!)\mid T'_{n,\,2^d} = n-k\Big]\Big]\right). \end{aligned}$$

The second term of the right-hand side in Equation (3.5.19) is of order $\log(k!)/k \sim \log(k)$ by Stirling's approximation. After withdrawing the terms of $R_{n,\,k}$ which are constant with respect to $k$, we obtain that

$$\begin{aligned} R_{n,\,k} \propto \frac{n(1-q_n)}{k}\left(\log(k!) - \mathbb{E}\Big[\mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T'_{n,j}!)\mid T'_{n,\,2^d} = n-k\Big]\Big]\right) \\ - \Big(nq_n\log(k/n) + n(1-q_n)\log(1-k/n)\Big). \end{aligned}$$

For the first term, we use the approximation $T'_{n,j} \approx kp_{n,\,j}$ and Stirling's approximation $\log(n!) \sim n\log(n) - n$ which entails that

$$\begin{aligned} \log(k!) - \mathbb{E}\Big[\mathbb{E}\Big[\sum_{j=1}^{2^d-1}\log(T'_{n,j}!)\mid T'_{n,\,2^d} = n-k\Big]\Big] &\approx k\log(k) - k - \sum_{j=1}^{2^d-1}\Big(kp_{n,\,j}\log(kp_{n,\,j}) - kp_{n,\,j}\Big) \\ &\approx k\log(k) - \sum_{j=1}^{2^d-1}kp_{n,\,j}\log(k) - \sum_{j=1}^{2^d-1}kp_{n,\,j}\log(p_{n,\,j}) \\ &\approx -k\sum_{j=1}^{2^d-1}p_{n,\,j}\log(p_{n,\,j}) \end{aligned}$$

where we use that the $p_{n,\,j}$ add up to 1. Thus, after multiplying this approximation by the ratio $(1-q_n)/k$, we can conclude that this terms are asymptotically constant. Finally, we assume that the real proportion of extreme values are small, which implies that $nq_n\log(k/n)$ can be neglected. The only term remaining is then $R_{n,\,k} \sim -n(1-q_n)\log(1-k/n)$.

All in all, we proved that

$$\mathbb{E}\Big[KL\Big(\mathbf{P}_n\Big\|\mathbf{M}'(n;\,\tilde{\mathbf{p}}')\Big)\Big|_{\widehat{\tilde{\mathbf{p}}}'}\Big] \propto \frac{1}{k}\Big(\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}};\mathbf{T}_n)\Big] + (s+1) - \Big)\,.$$

Practically speaking, we choose the parameters $k$ and $s$ which minimizes the quantity

$$\frac{1}{k}\Big( - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n) + (s+1) - k\log(1-k/n)\Big)\,. \tag{3.5.20}$$

where $\widehat{\tilde{\mathbf{p}}}$ denotes the estimator of the maximum likelihood of the model $\mathbf{M}(k;\tilde{\mathbf{p}})$ when the level $k$ is fixed.

## 3.6 Numerical results

We continue in this section the numerical study introduced in Section 2.5. We provide different examples to highlight the relevance of our findings. In particular, the results given in this chapter allows to tackle the two hyperparameter which appear in Algorithm 2. This leads to the following procedure to study dependence for extreme events. The code can be found at https://github.com/meyernicolas/phd_thesis/blob/master/chap_3.

**Data:** $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}_+^d$
**Result:** A list $\mathcal{S}(\mathbf{Z})$ of directions $\beta$ and an optimal level $k$
Compute $\pi(\mathbf{X}_j/t)$, $j = 1, \dots, n$ for different $t$;
Assign to each $\pi(\mathbf{X}_j/t)$ the subsets $C_\beta$ it belongs to;
Compute $\mathbf{T}_n$;
Compute the values of $k$ and $s$ which minimize the AIC Criterion given in Equation (3.5.20);
Define $\mathcal{S}(\mathbf{Z}) = \{\beta,\, T_{n,j}(\beta) > 0 \text{ for } j = 1, \dots, s\}$.

**Algorithm 4:** Tail inference for high-dimensional data.

**Remark 3.6.1.** In order to prove that our procedure to identify $k$ is robust, it may be interesting to still minimize the Kullback-Leibler divergence for different $k$ and to write down which $s = s(k)$ minimizes the divergence. Indeed, as already mentioned, even if our procedure leads to the choice of a unique $k$, it seems natural that all this approach is not too sensitive to this choice. In other words, if $k$ varies slightly around its optimal value, we should not observe huge variations of $s$. This is why we add the plot $k \mapsto s(k)$. Idealistically, we should observe a constant value of $s$ around the optimal value of $k$.

We generate data sets of size $n \in \{4 \cdot 10^3, 7 \cdot 10^3, \cdot 10^4\}$ and apply Algorithm 4. We repeat this procedure over $N = 100$ simulations. The results corresponds to the average number of the two types of errors among the $N$ simulations. Recall that these two types are: the selection of a feature $\beta$ while the distribution of $\mathbf{Z}$ does not place mass in this direction (Type 1), and the absence of a feature $\beta$ while this direction should appear theoretically (Type 2). Regarding the dimension, we choose a larger $d$ than in Chapter 2. On the contrary, compared to the examples in Section 2.5, we chose here lower values of $n$. As we will see, even for a low number of data, we obtain very promising results.

**An independent case** We slightly change the first example of 2.5 and consider i.i.d. vectors $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^{100}$. The first $d_{\text{ext}} = 30$ marginals of the $\mathbf{X}_j$ follow a Pareto(1) distribution while the last ones follow an Exponential(1) distribution. The idea of this choice is to force some coordinates to not contribute to the global extremal behavior of $\mathbf{X}$. In terms of the spectral measure, is is straightforward to see that it places mass on the axes $\mathbf{e}_j$ for $j = 1, \dots, d_{\text{ext}}$. Then, we obtain that $\mathbb{P}(\mathbf{Z} \in C_\beta) > 0$ if and only if $C_\beta = C_{\{j\}} = \mathbf{e}_j$, for $j = 1, \dots, d_{\text{ext}}$.

Table 3.1 summarizes the different outcomes of our algorithm. The first two columns detail the average errors of both types. The third one gives the average value of $s$. Recall that the theoretical value $s^* = 30$. Finally, the last two columns provides the average values of the level $k$ and the associated threshold $u$.

There are only very few errors of Type 1, regardless of $n$. This means that our procedure does not overestimate the number of $\beta$'s. On the other hand, the errors of Type 2 are more represented, especially when $n$ is not very large. There, the number of relevant features is underestimated. However, this issue gradually disappears when $n$ increases. In particular, both average errors for $n = n_3 = 10^4$ are very close to 0. Regarding the level $k$, it increases as expected with $n$. Besides, we notice that the ratio $k/n$ becomes smaller when $n$ increases. This fits with the standard EVT (and hence with our framework) in which it is customary to assume that $k \to \infty$ and $k/n \to 0$ when $n \to \infty$.

| | Errors of Type 1 | Errors of Type 2 | Average value of $s$ | Average value of the level $k$ | Average value of the threshold $u$ |
|---|---|---|---|---|---|
| $n_1 = 4 \cdot 10^3$ | 0.13 | 7.12 | 22.95 | 257 | 890 |
| $n_2 = 7 \cdot 10^3$ | 0.16 | 1.15 | 28.94 | 343 | 905 |
| $n_3 = 10^4$ | 0.13 | 0.45 | 29.64 | 390 | 1058 |

Table 3.1: Average number of errors in an independent case ($d = 100$).

Following Remark 3.6.1, we illustrate on an example the choice of $k$ and $s$. We keep the same example with $n = n_3 = 10^4$. There, we plot in Figure 3.4 the variations of the quantity in Equation (3.5.20), that is, up to some constant, an estimator of the Kullback-Leibler divergence $KL(\mathbf{P}_n \| \mathbf{M}'(n, \tilde{\mathbf{p}}'))$. On this simulation, the minimum in reached for an optimal value of $k = 350$. We notice that the plot of this estimator is quite sharp.

We plot in Figure 3.5 the variations of the choice of $s$ regarding $k$. For a large range of $k$, the value of $s$ chosen by the algorithm remains constant. This means that a slight variation in the choice of $k$ does not affect the optimal value of $s$.

**A dependent case** We consider a random vector $\mathbf{X} \in \mathbb{R}^{100}$ whose marginals are defined as follows. The first $d_{\text{indep}} = 10$ are independent Pareto(1)-distributed. Then, we consider a bivariate dependence for the $d_{\text{dep1}} = 5$ couples of marginals $(X_j, X_{j+1})$, with $X_j$ following a Pareto(1) distribution and $X_{j+1} = X_j + E_j$ where $E_j$ follows an Exponential(1) distribution independent of $X_j$. In others words, we have a strong dependence between the extremes of $X_j$ and the ones of $X_{j+1}$.

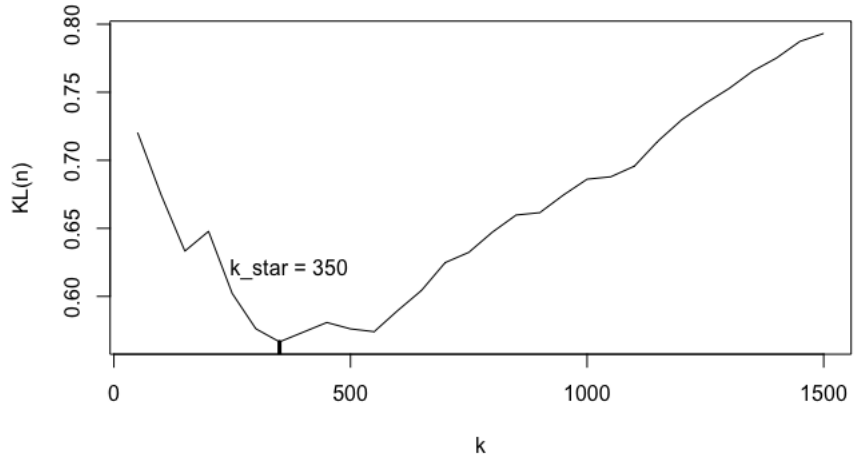



Figure 3.4: Evolution of the minimizer of $KL(n)$ in an independent case.

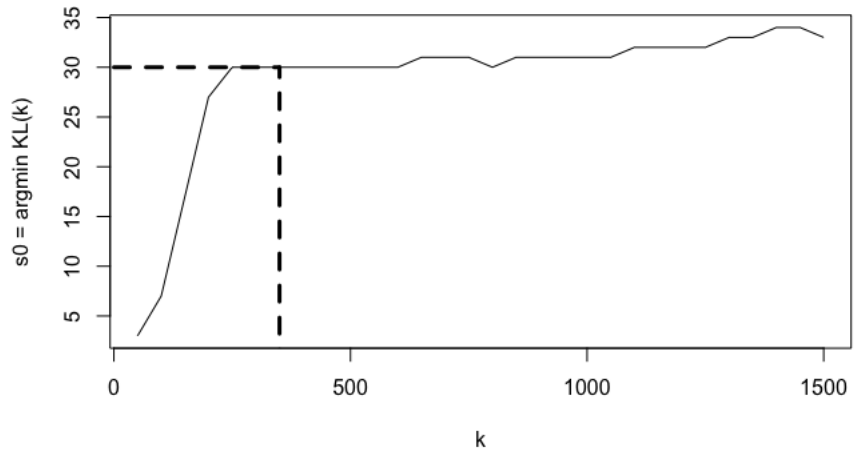



Figure 3.5: Evolution of the optimal value of $s$ in an independent case.

Similarly, we consider a three-dimensional dependence for the next $d_{\text{dep2}} = 5$ triplets of marginals. We consider a marginal $X_j$ with a Pareto(1) distribution, a marginal $X_{j+1} = X_j + E_j$ where $E_j$ follows a Exponential(1) distribution independent of $X_j$, and a marginal $X_{j+2} = X_j + E_{j+1}$ where $E_{j+1}$ follows a Exponential(1) distribution independent of $X_j$ and $E_j$. The last marginals are all independent and with an Exponential(1) distribution.

Let us summarize this example. The vector $\mathbf{X}$ satisfies

$$\begin{aligned} X_j &\sim \text{Pareto}(1)\,, \quad j = 1, \dots, 10\,, \\ (X_j, X_j + E_j) &\sim (\text{Pareto}(1), X_j + \text{Exp}(1))\,, \quad k = 11,\ 13,\ 15,\ 17,\ 19\,, \\ (X_j, X_j + E_j, X_j + E_{j+1}) &\sim (\text{Pareto}(1), X_j + \text{Exp}(1), X_j + \text{Exp}(1))\,, \quad j = 20,\ 23,\ 26,\ 29,\ 32\,, \\ X_j &\sim \text{Exp}(1)\,, \quad j = 36, \dots, 100\,. \end{aligned}$$

This implies that the spectral vector, and also the angular vector $\mathbf{Z}$, places mass on the following subsets:

$$\begin{aligned} C_{\{k\}} &= \mathbf{e}_k\,, \quad \text{for } k = 1, \dots, 10\,, \\ C_{\{k,k+1\}}\,, &\quad \text{for } k = 11, 13, 15, 17, 19\,, \\ C_{\{k,k+1,k+2\}}\,, &\quad \text{for } k = 20, 23, 26, 29, 32\,. \end{aligned}$$

Our goal is then to identify the $d_{\text{indep}} = 10$ one-dimensional subsets, the $d_{\text{dep1}} = 5$ two-dimensional subsets, and the $d_{\text{dep2}} = 5$ three-dimensional subsets. Thus in this example $s^* = 20$.

Table 3.2 summarizes the two types of errors averaged over the $N$ simulations, as well as the average number of relevant features $s$, the average level $k$ and the associated average threshold $u$. The errors of Type 2 decreases when $n$ increases, which makes sense: With only few data, our procedure fails to identify all relevant directions, but this issue vanishes when $n$ becomes large. For the errors of Type 1, it seems that their number slightly increases with $n$. If $n$ is large, it is possible to capture a direction that should not be taken into account. However, the average error is negligible regarding the total number of possible directions, that is, $2^d - 1 \sim 10^{30}$. In this example, we also observe that the chosen $k$ increases with $n$, while the ratio $k/n$ tends to decrease.

| | Errors of Type 1 | Errors of Type 2 | Average value of $s$ | Average value of the level $k$ | Average value of the threshold $u$ |
|---|---|---|---|---|---|
| $n_1 = 4 \cdot 10^3$ | 0.05 | 5.29 | 14.75 | 170 | 1368 |
| $n_2 = 7 \cdot 10^3$ | 0.09 | 1.66 | 18.42 | 262 | 1313 |
| $n_3 = 10^4$ | 0.25 | 0.82 | 19.42 | 304 | 1504 |

Table 3.2: Average number of errors in a dependent case ($d = 100$).

As for the independent case, and following Remark 3.6.1, we illustrate on an example the choice of $k$ and $s$ for this dependent case with $n = n_3 = 10^4$. There, we plot in Figure 3.6 the variations of the quantity in Equation (3.5.20), that is, up to some constant, an estimator of the Kullback-Leibler divergence $KL(\mathbf{P}_n \| \mathbf{M}'(n, \tilde{\mathbf{p}}'))$. This simulation leads to a choice of $k = 250$ and provides a very sharp graph. Figure 3.7 shows that the optimal value of $s$ remains constant around $k = 250$. As for the previous case, we conclude that a slight variation of $k$ does not impact the choice of $s$.

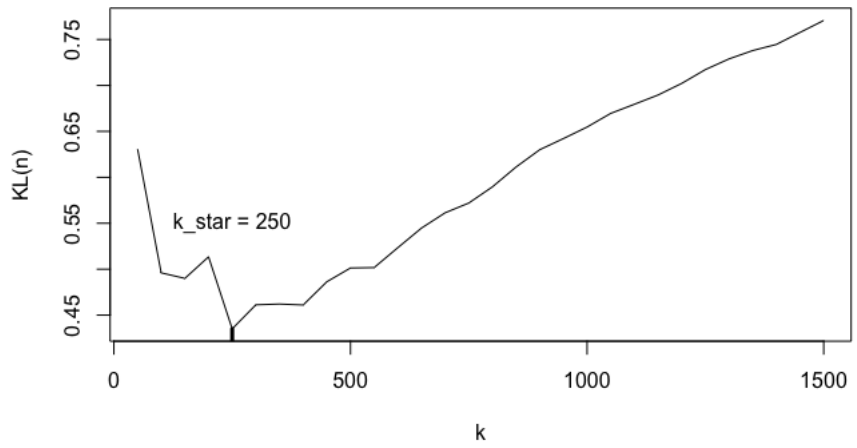



Figure 3.6: Evolution of the minimizer of $KL(n)$ in a dependent case.

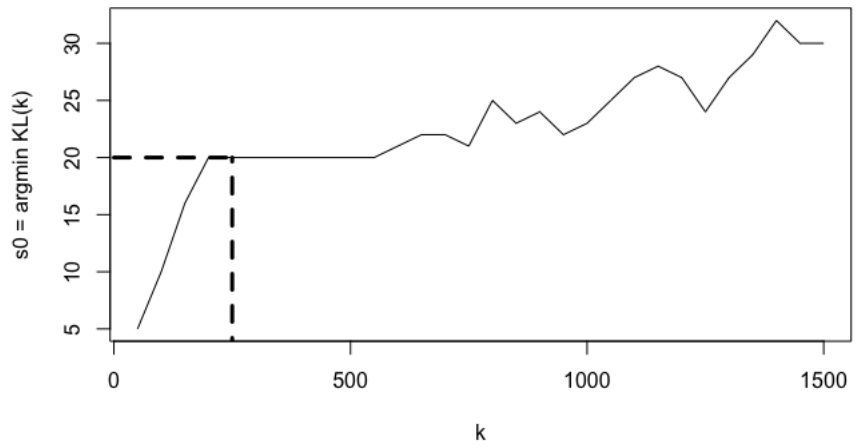



Figure 3.7: Evolution of the optimal value of $s$ in a dependent case.

## 3.7 Conclusion

This chapter develops a statistical framework based on the theoretical results obtained in Chapter 2. We provide a procedure whose main goals are to identify the features which gather extreme events and to exhibit a optimal level which corresponds to the number of extreme values among the observed sample. The different asymptotic results given in Proposition 3.3.1, Theorem 3.3.1 and Theorem 3.4.2 entail that the proposed estimators are accurate to deal with our problem.

The model selection proposed in Section 3.5 manages to tackle simultaneously two aforementioned major concerns in multivariate EVT. Regarding the dependence structure, our method highlights it with the identification of the most relevant subsets $C_\beta$. This selection is done with an AIC-type minimization whose penalization allows to reduce the number of selected subsets. Regarding the choice of an appropriate level $k$, which has been the subject of much attention in the literature, no theoretical-based procedure has been provided yet. Following the philosophical mantra of EVT, "let the tail speak for itself", we provide an ad hoc estimation procedure in which the data partitions itself into two categories, an extreme one and non-extreme one.

Regarding the numerical simulations, our algorithm provides promising results for independent cases but also for dependent cases. In particular, we manage to deal with high-dimensional data, at least compared to the examples given in the EVT literature so far. Besides, the large number of possible subsets $C_\beta$, that is, $2^{30}-1 \sim 10^{30}$, does not lead to any computational issue. The outcomes we obtain are all the more remarkable since we do not use large data sets, and even for moderate values of $n$ the empirical subsets which appear are close to the theoretical ones. For the choice of $k$, the main asset of our algorithm is the global consideration of the level and the subsets $C_\beta$. This is supported by the Euclidean projection which provides a balance choice between a interpretable sparsity and a sufficiently large threshold.

All in all, the notion of sparse regular variation introduced in Chapter 2 appears to have interesting statistical properties. It manages to tackles simultaneously the questions **(Q1)**, **(Q2)**, and **(Q3)** given in Chapter 1. This is why further researches should be conducted in this direction. The use of Theorem 3.4.1 may be a point to tackle.

As expected, the proposed method provides good results when the dimension $d$ is large. Paradoxically, it does not seem to be effective for small values of $d$. Indeed, some numerical results not exposed here show that in this case our algorithm does not manage to capture the theoretical subsets $C_\beta$ very well. A high variability appears in the choice of $s$ when $k$ varies and the procedure often chooses the largest $k$ as the optimal one. This issue arises mainly since there is no guarantee that the tail dependence is sparse if $d$ is moderate (say $d \sim 10$). In other words, it becomes likely that all directions are simultaneously large and hence that the most relevant subset (and maybe the only one) is $C_{\{1,\dots,d\}}$. In this case, other methods should be applied. This is the purpose of Chapter 4.

## 3.8 Proofs

*Proof of Proposition 3.3.1.* We use the Weak Law of Large Number for triangular array (see for instance Feller (1971)). For a Borel set $A \subset \mathbb{S}_+^{d-1}$ we set $Y_{j,n} = k_n^{-1}\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A, |\mathbf{X}_j| > u_n\}$. Then, in order to obtain the convergence in probability

$$\sum_{j=1}^{n}(Y_{j,n} - \mathbb{E}[Y_{j,n}]) \to 0\,, \quad n \to \infty\,,$$

it suffices to show that $\sup_{n,j} E[Y_{j,n}^2] < \infty$. Starting from the relation

$$E[Y_{j,n}^2] = \frac{\mathbb{P}(\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n)}{k_n^2} = \frac{p_n(A)}{nk_n}\,,$$

we obtain that $\sup_{n,j} E[Y_{j,n}^2] \leq k_n^{-1}$. Since $k_n \to \infty$, this implies the convergence in probability

$$\frac{T_n(A)}{k_n} - p_n(A) \to 0\,, \quad n \to \infty\,. \tag{3.8.1}$$

Finally, the convergence in Equation (3.3.6) is just a consequence of Equations (3.3.5) and (3.2.7). □

*Proof of Theorem 3.3.1.* For a Borel set $A$ of $\mathbb{S}_+^{d-1}$, we define

$$V_{j,n} = \sigma_n^{-1}\Big(\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\} - \frac{k_n}{n} p_n(A)\Big)\,,$$

where

$$\sigma_n^2 = n\mathbb{P}(\pi(\mathbf{X}/u_n) \in A,\, |\mathbf{X}| > u_n)[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in A,\, |\mathbf{X}| > u_n)]\,.$$

The variable $V_{j,n}$ satisfies the relations $\mathbb{E}[V_{j,n}] = 0$ and $\mathrm{Var}(V_{j,n}) = 1/n$. Then, in order to prove the convergence

$$\frac{\sum_{j=1}^n (V_{j,n} - \mathbb{E}[V_{j,n}])}{\sqrt{\sum_{j=1}^n \mathrm{Var}(V_{j,n})}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,,$$

it suffices to show that Lindeberg's condition holds:

$$\sum_{j=1}^n \mathbb{E}\Big[V_{j,n}^2 \mathbb{1}_{\{V_{j,n} > \epsilon\}}\Big] = n\mathbb{E}\Big[V_{1,n}^2 \mathbb{1}_{\{V_{1,n} > \epsilon\}}\Big] \to 0\,, \quad n \to \infty\,, \tag{3.8.2}$$

for all $\epsilon > 0$. Thus, fix $\epsilon > 0$. On the one hand, the variance $\sigma_n^2$ is equivalent to $(k_n p_n(A))^2$ which converges to $\infty$ by assumption. On the other hand, $|\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in A,\, |\mathbf{X}_j| > u_n\} - \frac{k_n}{n} p_n(A)|$ is always bounded by 1. Hence, for $n$ large enough, the inequality $V_{1,n} \leq \epsilon$ is always satisfied. This proves that the condition in (3.8.2) holds and then implies that

$$\frac{\sum_{j=1}^n (V_{j,n} - \mathbb{E}[V_{j,n}])}{\sqrt{\sum_{j=1}^n \mathrm{Var}(V_{j,n})}} = \frac{T_n(A) - k_n p_n(A)}{\sigma_n} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,.$$

Finally, Slutsky's theorem allows to replace $\sigma_n$ by $\sqrt{k_n p_n(A)}$, which yields to the following convergence

$$\sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} \xrightarrow{d} \mathcal{N}(0,1)\,, \quad n \to \infty\,.$$

For the convergence (3.3.8), the regularity assumption implies that $p_n(A) \to p(A) > 0$ when $n \to \infty$. Therefore, an application of Slutsky's theorem allows to conclude.

In order to prove (3.3.10), we decompose the previous ratio in the following way:

$$\sqrt{k_n}\frac{T_n(A)/k_n - p(A)}{\sqrt{p_n(A)}} = \sqrt{k_n}\frac{T_n(A)/k_n - p_n(A)}{\sqrt{p_n(A)}} + \sqrt{k_n}\frac{p_n(A) - p(A)}{\sqrt{p_n(A)}}\,.$$

It is then sufficient to show that the second term goes to 0 as $n \to \infty$. This is true thanks to the bias assumption (3.3.9) and since the denominator $\sqrt{p_n(A)}$ converges to a positive limit.

□

*Proof of Theorem 3.4.1.* We consider $x > 0$ and $\lambda > 0$ and define, for $\beta \in \mathcal{B}$,

$$V_n(\beta) = \frac{T_n(\beta)}{k_n} - p_n(\beta)\,.$$

Chernoff's inequality entails

$$\mathbb{P}\Big(\sup_{\beta\in\mathcal{B}} V_n(\beta) > x\Big) \le e^{-\lambda x}\mathbb{E}\big[e^{\lambda \sup_{\beta\in\mathcal{B}} V_n(\beta)}\big]\,.$$

Our aim is to bound the quantity $\mathbb{E}[e^{\lambda \sup_\beta V_n(\beta)}]$. To this end, we bound the supremum on $\beta$ by a sum:

$$\mathbb{E}\Big[e^{\lambda \sup_{\beta\in\mathcal{B}} V_n(\beta)}\Big] = \mathbb{E}\Big[\sup_{\beta\in\mathcal{B}} e^{\lambda V_n(\beta)}\Big] \le \mathbb{E}\Big[\sum_{\beta\in\mathcal{B}} e^{\lambda V_n(\beta)}\Big] = \sum_{\beta\in\mathcal{B}} \mathbb{E}\big[e^{\lambda V_n(\beta)}\big]\,.$$

Since $V_n(\beta)$ can be rewritten as

$$V_n(\beta) = \frac{1}{k_n}\sum_{j=1}^{n}\Big[\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\} - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n)\Big]\,,$$

and since the $\mathbf{X}_j$'s are i.i.d., we obtain that $\mathbb{E}\big[e^{\lambda V_n(\beta)}\big]$ can be expressed as

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{\lambda}{k_n}\sum_{j=1}^{n}\Big[\mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\} - \mathbb{P}(\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n)\Big]\Big)\Big] \\
&= \mathbb{E}\Big[\exp\Big(\frac{\lambda}{k_n}\Big[\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)\Big]\Big)\Big]^n\,.
\end{aligned}$$

The goal is now to bound this last expectation. We use Bernstein's inequality (see Lemma A.3.1) for $\lambda < k_n$:

$$\begin{aligned}
&\mathbb{E}\Big[\exp\Big(\frac{\lambda}{k_n}\Big[\mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)\Big]\Big)\Big]^n \\
&\le \exp\left[n\frac{\lambda^2\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)\big[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)\big]}{2k_n^2\Big(1 - \frac{\lambda}{3k_n}\Big)}\right] \\
&\le \exp\left[k_n\frac{\lambda^2 p_n(\beta)}{2k_n^2\big(1 - \frac{\lambda}{3k_n}\big)}\right]
\end{aligned}$$

$$= \exp\left(\frac{\lambda^2 p_n(\beta)}{2(k_n - \lambda/3)}\right).$$

Gathering all inequalities together, we obtain that

$$\mathbb{P}\Big(\max_{\beta\in\mathcal{B}} V_n(\beta) > x\Big) \leq e^{-\lambda x} \sum_{\beta\in\mathcal{B}} \exp\left(\frac{\lambda^2 p_n(\beta)}{2(k_n - \lambda/3)}\right). \tag{3.8.3}$$

After replacing $V_n(\beta)$ by $-V_n(\beta)$, a similar inequality can be obtained:

$$\mathbb{P}\Big(\max_{\beta\in\mathcal{B}} -V_n(\beta) > x\Big) \leq e^{-\lambda x} \sum_{\beta\in\mathcal{B}} \exp\left(\frac{\lambda^2 p_n(\beta)}{2(k_n - \lambda/3)}\right). \tag{3.8.4}$$

Then, we decompose the probability $\mathbb{P}(\max_{\beta\in\mathcal{B}} |V_n(\beta)| > x)$ as follows

$$\begin{aligned}\mathbb{P}\Big(\max_{\beta\in\mathcal{R}} |V_n(\beta)| > x\Big) &= \mathbb{P}\Big(\max_{\beta\in\mathcal{B}} V_n(\beta) > x \text{ or } \max_{\beta\in\mathcal{R}} -V_n(\beta) > x\Big) \\ &\leq \mathbb{P}\Big(\max_{\beta\in\mathcal{B}} V_n(\beta) > x\Big) + \mathbb{P}\Big(\max_{\beta\in\mathcal{B}} -V_n(\beta) > x\Big),\end{aligned}$$

and Equations (3.8.3) and (3.8.4) entail that

$$\mathbb{P}\Big(\max_{\beta\in\mathcal{R}} |V_n(\beta)| > x\Big) \leq 2e^{-\lambda x} \max_{\beta\in\mathcal{R}} \exp\left(\frac{\lambda^2 p_n(\beta)}{2(k_n - \lambda/3)}\right) \leq 2\exp\left(\log(b) - \lambda x + \frac{\lambda^2 c_n(\mathcal{B})}{2(k_n - \lambda/3)}\right), \tag{3.8.5}$$

where $b$ denotes the cardinality of $\mathcal{B} \subset \mathcal{P}_d^*$ and $c_n(\mathcal{B}) = \max_{\beta\in\mathcal{B}} p_n(\beta)$.

The last step is to optimize the previous quantity with respect to $\lambda$. We set $\varphi(\lambda) = -\lambda x + \lambda^2 c_n(\mathcal{B})/(k_n - \lambda)$ and we study this function for $\lambda \in (0, 3k_n)$. The derivative of $\varphi$ satisfies

$$\varphi'(\lambda) = -x + c_n(\mathcal{B}) \frac{(k_n - \lambda/3)\lambda + \lambda^2/6}{(k_n - \lambda/3)^2}.$$

The equation $\varphi'(\lambda) = 0$ has two solutions:

$$\lambda_\pm = 3k_n \pm \frac{3k_n}{\sqrt{\frac{2x}{3c_n(\mathcal{B})} + 1}},$$

and only $\lambda_-$ belongs to the interval $(0, 3k_n)$. Thus, the minimum of the function $\varphi$ in $(0, 3k_n)$ is

$$\varphi(\lambda_-) = -\frac{9}{2} k_n c_n(\mathcal{B}) \left(\sqrt{\frac{2x}{3c_n(\mathcal{B})} + 1} - 1\right)^2.$$

Hence, $\delta$ satisfies

$$\delta = \frac{9}{2} k_n c_n(\mathcal{B}) \left( \sqrt{\frac{2x}{3c_n(\mathcal{B})} + 1} - 1 \right)^2 - \log(b)\,,$$

if and only if $x$ satisfies

$$x = \frac{3c_n(\mathcal{B})}{2} \left[ \left( \frac{\sqrt{2}}{3} \sqrt{\frac{\log(b)+\delta}{k_n c_n(\mathcal{B})}} + 1 \right)^2 - 1 \right] = \sqrt{2}\sqrt{c_n(\mathcal{B})} \sqrt{\frac{\log(r)+\delta}{k_n}} + \frac{\log(b)+\delta}{3k_n} =: f(\delta, n, \mathcal{B})\,.$$

This leads to the desired inequality: for all $\delta > 0$,

$$\mathbb{P}\left( \max_{\beta \in \mathcal{B}} \left| \frac{T_n(C_\beta)}{k_n} - p_n(C_\beta) \right| > f(\delta, n, \mathcal{B}) \right) \leq 2e^{-\delta}\,.$$

□

*Proof of Lemma 3.4.1.* Recall that $T_n(\beta) = \sum_{j=1}^n \mathbb{1}\{\pi(\mathbf{X}_j/u_n) \in C_\beta,\, |\mathbf{X}_j| > u_n\}$ where the $\mathbf{X}_j$'s are i.i.d. This implies that

$$\begin{aligned}
\mathbb{P}(T_n(\beta) = 0) &= \mathbb{P}(\forall j = 1, \dots, n,\, \pi(\mathbf{X}_j/u_n) \notin C_\beta \text{ or } |\mathbf{X}_j| \leq u_n) \\
&= [1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| \leq u_n)]^n \\
&= \exp\left( n \log[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)] \right).
\end{aligned}$$

Hence, since $\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n) \to 0$, we obtain the Taylor expansion

$$\log[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)] \sim -\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n)\,, \quad n \to \infty\,.$$

Finally, we write $n\mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta\,, |\mathbf{X}| > u_n) = k_n \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta \mid |\mathbf{X}| > u_n)$ which gives the desired result. □

*Proof of Theorem 3.4.2.* We consider the vector $\mathbf{V}_{n,\mathcal{R}_k(\mathbf{z})} \in \mathbb{R}^{r^*}$ whose components are

$$V_{n,\beta} = \frac{1}{\sqrt{k_n p_n(\beta)}} \left( \mathbb{1}\{\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n\} - \frac{k_n}{n} p_n(\beta) \right).$$

This vector has null expectation. We denote by $\Sigma_n \in \mathcal{M}_{r^*}(\mathbb{R})$ its covariance matrix. First, the diagonal entries correspond to the variance of a Bernoulli distribution, i.e.

$$\Sigma_n(\beta,\beta) = \frac{1}{k_n p_n(\beta)} \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)[1 - \mathbb{P}(\pi(\mathbf{X}/u_n) \in C_\beta,\, |\mathbf{X}| > u_n)] = \frac{1}{n} - \frac{k_n}{n^2} p_n(\beta)\,.$$

Second, the non-diagonal entries can be computed as follows:

$$\begin{aligned}\Sigma_n(\beta,\beta') &= \mathbb{E}\big[V_{n,\beta}V_{n,\beta'}\big]\\ &= \frac{1}{\sqrt{k_n p_n(\beta)}}\frac{1}{\sqrt{k_n p_n(\beta')}}\Big(\mathbb{E}\big[\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_\beta,\,|\mathbf{X}|>u_n\}\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_{\beta'},\,|\mathbf{X}|>u_n\}\big]\\ &\quad - \frac{k_n}{n}p_n(\beta)\mathbb{E}\big[\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_{\beta'},\,|\mathbf{X}|>u_n\}\big] - \frac{k_n}{n}p_n(\beta')\mathbb{E}\big[\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_{\beta},\,|\mathbf{X}|>u_n\}\big]\\ &\qquad\qquad + \frac{k_n^2}{n^2}p_n(\beta)p_n(\beta')\Big)\\ &= -\frac{1}{k_n\sqrt{p_n(\beta)p_n(\beta')}}\frac{k_n^2}{n^2}p_n(\beta)p_n(\beta')\\ &= -\frac{k_n}{n^2}\sqrt{p_n(\beta)p_n(\beta')}\,.\end{aligned}$$

Hence, the covariance matrix $\Sigma_n$ can be written as

$$\Sigma_n = \frac{1}{n}Id_{r^*} - \frac{k_n}{n^2}\sqrt{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}}\cdot\sqrt{\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}}^{\top},$$

where the square root is meant componentwise. In particular, $n\Sigma_n \to Id_{r^*}$ when $n\to\infty$.

Consider now a triangular array $\mathbf{V}_{n,1},\dots,\mathbf{V}_{n,n}$ with the same distribution as $\mathbf{V}_{n,\mathcal{R}_k(\mathbf{Z})}$. We prove that this triangular array satisfies Lindeberg's condition:

$$\sum_{j=1}^{n}\mathbb{E}\left[\frac{1}{k_n}\max_{\beta}\frac{1}{p_n(\beta)}\left|\mathbb{1}\{\pi(\mathbf{X}_j/u_n)\in C_\beta,\,|\mathbf{X}_j|>u_n\} - \frac{k_n}{n}p_n(\beta)\right|^2\mathbb{1}_{\{\max_\beta|V_{n,j,\beta}|>\epsilon\}}\right]\to 0\,,\quad n\to\infty\,,$$

for all $\epsilon>0$, or equivalently that

$$\mathbb{E}\left[\frac{n}{k_n}\max_{\beta}\frac{1}{p_n(\beta)}\left|\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_\beta,\,|\mathbf{X}|>u_n\} - \frac{k_n}{n}p_n(\beta)\right|^2\mathbb{1}_{\{\max_\beta|V_{n,\beta}|>\epsilon\}}\right]\to 0\,,\quad n\to\infty\,. \tag{3.8.6}$$

Fix $\epsilon>0$. Recall that $\mathcal{R}_k(\mathbf{Z})$ gathers all features $\beta$ such that $k_n p_n(\beta)\to\infty$. Thus, there exists $n_0$ such that for all $n\geq n_0$,

$$\max_{\beta\in\mathcal{R}_k(\mathbf{Z})}\left|\mathbb{1}\{\pi(\mathbf{X}/u_n)\in C_\beta,\,|\mathbf{X}|>u_n\} - \frac{k_n}{n}p_n(\beta)\right| \leq \epsilon k_n \min_{\beta\in\mathcal{R}_k(\mathbf{Z})}p_n(\beta)\,,$$

since the term on the left-hand side is always bounded by 1. This implies that for $n$ large enough, the inequality $\max_\beta|V_{n,\beta}|>\epsilon$ is never satisfied. Hence, Lindeberg's condition in (3.8.6) holds and yields to the following convergence

$$\sum_{j=1}^{n}\mathbf{V}_{n,j}\xrightarrow{d}\mathcal{N}(0,Id_{r^*})\,,\quad n\to\infty\,.$$

This convergence can be rephrased as

$$\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})})^{-1/2}\left(\frac{\mathbf{T}_{n,\mathcal{R}_k(\mathbf{Z})}}{k_n}-\mathbf{p}_{n,\mathcal{R}_k(\mathbf{Z})}\right)\xrightarrow{d}\mathcal{N}(0,Id_{r^*})\,,\quad n\to\infty\,,$$

which proves (3.4.12).

To obtain the convergence in (3.4.13), it suffices to restrict the previous convergence to the coordinates $\beta\in\mathcal{S}(\mathbf{Z})$ and to notice that

$$\mathrm{Diag}(\mathbf{p}_{n,\mathcal{S}(\mathbf{Z})})^{1/2}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}\to Id_{s^*}\,,\quad n\to\infty\,.$$

Finally, to prove (3.4.15) it suffices to show that $\sqrt{k_n}\,\mathrm{Diag}(\mathbf{p}_{\mathcal{S}(\mathbf{Z})})^{-1/2}(\mathbf{p}_{n,\mathcal{S}(\mathbf{Z})}-\mathbf{p}_{\mathcal{S}(\mathbf{Z})})\to 0$ which is true under assumption (3.4.14). □

*Proof of Lemma 3.5.2.* Let $f$ be the function defined as $f(t)=h(t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*)$ for $t\in[0,1]$, where $h$ is defined as

$$h(\tilde{\mathbf{p}})=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)+\frac{\partial}{\partial\tilde{\mathbf{p}}}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}})\,.$$

Some short calculations give the following relations:

$$\begin{aligned}
f(1)&=h(\widehat{\tilde{\mathbf{p}}})=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}\,,\\
f(0)&=h(\tilde{\mathbf{p}}^*)=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}+\frac{\partial}{\partial\tilde{\mathbf{p}}}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}-\underbrace{\frac{\partial}{\partial\tilde{\mathbf{p}}}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}}_{=0\text{ by definition of }\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}\,,\\
f'(t)&=\frac{\partial h}{\partial\tilde{\mathbf{p}}}(t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=(\widehat{\tilde{\mathbf{p}}}-[t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*])^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\\
&=(1-t)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\Big]\Big|_{t\widehat{\tilde{\mathbf{p}}}+(1-t)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,.
\end{aligned}$$

We apply Lemma 3.5.1 to the function $f$ and $g:t\mapsto(t-1)^2$. There exists $c_1\in(0,1)$ such that $(f(1)-f(0))g'(c_1)=(g(1)-g(0))f'(c_1)$, i.e.

$$\begin{aligned}
&\left(KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\widehat{\tilde{\mathbf{p}}}}-KL\Big(\mathbf{P}_k\Big\|\mathbf{M}(k;\,\tilde{\mathbf{p}})\Big)\Big|_{\tilde{\mathbf{p}}=\tilde{\mathbf{p}}^*}\right)2(c_1-1)\\
&\qquad=(1-c_1)(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)^\top\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\Big[\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\mathbf{T}_n)\Big]\Big|_{c_1\widehat{\tilde{\mathbf{p}}}+(1-c_1)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}}-\tilde{\mathbf{p}}^*)\,.
\end{aligned}$$

Simplifying by $2(1-c_1)\neq 0$ gives the desired result. □

*Proof of Lemma 3.5.3.* Consider $f(t) = h(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}})$, for $t \in [0,1]$ where $h$ is defined as

$$h(\tilde{\mathbf{p}}) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n) + \frac{\partial}{\partial \tilde{\mathbf{p}}} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \tilde{\mathbf{p}})\,.$$

Some short calculations give the following relations:

$$\begin{aligned}
f(1) &= h(\tilde{\mathbf{p}}^*) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*; \mathbf{T}_n)\,,\\
f(0) &= h(\widehat{\tilde{\mathbf{p}}}) = \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n) + \underbrace{\frac{\partial}{\partial \tilde{\mathbf{p}}} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)}_{=0 \text{ by definition of } \widehat{\tilde{\mathbf{p}}}}(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,,\\
f'(t) &= \frac{\partial h}{\partial \tilde{\mathbf{p}}}(t\tilde{\mathbf{p}}^* + (1-t)t\widehat{\tilde{\mathbf{p}}})(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\\
&= (\tilde{\mathbf{p}}^* - [t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}])^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\\
&= (1-t)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(t\tilde{\mathbf{p}}^* + (1-t)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,.
\end{aligned}$$

We apply Lemma 3.5.1 to the function $f$ and $g : t \mapsto (t-1)^2$. There exists $c_2 \in (0,1)$ such that $f(1) - f(0))g'(c_2) = (g(1) - g(0))f'(c_2)$, i.e.

$$\begin{aligned}
\Big(\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}}^*; \mathbf{T}_n) - \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)\Big) 2(c_2 - 1)&\\
= -(1-c_2)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})^\top \frac{\partial^2}{\partial \tilde{\mathbf{p}}^2} \log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(c_2\tilde{\mathbf{p}}^* + (1-c_2)\widehat{\tilde{\mathbf{p}}}; \mathbf{T}_n)(\tilde{\mathbf{p}}^* - \widehat{\tilde{\mathbf{p}}})\,.&
\end{aligned}$$

By simplifying by $2(c_2 - 1) \neq 0$ leads to the desired result. □

For the two following Lemmas, we define the functions $\psi_j$ and $\psi$ as follows:

$$\psi_j(c) = c\widehat{\tilde{p}}_j + (1-c)\tilde{p}_j^* = c\frac{T_{n,j}}{k} + (1-c)p_{n,j}\,, \quad j = 1, \dots, s\,,$$

and

$$\psi(c) = c\widehat{\tilde{p}} + (1-c)\tilde{p}^* = \frac{1}{r-s}\sum_{j=s+1}^{r} \psi_j(c)\,.$$

Remark 3.3.1 yields to the following convergence in probability:

$$\frac{\psi_j(c)}{p_{n,j}} \to 1\,, \quad n \to \infty\,. \tag{3.8.7}$$

Besides, the functions $\psi_j$ and $\psi$ satisfy the relations

$$\begin{aligned}
\inf_{c\in(0,1)} \psi_j(c) = \widehat{\tilde{p}}_j \wedge \tilde{p}_j^* = m_j \quad \text{and} \quad \sup_{c\in(0,1)} \psi_j(c) = \widehat{\tilde{p}}_j \vee \tilde{p}_j^* = M_j\,, \quad j = 1, \dots, s\\
\inf_{c\in(0,1)} \psi(c) = \widehat{\tilde{p}} \wedge \tilde{p}^* =: m \quad \text{and} \quad \sup_{c\in(0,1)} \psi(c) = \widehat{\tilde{p}} \vee \tilde{p}^* =: M\,.
\end{aligned}$$

*Proof of Lemma 3.5.4.* We differentiate twice the expression in Equation (3.5.8) with respect to the vector $\tilde{\mathbf{p}}$. This leads to the following Hessian matrix:

$$-\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n) = \begin{pmatrix} \frac{T_{n,1}}{\tilde{p}_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{T_{n,2}}{\tilde{p}_2^2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \frac{\sum_{j=s+1}^r T_{n,j}}{\tilde{p}^2} \end{pmatrix}.$$

Then, our goal is to prove that

$$\forall j = 1,\dots,s\,, \qquad \sup_{(c,c')\in(0,1)^2}\left|\frac{T_{n,j}}{k\psi_j(c)^2} - \frac{p_{n,j}}{\psi_j(c')^2}\right| \to 0\,, \tag{3.8.8}$$

$$\text{and} \qquad \sup_{(c,c')\in(0,1)^2}\left|\frac{\sum_{j=s+1}^r T_{n,j}}{(r-s)k\psi(c)^2} - \frac{\sum_{j=s+1}^r p_{n,j}}{(r-s)\psi(c')^2}\right| \to 0\,. \tag{3.8.9}$$

Regarding (3.8.8), we write

$$\begin{aligned}\left|\frac{T_{n,j}}{k\psi_j(c)^2} - \frac{p_{n,j}}{\psi_j(c')^2}\right| &\le \frac{1}{\psi_j(c)^2}\left|\frac{T_{n,j}}{k} - p_{n,j}\right| + p_{n,j}\left|\frac{1}{\psi_j(c)^2} - \frac{1}{\psi_j(c')^2}\right| \\ &\le \frac{1}{m_j^2}\left|\frac{T_{n,j}}{k} - p_{n,j}\right| + \frac{p_{n,j}}{\psi_j(c)^2\psi_j(c')^2}\left|\psi_j(c')^2 - \psi_j(c)^2\right| \\ &\le \frac{1}{m_j^2}\left|\frac{T_{n,j}}{k} - p_{n,j}\right| + \frac{p_{n,j}}{m_j^4}\left|M_j^2 - m_j^2\right|,\end{aligned}$$

and thus we obtain that

$$\sup_{(c,c')\in(0,1)^2}\left|\frac{T_{n,j}}{k\psi_j(c)^2} - \frac{p_{n,j}}{\psi_j(c')^2}\right| \le \frac{1}{m_j^2}\left|\frac{T_{n,j}}{k} - p_{n,j}\right| + \frac{p_{n,j}}{m_j^4}\left|M_j^2 - m_j^2\right| \to 0\,, \quad n\to\infty\,,$$

where the convergence of both terms results from Assumption 3.5.1.

We move on to the term (3.8.9). For all $c\in(0,1)$ we have the following inequalities

$$\begin{aligned}\left|\frac{\sum_{j=s+1}^r T_{n,j}}{k(r-s)\psi(c)^2} - \frac{\sum_{j=s+1}^r p_{n,j}}{(r-s)\psi(c')^2}\right| &\le \frac{1}{\psi(c)^2}\left|\frac{\sum_{j=s+1}^r T_{n,j}}{k} - \sum_{j=s+1}^r p_{n,j}\right| + \frac{\sum_{j=s+1}^r p_{n,j}}{\psi(c)^2\psi(c')^2}\left|\psi(c')^2 - \psi(c)^2\right| \\ &\le \frac{1}{(\sum_{j=s+1}^r m_j)^2}\sum_{j=s+1}^r\left|\frac{T_{n,j}}{k} - p_{n,j}\right| + \frac{\sum_{j=s+1}^r p_{n,j}}{(\sum_{j=s+1}^r m_j)^4}\left|\Big(\sum_{j=s+1}^r M_j\Big)^2 - \Big(\sum_{j=s+1}^r m_j\Big)^2\right|,\end{aligned}$$

which converges to zero thanks to Assumption 3.5.1. □

*Proof of Lemma 3.5.5.* We start with Equation (3.5.8) and take the expectation of both sides:

$$\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)] = -\log(k!) + \sum_{j=1}^{2^d-1}\mathbb{E}[\log(T_j!)] - \sum_{j=1}^{s} kp_{n,j}\log(\tilde{p}_j) - \Big(\sum_{j=s+1}^{r} kp_{n,j}\Big)\log(\tilde{p})\,.$$

Then, differentiating twice this expression with respect to the vector $\tilde{\mathbf{p}}$ leads to the following Hessian matrix:

$$\frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)] = \begin{pmatrix} \frac{kp_{n,1}}{\tilde{p}_1^2} & 0 & 0 & \dots & 0 \\ 0 & \frac{kp_{n,2}}{\tilde{p}_2^2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \frac{\sum_{j=s+1}^{r} kp_{n,j}}{\tilde{p}^2} \end{pmatrix}.$$

Then, for $c \in (0,1)$, we write

$$\begin{aligned}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\big]\big|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)& \\ = \sum_{j=1}^{s}\frac{k(\widehat{\tilde{p}}_j - \tilde{p}_j^*)^2 p_{n,j}}{\psi_j(c)^2} + \frac{\sum_{j=s+1}^{r} k(\widehat{\tilde{p}} - \tilde{p}^*)^2 p_{n,j}}{\psi(c)^2} & \qquad (3.8.10)\\ = \sum_{j=1}^{s}\frac{k(T_{n,j}/k - p_{n,j})^2}{p_{n,j}}\frac{p_{n,j}^2}{\psi_j(c)^2} + k\frac{(\sum_{j=s+1}^{r} T_{n,j}/k - p_{n,j})^2}{\sum_{j=s+1}^{r} p_{n,j}}\frac{\sum_{j=s+1}^{r} p_{n,j}}{(r-s)^2\psi(c)^2}\,. & \qquad (3.8.11)\end{aligned}$$

Following Equation (3.8.7), we know that $\psi_j(c)/p_{n,j}$ and $\sum_{j=s+1}^{r} p_{n,j}/[(r-s)\psi(c)]$ converge to 1 when $n \to \infty$, and thus Equation (3.4.18) and Slutsky's theorem yield to the following convergence:

$$(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*)^\top \frac{\partial^2}{\partial\tilde{\mathbf{p}}^2}\mathbb{E}\big[-\log L_{\mathbf{M}(k;\tilde{\mathbf{p}})}(\tilde{\mathbf{p}};\mathbf{T}_n)\big]\big|_{c\widehat{\tilde{\mathbf{p}}}+(1-c)\tilde{\mathbf{p}}^*}(\widehat{\tilde{\mathbf{p}}} - \tilde{\mathbf{p}}^*) \xrightarrow{d} \chi^2(s+1)\,, \quad n \to \infty\,.$$

□

# Chapter 4

# Regular variation and conditional independence

**Abstract**

The multivariate Pareto distribution $\mathbf{Y}$ defined in terms of threshold exceedances (Rootzén and Tajvidi (2006)) summarizes the tail behavior of a regularly varying random vector. Since this vector does not take values in a product space, there is no natural way to introduce a concept of independence for its marginals. In a recent paper, Engelke and Hitz (2020) introduce an approach to define conditional independence for a multivariate Pareto distribution by restricting the support of $\mathbf{Y}$. This chapter consists in a discussion of this article. We analyze their different assumptions which lead us to develop another approach based on the minimum of the marginals of a regularly varying random vector. In this context, we establish some interpretable results for conditional independence regarding extreme values and compare our approach with the one of Engelke and Hitz (2020).

***Keywords***— multivariate extremes, multivariate Pareto distribution, regular variation, tail measure, threshold exceedances

**Regarding our questions**

**(Q1)** This chapter tackles the question of the dependence structure in EVT with a different angle as the two previous one. Indeed, we focus here only on extremal dependent data, that is, data for which it is likely that all marginals are simultaneously extreme. In this context, no tools regarding dimension reduction are provided since extreme events do not appear on lower-dimensional subsets.

**(Q2)** Our approach relies on the minimum of the marginals of a regularly varying random vector $\mathbf{X}$. We study the behavior of the vector $\mathbf{X}$ conditioned on the event that this minimum is large. There, we explore how conditional independence can be transpose to this setting.

**(Q3)** The regular variation assumption does not rely on a norm as it was the case until now. Here we replace the standard setting by a pseudo-norm, in our case the minimum of the marginals

and study the effect of this approach on the limit vector. The threshold condition used here is then $\min_{1\leq j\leq d} X_j > t$. We discuss the consequences of this choice only at a theoretical level.

**Contents**



## 4.1 Introduction

It is customary in Extreme Value Theory (EVT) to study the tail behavior of a random vector $\mathbf{X} \in \mathbb{R}_+^d$ through its threshold exceedances $\mathbf{X} \mid |\mathbf{X}| > t$. After a proper normalization this conditional distribution converges under some assumptions to a vector $\mathbf{Y}$ which summarizes the extreme structure of $\mathbf{X}$. When the chosen norm corresponds to the infinity norm the distribution of $\mathbf{Y}$ is called multivariate Pareto distribution (Rootzén and Tajvidi (2006)). Several results on this family of distributions have been established (Rootzén et al. (2018a), Kiriliouk et al. (2019)). In this context, the study of the tail behavior of $\mathbf{X}$ needs to properly take into account the dependence between the marginals $X_1, \ldots, X_d$. In particular, if all marginals are independent, it leads to asymptotic independence: The distribution of $\mathbf{Y}$ only places mass on the axes (see Section 1.2.3, and also De Haan and De Ronde (1998), Marshall and Olkin (1983), Ledford and Tawn (1996)). Conversely, asymptotic dependence arises when several marginals of $\mathbf{X}$ are likely to be simultaneously extreme (Coles et al. (1999)).

The notion of independence in EVT is therefore deeply linked to the dependence structure of the original vector $\mathbf{X}$. Regarding the limit vector $\mathbf{Y}$, its marginals are often strongly dependent in both asymptotic independence and asymptotic dependence cases. Moreover, studying the dependence structure of the marginals $Y_1, \ldots, Y_d$ is not straightforward since $\mathbf{Y}$ does not take values in a product space. In a recent paper, Engelke and Hitz (2020) introduce a concept of conditional independence to multivariate Pareto distributions under some assumptions on the support of $\mathbf{Y}$. It enables then to develop a theory of graphical models for extremes. Standard results on this topic like the pairwise Markov property or the Hammersley-Clifford theorem are transposed to the extreme case.

The aim of this chapter is twofold. The first step consists in a study of the approach developed by Engelke and Hitz (2020). A particular attention is paid to the assumption made on the limit vector $\mathbf{Y}$. This assumption is discussed in a context of regularly varying random vectors and leads to slightly different approach based on the minimum of the marginals of $\mathbf{X}$.

**Outline** General results on regular variation, conditional independence and graphical models are gathered in Section 4.2. Section 4.3 introduces the approach of Engelke and Hitz (2020) to study conditional independence for the marginals of a multivariate Pareto distribution. We particularly discuss the assumptions made on the limit vector $\mathbf{Y}$. This discussion highlights the central role played by the minimum of the marginals of a regularly varying random vector. In Section 4.4 we gather several results to characterize regular variation via the minimum of the marginals. Finally, Section 4.5 details the other approach of conditional independence for multivariate Pareto distributions.

## 4.2 Theoretical background

### 4.2.1 Regular variation

Let $\mathcal{E} = [0,\infty)^d \setminus \{\mathbf{0}\}$. We consider a regularly varying random vector $\mathbf{X} \in \mathcal{E}$: There exists $a_n \to \infty$ and a non-zero Radon measure on the Borel $\sigma$-field of $\mathcal{E}$ such that for any $\mu$-continuity set $A$ we have

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in A) \to \mu(A)\,, \quad n \to \infty\,. \tag{4.2.1}$$

In this case, there exist $\alpha > 0$ such that the limit measure $\mu$ is homogeneous with index $-\alpha$, i.e. $\mu(tA) = t^{-\alpha}\mu(A)$ for any $t > 0$ and any Borel set $A \subset \mathcal{E}$ (see Section 1.2.2.2).

The homogeneity property of $\mu$ implies that for $u > 0$ the complementary of the product set $[0,u]^d$ in $\mathcal{E}$ is a $\mu$-continuity set. Indeed, the boundary of this set is equal to $\{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty = u\}$. Thus, following Equation (4.2.1) we write

$$\begin{aligned}
\mu\big(\{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty = u\}\big) &= \lim_{\epsilon \to 0} \big[\mu\big(\{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty \geq (1-\epsilon)u\}\big) - \mu\big(\{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty \geq (1+\epsilon)u\}\big)\big] \\
&= \lim_{\epsilon \to 0} \big[(1-\epsilon)^{-\alpha} - (1+\epsilon)^{-\alpha}\big]\mu\big(\{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty \geq u\}\big) \\
&= 0\,,
\end{aligned}$$

which gives the desired result.

**Remark 4.2.1.** With $u = 1$ we obtain the convergence

$$n\mathbb{P}(|\mathbf{X}|_\infty > a_n) = n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{1}]^c) \to \mu([\mathbf{0},\mathbf{1}]^c)\,, \quad n \to \infty\,.$$

As explained in Section 1.2.2.2, it is sometimes convenient to choose a sequence $(a_n)$ such that $n\mathbb{P}(|\mathbf{X}|_\infty > a_n) \to 1$ when $n \to \infty$ so that $\mu([\mathbf{0},\mathbf{1}]^c) = 1$.

Actually, we can extend the previous property to sets of the form $[\mathbf{0},\mathbf{z}]^c$ for $\mathbf{z} > 0$. Indeed, if $\mathbf{z} > \mathbf{0}$, then the boundary of the set $[\mathbf{0},\mathbf{z}]^c$ corresponds to the union $\cup_{j=1}^d A_j(\mathbf{z})$ where the sets $A_j(\mathbf{z})$ are defined as

$$A_j(\mathbf{z}) = \prod_{l=1}^{j-1}[0,z_l] \times \{z_j\} \times \prod_{l=j+1}^{d}[0,z_l]\,, \quad j = 1,\ldots,d\,.$$

For $j \in \{1,\ldots,d\}$, the set $A_j(\mathbf{z})$ is included in the set $\{\mathbf{x} \in \mathcal{E},\, x_j = z_j\}$. Therefore, using the homogeneity property of $\mu$ leads to the inequality

$$\begin{aligned}
\mu(A_j(\mathbf{z})) &\leq \mu(\{\mathbf{x} \in \mathcal{E},\, x_j = z_j\}) \\
&= \lim_{\epsilon\to 0}\Big[\mu\big(\{\mathbf{x} \in \mathcal{E},\, x_j \geq (1-\epsilon)z_j\}\big) - \mu\big(\{\mathbf{x} \in \mathcal{E},\, x_j \geq (1+\epsilon)z_j\}\big)\Big] \\
&= \lim_{\epsilon\to 0}\Big[(1-\epsilon)^{-\alpha} - (1+\epsilon)^{-\alpha}\Big]\mu\big(\{\mathbf{x} \in \mathcal{E},\, x_j \geq z_j\}\big) \\
&= 0\,,
\end{aligned}$$

since $\mu\big(\{\mathbf{x} \in \mathcal{E},\, x_j \geq z_j\}\big)$ is finite, see Section 1.2.2.2. We then conclude that $\mu(\partial[\mathbf{0},\mathbf{z}]^c) \leq \mu(A_1(\mathbf{z})) + \ldots + \mu(A_d(\mathbf{z})) = 0$ which proves that $[\mathbf{0},\mathbf{z}]^c$ is a $\mu$-continuity set. In particular it means that

$$n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}]^c) \to \mu([\mathbf{0},\mathbf{z}]^c)\,, \quad n \to \infty\,,$$

for all $\mathbf{z} > \mathbf{0}$.

We rephrase the convergence in Equation (4.2.1) in terms of threshold exceedances. To this end, consider $\mathbf{z} > \mathbf{0}$ and observe that

$$\begin{aligned}
\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}] \mid |\mathbf{X}|_\infty > a_n) &= \frac{\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}],\, |\mathbf{X}|_\infty > a_n)}{\mathbb{P}(|\mathbf{X}|_\infty > a_n)} \\
&= \frac{\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}] \cap [\mathbf{0},\mathbf{1}]^c)}{\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{1}]^c)} \\
&= \frac{n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}\wedge\mathbf{1}]^c \setminus [\mathbf{0},\mathbf{z}]^c)}{n\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{1}]^c)}\,.
\end{aligned}$$

Since the sets $[\mathbf{0},\mathbf{1}]^c$, $[\mathbf{0},\mathbf{z}]^c$, and $[\mathbf{0},\mathbf{z}\wedge\mathbf{1}]^c$ are $\mu$-continuity sets, we obtain the convergence

$$\lim_{n\to\infty}\mathbb{P}(a_n^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}] \mid |\mathbf{X}|_\infty > a_n) = \frac{\mu([\mathbf{0},\mathbf{z}\wedge\mathbf{1}]^c) - \mu([\mathbf{0},\mathbf{z}]^c)}{\mu([\mathbf{0},\mathbf{1}]^c)}, \quad n \to \infty\,. \tag{4.2.2}$$

It is possible to obtain a continuous version of this convergence replacing $a_n \to \infty$ by $u \to \infty$ (see Resnick (1987), Theorem 6.1). This leads then to the definition of the limit distribution of the threshold exceedances of $\mathbf{X}$:

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = \lim_{u\to\infty}\mathbb{P}(u^{-1}\mathbf{X} \leq \mathbf{z} \mid |\mathbf{X}|_\infty > u)\,, \quad \mathbf{z} > \mathbf{0}\,. \tag{4.2.3}$$

The distribution of $\mathbf{Y}$ is called *multivariate Pareto distribution* (Rootzén and Tajvidi (2006)).

**Remark 4.2.2.** Since the family of sets $\{[\mathbf{0},\mathbf{z}],\, \mathbf{z} > \mathbf{0}\}$ generates the $\sigma$-algebra of $(0,\infty)^d$, the previous convergence can be extended to any Borel set $A$ of $(0,\infty)^d$. In particular, with the $\mu$-continuity set $A_t = \{\mathbf{x} \in \mathcal{E},\, |\mathbf{x}|_\infty = t\}$, for $t \geq 1$, we obtain that

$$\mathbb{P}(\mathbf{Y} \in A_t) = \lim_{n\to\infty} \mathbb{P}(a_n^{-1}\mathbf{X} \in A_t \mid |\mathbf{X}|_\infty > a_n) = \lim_{n\to\infty} \frac{n\mathbb{P}(a_n^{-1}\mathbf{X} \in A_t)}{n\mathbb{P}(a_n^{-1}\mathbf{X} \in A_1)} = \frac{\mu(A_t)}{\mu(A_1)} = t^{-\alpha}\,,$$

by the homogeneity property of $\mu$. This proves that $\mathbb{P}(|\mathbf{Y}|_\infty > t) = t^{-\alpha}$, which means that the $|\mathbf{Y}|_\infty$ follows a Pareto($\alpha$) distribution.

**Remark 4.2.3.** If we assume that $\mu([\mathbf{0},\mathbf{1}]^c) = 1$ (see Remark 4.2.1), then combining Equations(4.2.2) and (4.2.3) leads to the relation

$$\mathbb{P}(\mathbf{Y} \in [\mathbf{0},\mathbf{z}]^c) = \mu([\mathbf{0},\mathbf{z}]^c)\,,$$

for all $\mathbf{z} \in [1,\infty)^d$.

It is possible to extend the convergence in Equation (4.2.3) to the whole set $\mathcal{E}$. To this end, it suffices to prove that this convergence holds for all $\mathbf{z} \in \mathcal{E}$. We then conclude by using the fact that the family of sets $\{[\mathbf{0},\mathbf{z}],\, \mathbf{z} \in \mathcal{E}\}$ generates the $\sigma$-algebra of $\mathcal{E}$. Consider a vector $\mathbf{z} \in \mathcal{E}$ with at least one null coordinate, say $z_k = 0$. If $\mathbb{P}(\mathbf{Y} \in \partial[\mathbf{0},\mathbf{z}]) = 0$, then $\mathbb{P}(\mathbf{Y} \in [\mathbf{0},\mathbf{z}]) = 0$. As soon as $\mathbf{X}$ has non-degenerate marginals (which we assume), we obtain the inequality $\mathbb{P}(u^{-1}\mathbf{X} \in [\mathbf{0},\mathbf{z}]) \leq \mathbb{P}(X_k = 0) = 0$. Hence, Equation (4.2.3) holds for this $\mathbf{z}$. All in all, we obtain the following convergence in distribution in $\mathcal{E}$:

$$\mathbf{X}/u \mid |\mathbf{X}|_\infty > u \xrightarrow{d} \mathbf{Y}\,, \quad u \to \infty\,. \tag{4.2.4}$$

Equation (4.2.4) has a natural interpretation in terms of extreme values since it defines a multivariate Pareto distribution as the limit of a regularly varying random vector $\mathbf{X}$ conditioned on the event that at least one component of $\mathbf{X}$ exceeds a high threshold. Several properties of this family of distributions have been established by Rootzén et al. (2018a) and Kiriliouk et al. (2019). Different models for this kind of distribution have been provided by Rootzén et al. (2018b). The choice of the infinity norm in (4.2.4) implies that the vector $\mathbf{Y}$ belongs to the space

$$\mathcal{L} = \{\mathbf{x} \in \mathbb{R}_+^d,\, |\mathbf{x}|_\infty > 1\} = \{\mathbf{x} \in \mathbb{R}_+^d,\, \exists k = 1,\ldots,d,\, x_k > 1\} = \mathbb{R}_+^d \setminus [0,1]^d\,. \tag{4.2.5}$$

since $\mathcal{L} = [\mathbf{0},\mathbf{1}]^c$ is a $\mu$-continuity set and $\mathbb{P}(\mathbf{X}/u \in \mathcal{L} \mid |\mathbf{X}|_\infty > u) = 1$.

### 4.2.2 Independence and conditional independence

Conditional independence is an extension of independence to conditional probabilities and distributions. It highlights the fact that two independent events can be dependent as soon as a third event occurs. Given three random variables $X$, $Y$, and $Z$ with support $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$, we say that $X$ is

independent of $Y$ given $Z$, and we write $X \perp Y \mid Z$, if

$$\mathbb{P}(X \in A,\, Y \in B \mid Z \in C) = \mathbb{P}(X \in A \mid Z \in C)\mathbb{P}(Y \in B \mid Z \in C)\,, \tag{4.2.6}$$

for all sets $A \subset \mathcal{X}$, $B \subset \mathcal{Y}$, and $C \subset \mathcal{Z}$ such that $\mathbb{P}(Z \in C) > 0$. If $Z$ is deterministic, then Equation (4.2.6) boils down to standard independence between $X$ and $Y$.

**Example 4.2.1.** Consider a triplet of random variables $(X, Y, Z)$ with a joint distribution $p(X, Y, Z)$ which factorizes as

$$p(X, Y, Z) = p(X \mid Z)p(Y \mid Z)p(Z)\,.$$

This factorization implies that $p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$. Therefore, the random variables $X$ and $Y$ are conditionally independent given $Z$. On the contrary, the joint distribution of $(X, Y)$ is given by

$$p(X, Y) = \sum_z p(X, Y, Z = z) = \sum_z p(X \mid Z = z)p(Y \mid Z = z)p(Z = z)\,,$$

which does not factorize in general into the product $p(X)p(Y)$. Hence the random variables $X$ and $Y$ are in general not independent.

Conditional independence is often used in a multivariate framework to study the dependence structure of the marginals of a random vector. Let $\mathbf{X}$ be a random vector taking values in a Cartesian product $E = E_1 \times \ldots \times E_d \subset \mathbb{R}^d$ and consider a partition $A \sqcup B \sqcup C$ of $\{1, \ldots, d\}$. Then, in a lot of situations, we are willing to study the conditional independence of $\mathbf{X}_A$ and $\mathbf{X}_C$ given $\mathbf{X}_B$. If we assume that $\mathbf{X}$ has a positive and continuous density $f_{\mathbf{X}}$ on $E$, then $\mathbf{X}_A \perp \mathbf{X}_C \mid \mathbf{X}_B$ if and only if the density factorizes as

$$f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{X},\, B}(\mathbf{x}_B) = f_{\mathbf{X},\, A\cup B}(\mathbf{x}_{A\cup B}) f_{\mathbf{X},\, B\cup C}(\mathbf{x}_{B\cup C})\,, \quad \mathbf{x} \in E\,. \tag{4.2.7}$$

In particular if $B = \emptyset$ it boils down to standard independence of $\mathbf{X}_A$ and $\mathbf{X}_C$.

Note that as for standard independence, conditional independence requires distributions supported on product spaces.

**Graphical Models** In the context of graphical models, conditional independence of the marginals of a random vector in $\mathbb{R}^d$ can be represented via a graph whose set of vertices is $V = \{1, \ldots, d\}$. Indeed, for an undirected graph $\mathcal{G} = (E, V)$ with vertices $V = \{1, \ldots, d\}$ and a set of edges $E \subset V \times V$, the idea behind graphical models is to associate to each vertice $k \in V$ a random variable $X_k$ taking values in $\mathcal{X}_k \subset \mathbb{R}$ and to build the vector $\mathbf{X} = (X_1, \ldots, X_d)^\top \in \mathcal{X} = \prod_{k=1}^d \mathcal{X}_k$. If $\mathbf{X}$ admits a positive and continuous density $f_{\mathbf{X}}$ on $\mathcal{X}$, then it is possible to characterize conditional independence through the pairwise Markov property relative to $\mathcal{G}$:

$$\forall (i, j) \notin E\,, \quad X_i \perp X_j \mid \mathbf{X}_{\setminus\{i,j\}}\,. \tag{4.2.8}$$

A random vector $\mathbf{X}$ is then called *probabilistic graphical model* on the graph $\mathcal{G}$ if its distribution satisfies the pairwise Markov property. Hence, the conditional dependence structure of a probabilistic graphical model can be easily visualized through the associated graph. Note that conditional independence also arises with directed graphs. This is for instance the case in Bayesian Network Theory in which the dependence relations rely on the concepts of parents variables (see for instance the monograph of Jensen et al. (1996)).

**Example 4.2.2** (Continuation of Example 4.2.1)**.** The vector $\mathbf{X} \in \mathbb{R}^3$ with marginals $X$, $Y$, and $Z$ defined in Example 4.2.1 is a probabilistic graphical model on the graph $\mathcal{G}$ given in Figure 4.1.

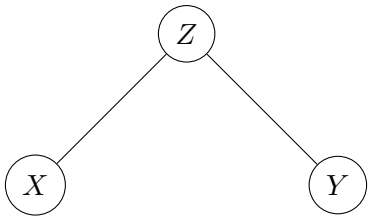



Figure 4.1: A undirected graph $\mathcal{G}$ with three vertices.

**Example 4.2.3.** Consider the graph $\mathcal{G}'$ given in Figure 4.2. The random vector $\mathbf{X}$ with marginals $X_j$ is a probabilistic graphical model on $\mathcal{G}'$ if

$$X_1 \perp X_3 \mid \mathbf{X}_{\{2,4\}} \quad \text{and} \quad X_2 \perp X_4 \mid \mathbf{X}_{\{1,3\}}\,.$$

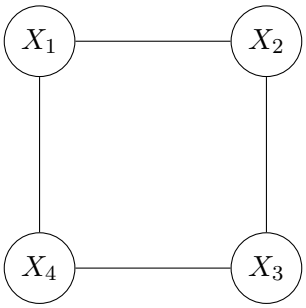



Figure 4.2: A undirected graph $\mathcal{G}'$ with four vertices.

Recently, graphical representations of probabilistic relations has been studied in a statistical framework (Dawid (1979) and Dawid (1980)). The monograph of Lauritzen (1996) gathers all useful results on graphical models and a particular attention is paid to conditional independence. One of the advantages of graphical models and conditional independence is their interpretability in terms of probabilistic structure. Indeed, these models often lead to sparse pattern in multivariate random vectors as explained in Wainwright and Jordan (2008).

## 4.3 Conditional independence according to Engelke and Hitz (2020)

We consider a vector $\mathbf{Y}$ with a multivariate Pareto distribution as defined in Equation (4.2.4). The choice of the infinity norm and the condition $\{|\mathbf{X}|_\infty > u\}$ lead to a natural interpretation of $\mathbf{Y}$ in terms of the extreme behavior of $\mathbf{X}$ (at least one of the components of $\mathbf{X}$ is large). However, this choice has the drawback to provide a limit $\mathbf{Y}$ which belongs to a subset that is not a product space (see Figure 4.3a). In order to exhibit a product space we rewrite the set $\mathcal{L}$ introduced in Equation (4.2.5) as the union $\mathcal{L} = \cup_{k=1}^d \mathcal{L}^k$ with

$$\mathcal{L}^k = \{\mathbf{x} \in \mathcal{E},\, x_k > 1\}\,, \quad k = 1, \dots, d\,.$$

For a fixed $k \in \{1, \dots, d\}$ the subset $\mathcal{L}^k$ is a product set. Moreover, if we condition the limit vector $\mathbf{Y}$ on the event that $\{Y_k > 1\}$, then we obtain a conditional vector $\mathbf{Y}^k = \mathbf{Y} \mid Y_k > 1$ whose support is included in $\mathcal{L}^k$ (see Figure 4.3).

A particular attention should be paid on the condition $\{Y_k > 1\}$. Indeed, since $\mathbf{Y} \in \mathcal{E}$ there exists $k \in \{1, \dots, d\}$ such that $\mathbb{P}(Y_k > 1) > 0$. But for the moment there is no guarantee that the inequality $\mathbb{P}(Y_k > 1) > 0$ holds for all $k$. To tackle this issue (and also to avoid other technical difficulties), Engelke and Hitz (2020) assume that the limit distribution $\mathbf{Y}$ does not place any mass on lower-dimensional faces of $\mathcal{E}$. It is equivalent to assume that

$$\mathbb{P}\big(\mathbf{Y} \in \tilde{\mathcal{E}}\big) = 1\,, \tag{4.3.1}$$

where $\tilde{\mathcal{E}} = (0, \infty)^d$. Regarding the subsets $C_\beta$ defined in (1.4.4), it means that the only subset on which $\mathbf{Y}$ places mass is the central one $C_{\{1,\dots,d\}}$. In particular, this assumption implies that the distribution $\mathbf{Y}$ is a model for asymptotic extremal dependence. Under assumption (4.3.1), we have the equality $\mathbb{P}(Y_k > 0) = 1$. By homogeneity of the underlying measure $\mu$, this leads to the inequality $\mathbb{P}(Y_k > 1) > 0$. Hence, the conditional vector $\mathbf{Y}^k$ is now well defined for all $k = 1, \dots, d$.

**Remark 4.3.1.** The models studied in this framework are used for data with a strong tail dependence between the components. This differs from the approach used in Chapter 2 and Chapter 3 in which we assumed that large events appear due to the extreme behavior of only few coordinates. Hence, sparse regular variation is not helpful in this framework.

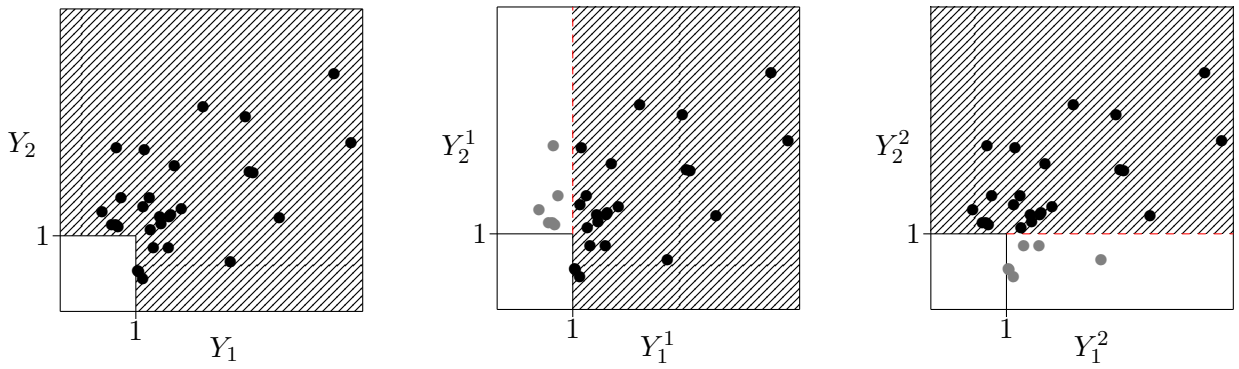
**Remark 4.3.2.** Under assumption (4.3.1), we obtain that the convergence

$$\mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = \lim_{u \to \infty} \mathbb{P}(u^{-1}\mathbf{X} \leq \mathbf{z} \mid |\mathbf{X}|_\infty > u)\,, \tag{4.3.2}$$

which appears in Equation (4.2.3) holds for all $\mathbf{z} \in \mathcal{E}$. If $\mathbf{X}$ has non-degenerate marginals, then the converse is true. Indeed, if we consider a vector $\mathbf{z} \in \mathcal{E}$ with $z_k = 0$, then Equation (4.3.2) implies that

$$0 = \mathbb{P}(X_k = 0 \mid |\mathbf{X}|_\infty > u) \geq \mathbb{P}(u^{-1}\mathbf{X} \leq \mathbf{z} \mid |\mathbf{X}|_\infty > u) \to \mathbb{P}(\mathbf{Y} \leq \mathbf{z})\,, \quad u \to \infty\,.$$

Hence, $\mathbb{P}(\mathbf{Y} \leq \mathbf{z}) = 0$ for all $\mathbf{z} \in \mathcal{E}$ such that $z_k = 0$. This implies that $\mathbb{P}(Y_k = 0) = 0$. Since this holds for any $k = 1, \ldots, d$, it follows that $\mathbf{Y}$ does not place any mass on lower-dimensional subspaces, hence (4.3.1) holds.




(a) (b) (c)

Figure 4.3: A data set and the supports of the different vectors $\mathbf{Y}$, $\mathbf{Y}^1$, $\mathbf{Y}^2$. The shaded areas correspond to (a) the support of $\mathbf{Y}$, (b) the support of $\mathbf{Y}^1$ and (c) the support of $\mathbf{Y}^2$.

If $\mathbf{Y}$ admits a density $\mathrm{f}_{\mathbf{Y}}$ on $\mathcal{L}$, then the density of $\mathbf{Y}^k$ is given by

$$f^k(\mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{\int_{\mathcal{L}^k} f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}}, \quad \mathbf{y} \in \mathcal{L}^k\,. \tag{4.3.3}$$

Note that the existence of a density for $\mathbf{Y}$ also requires to assume that (4.3.1) holds.

In this context, Engelke and Hitz (2020) define a notion of conditional independence for extremes as follows. Consider $A, B, C \subset \{1, \ldots, d\}$ non-empty disjoint subsets whose union is $\{1, \ldots, d\}$. Then, $\mathbf{Y}_A$ is said to be *conditionally independent* of $\mathbf{Y}_C$ given $\mathbf{Y}_B$ if

$$\forall k = 1, \ldots, d\,, \quad \mathbf{Y}_A^k \perp \mathbf{Y}_C^k \mid \mathbf{Y}_B^k\,. \tag{4.3.4}$$

If such a condition holds, they denote it by $\mathbf{Y}_A^k \perp_e \mathbf{Y}_C^k \mid \mathbf{Y}_B^k$. In terms of the density $f_{\mathbf{Y}}^k$ of $\mathbf{Y}^k$, Equation (4.3.4) is equivalent to the factorization

$$\forall k = 1, \ldots, d, \quad f_{\mathbf{Y}}^k(\mathbf{y}) f_{\mathbf{Y},B}^k(\mathbf{y}_B) = f_{\mathbf{Y},A\cup B}^k(\mathbf{y}_{A\cup B}) f_{\mathbf{Y},B\cup C}^k(\mathbf{y}_{\mathbf{Y},B\cup C})\,, \quad \mathbf{y} \in \mathcal{L}^k\,.$$

Actually it sufficies to check that the condition in (4.3.4) holds for one $k \in B$ (see Engelke and Hitz (2020) Proposition 1).

**Graphical models** In the context of graphical models, Engelke and Hitz (2020) adapt the pairwise Markov property given in (4.2.8) to the definition given by (4.3.4). The multivariate Pareto distribution satisfies the pairwise Markov property relative to a graph $\mathcal{G} = (V, E)$ with $V = \{1, \ldots, d\}$ if

$$\forall (i,j) \notin E, \quad Y_i \perp_e Y_j \mid \mathbf{Y}_{\{i,j\}}\,.$$

If such a property holds, the vector $\mathbf{Y}$ is called an *extremal graphical model* with respect to the graph $\mathcal{G}$. Several characterizations of extremal graphical models are established by Engelke and Hitz (2020). The general idea is that the standard results in the theory of graphical models and conditional independence can be transposed to the extremal case. We do not insist more on the concept of graphical models and rather focus on an another approach to build a similar notion of conditional independence for $\mathbf{Y}$.

## 4.4 Regular variation via the minimum of the marginals

### 4.4.1 Considering the minimum

In all this section we consider a random vector $\mathbf{X} \in \mathcal{E}$. If $\mathbf{X}$ is regularly varying with multivariate Pareto distribution $\mathbf{Y}$ (see Equation (4.2.4)), then the assumption (4.3.1) on $\mathbf{Y}$ done by Engelke and Hitz (2020) can be reformulated in terms of the minimum of the marginals of $\mathbf{Y}$:

$$\mathbb{P}\Big(\min_{1\leq k\leq d} Y_k > 0\Big) = \mathbb{P}\big(\mathbf{Y} \in \tilde{\mathcal{E}}\big) = 1\,. \tag{4.4.1}$$

With Equation (4.4.1) we introduce two aspects which are of constant use in what follows.

The first one is the use of the function $\min : \mathbb{R}_+^d \to [0,\infty)$. Indeed, the probability $\mathbb{P}(\min_{1\leq k\leq d} Y_k > 0)$ encourages to characterize regularly varying random vectors via the minimum of their marginals. The goal is then to replace $|\mathbf{X}|_\infty$ in (4.2.4) by $\min_{1\leq k\leq d} X_k$. This approach requires some precautions since min is not a norm even it satisfies some homogeneity properties. Besides, conditioning on the event that $\{\min_{1\leq k\leq d} X_k > u\}$ needs to place ourselves in an appropriate subspace. Indeed, just as we have to restrict the study of regular variation to the space $\mathbb{R}_+^d \setminus \{\mathbf{0}\}$ when we use a arbitrary norm $|\cdot|$, it is necessary to remove the sets on which the function min vanishes, that is, the axes. A regular variation property which involves the minimum of the marginals must therefore be define on the restricted subspace $\tilde{\mathcal{E}} = (0,\infty)^d$. This leads to the second key aspects that appears with (4.4.1): The set $\tilde{\mathcal{E}} = (0,\infty)^d$ is a product space which enables to consider conditional independence on this set.

Recall that the vector $\mathbf{Y}$ can be expressed as the limit of the excedeences of $\mathbf{X}$ in a sequential form as in Equation (4.2.2) or in a continuous form as in Equation (4.2.4). Equivalently, the regular variation property (4.2.1) based on a sequence $(a_n)$ can be expressed in a continuous fashion. A random vector $\mathbf{X} \in \mathcal{E}$ is regularly varying with tail index $\alpha > 0$ if there exists a regularly varying function $R$ with index $-\alpha$ and a non-null Radon measure $\mu$ on the Borel $\sigma$-field of $\mathcal{E}$ such that for any $\mu$-continuity set $A$ we have

$$\frac{1}{R(u)}\mathbb{P}(u^{-1}\mathbf{X} \in A) \to \mu(A)\,, \quad u \to \infty\,, \tag{4.4.2}$$

see Hult and Lindskog (2006), Theorem 3.1 for more details. A standard choice for the regularly varying function is $R(u) = \mathbb{P}(|\mathbf{X}| > u)$. The idea is now to extend the standard notion of regular

variation which relies on a fixed norm to the minimum of the marginals or more generally to a norm-like function called *modulus* $\rho$. Starting from Equation (4.4.2), the central point is to prove that it is possible to replace $R(u) = \mathbb{P}(|\mathbf{X}| > u)$ by $R'(u) = \mathbb{P}(\rho(\mathbf{X}) > u)$. This issue is addressed by Segers et al. (2017), Lemma 3.1. The function $\rho : \mathbb{R}_+^d \to [0, \infty)$ has to satisfy some properties closed to the one satisfied by a norm (see in Segers et al. (2017), Definition 2.2). It is the case of the minimum of the components $\min : \tilde{\mathcal{E}} \to (0, \infty)$ which leads to the following characterization (see Segers et al. (2017), Proposition 3.1). A random vector $\mathbf{X}$ is regularly varying on the space $\tilde{\mathcal{E}}$ if and only if there exists a random vector $\mathbf{Y}'$ such that

$$\min_{1\le k\le d} X_k \text{ is regularly varying} \quad \text{and} \quad \big(u^{-1}\mathbf{X} \mid \min_{1\le k\le d} X_k > u\big) \xrightarrow{d} \mathbf{Y}', \quad u \to \infty\,. \tag{4.4.3}$$

In this case, there exists a non-null Radon measure $\mu'$ on the Borel $\sigma$-field of $\tilde{\mathcal{E}}$ such that for any $\mu'$-continuity set $A$ we have

$$\frac{1}{\mathbb{P}(\min_{1\le k\le d} X_k > u)}\mathbb{P}(u^{-1}\mathbf{X} \in A) \to \mu'(A)\,, \quad u \to \infty\,.$$

As already explained, removing the axes and hence restricting the regular variation condition of $\mathbf{X}$ to the space $\tilde{\mathcal{E}}$ is necessary to capture the asymptotic behavior of $\mathbf{X}$ through the events $\{\min_{1\le k\le d} X_k > u\}$. Regarding the vector $\mathbf{X}$, we already mentioned that the definition of $\mathbf{Y}$ in Equation (4.2.4) has a natural interpretation in terms of threshold exceedances. A multivariate Pareto distribution corresponds to the distribution of $\mathbf{X}$ conditioned on the event that at least one marginal is large. On the other hand, the random vector $\mathbf{Y}'$ defined in (4.4.3) correspond to the distribution of $\mathbf{X}$ conditioned on the event that all marginals are simultaneously large. Although this approach seems more restrictive (all marginals have to be simultaneously large), it is actually not the case under the additional assumption (4.4.1). A detailed comparison of both approaches is provided at the end of this section.

A short computation shows that the support of $\mathbf{Y}'$ is included in

$$\big\{\mathbf{x} \in \mathcal{E},\ \min_{1\le k\le d} x_k > 1\big\} = (1, \infty)^d = \bigcap_{1\le k\le d} \mathcal{L}^k\,.$$

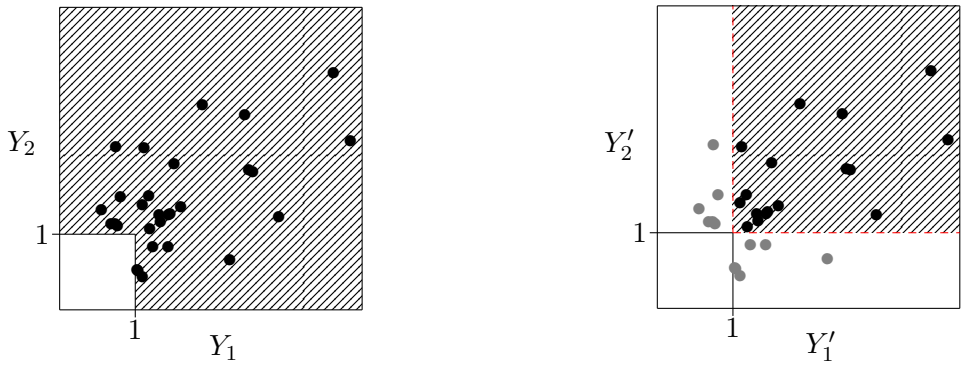
Indeed, the boundary of the set $(1, \infty)^d$ in $\tilde{\mathcal{E}}$ corresponds to the set $\{\mathbf{x} \in \tilde{\mathcal{E}}, \min_{1\le k\le d} x_k = 1\}$. Then, a short calculation gives that

$$\begin{aligned}
\mathbb{P}(\mathbf{Y}' \in \partial(1, \infty)^d) &= \mathbb{P}\big(\min_{1\le k\le d} Y'_k = 1\big) \\
&= \lim_{\epsilon\to 0}\big[\mathbb{P}\big(\min_{1\le k\le d} Y'_k \ge 1 - \epsilon\big) - \mathbb{P}\big(\min_{1\le k\le d} Y'_k \ge 1 + \epsilon\big)\big] \\
&= \lim_{\epsilon\to 0}\big[(1-\epsilon)^{-\alpha} - (1+\epsilon)^{-\alpha}\big]\mathbb{P}\big(\min_{1\le k\le d} Y'_k \ge 1\big) \\
&= 0\,,
\end{aligned}$$

where for the third equality we have also used the homogeneity of the minimum. Then, following Portmanteau Theorem we can conclude that

$$\mathbb{P}(\mathbf{Y}' \in (1,\infty)^d) = \lim_{u\to\infty} \mathbb{P}(u^{-1}\mathbf{X} \in (1,\infty)^d \mid \min_{1\le k\le d} X_k > u) = \lim_{u\to\infty} 1 = 1\,.$$

A illustration of the set $(1,\infty)^d$ is given in Figure 4.4b.




(a) (b)

Figure 4.4: A data set and the supports of $\mathbf{Y}$ and $\mathbf{Y}'$. The shaded areas correspond to (a) the support of $\mathbf{Y}$ and (b) the support of $\mathbf{Y}'$.

### 4.4.2 Comparison of both approaches

Let us recap the different approaches we have introduced. For $\mathbf{X} \in \mathcal{E}$, we say that $\mathbf{X}$ is regularly varying on $\mathcal{E}$ if there exists a random vector $\mathbf{Y}$ such that distribution of exceedances converges weakly:

$$u^{-1}\mathbf{X} \mid |\mathbf{X}|_\infty > u \xrightarrow{d} \mathbf{Y}\,, \quad u\to\infty\,. \tag{RV}$$

Engelke and Hitz (2020) add to this convergence an assumption on the support of the limit vector $\mathbf{Y}$:

$$\mathbb{P}(\mathbf{Y} \in \tilde{\mathcal{E}}) = 1\,. \tag{E-H}$$

On the other hand, we place ourselves in the context of a random vector $\mathbf{X}$ which is regularly varying on $\tilde{\mathcal{E}}$, which corresponds to

$$\min_{1\le k\le d} X_k \text{ is regularly varying}\,, \tag{Min-1}$$

and

$$\big(u^{-1}\mathbf{X} \mid \min_{1\le k\le d} X_k > u\big) \xrightarrow{d} \mathbf{Y}'\,, \quad u\to\infty\,. \tag{Min-2}$$

**Proposition 4.4.1.** *Assume that* (RV) *and* (E-H) *hold. Then* (Min-1) *and* (Min-2) *hold.*

*Proof.* For $u > 0$, we write

$$\mathbb{P}\Big(\min_{1\le k\le d} X_k > u\Big) = \mathbb{P}(u^{-1}\mathbf{X} \in (1,\infty)^d) = \mathbb{P}(u^{-1}\mathbf{X} \in (1,\infty)^d \mid |\mathbf{X}|_\infty > u)\mathbb{P}(|\mathbf{X}|_\infty > u)\,,$$

which implies that

$$\frac{\mathbb{P}\Big(\min_{1\le k\le d} X_k > tu\Big)}{\mathbb{P}\Big(\min_{1\le k\le d} X_k > u\Big)} = \frac{\mathbb{P}(u^{-1}\mathbf{X}\in(t,\infty)^d \mid |\mathbf{X}|_\infty > u)}{\mathbb{P}(u^{-1}\mathbf{X}\in(1,\infty)^d \mid |\mathbf{X}|_\infty > u)}\,.$$

Since $\mathbb{P}(\mathbf{Y}\in\partial(1,\infty)^d)=0$, this ratio converges to $\mathbb{P}(\mathbf{Y}\in(t,\infty)^d)/\mathbb{P}(\mathbf{Y}\in(1,\infty)^d) = t^{-\alpha}$. Hence the minimum of the marginals is regularly varying with tail index $\alpha$ which proves (Min-1).

We now prove (Min-2). For $\mathbf{z}\in\tilde{\mathcal{E}}$ we write

$$\begin{aligned}\mathbb{P}\big(\mathbf{X}/u\le\mathbf{z} \mid \min_{1\le k\le d} X_k > u\big) &= \frac{\mathbb{P}\big(\mathbf{X}/u\le\mathbf{z},\, \min_{1\le k\le d} X_k > u\big)}{\mathbb{P}\big(\min_{1\le k\le d} X_k > u\big)}\\ &= \frac{\mathbb{P}\big(\mathbf{X}/u\le\mathbf{z},\, \min_{1\le k\le d} X_k > u \mid |\mathbf{X}|_\infty > u\big)}{\mathbb{P}\big(\min_{1\le k\le d} X_k > u \mid |\mathbf{X}|_\infty > u\big)}\\ &= \frac{\mathbb{P}\big(\mathbf{X}/u\in[\mathbf{0},\mathbf{z}]\cap(1,\infty)^d \mid |\mathbf{X}|_\infty > u\big)}{\mathbb{P}\big(\mathbf{X}/u\in(1,\infty)^d \mid |\mathbf{X}|_\infty > u\big)}\,.\end{aligned}$$

Following (RV), we obtain

$$\lim_{u\to\infty}\mathbb{P}\big(\mathbf{X}/u\le\mathbf{z} \mid \min_{1\le k\le d} X_k > u\big) = \frac{\mathbb{P}(\mathbf{Y}\in[\mathbf{0},\mathbf{z}]\cap(1,\infty)^d)}{\mathbb{P}(\mathbf{Y}\in(1,\infty)^d)} = \mathbb{P}(\mathbf{Y}\in[\mathbf{0},\mathbf{z}] \mid \mathbf{Y}\in(1,\infty)^d)\,. \quad (4.4.4)$$

Hence the limit vector $\mathbf{Y}'$ corresponds to the vector $\mathbf{Y}$ conditioned on the event that $\{\mathbf{Y}\in(1,\infty)^d\} = \{\min_{1\le k\le d} Y_k > 1\}$. Therefore, condition (Min-2) holds with $\mathbf{Y}' = \mathbf{Y} \mid \mathbf{Y}\in\tilde{\mathcal{E}}$. □

Proposition 4.4.1 states that the framework of (4.4.3) includes the one of Engelke and Hitz (2020). In particular, Equation (4.4.4) implies that if $\mathbf{Y}$ admits a density $f_{\mathbf{Y}}$, then $\mathbf{Y}'$ admits a density $f_{\mathbf{Y}'}$ on $\tilde{\mathcal{E}}$ satisfying the relation

$$f_{\mathbf{Y}'}(\mathbf{y}) = \frac{f_{\mathbf{Y}}(\mathbf{y})}{\mathbb{P}(\mathbf{Y}>1)}\,, \quad \mathbf{y}\in(1,\infty)^d\,. \quad (4.4.5)$$

Conversely, the conditions (Min-1) and (Min-2) are more general than the setting of Engelke and Hitz (2020), i.e. (RV) and (E-H). Indeed, assume for instance that $\mathbf{X}$ is regularly varying on $\mathcal{E}$ with independent marginals. On the one hand, the limit vector $\mathbf{Y}$ has a distribution which only places mass on the axes (see Section 1.2.3). Therefore, assumption (4.3.1) does not hold in this case. On the other hand, regarding the minimum of the marginals, their independence implies that for $x>0$,

$$\frac{\mathbb{P}\big(\min_{1\le k\le d} X_k > xt\big)}{\mathbb{P}\big(\min_{1\le k\le d} X_k > t\big)} = \frac{\mathbb{P}\big(\forall k=1,\dots,d,\, X_k > xt\big)}{\mathbb{P}\big(\forall k=1,\dots,d,\, X_k > t\big)} = \prod_{k=1}^d \frac{\mathbb{P}\big(X_k > xt\big)}{\mathbb{P}\big(X_k > t\big)} \to \prod_{k=1}^d x_k^{-\alpha_k}\,,$$

when $t\to\infty$ and where $\alpha_k>0$ denotes the tail index of the regularly varying random variable $X_k$. This proves that the minimum is regularly varying with tail index $\alpha_1+\ldots+\alpha_d$, hence (Min-1) is true. In particular if all marginals have the same tail index $\alpha$ (for instance after a rank transform),

then the tail index of the minimum is $d\alpha$ (see Jessen and Mikosch (2006), Section 5.4 for more details).

Moreover, the independence of the marginals of $\mathbf{X}$ implies that for all $\mathbf{z} > \mathbf{0}$,

$$\mathbb{P}\big(\mathbf{X}/u > \mathbf{z} \mid \min_{1\le k\le d} X_k > u\big) = \frac{\mathbb{P}\big(\mathbf{X}/u > \mathbf{1} \vee \mathbf{z}\big)}{\mathbb{P}\big(\mathbf{X}/u > 1\big)} = \frac{\prod_{k=1}^d \mathbb{P}\big(X_k/u > 1 \vee z_k\big)}{\prod_{k=1}^d \mathbb{P}\big(X_k/u > 1\big)}\,,$$

which converges to $\prod_{k=1}^d (1\vee z_k)^{-\alpha_k}$ when $u \to \infty$. Since the family of sets $\{[\mathbf{z}, \boldsymbol{\infty}), \mathbf{z} > \mathbf{0}\}$ generates the convergence in $\tilde{\mathcal{E}}$ it proves (Min-2).

These considerations imply that the conditions (Min-1) and (Min-2) are more general than the framework of Engelke and Hitz (2020). Based on the setting of Equation (4.4.3), we introduce a new definition of conditional independence for multivariate Pareto distributions.

## 4.5 Another notion of conditional independence

In this section we consider a random vector $\mathbf{X} \in \mathcal{E}$ which satisfies (4.4.3). The limit vector $\mathbf{Y}'$ is defined on the product space $(1,\infty)^d$ so that it is possible to define a notion of conditional independence for $\mathbf{Y}'$ by

$$\mathbf{Y}'_A \perp \mathbf{Y}'_C \mid \mathbf{Y}'_B\,, \tag{4.5.1}$$

for three non-empty disjoint subsets $A, B, C \subset \{1,\dots,d\}$ whose union is $\{1,\dots,d\}$.

If $\mathbf{Y}'$ admits a positive and continuous density $f_{\mathbf{Y}'}$ it is equivalent to say that the density $f_{\mathbf{Y}'}$ factorizes as

$$f_{\mathbf{Y}'}(\mathbf{y}) f_{\mathbf{Y}', B}(\mathbf{y}_B) = f_{\mathbf{Y}', A\cup B}(\mathbf{y}_{A\cup B}) f_{\mathbf{Y}', B\cup C}(\mathbf{y}_{B\cup C}), \quad \mathbf{y} \in (1,\infty)^d\,. \tag{4.5.2}$$

**Proposition 4.5.1.** *Consider a random vector $\mathbf{X}$ satisfying (RV) and (E-H) and assume that the limit vector $\mathbf{Y}$ admits a density $f_{\mathbf{Y}}$ on $\mathcal{L}$. Consider three non-empty disjoint subsets $A, B, C \subset \{1,\dots,d\}$ whose union is $\{1,\dots,d\}$. If $\mathbf{Y}_A \perp_e \mathbf{Y}_C \mid \mathbf{Y}_B$, then $\mathbf{Y}'_A \perp \mathbf{Y}'_C \mid \mathbf{Y}'_B$.*

*Proof.* Recall that if $\mathbf{Y}$ satisfies (E-H) then by Proposition 4.4.1 the vector $\mathbf{Y}'$ exists and the density of $\mathbf{Y}'$ is given by (4.4.5).

From the definition of conditional independence for multivariate Pareto distributions, the relation $\mathbf{Y}_A \perp_e \mathbf{Y}_C \mid \mathbf{Y}_B$ is equivalent to

$$\forall k = 1,\dots,d, \quad f^k_{\mathbf{Y}}(\mathbf{y}) f^k_{\mathbf{Y}, B}(\mathbf{y}_B) = f^k_{\mathbf{Y}, A\cup B}(\mathbf{y}_{A\cup B}) f^k_{\mathbf{Y}, B\cup C}(\mathbf{y}_{\mathbf{Y}, B\cup C}), \quad \mathbf{y} \in \mathcal{L}^k\,.$$

With Equation (4.3.3) it follows that for all $\mathbf{y} \in (1,\infty)^d$

$$f_{\mathbf{Y}}(\mathbf{y}) f_{\mathbf{Y}, B}(\mathbf{y}_B) = f_{\mathbf{Y}, A\cup B}(\mathbf{y}_{A\cup B}) f_{\mathbf{Y}, B\cup C}(\mathbf{y}_{\mathbf{Y}, B\cup C})\,.$$

We then conclude by using the relation between the density $f_{\mathbf{Y}}$ and $f_{\mathbf{Y}'}$ which leads to (4.5.2). Hence, we have proved that $\mathbf{Y}'_A \perp \mathbf{Y}'_C \mid \mathbf{Y}'_B$. □

This proves once again that the notion of conditional independence for $\mathbf{Y}'$ is an extension of the one regarding $\mathbf{Y}$.

Regarding the original vector $\mathbf{X}$, the study of $\mathbf{Y}'$ instead of $\mathbf{Y}$ provides two advantages. First, the vector $\mathbf{Y}'$ models the extreme behavior of $\mathbf{X}$ when all its marginals are simultaneously large. It therefore provides accurate models for extremal dependent data. Note that if the strong condition that all marginals are extreme together is not satisfied, then (4.4.3), and thus (4.4.1), is very unlikely.

The second advantage is that conditional independence of $\mathbf{Y}'$ can be interpreted in terms of $\mathbf{X}$, as stated in the following proposition.

**Proposition 4.5.2.** *Consider a random vector $\mathbf{X}$ satisfying (Min-1) and (Min-2). Let $A, B, C \subset \{1, \dots, d\}$ be three non-empty disjoint subsets whose union is $\{1, \dots, d\}$. If $\mathbf{X}_A \perp \mathbf{X}_C \mid \mathbf{X}_B$, then $\mathbf{Y}'_A \perp \mathbf{Y}'_C \mid \mathbf{Y}'_B$.*

*Proof.* Denote by $a$ (respectively $b$ and $c$) the number of elements of $A$ (respectively $B$ and $C$). We consider a subset $E_A$ (respectively $E_B$ and $E_C$) of $(1,\infty)^a$ (respectively $(1,\infty)^b$ and $(1,\infty)^c$). The conditional independence of $\mathbf{X}_A$ and $\mathbf{X}_C$ given $\mathbf{X}_B$ implies that

$$\mathbb{P}\big(x^{-1}\mathbf{X} \in E_A \times E_B \times (1,\infty)^c\big) = \mathbb{P}\big(x^{-1}\mathbf{X}_{A\cup B} \in E_A \times E_B\big)\mathbb{P}\big(x^{-1}\mathbf{X}_{B\cup C} \in E_B \times (1,\infty)^c\big)\,, \quad (4.5.3)$$

and

$$\mathbb{P}\big(x^{-1}\mathbf{X} \in (1,\infty)^a \times E_B \times E_C\big) = \mathbb{P}\big(x^{-1}\mathbf{X}_{A\cup B} \in (1,\infty)^a \times E_B\big)\mathbb{P}\big(x^{-1}\mathbf{X}_{B\cup C} \in E_B \times E_C\big)\,. \quad (4.5.4)$$

Multiplying each side of Equations (4.5.3) and (4.5.4) together, and using again the conditional independence of $\mathbf{X}_A$ and $\mathbf{X}_C$ given $\mathbf{X}_B$ gives the following equality:

$$\begin{aligned}\mathbb{P}\big(x^{-1}\mathbf{X} \in E_A \times E_B \times (1,\infty)^c\big)&\mathbb{P}\big(x^{-1}\mathbf{X} \in (1,\infty)^a \times E_B \times E_C\big) \quad (4.5.5)\\ &= \mathbb{P}\big(x^{-1}\mathbf{X} \in E_A \times E_B \times E_C\big)\mathbb{P}\big(x^{-1}\mathbf{X} \in (1,\infty)^a \times E_B \times (1,\infty)^c\big)\,.\end{aligned}$$

Then, dividing both sides of Equation (4.5.5) by $\mathbb{P}(\min \mathbf{X} > \mathbf{x})^2$ and letting $x \to \infty$ lead to the desired result:

$$\mathbb{P}\big(\mathbf{Y}'_{A\cup B} \in E_A \times E_B\big)\mathbb{P}\big(\mathbf{Y}'_{B\cup C} \in E_B \times E_C\big) = \mathbb{P}\big(\mathbf{Y}' \in E_A \times E_B \times E_C\big)\mathbb{P}\big(\mathbf{Y}'_B \in E_B\big)\,,$$

which proves that $\mathbf{Y}'_A$ is conditionally independent of $\mathbf{Y}'_C$ given $\mathbf{Y}'_B$. □

If $B = \emptyset$, then we obtain independence of $\mathbf{X}_A$ and $\mathbf{X}_C$ and therefore of $\mathbf{Y}'_A$ and $\mathbf{Y}'_C$. In this case, if $\mathbf{Y}$ exists, then its support is included in $\mathcal{E} \setminus \tilde{\mathcal{E}}$. There, (4.4.3) is satisfied but not (4.4.1).

## 4.6 Conclusion

Essentially, conditional independence allows a better understanding of the dependence structure of the marginals of a random vectors while it has a natural representation through graphical models. Although it has been widely studied in different contexts, this concept has received little attention in multivariate EVT so far. It is therefore a challenging issue to tackle conditional independence for a multivariate Pareto distribution because of the structure of its support.

Based on the conditional distribution, Engelke and Hitz (2020) adapt conditional independence, graphical models, and tree structures to extreme events. The authors show that the latter concepts naturally arise with the expected properties when the distribution studied is the multivariate Pareto distribution $\mathbf{Y}$ conditioned on the event that at least one marginal exceeds 1. This approach requires to assume that $\mathbf{Y}$ does not place any mass on lower dimensional subspaces. Hence, this restrict our study to extremal dependent data.

In this context, our approach consists in studying extreme values via the minimum of the marginals instead of the infinity norm, see Equation (4.4.3). This approach is possible thanks to different results on the minimum of the marginals of regularly varying random vectors (Jessen and Mikosch (2006), Segers et al. (2017)). In particular, we proved that our setting includes the one of Engelke and Hitz (2020), see Proposition 4.4.1.

We also proved that in terms of conditional independence our approach has a direct link with the original regularly varying random vector $\mathbf{X}$. However, no relation with the associated max-stable distribution has been established yet. This may be a starting point for future work which could also tackle extremal tree structure, as stated in Engelke and Hitz (2020).

# Appendix A

## A.1 Convergence to Types

For two random variables $X$ and $Y$ taking values in a general measurable space, we say that $X$ and $Y$ are of the same type if there exist constant $a > 0$ and $b \in \mathbb{R}$ such that

$$X = aY + b\,.$$

In other words, $X$ and $Y$ have the same distribution up to a location and a scale parameter.

The following result is widely used in EVT. It ensures the unique limit distribution for the normalized maximum and gives the possible normalization sequences. A proof can be found in Resnick (1987), p. 7.

**Theorem A.1.1** (Convergence to types theorem)**.** *Let $(X_n)$ be a sequence of random variables, $X$ and $Y$ be two random variables, and $a_n > 0$, $a'_n > 0$, $b_n$, $b'_n$ be constants. Suppose that*

$$\frac{X_n - b_n}{a_n} \xrightarrow{d} X\,.$$

*Then the relation*

$$\frac{X_n - b'_n}{a'_n} \xrightarrow{d} Y$$

*holds if and only if*

$$a_n/a'_n \to a \geq 0\,, \quad \text{and} \quad (b_n - b'_n)/a'_n \to b \in \mathbb{R}\,, \quad n \to \infty\,.$$

*In this case, $Y = aX + b$, and $a, b$ are the unique constants for which this holds.*

*Moreover, $X$ is non-degenerate if and only if $a > 0$ and in this case, $X$ and $Y$ are of the same type.*

## A.2 Dynkin's Theorem

Consider a measurable space $(E, \mathcal{E})$. In many cases, we are willing to study some properties of the subsets in $\mathcal{E}$ by focusing only on some particular subsets. Dynkin's theorem ensures that is it

possible to do so as soon as the subsets studied satisfy some stability properties. We recall here two notions of measure theory.

1. A subset $\mathcal{T} \subset \mathcal{E}$ is called a $\Pi$-system if $\mathcal{T}$ is closed under finite intersection, i.e. if $A, B \in \mathcal{T}$, then $A \cap B \in \mathcal{T}$.

2. A subset $\mathcal{L} \subset \mathcal{E}$ is called a $\lambda$-system if

   - $E \in \mathcal{L}$,
   - if $A, B \in \mathcal{L}$ with $A \subset B$, then $B \setminus A \in \mathcal{L}$,
   - if $(A_n)$ is a sequence of sets in $\mathcal{L}$ with $A_n \subset A_{n+1}$, then $\lim_{n\to\infty} \uparrow A_n \in \mathcal{L}$.

Finally, for $\mathcal{E}' \subset \mathcal{E}$, denote by $\sigma(\mathcal{E}')$ the smallest $\sigma$-algebra generated by $\mathcal{E}'$.

**Theorem A.2.1** (Dynkin)**.** *If $\mathcal{T}$ is a $\Pi$-system and $\mathcal{L}$ a $\lambda$-system satisfying $\mathcal{T} \subset \mathcal{L}$, then $\sigma(\mathcal{T}) \subset \mathcal{L}$.*

A proof of this theorem can be found in Kallenberg (2006). In practice, the following corollary is often used.

**Corollary A.2.1.** *If two measures are equal on a $\Pi$-system which generates the $\sigma$-algebra, then they are equal on the $\sigma$-algebra.*

## A.3 Bernstein's inequality

The proof of Theorem 3.4.1 relies on Bernstein's inequality. This inequality is widely used in concentration theory and several different versions have been established. We state here the one used in the proof of Theorem 3.4.1.

**Lemma A.3.1** (Bernstein's inequality)**.** *If $X$ is a random variable such that $\mu = \mathbb{E}[X]$ is finite and for all $j \geq 2$, $\mathbb{E}[|X-\mu|^j] \leq \mathrm{Var}(X) < \infty$, then for all $\lambda \in (-1, 1)$, the following inequality holds:*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\lambda^2 \,\mathrm{Var}(X)}{2(1-\lambda/3)}\right). \tag{A.3.1}$$

*Proof.* For $\lambda \in (-1, 1)$ we write

$$e^{\lambda(X-\mu)} = 1 + \lambda(X-\mu) + \lambda^2 \sum_{j=2}^{\infty} \lambda^{j-2} \frac{(X-\mu)^j}{j!}.$$

Then, the Monotone Convergence Theorem entails

$$\mathbb{E}\left[e^{(\lambda(X-\mu)}\right] = 1 + \lambda^2 \sum_{j=2}^{\infty} \lambda^{j-2} \frac{\mathbb{E}\left[(X-\mu)^j\right]}{j!}.$$

Finally, we use the assumption on the moments of $\mathbf{X}$, and the inequalities $j! \geq 2 \cdot 3^{j-2}$, for $j \geq 2$, and $e^x \geq 1 + x$, and obtain that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq 1 + \lambda^2 \sum_{j=2}^{\infty} \frac{\lambda^{j-2}}{2 \cdot 3^{j-2}} \operatorname{Var}(X) = 1 + \frac{\lambda^2 \operatorname{Var}(X)}{2(1-\lambda/3)} \leq \exp\left(\frac{\lambda^2 \operatorname{Var}(X)}{2(1-\lambda/3)}\right).$$

□

# Bibliography


1. Abdous, B. and Ghoudi, K. (2005). Non-parametric estimators of multivariate extreme dependence functions. *Nonparametric Statistics*, 17(8):915–935.

2. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 267–281, Budapest. Akademia Kiado.

3. Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *The Annals of Mathematical Statistics*, 34(1):122–148.

4. Bahadur, R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā: The Indian Journal of Statistics*, 20(3-4):207–210.

5. Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, 2(5):792–804.

6. Basrak, B. (2000). *The sample autocorrelation function of non-linear time series*. PhD thesis, University of Groningen.

7. Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). *Statistics of Extremes: Theory and Applications*. John Wiley & Sons Ltd., Chichester.

8. Beirlant, J. and Teugels, J. L. (1992). Modeling large claims in non-life insurance. *Insurance: Mathematics and Economics*, 11(1):17–29.

9. Beirlant, J., Teugels, J. L., and Vynckier, P. (1996a). *Practical Analysis of Extreme Values*. Leuven University Press, Leuven.

10. Beirlant, J., Vynckier, P., and Teugels, J. L. (1996b). Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318.

11. Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, USA.

12. Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463.

13. Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, second edition.

14. Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular variation*. Cambridge University Press, Cambridge.

15. Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012). Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379.

16. Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97(1-2):113–150.

17. Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

18. Blanchard, G., Bousquet, O., and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, 66(2-3):259–294.

19. Boldi, M.-O. and Davison, A. (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):217–229.

20. Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12,267–12,272.

21. Bruun, J. T. and Tawn, J. A. (1998). Comparison of approaches for estimating the probability of coastal flooding. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):405–423.

22. Caeiro, F. and Gomes, M. I. (2015). Threshold selection in extreme value analysis. *Extreme Value Modeling and Risk Analysis: Methods and Applications*, pages 69–86.

23. Chautru, E. (2015). Dimension reduction in multivariate extreme value analysis. *Electronic Journal of Statistics*, 9(1):383–418.

24. Chiapino, M. and Sabourin, A. (2016). Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer.

25. Chiapino, M., Sabourin, A., and Segers, J. (2019). Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222.

26. Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, New-York.

27. Coles, S., Heffernan, J., and Tawn, J. (1999). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

28. Coles, S. G. and Tawn, J. A. (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):377–392.

29. Condat, L. (2016). Fast projection onto the simplex and the $\ell_1$ ball. *Mathematical Programming*, 158(1):575–585.

30. Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

31. Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to algorithms*. MIT press, Cambrige.

32. Crammer, K. and Singer, Y. (2002). On the learnability and design of output codes for multiclass problems. *Machine learning*, 47(2-3):201–233.

33. Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1):1–31.

34. Dawid, A. P. (1980). Conditional independence for statistical operations. *The Annals of Statistics*, 8(3):598–617.

35. De Fondeville, R. and Davison, A. (2016). High-dimensional peaks-over-threshold inference. *Biometrika*, 105(3):575–592.

36. De Haan, L. and De Ronde, J. (1998). Sea and wind: multivariate extremes at work. *Extremes*, 1(1):7–45.

37. de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New-York.

38. de Haan, L. and Resnick, S. I. (1977). Limit theory for multivariate sample extremes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 40(4):317–337.

39. Dekkers, A. L., Einmahl, J. H., De Haan, L., et al. (1989). A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855.

40. Devroye, L. (1989). The double kernel method in density estimation. *Annales de l'Institut Poincaré, section B*, 25(4):533–580.

41. Dhillon, I. S., Guan, Y., and Kogan, J. (2002). Iterative clustering of high dimensional text data augmented by local search. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 131–138.

42. Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2):143–175.

43. Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS Conference on Mathematical Challenges of the 21st Century*, pages 1–32.

44. Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. (2008). Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th Annual International Conference on Machine learning*, pages 272–279.

45. Einmahl, J. H., De Haan, L., Piterbarg, V. I., et al. (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics*, 29(5):1401–1423.

46. Einmahl, J. H., de Haan, L., and Sinha, A. K. (1997). Estimating the spectral measure of an extreme value distribution. *Stochastic Processes and their Applications*, 70(2):143–171.

47. Einmahl, J. H., Dehaan, L., and Huang, X. (1993). Estimating a multidimensional extreme-value distribution. *Journal of Multivariate Analysis*, 47(1):35–47.

48. Einmahl, J. H. and Segers, J. (2009). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 37(5B):2953–2989.

49. Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin.

50. Engelke, S. and Hitz, A. S. (2020). Graphical models for extremes. *arXiv:1812.01734*.

51. Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, volume II. Wiley, New-York, second edition.

52. Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

53. Fougères, A.-L. (2004). Multivariate extremes. In *Extreme Values in Finance, Telecommunications, and the Environment*, pages 373–388. Chapman and Hall.

54. Fougères, A.-L. and Soulier, P. (2010). Limit conditional distributions for bivariate vectors with polar representation. *Stochastic models*, 26(1):54–77.

55. Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Annales de la Société Polonaise de Mathématique*, 6(1):93–116.

56. Gafni, E. M. and Bertsekas, D. P. (1984). Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964.

57. Galambos, J. (1978). *The asymptotic theory of extreme order statistics*. Kreiger, Malabar.

58. Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

59. Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall, Boca Raton.

60. Gissibl, N., Klüppelberg, C., et al. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24(4A):2693–2720.

61. Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of mathematics*, 44(3):423–453.

62. Goix, N., Sabourin, A., and Clémençon, S. (2016). Sparse representation of multivariate extremes with applications to anomaly ranking. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pages 75–83.

63. 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.

64. Gomes, M. I. and Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: a review. *International statistical review*, 83(2):263–292.

65. Guillotte, S., Perron, F., and Segers, J. (2011). Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):377–406.

66. Gumbel, E. J. (1958). *Statistics of Extremes*. Columbia University Press, New-York.

67. Hans-Hermann, B. (2008). Origins and extensions of the $k$-means algorithm in cluster analysis. *Journal Électronique d'Histoire des Probabilités et de la Statistique*, 4(2).

68. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New-York.

69. Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall/CRC, Boca Raton.

70. Hazan, E. (2006). Approximate convex optimization by online game playing. *arXiv: 0610.00119*.

71. Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):497–546.

72. Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174.

73. Hille, E. (1964). *Analysis*, volume 1. Blaisdell, New-York.

74. Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

75. Hult, H. and Lindskog, F. (2006). Regular variation for measures on metric spaces. *Publications de l'Institut Mathématique*, 80(94):121–140.

76. Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

77. Janßen, A. and Wan, P. (2020). $k$-means clustering for extremes. *Electronic Journal of Statistics*, 14(1):1211–1233.

78. Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171.

79. Jensen, F. V. et al. (1996). *An introduction to Bayesian networks*, volume 210. Springer, Berlin.

80. Jessen, A. H. and Mikosch, T. (2006). Regularly varying functions. *Publications de l'institut Mathématique*, 80(94):171–192.

81. Joe, H., Smith, R. L., and Weissman, I. (1992). Bivariate threshold methods for extremes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):171–183.

82. Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, New-York.

83. Jung, S., Dryden, I. L., and Marron, J. (2012). Analysis of principal nested spheres. *Biometrika*, 99(3):551–568.

84. Kallenberg, O. (2006). *Foundations of modern probability*. Springer, New-York.

85. Karamata, J. (1933). Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, 61:55–62.

86. Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8-12):1287–1304.

87. Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135.

88. Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167.

89. Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions. Theory and Applications.* Imperial College Press, London.

90. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

91. Kyrillidis, A., Becker, S., Cevher, V., and Koch, C. (2013). Sparse projections onto the simplex. In *Proceedings of the 30th Annual International Conference on Machine Learning*, pages 235–243.

92. Lauritzen, S. L. (1996). *Graphical models.* Clarendon Press, Oxford.

93. Lawless, J. F. (2011). *Statistical models and methods for lifetime data.* John Wiley & Sons, Hoboken, second edition.

94. Leadbetter, M. R. (1991). On a basis for 'Peaks over Threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362.

95. Ledford, A. W. and Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

96. Lee, J., Fan, Y., and Sisson, S. A. (2015). Bayesian threshold selection for extremal models using measures of surprise. *Computational Statistics & Data Analysis*, 85:84–99.

97. Lehtomaa, J. and Resnick, S. (2019). Asymptotic independence and support detection techniques for heavy-tailed multivariate data. *arXiv: 1904.00917.*

98. Lellmann, J., Kappes, J., Yuan, J., Becker, F., and Schnörr, C. (2009). Convex multi-class image labeling by simplex-constrained total variation. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 150–162.

99. Liu, J. and Ye, J. (2009). Efficient euclidean projections in linear time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 657–664.

100. Longin, F. M. (2000). From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance*, 24(7):1097–1130.

101. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297.

102. Marshall, A. W. and Olkin, I. (1983). Domains of attraction of multivariate extreme value distributions. *The Annals of Probability*, 11(1):168–177.

103. Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *The Annals of Probability*, 17(1):266–291.

104. Massart, P. (2007). *Concentration inequalities and model selection.* Springer, Berlin.

105. Meyer, N. and Wintenberger, O. (2019). Sparse regular variation. *arXiv:1907.00686.*

106. Mikosch, T. (1999). Regular variation, subexponentiality and their applications in probability theory. In *EURANDOM report*, volume 99. Eindhoven University of Technology.

107. Papastathopoulos, I. and Strokorb, K. (2016). Conditional independence among max-stable laws. *Statistics & Probability Letters*, 108:9–15.

108. Pelleg, D. and Moore, A. (1999). Accelerating exact $k$-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–281.

109. Pericchi, L. R. and Rodríguez-Iturbe, I. (1985). On the statistical analysis of floods. In *A celebration of statistics*, pages 511–541. Springer, New-York.

110. Pickands, J. I. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

111. Ramos, A. and Ledford, A. (2009). A new class of models for bivariate joint tails. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):219–241.

112. Reiss, R.-D. (1989). *Approximate distributions of order statistics: with applications to non-parametric statistics.* Springer, New-York.

113. Resnick, S. I. (1986). Point processes, regular variation and weak convergence. *Advances in Applied Probability*, 18(1):66–138.

114. Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes.* Springer, New-York.

115. Resnick, S. I. (1997). Discussion of the Danish data on large fire insurance losses. *ASTIN Bulletin*, 27(1):139–151.

116. Resnick, S. I. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336.

117. Resnick, S. I. (2007). *Heavy-Tail Phenomena. Probabilistic and Statistical Modeling.* Springer, New-York.

118. Ribatet, M. (2013). Spatial extremes: Max-stable processes at work. *Journal de la Société Française de Statistique*, 154(2):156–177.

119. Rootzén, H., Segers, J., and Wadsworth, J. L. (2018a). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis*, 165:117–131.

120. Rootzén, H., Segers, J., and Wadsworth, J. L. (2018b). Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145.

121. Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli*, 12(5):917–930.

122. Sabourin, A. and Drees, H. (2019). Principal component analysis for multivariate extremes. *arXiv:1906.11043*.

123. Sabourin, A. and Naveau, P. (2014). Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis*, 71:542–567.

124. Sabourin, A., Naveau, P., and Fougeres, A.-L. (2013). Bayesian model averaging for multivariate extremes. *Extremes*, 16(3):325–350.

125. Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT - Statistical Journal*, 10(1):33–60.

126. Schmidt, R. and Stadtmüller, U. (2006). Non-parametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33(2):307–335.

127. Segers, J. (2012). Max-stable models for multivariate extremes. *Revstat Statistical Journal*, 10(1):61–82.

128. Segers, J., Zhao, Y., and Meinguet, T. (2017). Polar decomposition of regularly varying time series in star-shaped metric spaces. *Extremes*, 20(3):539–566.

129. Seghouane, A.-K. and Amari, S.-I. (2007). The AIC criterion and symmetrizing the Kullback–Leibler divergence. *IEEE Transactions on Neural Networks*, 18(1):97–106.

130. Seghouane, A.-K. and Bekara, M. (2004). A small sample model selection criterion based on kullback's symmetric divergence. *IEEE transactions on signal processing*, 52(12):3314–3323.

131. Sibuya, M. (1960). Bivariate extreme statistics I. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.

132. Simpson, E., Wadsworth, J., and Tawn, J. (2019). Determining the dependence structure of multivariate extremes. *arXiv:1809.01606*.

133. Smith, R. L. (1987). Estimating tails of probability distributions. *The Annals of Statistics*, 15(3):1174–1207.

134. Smith, R. L. (1994). Multivariate threshold methods. In *Extreme Value Theory and Applications*, pages 225–248. Springer, New-York.

135. Smith, R. L., Tawn, J. A., and Yuen, H. K. (1990). Statistics of multivariate extremes. *International Statistical Review / Revue Internationale de Statistique*, 58(1):47–58.

136. Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415.

137. Tawn, J. A. (1990). Modelling multivariate extreme value distributions. *Biometrika*, 77(2):245–253.

138. Tawn, J. A. (1992). Estimating probabilities of extreme sea-levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(1):77–93.

139. Tawn, J. A. (1994). Applications of multivariate extremes. In *Extreme value theory and applications*, pages 249–268. Springer, New-York.

140. Teugels, J. L. (1984). Extreme values in insurance mathematics. In *Statistical Extremes and Applications*, pages 253–259. Springer, Dordrecht.

141. Thom, H. (1954). Frequency of maximum wind speed. In *Proceedings of the American Society of Civil Engineers*, volume 80, pages 104–114.

142. Tiago De Oliveira, J. (1958). Extremal distributions. *Rev. Fac. Cienc. Lisboa Ser. A*, 7:215–227.

143. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

144. Verleysen, M. (2003). Learning high-dimensional data. In *Limitations and Future Trends in Neural Computation*, pages 141–162. IOS Press.

145. Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

146. Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In *Advances in Neural Information Processing Systems*, pages 1649–1656.

# Résumé

L'étude de la dépendance des événements extrêmes a été jusqu'à présent essentiellement traiéte en petite dimension. Le but de cette thèse est de développer une approche statistique permettant d'apprendre la structure de dépendance des extrêmes dans un contexte de grande dimension. Le premier chapitre rassemble les résultats principaux concernant la théorie des valeurs extrêmes multivariées. Les différents outils nécessaires aux chapitres suivants sont présentés, notamment le concept de vecteurs aléatoires à variation régulière, mais également les notions d'apprentissage statistique, de statistique en grande dimension et de sélection de modèles. Le deuxième chapitre est un travail conjoint avec Olivier Wintenberger. Il expose le concept de variation régulière parcimonieuse, définie via la projection euclidienne sur le simplexe, et qui étend la notion standard de variation régulière. Cette approche introduit de la parcimonie dans l'étude des extrêmes multivariés et réduit ainsi la dimension. Le troisième chapitre est un travail en cours avec Olivier Wintenberger sur une approche statistique des vecteurs aléatoires à variation régulière parcimonieuse. L'idé de ce chapitre est de proposer une méthode qui permet d'identifier les sous-ensembles de $\mathbb{R}^d$ sur lesquels les valeurs extrêmes se concentrent. Basée sur la sélection de modèles multinomiaux, cette méthode de détection des extrêmes permet également d'identifier de manière ad hoc le seuil optimal au-delà duquel les données sont considérées comme extrêmes. Au niveau des simulations, nous fournissons des preuves numériques de nos résultats théoriques et nous illustrons notre approche sur quelques exemples basés sur des données réelles. Enfin, le quatrième chapitre propose une discussion de l'article de Engelke and Hitz (2020) dans lequel les auteurs définissent une notion d'indépendance conditionnelle pour une loi de Pareto multivariée. On étend leur approache en s'appuyant sur l'étude du minimum des marginales d'un vecteur aléatoire à variation régulière.

***Mots-clés***— Extrêmes multivariés, mesure spectrale, projection euclidienne sur le simplexe, sélection de modèle, statistique en grande dimension, variation régulière, variation régulière parcimonieuse

# Abstract

Studying the dependence of extreme events has so far only been dealt in low dimension. The aim of this thesis is to develop a statistical approach to learn the dependence structure of extremes in a high-dimensional setting. The first chapter brings together the main results concerning multivariate extreme value theory. The different tools needed in the following chapters are introduced, notably the concept of regularly varying random vectors, but also the notions of statistical learning, high-dimensional statistics and model selection. The second chapter is a joint work with Olivier Wintenberger. It outlines the concept of sparse regular variation defined via the Euclidean projection onto the simplex and which extends the standard notion of regular variation. This approach introduces sparsity into the study of multivariate extremes and thus reduces the dimension. The third chapter presents a work in progress with Olivier Wintenberger on a statistical approach of sparsely regularly varying random vectors. The idea of this chapter is to bring out a method which allows us to identify the subsets of $\mathbb{R}^d$ on which extremes concentrate. Based on a multinomial model selection, this method of extremes' detection also provides a way to identify the optimal threshold above which the data are considered to be extreme. In simulations, we provide numerical evidence of our theoretical findings and illustrate our approach on some data-driven examples. Finally, the fourth chapter discusses the article by Engelke and Hitz (2020) in which the authors define a notion of conditional independence for a multivariate Pareto distribution. We extend their approach with the study of the minimum of the marginals of a regularly varying random vector.

***Keywords***— Euclidean projection onto the simplex, high-dimensional statistics, model selection, multivariate extremes, regular variation, sparse regular variation, spectral measure