Statistical learning for multivariate and functional extremes

Nathan Huet

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT031

Thèse de doctorat

Statistical Learning for Multivariate and Functional Extremes

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 15/11/2024, par

NATHAN HUET

Composition du Jury :

Pavlo Mozharovskyi Professeur, Télécom Paris (LTCI)	Président/Examinateur
Clément Dombry Professeur, Université de Franche-Comté (LmB)	Rapporteur
Céline Duval Professeure, Sorbonne Université (LPSM)	Rapportrice
Aurélie Fischer Professeure, Université Paris Cité (LPSM)	Examinatrice
Stéphan Clémençon Professeur, Télécom Paris (LTCI)	Directeur de thèse
Anne Sabourin Professeure, Université Paris Cité (MAP5)	Co-directrice de thèse

Contents

Remerciements

Je tiens à remercier toutes les personnes qui m'ont accompagné de près comme de loin tout au long de ma route.

Tout d'abord, un grand merci à Anne et Stephan. Merci Anne d'avoir donné ce cours sur les extrêmes à Orsay et de m'avoir permis de vivre ces trois belles années avec vous. Merci à tous les deux de m'avoir fait découvrir, de m'avoir appris et de m'avoir fait aimer le monde de la recherche. Si je continue sur cette voie, c'est en grande partie grâce à vous. Un grand merci également à Philippe pour le travail ensemble durant cette dernière année. Tu m'as également beaucoup appris.

Merci Clément Dombry et Céline Duval pour avoir accepté de relire mon manuscrit de thèse et de m'avoir fourni des commentaires qui m'ont permis de l'améliorer. Merci Aurélie Fischer et Pavlo Mozharovskyi d'avoir fait partie de mon jury.

Merci à toutes les personnes qui ont partagé mon quotidien à Télécom durant ces trois ans. Merci beaucoup Anass d'avoir été là à Orsay, à Télécom et à Cité U. Savoir qu'on travaillera ensemble m'a donné la motivation quand j'en manquais; toutes nos discussions n'ont rendu que plus léger ce long périple. Merci Emilia d'avoir illuminé le labo par ta bonne humeur, ta gentillesse et ta bienveillance. A presto! Merci Junjie pour toutes ces discussions endiablées, pour ton rire communicatif et pour m'avoir fait découvrir ce cinéma que j'aime tant. Merci Mathilde d'avoir été là durant cette fin de thèse, ton humour et ta joie de vivre m'ont beaucoup apporté! Viareggio restera à jamais gravé dans ma mémoire grâce (à cause?) de toi. Plus que des collègues, j'ai rencontré quatre amis chers.

Merci à toutes les personnes avec qui j'ai pu partager ne serait-ce qu'un instant. Merci Tamim, Arturo, Joël, Jérémy, Dimitri, Lilian, Ikhlas, Iyad, Louise, Paul, Quentin, Aina, Marc, Amaury, Guillaume, et tous ceux que j'oublie. Bon courage à tous ceux qui n'ont pas encore pu écrire de remerciements, vous y êtes presque!

Merci à tous mes amis pour tous les moments passés ensemble. Merci Martin d'avoir été à mes cotés depuis si longtemps, toujours disponible, toujours souriant, toujours bienveillant, toujours toi. Merci Paolini pour ces coriaces parties de tennis, de foot, de ping-pong, de squash... c'est grâce à toi que j'ai gardé la forme durant ces trois ans. Merci Élise, Bastien, Hugo, Yannick, Michael G. Scott, Charles, Hippolyte et Robin de m'avoir si bravement soutenu.

Enfin, merci Papa, Maman, Léo, Papy et Mamy pour votre soutien sans faille. Voir la plus faible lueur de fierté dans vos yeux m'a procuré la plus grande motivation durant toute ma vie. Si je suis ici aujourd'hui, c'est entièrement grâce à vous, vous m'êtes tout.

Abstract

In a world where climate change is causing increasingly severe extreme weather events, the study of such phenomena has become essential for risk management in many fields. From climate sciences, with heavy rainfall and heatwaves to finance, with stock market crashes, extremes are omnipresent. Specifically, the *Extreme Value Theory* allows for modeling rare, previously unobserved events by extrapolating from the largest observed data. For instance, the following application is crucial for constructing appropriate coastal defenses against marine submersion : relying solely on past high sea levels, without planning even higher levels, would be naive. Here, the "Peaks-over-Threshold" perspective is adopted, meaning an observation is considered extreme if it exceeds, in some sense, a high threshold. This thesis aims to enhance statistical methods related to the prediction and modeling of extreme data using tools from statistical learning. It is divided into two main parts.

First, motivated by the continuous improvement of measurement devices providing increasingly precise temporal or spatial data, we study functional extremes, *i.e.*, extremes of data explicitly dependent on a continuous variable such as time. To develop a general viewpoint, we work within a separable Hilbert space, focusing on the $L^2[0,1]$ space of square-integrable functions over $[0,1]$. We establish results concerning *regular variation* in this space, a fundamental assumption at the core of extreme value theory. We propose characterizations involving only finite-dimensional objects and non-trivial examples of random elements satisfying these assumptions. A second aspect of this work involves developing probabilistic and statistical guarantees for optimal finite-dimensional representation of functional extreme data through principal component analysis. Experiments conducted on simulated and real datasets validate the effectiveness of our dimensionality reduction procedure.

Second, we focus on the task of prediction in extreme regions. We construct a probabilistic framework suitable for regression where the input variable can be extreme but not the output variable, contrasting with existing works that typically consider extremes of the variable to be predicted. Specifically, we work within the context of the regular variation with respect to a component. We outline several properties, usual to this type of hypothesis, and provide examples through common regression scenarios that satisfy this hypothesis, thereby demonstrating its relevance. From this established framework, we derive results concerning risks and regression functions in extremes, leading to the development of an algorithm for regression in extreme regions. We demonstrate that an optimal regression function in regions far from the origin enjoys desirable properties such as radial invarariance. We illustrate the strength of our algorithmic approach on several simulated and real datasets, comparing it with standard regression methods. After establishing the effectiveness of this method, we apply it to the study

of skew surges and sea levels in Brittany. We aim to predict maritime extremes at a Breton station with a short temporal record using data from stations with long-range data histories. This procedure aims to augment historical extreme data to reduce uncertainties in extreme estimations at this station. Alongside our regression method, another multivariate extreme value modeling method is implemented to, for instance, generate samples of extreme sea levels or skew surges.

Résumé

Dans un monde où le réchauffement climatique provoque de plus en plus de phénomènes météorologiques extrêmes d'ampleurs croissantes, l'étude de tels événements devient indispensable à la gestion des risques dans de nombreuses applications. Des sciences du climat, avec les fortes précipitations et les vagues de chaleur, à la finance, avec les krachs boursiers, les extrêmes sont omniprésents. Précisément, la *théorie des valeurs extrêmes* permet de modéliser des événements rares, jusqu'alors jamais rencontrés, en extrapolant à partir des plus grandes données observées par le passé. Par exemple, l'utilisation de la théorie des valeurs extrêmes est cruciale pour la construction de défenses littorales adaptées contre la submersion marine : considérer uniquement les grandes hauteurs de mer passées, sachant qu'il en surviendra de plus en plus élevées, serait naïf. Ici, le point de vue dit de "Dépassement d'un Seuil" est adopté, c'est-à-dire qu'une observation est déclarée extrême si elle dépasse, en un certain sens, un seuil important. Cette thèse vise à enrichir les méthodes statistiques liées à la prédiction et à la modélisation des données extrêmes à partir d'outils provenant de l'apprentissage statistique. Elle se divise en deux grandes parties.

Dans un premier temps, motivés par l'amélioration perpétuelle des appareils de mesure fournissant des données temporelles ou spatiales de plus en plus précises, nous étudions les extrêmes fonctionnels, c'est-à-dire les extrêmes de données dépendant explicitement d'une variable continue comme le temps. Afin de développer un point de vue le plus général possible, nous nous plaçons dans un espace de Hilbert séparable, avec en vue l'espace $L^2[0,1]$ des fonctions de carrés intégrables sur $[0,1]$. Nous développons des résultats portant sur la *variation régulière* dans cet espace, hypothèse fondamentale au cœur de la théorie des valeurs extrêmes. Nous proposons des caractérisations n'impliquant que des objets de dimension finie, ainsi que des exemples non triviaux d'éléments aléatoires satisfaisant ces hypothèses. Un second pan de ce travail réside dans l'élaboration de garanties probabilistes et statistiques permettant une représentation optimale en dimension finie de données fonctionnelles extrêmes, à travers leurs analyses en composantes principales. Des expériences sur des jeux de données simulées et réelles témoignent de la légitimité de notre procédure de réduction de dimension.

Dans un second temps, nous nous intéressons à la tâche de prédiction dans les régions extrêmes. Nous construisons un cadre probabiliste adapté à la régression dans lequel la variable d'entrée peut être extrême mais pas la variable de sortie, prenant à contrepied les travaux existants qui considèrent communément les extrêmes de la variable à prédire. Précisément, nous travaillons dans le contexte de la variation régulière par rapport à une composante. Nous développons les propriétés usuelles liées à ce type d'hypothèses et proposons des exemples à travers des scénarios classiques de régression

s'inscrivant dans ce cadre de travail, justifiant ainsi son bien-fondé. À partir de ce cadre ainsi construit, nous établissons des résultats concernant les risques et les fonctions de régression dans les extrêmes, permettant l'élaboration d'un algorithme de régression dans les extrêmes. Nous démontrons qu'une fonction de régression optimale dans des régions éloignées de l'origine admet certains attributs profitables. Nous illustrons la puissance de notre approche algorithmique sur plusieurs jeux de données simulées et réelles en le comparant à des méthodes de régression usuelles. Une fois la pertinence de cette méthode prouvée, nous l'appliquons à l'étude de la surcote et des hauteurs d'eau en Bretagne. Nous cherchons à prédire les extrêmes maritimes à une station bretonne avec une courte profondeur temporelle à partir de stations présentant un grand historique de données. Cette procédure a pour objectif de compléter les données extrêmes passées pour réduire les incertitudes liées aux estimations extrêmes à cette station. En plus de notre méthode de régression, une autre méthode de modélisation des valeurs extrêmes multivariées est mise en place pour permettre, entre autres, d'obtenir un générateur de hauteurs d'eau ou de surcotes extrêmes.

Notation

$\mathcal{C}[0,1]$	space of real-valued continuous functions over $[0,1]$
$\mathcal{C}(M,I)$	space of continuous functions from M to I
$\mathcal{C}_0(M)$	set of real-valued continuous functions on $M\backslash\{0_M\}$
$\mathbb{D}[0,1]$	space of real-valued *càdlàg* functions over $[0,1]$
ℓ^2	space of square summable real-valued sequences
$L^2[0,1]$	space of square-integrable real-valued functions over $[0,1]$
$\mathbb{H}$	a separable Hilbert space
$\mathbb{S}, \mathbb{S}^{d-1}$	unit sphere, unit sphere of $\mathbb{R}^d$
$\mathbb{B}$	unit ball
$\mathbb{B}(0,r)$	ball of radius r
∂A	boundary of A
A^c	complement space of A
$\bar{A}$	closure of A
tA	$\{ta, a \in A\}$
$\|\cdot\|, \|\cdot\|_p$	norm/L^p-norm in $\mathbb{H}$ or $\mathbb{R}^d$
$\langle\cdot,\cdot\rangle$	scalar product in $\mathbb{H}$
$\|\cdot\|_{op}, \|\cdot\|_{HS(\mathbb{H})}, \|\cdot\|_{tr}$	operator/Hilbert-Schmidt/trace-class norm
$HS(\mathbb{H})$	space of Hilbert-Schmidt operators of $\mathbb{H}$
$h_1 \otimes h_2(h)$	$\langle h_1, h\rangle h_2$
$\pi_N(x)$	$(\langle e_1, x\rangle, ..., \langle e_N, x\rangle)$ for $(e_i)_{i\geq 1}$ a Hilbert basis
Π_V	projection on V
$\rho(V,W)$	$\|\Pi_V - \Pi_W\|_{HS(\mathbb{H})}$

$\mathcal{L}(X)$	distribution of X
$\mathcal{L}(Y \mid X)$	conditional distribution of Y given X
$\mathcal{L}(X_n) \to \mathcal{L}(X)$ $X_n \xrightarrow{d} X$ $X_n \xrightarrow{w} X$	X_n converges in distribution to X
Θ	$X/\|\|X\|\|$
$X_{(k)}$	k-th order statistic
R	$\|\|X\|\|$
$X \sim P$	X is distributed according to P
$Pareto(\alpha)$	Pareto distribution with parameter α
$\mathcal{B}(M)$	Borel σ-algebra of M
$G_{\mu,\sigma,\xi}$	GEV cdf with parameters (μ, σ, ξ)
$H_{\mu,\sigma,\xi}$	GP cdf with parameters (μ, σ, ξ)
$F_{\sigma,\xi,\kappa}$	EGP cdf with parameters (σ, ξ, κ)
RV_ρ	Set of regularly varying real-valued functions and random variables with index ρ
$RV_\rho(M)$	Set of regularly varying M-valued functions and random variables with index ρ
μ	exponent measure
Φ	angular/spectral measure
$\mathbb{M}_0$	set of Borel measures on $M \backslash \{0_M\}$
$\mu_n \xrightarrow{\mathbb{M}_0} \mu$	μ_n converges to μ in $\mathbb{M}_0$
$\mathbf{a} \geq \mathbf{b}$	$\forall j \in \{1, ..., d\}, a_j \geq b_j$
$\mathbf{a} \nleqslant \mathbf{b}$	$\exists j \in \{1, ..., d\}, a_j > b_j$
x_+	$max(x, 0)$
$\lfloor x \rfloor$	integer part of x
$\boldsymbol{\theta}(\mathbf{x}), \theta(x)$	$\mathbf{x}/\|\|\mathbf{x}\|\|$, $x/\|\|x\|\|$
$z_{i:j}$	$z_i, ..., z_j$
$S_{\mathcal{A}}(n)$	shatter coefficient of $\mathcal{A}$ of size n
$V_{\mathcal{A}}$	VC-dimension of $\mathcal{A}$

Abbreviation

a.s.	almost surely
cdf	cumulative distribution function
EGP	Extended Generalized Pareto
ERM	Empirical Risk Minimization
EVA	Extreme Value Analysis
EVT	Extreme Value Theory
FDA	Functional Data Analysis
GEV	Generalized Extreme Value
GP	Generalized Pareto
HS	Hilbert-Schmidt
MAE	Mean Absolute Error
MGP	Multivariate Generalized Pareto
MSE	Mean Square Error
OLS	Ordinary Least Square
PCA	Principal Component Analysis *or* Decomposition
PoT	Peaks-over-Threshold
RF	Random Forest
RMSE	Root Mean Square Error
RV	Regular Variation
SL	Sea Level
SS	Skew Surge
SVR	Support Vector Regression
VC	Vapnik-Chervonenkis
w.r.t.	with respect to

Chapter 1

Introduction

Contents

1.1 Motivations

On February 1st, 1953, a catastrophic storm struck Northern Europe, affecting the Netherlands and the United Kingdom. The storm overwhelmed most sea defenses, leading to an unprecedented flood that claimed over 2,000 lives, with more than 1,800 fatalities in the Netherlands alone. Following this tragic event, the Dutch government faced a crucial question: how high should the new dikes be built to mitigate economic impact and prevent future disasters of this magnitude? This question hinges on determining the maximum sea levels that can be expected over the next hundred or thousand years. Traditional statistical methods fall short in addressing this issue because they require making inferences over a longer period than the available observational data.

Extreme value theory provides the necessary statistical tools to investigate such rare events. This theory focuses on understanding events with low probabilities that lie outside the bulk of the distribution, yet hold significant importance across various fields. These events, which lie outside the bulk of the distribution, could, however, be crucially important in a wide range of applications, from risk management in finance or insurance to extreme event modeling in climate science, such as heavy rainfalls or heatwaves, by predicting extreme air pollution levels or extreme loads on network traffic in health sciences or telecommunications.

Figure 1.1: North Sea flood in Netherlands, 1953 (photo from *Watersnoodmuseum*).

In this thesis, we explore the intersection of extreme value theory and statistical learning, a branch of statistics aimed at predicting and modeling data patterns. Our focus is on two main areas of statistical learning: functional data analysis and regression. Functional data analysis deals with data that are functions, depending on continuous variables like time or space. With advancements in sensor technology, providing massive and increasingly granular measurements, modeling functional extremes, such as large energy loads or significant precipitation over time, has become essential. Regression, one of the most fundamental tasks in statistical learning, involves learning predictive functions from labeled examples to make predictions on new, unlabeled data. While predictive functions typically target the bulk of the distribution, it is of vital interest in many applications to develop models that specifically address examples outside the core distribution, particularly those of an extreme nature.

1.2 State-of-the-art

Extreme value theory (EVT) and statistical learning are two branches of statistics that have been actively researched for many decades. EVT focuses on modeling rare events, while statistical learning encompasses methods for learning patterns and features from data. Recently, there has been growing interest in applying statistical learning tools to improve the study of extremes, particularly in unsupervised learning contexts. Examples include dimensionality reduction through multiple subspace clustering in Goix et al. (2016, 2017); Chiapino et al. (2019); Simpson et al. (2020); Meyer and Wintenberger (2021, 2023), as well as Principal Component Analysis (PCA) in Cooley and Thibaud (2019); Drees and Sabourin (2021). Central to Chapter 6, dimension reduction for extremes are broadly presented in Section 1.2.2. In addition, there has been notable exploration in clustering methods Janßen and Wan (2020); Vignotto et al. (2021), graphical models Engelke and Hitz (2020), and applications such as anomaly detection Chiapino et al. (2020); Vignotto and Engelke (2020) (see Section 1.2.2 for additional references). In the supervised setting, the predominant focus in the literature revolves around predicting extreme values of the target variable Y Aghbalou et al.

(2024a) or tackling extreme quantile regression through methods such as gradient boosting Velthoen et al. (2023) or random forests Gnecco et al. (2024).

To our knowledge, the only work that addresses predicting a target variable Y based on extremes values of the input variable $\mathbf{X}$[1] is by Jalalzai et al. (2018). This study develops an Empirical Risk Minimization (ERM) framework for binary classification with extreme covariates, assuming that the conditional distributions of $\mathbf{X}$ given $Y = \pm 1$ are regularly varying (see Chapter 2 for more details). The authors construct then a regression function adapted to the ERM problem $\min_g L_t(g) = \mathbb{P}(Y \neq g(\mathbf{X}) \mid \|\mathbf{X}\| \geq t)$ for some norm $\|\cdot\|$. Part III of this thesis aims to extend these results to the regression problem, establishing sufficient and reasonable conditions for statistical regression with a continuous target and an appropriate real-valued loss. Specifically, we seek to extend the non-asymptotic statistical guarantees provided for extreme classifiers to extreme regression functions.

Recent developments in concentration inequalities within extreme settings are noteworthy. To our knowledge, the first of this kind is due to Boucheron and Thomas (2012) (see also Boucheron and Thomas (2015)), which proves concentration bounds for extreme order statistics. In a different approach, another pioneering work that has influenced many subsequent studies is Goix et al. (2015), which presents general concentration inequalities for low-probability events and applies them to classification settings. These results form the basis for the non-asymptotic work in Jalalzai et al. (2018). Further, the authors in Clémençon et al. (2023) provide statistical bounds on using the empirical marginal standardization (see Equation (2.10) and Remark 7.3 for more details) instead of the true (unknown) marginal standardization in the classification procedure. Concentrations inequalities are also for extreme cross-validation problem Aghbalou et al. (2023) (also based on Goix et al. (2015)) and for imbalanced classification Aghbalou et al. (2024b), where the minority class corresponds to extremes. General concentration inequalities, part of the Vapnik-Chervonenkis (VC) theory, for extremes have also been broadly developed in Lhaut et al. (2022) and Lhaut and Segers (2021). In Chapter 6, concentration inequalities are used extensively to control the reconstruction error related to the Principal Component Decomposition (PCA) of a Hilbert extreme random element and to bound the maximal deviation between an extreme regression risk and its empirical counterpart in Chapter 8. More details on concentration inequalities can be found in Chapter 4.

In the rest of this section, we delve deeper into two particularly active lines of research at the intersection of statistical learning and EVT. Section 1.2.1 discusses functional approaches for EVT, specifically covering the general theory in general metric spaces and in the space of continuous functions over $[0,1]$. In Section 1.2.2, dimension reduction techniques, such as clustering or PCA, for extremes are presented, with a focus on anomaly detection. In a final section, we review existing research on a critical applied fields for EVT: the modeling of extreme sea levels and the crucial estimation of return periods.

[1] For clarity, throughout this thesis, multivariate quantities are bolded when necessary, such as $\mathbf{x} \in \mathbb{R}^d$, to distinguish sample observations from vector coordinates. Univariate or Hilbert quantities are denoted in the traditional manner, such as $x \in \mathbb{R}$ or $h \in \mathbb{H}$, as no confusion is likely to arise in these cases.

1.2.1 Functional extremes

The ubiquity of sensors providing ever more precise massive measurements of time - or space - dependent quantities has highlighted the importance of understanding continuous data, known as functional data. Functional Data Analysis (FDA) is a specialized branch of statistical studying infinite-dimensional data which has garnered the interest of research for many years. The monographs Hsing and Eubank (2015), Horváth and Kokoszka (2012) and Ramsay and Silverman (2005), offers a comprehensive overview of this fields from the theoretical foundations to the various applications of FDA. The increasing availability of data of functional nature opens new roads of research, such as exploring functional extremes. This area of study is a well-established and active area of research in spatial statistics, as highlighted by the recent review by Huser and Wadsworth (2022).

Most existing studies on functional extremes focus on the continuous case, following in the footsteps of seminal works on Max-stable processes (De Haan (1984); De Haan and Ferreira (2006)): the random objects under study are random functions in the space $\mathcal{C}[0,1]$, *i.e.*, the space of continuous functions on $[0,1]$ endowed with the supremum norm. In the Peaks-over-Threshold (PoT) framework, the focus lies on the asymptotic distribution of rescaled observations, conditioned on their norm exceeding a threshold, as this threshold tends to infinity. The extremality of an observation is measure using its supremum norm. The resulting limit process in this context is a Generalized Pareto process (see, *e.g.*, Ferreira and de Haan (2014)). Unlike finite-dimensional contexts, defining extremes in infinite-dimensional spaces necessitates selecting a specific norm due to non-equivalence among norms. This choice holds significant practical relevance; for example, in flood risk assessment, it might be more pertinent to analyze total daily precipitation rather than maximum daily precipitation over a short period. This critical norm selection motivates the research of Dombry and Ribatet (2015), who explore alternative definitions of extreme events through a homogeneous cost functional, leading to the development of r-Pareto processes. Additional details and precise definitions about extremes in $\mathcal{C}[0,1]$ are provided in the dedicated Section 2.2.2.

Some exceptions to the continuous case exist, *e.g.*, the functional Skorokhod space $\mathbb{D}[0,1]$ equipped with the J_1- topology has been considered in several works (see Davis and Mikosch (2008); Hult and Lindskog (2005) and the references therein), and upper-semicontinuous functions equipped with the Fell topology are considered in Resnick and Roy (1991); Molchanov and Strokorb (2016); Sabourin and Segers (2017); Samorodnitsky and Wang (2019).

A classic assumption in EVT suited for the PoT framework is to assume that the observed random variable X is regularly varying, that is that the rescaled distribution $t^{-1}X$, conditioned on a excess of its norm above a threshold $\|X\| \geq t$ converges to some limit random variable X_∞, as the threshold grows to infinity, *i.e.*, $\mathcal{L}(t^{-1}X \mid \|X\| \geq t) \to \mathcal{L}(X_\infty)$ as $t \to +\infty$ (see the monographs Resnick (1987, 2007) for a comprehensive presentation of Regular Variation (RV) in the multivariate case). Hult and Lindskog (2006b) extend the notion of RV, initially defined on a Euclidean space, to measures on complete and separable metrics spaces. In this context, the authors characterizes the RV of a random element X through two conditions: the real RV of its norm $\|X\|$ and the weak convergence of its angle $\Theta = \|X\|^{-1}X$ given that $\|X\|$ exceeds a threshold $\|X\| \geq t$ towards a limit angular random element Θ_∞ as the threshold tends to infinity, $\mathcal{L}(\Theta \mid \|X\| \geq t) \to \mathcal{L}(\Theta_\infty)$ as $t \to +\infty$ (see, *e.g.*, Segers et al. (2017); Davis and Mikosch (2008)).

While RV theory has been extensively studied in $\mathcal{C}[0,1]$ and has strong theoretical bases in general metric spaces, it has received far less attention in $L^2[0,1]$, the space of square-integrable real-valued function over $[0,1]$, and more generally, in general separable Hilbert spaces. We propose in Chapter 5 to formalize this concept, in the framework of Hult and Lindskog (2006b).

One of the primary interests for working in a separable Hilbert space is to consider the principal component decomposition of a random element (see Section 3.3 in Chapter 3 for details). Extreme Value Analysis (EVA) of functional PCA with L^2-valued random functions has already been considered in the literature, from a quite different perspective however, leaving some questions unanswered. In Kokoszka and Xiong (2018), the authors assume RV of the scores of a principal component decomposition, (*i.e.*, the random coordinates of the observations projected onto a L^2-orthogonal family) and they investigate the extremal behavior of their empirical counterparts. In Kokoszka et al. (2019) and Kokoszka and Kulik (2023), RV is assumed and various convergence results regarding the empirical covariance operators of the random function X (not the angular component Θ) are established, under the condition that the RV index belongs to some restricted interval, respectively $2 < \alpha < 4$ and $0 < \alpha < 2$. In Kim and Kokoszka (2022), extremal dependence between the scores of the functional PCA of X is investigated. They prove on this occasion that RV in $L^2[0,1]$ implies multivariate RV of finite-dimensional projections of X. However, the converse of this conditional statement is not investigated. Kim and Kokoszka (2024) generalize the notion of correlation coefficient for functional extremes.

The aforementioned works involve PCA for extremes of $L^2[0,1]$-random function, in one way or another, but there has been limited exploration of the principal component decomposition of a regularly varying element. In Chapter 6, under RV of X, we propose a investigation of the convergence of the PCA associated with Θ towards the PCA of Θ_∞. Herein, the value of the RV index is unimportant as the PCA that we consider is that of the *angular component* Θ of the random functions.

1.2.2 Dimension reduction for extremes

The advancement in data acquisition devices has led to an increase in the availability of massive measurements, which both motivates the development of statistics for functional data and presents challenges. On one hand, more available data allows for more precise studies. On the other hand, analyzing high-dimensional data is challenging due to the difficulty of identifying the informative parts and the heavy computational demands of processing these data in complex machine learning tasks. This ambivalence in high-dimensional statistics is often referred to as the "curse of dimensionality" (Giraud (2021)). In applications such as neuroscience and image processing, where data dimensions can explode, it becomes crucial to reduce the dimensionality to retain only essential characteristics.

In recent years, interests in high dimensional EVT problems have surged. Since EVA focuses on a restricted part of the data, the effective size of the dataset used for inference can be relatively limited, highlighting the importance of dimension reduction techniques adapted to extreme settings. An active line of research concerns unsupervised dimension reduction for which a variety of methods have been proposed over the past few years, some of them assorted with non-asymptotic statistical guarantees relying on suitable concentration inequalities. Examples of such strategies include

identification of a sparse support for the limiting distribution of appropriately rescaled extreme observations (Goix et al. (2017); Simpson et al. (2020); Meyer and Wintenberger (2021); Drees and Sabourin (2021); Cooley and Thibaud (2019); Medina et al. (2021)), graphical modeling and causal inference based on the notion of tail conditional independence (Hitz and Evans (2016); Segers (2020); Gnecco et al. (2021)), clustering (Chautru (2015); Janßen and Wan (2020); Chiapino et al. (2020)), see also the review paper Engelke and Ivanovs (2021). In these works, the dimension of the sample space, although potentially high, is finite, and dimension reduction is a key step, if not the main purpose, of the analysis.

EVA characterizes the behavior of extreme data, which are far from the center of the distribution. This makes EVA tools naturally suitable for developing anomaly detection procedures, as outliers also lie outside the bulk of the distribution. Dimension reduction for extremes aims to uncover regions that encapsulate the essence of large data. Algorithms like DAMEX (Goix et al. (2016, 2017)) and CLEF (Chiapino et al. (2020); Chiapino and Sabourin (2016)) identify subspaces where components of the observed vector can be large together. An anomaly is thus detected when data points do not belong to these subspaces despite having a large norm. Another approach characterizes outliers as observations lying outside extreme MV-sets (which can be sought among outcome spaces of the CLEF or DAMEX algorithms), that are small volumes but large masses (Thomas et al. (2017)).

A classic dimension reduction technique in signal processing involves decomposing data onto a basis selected according to the problem at hand and then retaining the most important components. Common base families include Fourier (Example 3.4) and wavelets bases. The reader is invited to consult the non-exhaustive and easy-to-read book Mallat (1999) for more details about signal processing. The beneficial properties of those bases are wide and various but choosing a precise base tailored for a problem sometimes feels like looking for a needle in a haystack. Hence, for tasks where the data structure is linear, or requires efficient dimensionality reduction without losing important features, PCA can be particularly advantageous. PCA automatically determines a set of orthogonal components that capture the maximum variance in the data, providing a simplified representation that often aligns well with the underlying structure of the data (refer to Mallat (1999) for comparison of PCA, Fourier and wavelets bases and Hsing and Eubank (2015) or Section 3.3 for backgrounds on PCA).

Several works have applied PCA to EVT across the years. In infinite-dimensional settings, studies such as Kokoszka and Xiong (2018); Kokoszka et al. (2019); Kokoszka and Kulik (2023); Kim and Kokoszka (2022) have explored PCA for functional extremes but none propose a method to apply PCA specifically to extreme data. These works have already been mentioned and presented in the previous section. To the best of our knowledge, only two works involve a PCA for extremes in finite-dimensional settings. In Cooley and Thibaud (2019), the authors propose a PCA of a matrix composed by pairwise tail dependency coefficients of a positive-orthant-valued regularly varying random vector, outcome of a transformation of a regularly varying random vector valued in the whole ambient space. In Drees and Sabourin (2021), authors investigate the relationships between the PCA of the random angle Θ, the angular component of a $\mathbb{R}^d$-valued regularly varying random vector X, and the PCA of its extreme limit Θ_∞, since the RV of X implies $\mathcal{L}(\Theta \mid \|X\| \geq t) \to \mathcal{L}(\Theta_\infty)$ as $t \to +\infty$. A key argument in their proof is that Θ belongs to the unit sphere of $\mathbb{R}^d$ which is a compact set. Following Riesz's lemma, their proof techniques cannot be extended to infinite-dimensional space

while the mathematical objects involved in this article are defined *mutatis mutandis* in general separable Hilbert space. The purpose of Chapter 6 aims to extend their results to non-finite-dimensional spaces by circumventing the compactness argument by proving that the eigenstructure of Θ converges towards the eigenstructure of Θ_∞ under the RV assumption of the random element X in a separable Hilbert space.

1.2.3 Extremes for sea levels

Sea levels can be decomposed into a deterministic tidal component and a stochastic non-tidal component, which corresponds to storm surges. Storm surges are defined as the instantaneous differences between astronomically predicted tides and observed sea levels. Large storm surges are caused by low atmospheric pressure and strong wind conditions (intensity or direction). When these meteorological conditions coincide with the high water levels of spring tides, they can lead to devastating floods. A notable example is the North Sea flood of 1953, known in Dutch as the *Watersnoodramp*, which resulted in over 2000 deaths across Northern Europe (McRobie et al. (2005)). Studying such events to infer their intensity and frequency is therefore a crucial challenge in coastal risk monitoring, aiming to prevent significant human and material losses (Genovese and Przyluski (2013); Chadenas et al. (2014); Karamouz et al. (2019)). This task is even more critical in the context of global warming, which increases both the frequency and the amplitude of such extreme events (see Seneviratne et al. (2021)).

The study of extreme sea levels has been and continues to be an active research area for decades. A central concept in this field is the inference of return levels, which correspond to the maximum levels expected over a specified period ("inverse" of the return period, as detailed in Coles et al. (2001)). Two pioneering studies in this domain are Lennon et al. (1963) and Suthons (1963), which utilize annual maxima methods to infer them. Since relying solely on annual maxima restricts the data available, new extreme methods have been developed. Smith (1986) and Tawn (1988) introduce the use of r-annual maxima, while Davison and Smith (1990) are the first to consider exceedances over a threshold as extremes. These studies employ direct methods, which involve analyzing the sea levels directly, without accounting for their structure into a deterministic tidal component and a stochastic surge component.

Indirect methods involve separately analyzing the tidal and surge components. These methods are often preferred over direct approaches because they require fewer data to efficiently perform an extreme value study. Convolution methods, for instance, allow the consideration of extreme sea levels by combining extreme storm surges with extreme tidal levels. As highlighted in Dixon and Tawn (1999), direct methods can induce additional estimation errors. Early work in this field includes Pugh and Vassie (1978) and Pugh and Vassie (1980), which introduced the joint probabilities method to combine storm surges to sea levels by means of convolution. However, these studies assumed that hourly surges were independent, a notion deemed unrealistic by Tawn et al. (1989). To address this, Tawn (1992) proposed the revised joint probabilities method, incorporating the extremal index (Leadbetter (1982)) to account for temporal dependence. For a comprehensive comparison between direct and indirect methods, refer to Haigh et al. (2010).

Modeling the tide-surge dependence in indirect methods can be challenging (Idier et al. (2012)). Consequently, skew surges, defined as the difference between the maximum observed sea levels during a tide and the maximum predicted astronomical sea levels

during the same tide, are often used instead. This approach is advantageous because high tides generally do not impact skew surges (see Williams et al. (2016)). In this line of thoughts, Batstone et al. (2013) proposed the skew surge joint probabilities method to circumvent modeling the tide-surge interaction. Note that the independence between high tide and skew surge has been empirical proven for most of the French coast station, with the exception of the Saint-Malo station (see Kergadallan et al. (2014); Kergadallan (2022)).

All the aforementioned methods are applied individually to each measurement station, called *tide gauge*, ignoring the spatial dependence between stations. This can be a significant limitation, as the occurrence of an extreme event at one location increases the likelihood of another extreme event at a nearby station. The extreme value community has been interested in modeling the multivariate dependence structure for years. The literature in this domain typically focuses on models for either asymptotically dependent or asymptotically independent data. Asymptotic dependence is evaluated using the dependence measure (Coles et al. (1999)). Broadly speaking, extremes in asymptotically independent regimes tend to occur separately, while extremes in asymptotically dependent regimes are likely to occur simultaneously. For asymptotically dependent data, the seminal work is the celebrated conditional model of Heffernan and Tawn (2004), which characterizes the distribution of a random vector given that one of its components is extreme. This work has been refined over the years, leading to numerous conditional models, such as those proposed by Keef et al. (2013), Tawn et al. (2018), and Shooter et al. (2021).

While some of the aforementioned models also apply to asymptotically independent data, other models are better suited to capture strong connections between components, such as the hierarchical max-stable model of Reich and Shaby (2012), the generalized Pareto process of Ferreira and de Haan (2014), and the multivariate generalized Pareto distribution of Rootzén and Tajvidi (2006) (see also Kiriliouk et al. (2019); Rootzén et al. (2018)). The above list of multivariate extreme value models is not exhaustive, given the extensive number of existing models. Additional references include Davison et al. (2012); Huser and Wadsworth (2022) for advances in spatial extremes, Engelke and Ivanovs (2021) for sparse structures, Hao et al. (2018) for compound extremes, and de Carvalho and Ramos (2012) for bivariate asymptotically independent data.

1.3 Summary of contributions

This section aims to summarize the main results of the thesis, leaving the motivation and contextualization to Section 1.2 and the introductory sections at the beginning of each chapter.

Chapter 5 focuses on characterizing the Regular Variation in a Hilbert space. Chapter 6 relies upon the formalism introduced in Chapter 5 to obtain consistency and statistical guarantees for the PCA of regularly varying elements in a Hilbert space. The materials of Chapters 5 and 6 have been published in the peer-reviewed journal *Stochastic Processes and their Applications* (see Clémençon et al. (2024)). The main findings of this research project are summarized in Sections 1.3.1 and 1.3.2.

Chapter 7 proposes a novel regularly varying framework, namely regular variation with respect to a component, integral to the formalization of a regression setup for extremes. In Chapter 8, we develop results for regression in extremes, proving the

optimality of a regression function depending solely on the angle of the input, the consistency of this estimator, and statistical guarantees on the error associated with this estimator. The material presented in Chapters 7 and 8 has been the subject of a pre-publication Huet et al. (2023) which is under review in a peer-reviewed journal. A summary of the main findings is provided in Sections 1.3.3 and 1.3.4.

Part IV proposes two approaches to predicting extreme sea levels: the first approach is based on the regression framework developed in Chapter 8, and the second approach is based on a multivariate generalized Pareto density model. The material of Chapter 9 is the subject of an ongoing journal submission and is summarized in Section 1.3.5.

1.3.1 Regular variation in Hilbert spaces

The main purpose Chapter 5 is to develop a general probabilistic framework for extremes of regularly varying element in a separable Hilbert space $\mathbb{H}$, as the space $L^2[0,1]$, the Hilbert space of square-integrable, real-valued functions over $[0,1]$, with immediate possible generalization to other compact domains, *e.g.*, spatial ones.

In the present work we place ourselves in the general RV context defined through $\mathbb{M}_0$-convergence in Hult and Lindskog (2006b), and we focus our analysis on random functions valued in the Hilbert space $L^2[0,1]$, which has received far less attention (at least in EVT) than the spaces of continuous, semi-continuous or *càdlàg* functions. One main advantage of the proposed framework, in addition to allowing for rough function paths, is to pave the way for dimension reduction of the observations *via* functional PCA of the *angular* component Θ (see Chapter 6).

Several questions arise. First, when dealing with functional observations, the choice of the norm (thus of a functional space) is not indifferent, since not all norms are equivalent. In particular, there is no reason why RV in one functional space (say, $\mathcal{C}[0,1]$) would be equivalent to RV in a larger space such as $L^2[0,1]$. Also a recurrent issue in the context of weak convergence of stochastic processes is to verify tightness conditions in addition to weak convergence of finite-dimensional projections, in order to ensure weak convergence of the process as a whole. The case of Hilbert valued random variables makes no exception (see, *e.g.*, Chapter 1.8 in van der Vaart and Wellner (1996)). A natural question to ask is then: 'What concrete conditions regarding the angular and radial components $(\Theta, \|X\|)$ in a PoT framework, which may be verified in practice on specific generative examples or even on real data, are sufficient in order to ensure tightness and thus RV?'.

First, to address these questions, we present a comprehensive description of the notion of RV in a separable Hilbert space which fits into the framework of Hult and Lindskog (2006b). Specifically, we propose characterizations of RV involving finite-dimensional projections and moments of the angular variable Θ through the following first main result (Theorem 5.8 in Chapter 5).

Theorem. *Let X be a random element in $\mathbb{H}$ and let Θ_t be a random element in $\mathbb{H}$ distributed on the sphere $\mathbb{S}$ according to the conditional angular distribution $P_{\Theta,t} := \mathcal{L}(X/\|X\| \,|\, \|X\| \geq t)$. Let $P_{\Theta,\infty}$ denote a probability measure on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ and let Θ_∞ be a random element distributed according to $P_{\Theta,\infty}$. The following statements are equivalent.*

1. *X is regularly varying with index α with limit angular measure $P_{\Theta,\infty}$, so that $P_{\Theta,t} \xrightarrow{w} P_{\Theta,\infty}$.*

2. $\|X\|$ *is regularly varying in* $\mathbb{R}$ *with index* α, *and*

$$\forall h \in \mathbb{H}, \langle \Theta_t, h \rangle \xrightarrow{w} \langle \Theta_\infty, h \rangle \quad \text{as } t \to +\infty.$$

3. $\|X\|$ *is regularly varying in* $\mathbb{R}$ *with index* α, *and*

$$\forall N \geq 1, \pi_N(\Theta_t) \xrightarrow{w} \pi_N(\Theta_\infty) \quad \text{as } t \to +\infty,$$

with $\pi_N : \mathbb{H} \to \mathbb{R}^N$ *the projection onto the* N *first elements of a basis* $(e_i)_{i \geq 1}$.

To validate this framework, we provide several examples of regularly varying random elements in $\mathbb{H}$, such as random sums $\sum_{i=1}^{D} Z_i A_i$ where the Z_i's are real-valued regularly varying random variables, the A_i's are random elements in $\mathbb{H}$ and D a constant or a real-valued random valued with finite first moment (Propositions 5.1 and 5.2). We emphasize the necessity of tightness conditions for achieving global RV by constructing a random element in $\mathbb{H}$ with regularly varying finite-dimensional projections and norm, which is not regularly varying in $\mathbb{H}$ (Proposition 5.4). In the final Section of the chapter, we discuss the relationships between RV in $\mathcal{C}[0,1]$ and RV in $L^2[0,1]$. We demonstrate that RV in $\mathcal{C}[0,1]$ implies RV in $L^2[0,1]$ and we show that the limit random variables in both settings can be connected through an explicit formula, as per the results in Dombry and Ribatet (2015) (Proposition 5.9). The converse, however, is not true, as we illustrate by constructing a regularly varying random function in $L^2[0,1]$ that is not regularly varying in $\mathcal{C}[0,1]$ (Proposition 5.10).

1.3.2 PCA for functional extremes

A major feature of the proposed framework in Chapter 5 is the possibility to project the observations onto a finite-dimensional functional space, *via* a modification of the standard functional PCA which is suitable for heavy-tailed observations, for which second (or first) moments may not exist.

In this respect the dimension reduction strategy that we propose may be seen as an extension of Drees and Sabourin (2021), who worked in the finite-dimensional setting and derived finite sample guarantees regarding the eigenspaces of the empirical covariance operator for Θ. However their techniques of proof cannot be leveraged in the present context because they crucially rely on the compactness of the unit sphere in $\mathbb{R}^d$, while the unit sphere in an infinite-dimensional Hilbert space is not compact.

Regarding the PCA of the angular distribution, the natural extension of the finite-dimensional covariance matrix of extreme angles $C_{t,\mathbb{R}^d} = \mathbb{E}\left[\Theta\Theta^\top \mid \|\mathbf{X}\| > t\right]$ in Drees and Sabourin (2021) where $X \in \mathbb{R}^d$, is the covariance operator $C_t = \mathbb{E}[\Theta \otimes \Theta \mid \|X\| > t]$ when $X \in \mathbb{H}$, see Sections 3.1 and 3.3 of Chapter 3 for minimal background regarding probability in Hilbert spaces and covariance operators. Under RV of X, so that $P_{\Theta,t} := \mathcal{L}(\Theta \mid \|X\| \geq t) \to \mathcal{L}(\Theta_\infty) =: P_{\Theta,\infty}$, one may wonder whether the eigenstructure of C_t indeed converges as $t \to +\infty$ to that of $C_\infty = \mathbb{E}[\Theta_\infty \otimes \Theta_\infty]$, where $\Theta_\infty \sim P_{\Theta,\infty}$, and whether the results of Drees and Sabourin (2021) regarding concentration of the empirical eigenspaces indeed extend to the infinite-dimensional Hilbert space setting. We make a first step to answer these interrogations by proving the following approximation result (Theorem 6.1 in Chapter 6).

Theorem. *The following convergence in the Hilbert-Schmidt norm holds true,*

$$\|C_t - C_\infty\|_{HS(\mathbb{H})} \to 0,$$

as $t \to +\infty$.

Using Theorem 3 in Zwald and Blanchard (2005) (Theorem 3.19) and the Weyl's inequality (Theorem 3.11), we prove that, under unequivocal definition, the eigenvalues and the eigenspaces of C_t converge to those of C_∞ (Corollary 6.3).

Secondly, we investigate the convergence of the empirical counterpart of the non-asymptotic covariance operator associated with distribution the $P_{\Theta,t}$. In the situation where $n \geq 1$ independent realizations $X_1, ..., X_n$ of the random function X are observed, we aim to estimate the sub asymptotic covariance operator associated with a radial threshold $t_{n,k}$, that is a quantile of the radial variable $\|X\|$ at level $1 - k/n$, given by $C_{t_{n,k}} := \mathbb{E}[\Theta \otimes \Theta \mid \|X\| \geq t_{n,k}]$. To do so, we consider the empirical covariance operator

$$\widehat{C}_k := \frac{1}{k} \sum_{i=1}^{n} \Theta_i \otimes \Theta_i \mathbb{1}\{\|X_i\| \geq \hat{t}_{n,k}\},$$

where $\hat{t}_{n,k}$ is the k-th larger norm of the $\|X_i\|$'s, that is an empirical counterpart of the quantile $t_{n,k}$. We provide statistical guarantees in the form of concentration inequalities regarding the Hilbert-Schmidt norm of the estimation error, which leading terms involve the number $k \leq n$ of extreme order statistics considered to compute the estimator. This is given by the following estimation result (Theorem 6.8 in Chapter 6).

Theorem. *Let* $\delta \in (0,1)$. *With probability larger than* $1 - \delta$, *we have*

$$\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})} \leq C(\delta)/\sqrt{k} + +o(1/\sqrt{k}),$$

with $C(\delta)$ *is a constant depending only on* δ.

These bounds, combined with regular variation of the observed random function X and the results from the preceding section ensure in particular the consistency of the empirical estimation procedure (Corollary 6.10 in Chapter 6).

In the final section of this chapter, we present experimental results using both real and simulated data. Specifically, we analyze an electricity demand dataset and simulated data, as detailed in Chapter 5. These experiments demonstrate the relevance of the proposed dimension reduction framework by comparing its performance with the closest alternative, namely, the PCA applied to the full sample (not limited to extreme observations).

1.3.3 A Regular Variation Framework for Regression on Extremes

In Chapter 7, we introduce a probabilistic framework, specifically regular variation with respect to a component, to address Regression on Extremes. We also propose a dedicated algorithmic approach, which is further analyzed in Chapter 8.

To motivate the subsequent analysis, the Regression On eXtreme ANglEs (ROXANE) algorithm is introduced at the beginning of Part III. This method addresses the regression problem for the input/output pair $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ in extreme regions, specifically where $\|\mathbf{X}\| \gg 1$. The algorithm's core objective is to minimize an empirical extreme quadratic risk using only the angle of the input variable $\mathbf{X}/\|\mathbf{X}\|$. This is achieved without loss of information under a specific RV assumption, as justified in Chapter 8. For simplicity, we assume the output is bounded, *i.e.*, there exists $M > 0$ such that $Y \in I := [-M, M]$.

Our central hypothesis, which underpins the ROXANE algorithm, is RV with respect to the first component of $(\mathbf{X}, Y)$. This modified RV assumption, where the extremality of the random vector is defined solely with respect to the input variable, is detailed in the following Definition (Assumption 7.2 in Chapter 7). Let $E := \mathbb{R}^d \backslash \{\mathbf{0}\}$.

Definition (Regular variation w.r.t. the first component)**.** *A random vector* $(\mathbf{X}, Y) \in \mathbb{O} := E \times I$ *is regularly varying w.r.t. the first component with index* $\alpha > 0$*, if there exist a regularly varying function* b *with index* α *and a nonzero Borel measure* μ *on* $\mathbb{O}$*, on all Borel subsets of* $\mathbb{O}$ *bounded away from* $\mathbb{C} = \{\mathbf{0}\} \times I$*, such that*

$$\lim_{t \to +\infty} b(t) \mathbb{P}\Big(t^{-1}\mathbf{X} \in A, Y \in C\Big) = \mu(A \times C),$$

for all $A \in \mathcal{B}(E)$ *bounded away from zero and* $C \in \mathcal{B}(I)$ *such that* $\mu(\partial(A \times C)) = 0$.

This assumption is a particular case of the theory developed in Lindskog et al. (2014) for regularly varying measures on separable metric spaces with a closed set $\mathbb{C}$ removed; in our context, $\mathbb{C} = \{\mathbf{0}\} \times I$. We clarify this connection in Section 7.5 with equivalent statements of RV with respect to the first component. Similar implications as those of classic RV are demonstrated following Assumption 7.2, such as the homogeneity of order $-\alpha$ of the limit measure μ with respect to the first component, which leads to a decomposition of the measure:

$$\mu(\{\mathbf{X} \in E : \|\mathbf{x}\| \geq r, \boldsymbol{\theta}(\mathbf{x}) \in B\} \times C) = r^{-\alpha} \Phi(B \times C),$$

with $\boldsymbol{\theta}(\mathbf{x}) := \mathbf{x}/\|\mathbf{x}\|$ and for all $C \in \mathcal{B}(I), B \in \mathcal{B}(\mathbb{S})$, $r > 0$. This entails the existence of a pair of limit random variables $(\mathbf{X}_\infty, Y_\infty)$

$$\mathcal{L}(t^{-1}\mathbf{X}, Y \mid \|\mathbf{X}\| \geq t) \to \mathcal{L}(\mathbf{X}_\infty, Y_\infty),$$

as $t \to +\infty$. We further assume the convergence of the Bayes regression function $f^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ towards the limit Bayes regression function $\mathbb{E}[Y_\infty \mid \mathbf{X}_\infty]$ that is (Assumption 7.5)

$$\mathbb{E}\Big[|f^*(\mathbf{X}) - f^*_{P_\infty}(\mathbf{X})| \mid \|\mathbf{X}\| \geq t\Big] \to 0. \tag{1.1}$$

Three conditions implying this assumption, such as the uniform convergence of densities proposed in De Haan and Resnick (1987), are provided to support the validity of Equation (1.1) (Proposition 7.6 in Chapter 7). Finally, we propose four practical scenarios that satisfy all the assumptions in Section 7.4, including an example that reliably predicts a missing extreme component in a regularly varying random vector (Proposition 7.10 in Chapter 7).

1.3.4 Regression on extremes

Chapter 8 aims to develop a regression framework for extreme covariates, under the assumptions discussed in Chapter 7, such as RV w.r.t. the first component of input/output pair $(\mathbf{X}, Y)$. The underlying goal is then to legitimize the ROXANE algorithm, that is to prove that a regression function can be optimally constructed in extreme regions, using only the angle of the input variable.

Regression is a crucial predictive problem in statistical learning, encompassing a wide variety of applications. In the standard setup, the predictive learning problem consists in building, from a training dataset $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ composed of $n \geq 1$

independent copies of two random variables $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$, a mapping $f : \mathcal{X} \to \mathbb{R}$ in order to produce a 'good' prediction $f(\mathbf{X})$ for Y, with the quadratic risk

$$R_P(f) = \mathbb{E}\Big[(Y - f(\mathbf{X}))^2\Big] \tag{1.2}$$

as close as possible to that of the Bayes regression function $f^*(X) = \mathbb{E}[Y \mid \mathbf{X}]$, which minimizes (1.2).

A natural strategy consists in solving the Empirical Risk Minimization problem (ERM in abbreviated form) $\min_{f \in \mathcal{F}} R_{\hat{P}_n}(f)$, where $\mathcal{F}$ is a class of functions sufficiently rich to include a reasonable approximant of f^* and $\hat{P}_n$ is an empirical version of P based on $\mathcal{D}_n$.

This chapter addresses regression in extreme regions, where the input variable is extreme. Covariates are considered extreme when their norm $\|\mathbf{X}\|$ exceeds an (asymptotically) large threshold $t > 0$ (see Chapter 7). The choice of the norm is typically determined by the application context.

The threshold t depends on the observations, as 'large' should be understood relative to the majority of observed data. Consequently, extreme observations are rare and underrepresented in the training dataset. As a result, prediction errors in extreme regions generally have a negligible impact on the global regression error of $\hat{f}$. Indeed, the law of total probability yields:

$$R_P(f) = \mathbb{P}(\|\mathbf{X}\| \geq t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big] + \mathbb{P}(\|\mathbf{X}\| < t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| < t\Big].$$

The above decomposition involves a conditional error term relative to excesses of $\|\mathbf{X}\|$ above t, which we term *conditional quadratic risk* (or simply *conditional risk*)

$$R_t(f) := \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big].$$

It is the purpose of the subsequent analysis to construct a predictive function $\hat{f}$ that (approximately) minimizes $R_t(f)$ for all $t > t_0$, with t_0 being a large threshold. It is important to note that an approximate minimizer of R_t might not be suitable for minimizing $R_{t'}$ when $t' > t$. To ensure robust extrapolation performance for our learned function, we focus on obtaining a prediction function, $\hat{f}$, that minimizes the *asymptotic conditional quadratic risk* defined as

$$R_\infty(f) := \limsup_{t \to +\infty} R_t(f) = \limsup_{t \to +\infty} \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big].$$

Thus, the objective is to establish connections between the risks R_t and R_∞ and their respective minimizers, leveraging the advantageous properties of the RV assumption. The following theorem (part of Theorem 8.2 in Chapter 8) provides initial motivations for the ROXANE algorithm. Denote $\boldsymbol{\theta}(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$.

Theorem. *Under Assumptions 7.1, 7.2 and 7.5, the two following statements hold true.*

1. *As $t \to +\infty$, the minimum value of R_t converges to that of R_∞, i.e.,* $\inf_f R_t(f) \to \inf_f R_\infty(f)$.
2. *the infimum of R_∞ over all measurable function is equal to its infimum over all angular measurable functions, i.e.,* $\inf_f R_\infty(f) = \inf_h R_\infty(h \circ \boldsymbol{\theta})$.

The use of RV w.r.t. the first component is crucial for proving this result. This key concept enables the connection between the two risks, R_t and R_∞ and offers the angular nature of a minimizer of R_∞. Consequently, it is reasonable to restrict the search for a minimizer in extreme regions to angular functions, as suggested by the ROXANE algorithm. Our strategy involves solving the ERM optimization problem associated with $\min_{h\in\mathcal{H}} R_{t_{n,k}}(h\circ\boldsymbol{\theta})$ with $t_{n,k}$, a quantile of the radial variable $\|\mathbf{X}\|$ at level $1-k/n$. To achieve this, we consider its empirical counterpart

$$\hat{R}_k(f) = \frac{1}{k}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2\,\mathbb{1}\{\|\mathbf{X}_i\| \geq \hat{t}_{n,k}\},$$

where $\hat{t}_{n,k}$ is the k-th larger norm of the $\|\mathbf{X}_i\|$'s, serving as an empirical counterpart of the quantile $t_{n,k}$. As with traditional ERM, we investigate the minimization of the empirical risk over a class of functions with controlled complexity. Let $\mathcal{H}$ be a class of continuous, real-valued, angular and uniformly bounded by M functions $f \in \mathcal{C}(\mathbb{S}, I)$. To fully validate the empirical strategy of the ROXANE algorithm, we provide a non-asymptotic bound on the maximal deviation between $R_{t_{n,k}}$ and $\hat{R}_k$ over $\mathcal{H}$, as given in the following theorem (Theorem 8.4 in Chapter 8).

Theorem. *Suppose that Assumptions 7.1 and 8.3 are satisfied. Let $\delta \in (0,1)$. We have with probability larger than $1-\delta$*

$$\sup_{h\in\mathcal{H}}\left|\hat{R}_k(h\circ\boldsymbol{\theta}) - R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right| \leq C(\mathcal{H}, M, \delta)/\sqrt{k} + o(1/\sqrt{k}),$$

where $C(\mathcal{H}, M, \delta)$ is a constant depending on $\mathcal{H}, M$ and δ.

Additionally, with an extra assumption about the class of functions $\mathcal{H}$, which is satisfied in particular by regression functions involved in constrained Ridge and Lasso regression (Remark 8.6 in Chapter 8), we show that the bias term $\sup_{h\in\mathcal{H}}|R_{t_{n,k}}(h\circ\boldsymbol{\theta}) - R_\infty(h\circ\boldsymbol{\theta})|$ converges to zero as $n \to +\infty$ (Proposition 8.5 in Chapter 8). This leads to a maximal control over the excess of R_∞-risk of a regression function produced by the ROXANE algorithm (Corollary 8.8 in Chapter 8).

1.3.5 Modeling and Prediction of Extreme Sea Levels

Chapter 9 focuses on the task of predicting extreme sea levels and skew surges at tide gauges along the French Atlantic coast. We propose to learn the extreme spatial dependence structure among observations from various stations over their common time range. This learned model is then utilized to reconstruct sea levels and skew surges at locations with limited historical records, based on extreme observations from nearby stations with extensive temporal records. Specifically, we aim to predict values at the Port-Tudy station given extreme values from the Brest and Saint-Nazaire stations (see Figure 9.1). An observation is declared extreme if at least one of its input components is extreme, since a single large sea level or skew surge can cause flooding at the recorded station, regardless of conditions at the other stations. We outline two different procedures for learning the extreme dependence structure.

In the first method, we fit an appropriate extreme distribution to the data. Given the clear asymptotic dependence observed in the data (see Figures 9.3 and 9.4), we model the observations using a Multivariate Generalized Pareto (MGP) distribution H

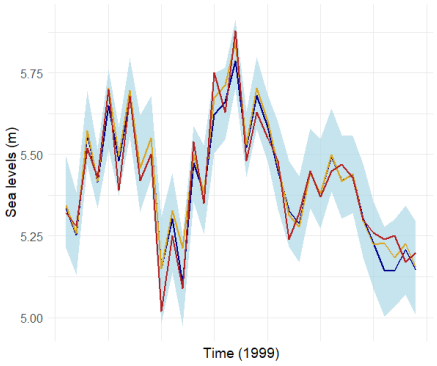


Figure 1.2: Predicted sea levels at the Port-Tudy station for the year 1999. The red curve represents the true values on the test set; the orange curve represents the predicted values by the ROXANE procedure with OLS algorithm; the blue curve represents the predicted values by the MGP procedure with bootstrap 0.95 confidence intervals (light blue).

(Rootzén and Tajvidi (2006); Rootzén et al. (2018)), defined as

$$H(\mathbf{x}) = \frac{\log G(\mathbf{x} \wedge \mathbf{0}) - \log G(\mathbf{x})}{\log G(\mathbf{0})},$$

where G is a multivariate extreme value distribution (Definition 2.10). Specifically, we follow the parametric fitting procedure of Kiriliouk et al. (2019), leveraging the convenient decomposition of the MGP distribution (2.14). The final predictions are obtained by averaging Monte Carlo simulations generated from the conditional density given the two input values.

In the second approach, we use the regression procedure proposed in Part III (Huet et al. (2023)), which is designed for extreme value prediction problems where the extremality is measured w.r.t. covariates - values from the long-term stations. We learn a prediction function using the ROXANE algorithm (Algorithm 7.1) over the common time range of the data, which predicts values at the output stations based on extreme values at the input stations.

Both procedures require rescaling of the marginal observations to a common scale: unit exponential scales for the MGP procedure and unit Pareto distributions for the ROXANE procedure. Following Legrand et al. (2023), we use an Extended Generalized Pareto (EGP) distribution as our marginal model to meet the different requirements described in the introductory part of Chapter 9. Specifically, we consider the model EGPD3 from Papastathopoulos and Tawn (2013) with cdf

$$F_{\sigma,\xi,\kappa}(x) = \left(1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}\right)^{\kappa},$$

with $\sigma > 0$, $\xi \in \mathbb{R}$, $\kappa \in \mathbb{R}$ and $x \in [0, +\infty[$ if $\xi \geq 0$ and $x \in [0, -\sigma/\xi]$ otherwise.

The multivariate prediction procedures are synthesized into two comprehensive algorithms, Algorithm 9.2 and 9.3. In addition to the marginal pre-processing steps, Algorithm 9.1 introduces a novel method for selecting appropriate thresholds within the extreme value studies, based on properties of the EGP distribution.

The proposed methods are applied to the sea level data and their performance is evaluated in terms of Root Mean Square Error and Mean Absolute Error on a test set consisting of the earliest extreme observations. Both multivariate prediction procedures yield valid results of significant importance for practitioners, each offering distinct advantages: one provides better point estimates, while the other offers a robust generative model. In particular, Figure 1.2 shows a visual assessment of the prediction quality of both methods (Table 9.3), and Figure 9.5 presents QQ-plots for further validation. Finally, similar studies carried out for the Concarneau and Le Crouesty stations are shown in Appendix 9.A.

1.4 Outline of the thesis

The thesis manuscript is organized as follow.

Chapter 1 provides a summary of the state-of-the-art and the contributions of this thesis.

Part I introduces the necessary background for understanding and proving the thesis results.

Chapter 2 covers basic notions of Extreme Value Theory, ranging from univariate to infinite-dimensional extremes, including multivariate extremes.

Chapter 3 discusses Functional Data Analysis concepts, encompassing operators and probability theory on Hilbert spaces.

Chapter 4 outlines the basics of Statistical Learning, with a particular focus on its applications to extremes.

Part II concerns Hilbertian extremes.

Chapter 5 develops the theory of Regular Variation in separable Hilbert spaces.

Chapter 6 uses the rationale of Chapter 5 to establish consistency and concentration results for PCA in the context of functional extremes.

Part III studies the task of regression in extreme regions.

Chapter 7 presents a novel regularly varying framework to handle extremes w.r.t. some component.

Chapter 8 exploits the framework of Chapter 7 to develop a novel framework suitable for regression in extreme regions.

Part IV is an application to reconstruction of extreme sea levels.

Chapter 9 applies the extreme regression procedure of Chapter 8 and an extreme modeling procedure to extreme sea level data from tide gauges along the French Atlantic coast.

The manuscript ends with a discussion on the global conclusions and perspectives of the results developed in this thesis, followed by an appendix section that includes technical proofs and an introduction in french.

The materiel of this thesis relies on the following works

Part II: Clémençon, S., Huet, N., and Sabourin, A., (2024) Regular Variation in Hilbert Spaces and Principal Component Analysis for Functional Extremes, *Stochastic Processes and their Applications*, 174, 104375;

Part III: Huet, N., Clémençon, S., and Sabourin, A., (2024) On Regression in Extreme Regions, arXiv:2303.03084 (submitted).

Part I

Background and Preliminaries

Chapter 2

Extreme Value Theory

Contents

Extreme Value Theory (EVT) is a branch of statistics that focuses on the study of rare events, known as *extremes*. These extremes are unusual observations that significantly deviate from the majority of other observations. The normality of an observation can be measured by any norm, and an observation is considered extreme if its norm exceeds a large threshold. In practical scenarios, there may be few or no observations in the extreme region of interest. EVT addresses this issue by developing extrapolation models, which enable inferences about unseen phenomena. These models are crucial for risk monitoring and are widely applied in various fields such as finance, insurance, environmental sciences, and climatology. The findings presented in this chapter draw primarily from several key books and articles: Resnick (1987), De Haan and Ferreira (2006), Hult and Lindskog (2006b) and Resnick (2007). For a comprehensive and accessible introduction to EVT, we particularly recommend the two books by Resnick (1987, 2007). EVT is central to this thesis and plays a crucial role in all the results presented in the contributions.

This chapter is structured as follows. Section 2.1 addresses finite-dimensional extremes. Specifically, Section 2.1.1 presents the Generalized Extreme Value distribution and the concept of regular variation in $\mathbb{R}$. These concepts are extended to the multivariate case $\mathbb{R}^d$ in Section 2.1.2, where the notions of *exponent* and *angular* measures are introduced. Section 2.2 then explores EVT in infinite dimensions, a less extensively studied area. Section 2.2.1 examines regular variation in general metric spaces, focusing on the notion of $\mathbb{M}_0$-convergence as developed by Hult and Lindskog (2006b). Finally, Section 2.2.2 takes a detailed look at extreme random processes with sample paths in $\mathcal{C}[0,1]$, which are the most studied functional extremes.

2.1 Finite-dimensional Extremes

2.1.1 Univariate extremes

The results presented in this section are borrowed from Sections 0.2 and 0.3 in Resnick (1987). One of the main purpose of univariate EVT is to characterize the limit of the maxima of a random variable. Let X be a random variable distributed according to F and let $(X_i)_{i\geq 1}$ be independent and identically distributed copies of X. Set $M_n = \bigvee_{i=1}^n X_i$. The distribution of M_n is given by F^n since

$$\mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x, ..., X_n \leq x) = F^n(x).$$

Let $x_0 = \sup\{x \in \mathbb{R}, F(x) < 1\}$ be the right-endpoint of F. It is then easy to see that

$$\lim_{n\to+\infty} \mathbb{P}(M_n \leq x) = \begin{cases} 0 & \text{if } x < x_0 \\ 1 & \text{if } x \geq x_0. \end{cases}$$

In view of this equation, an affine normalized version of the maximum of the sample has to be considered to obtain a non-degenerate limit distribution.

Definition 2.1 (Maximum Domain of Attraction)**.** *Suppose there exist sequences $(b_n)_{n\geq 1}$ and $(a_n)_{n\geq 1}$, with $a_n > 0$, and a random variable Z distributed according to G, so that*

$$\mathcal{L}\left(\frac{M_n - b_n}{a_n}\right) \to \mathcal{L}(Z),$$

or equivalently in terms of distributions,

$$F^n(a_n x + b_n) \underset{n\to+\infty}{\longrightarrow} G(x). \tag{2.1}$$

In this case, X (or F) is said to belong to the maximum domain of attraction of Z (or G).

The Maximum Domain of Attraction is abbreviated as MDA. The elegance of EVT lies in the fact that if a distribution G fulfills such a condition, its form is determined. This fundamental result of EVT is known as the Fisher-Tipett-Gnedenko theorem (Fisher and Tippett (1928) and Gnedenko (1943)).

Theorem 2.2 (Fisher-Tippett-Gnedenko)**.** *Suppose there exist sequences $(b_n)_{n\geq 1}$ and $(a_n)_{n\geq 1}$, with $a_n > 0$, and a non-degenerate distribution G such that*

$$F^n(a_n x + b_n) \underset{n\to+\infty}{\longrightarrow} G(x),$$

then G is one of the following three forms (up to a rescaling input factor)

1. *Fréchet:* $G(x) = \begin{cases} 0 & \text{if } x < 0 \\ \exp(-x^{-\alpha}) & \text{if } x \geq x_0 \end{cases}$, *for some* $\alpha > 0$.
2. *Weibull:* $G(x) = \begin{cases} \exp(-(-x)^{-\alpha}) & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$, *for some* $\alpha > 0$.
3. *Gumbel:* $G(x) = \exp(-\exp(-x))$ *for* $x \in \mathbb{R}$.

These distributions are called the extreme value distributions.

The three extreme value distributions can be encapsulated into one common distribution which is the only possible distribution at the limit in Equation (2.1), namely the Generalized Extreme Value (GEV) distribution

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

for some $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$, where the case $\xi = 0$ has be understood as its limit as $\xi \to 0$. In the light of Theorem 2.2, if $\xi = 0$, $G_{\mu,\sigma,\xi}$ is of Gumbel type, if $\xi > 0$, $G_{\mu,\sigma,\xi}$ is of Fréchet type, and if $\xi < 0$, $G_{\mu,\sigma,\xi}$ is of Weibull type. μ referred to as a location parameter; σ referred to as a scale parameter; ξ referred to as a shape parameter. Figure 2.1 illustrates their forms for fixed parameters $\mu = 0$ and $\sigma = 1$.

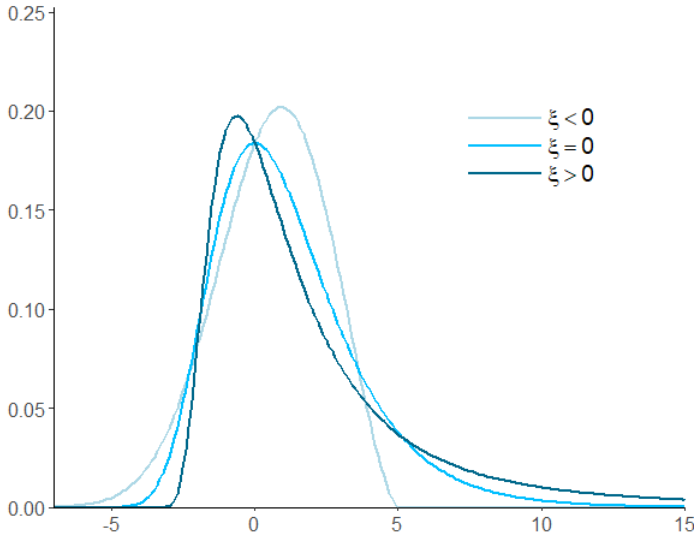


Figure 2.1: GEV probability density functions for $\mu = 0$ and $\sigma = 1$.

There are two possible perspectives in EVT. The first one, aligned the aforementioned results, which is to consider extremes as maxima over pre-defined blocks, called *Block maxima* approach. The second perspective is to consider extremes as observations exceeding a large threshold, called *Peaks-over-Threshold* (PoT) approach. The two points of view are essentially equivalent. This relationship is expressed by the fact that the MDA-assumption in Equation (2.1) is first equivalent to the following statement

$$\lim_{n\to+\infty} n\mathbb{P}\left(\frac{X-b_n}{a_n} \geq x\right) = -\log(G(x)), \tag{2.2}$$

which is equivalent to the convergence of the conditional distribution of excesses above a threshold, through the following theorem (see Balkema and De Haan (1974)).

Theorem 2.3. *The following statements are equivalent.*

1. *there exist* $(b_n)_{n\geq 1}$ *and* $(a_n)_{n\geq 1}$ *with* $a_n > 0$ *such that* $F^n(a_n x + b_n) \to \exp(-(1+\xi x)_+^{-1/\xi}$, *as* $n \to +\infty$.

2. *there exists a function* $\sigma : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ *such that*

$$\mathbb{P}\left(\frac{X-t}{\sigma(t)} \geq x \mid X \geq t\right) \to (1+\xi x)_+^{-1/\xi},$$

as $t \to x_0$, *with* x_0 *the upper end-point of the distribution* F.

Figure 2.2 illustrates the PoT and the Block Maxima approaches on a dataset, provided by Météo-France, composed by temperatures at Orly Airport from 2020 to 2022.

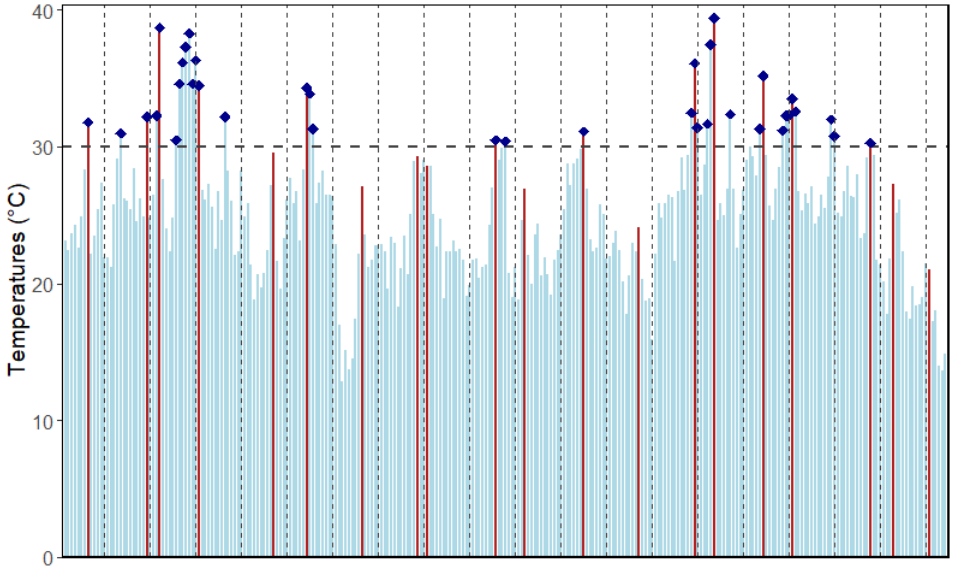


Figure 2.2: Block Maxima model *vs* Peaks-over-Threshold model: red bars represent the maximum observations within 14-day blocks; blue points indicate observations exceeding the 30°C threshold.

From this convergence, observe that a new central distribution appears at the limit, namely the *Generalized Pareto* (GP) distribution, with cdf

$$H_{\mu,\sigma,\xi}(x) = 1 - \left(1 + \xi \frac{x-\mu}{\sigma}\right)_+^{-1/\xi}, \tag{2.3}$$

with same parameters than the GEV distribution. Figure 2.3 illustrates their forms for fixed parameters $\mu = 0$ and $\sigma = 1$.

To fully understand the study of right-tail points of a distribution, it is essential to use tools provided by a theory closely related to EVT, known as the theory of Regular Variation (RV). For a comprehensive exposition of the notion of RV, the reader is referred to Bingham et al. (1989).

Definition 2.4 (Regularly varying function)**.** *A measurable function* $f : \mathbb{R}_+ \to \mathbb{R}_+$ *is regularly varying with index* $\rho \in \mathbb{R}$, *written* $f \in RV_\rho$, *if for all* $x > 0$

$$\lim_{t \to +\infty} \frac{f(tx)}{f(t)} \to x^\rho.$$

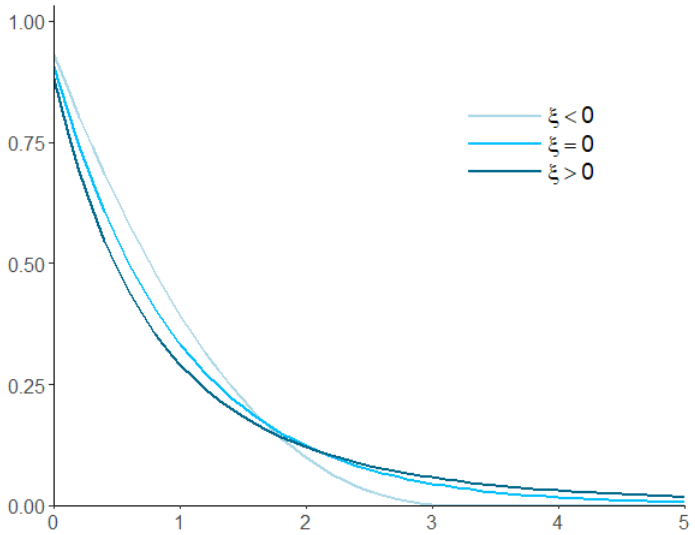


Figure 2.3: GP probability density functions for $\mu = 0$ and $\sigma = 1$.

In the case where $\rho = 0$, the function f is *slowly varying*. It is easy to see that every regularly varying function $f \in RV_\rho$ can be represented as $f(x) = x^\rho L(x)$ where L is a slowly varying function. The most fundamental theorem of this theory is called the Karamata's theorem Karamata (1933).

Theorem 2.5 (Theorem 2.1 in Resnick (2007))**.** *(a) Suppose $\rho \geq -1$ and $f \in RV_\rho$. Then $\int_0^t f(s)ds \in RV_{\rho+1}$ and*

$$\lim_{t\to+\infty} \frac{tf(t)}{\int_0^t f(s)ds} = \rho + 1.$$

If $\rho < -1$ (or if $\rho = -1$ and $\int_t^{+\infty} f(s)ds < +\infty$), then $f \in RV_\rho$ implies that $\int_t^{+\infty} f(s)ds$ is finite, $\int_t^{\infty} f(s)ds \in RV_{\rho+1}$, and

$$\lim_{t\to+\infty} \frac{tf(t)}{\int_t^{+\infty} f(s)ds} = -\rho - 1.$$

(b) If f satisfies

$$\lim_{t\to\infty} \frac{tf(t)}{\int_0^t f(s)ds} = \lambda \in (0, +\infty),$$

then $f \in RV_{\lambda-1}$. If $\int_t^{+\infty} f(s)ds < +\infty$ and

$$\lim_{t\to+\infty} \frac{tf(t)}{\int_t^{+\infty} f(s)ds} = \lambda \in (0, +\infty)$$

then $f \in RV_{-\lambda-1}$.

The theory of RV finds several links and application in the probabilistic theory (see Chapter VIII in Feller (1991)). This concept extends to random variables through RV of its survival function.

Definition 2.6 (Regularly varying univariate random variable)**.** *A nonnegative random variable X distributed according to F is regularly varying with index $\alpha \geq 0$, written $X \in RV_{-\alpha}$, if its survival function $1-F$ is regularly varying with index $-\alpha$, i.e., for all $x > 0$*

$$\lim_{t \to +\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\alpha}.$$

The index α in the above definition is referred to as the *tail index* of X. The relationship between the MDA and RV conditions is illustrated, for instance, by the fact that the RV of a random variable is equivalent to belonging to the MDA of a Fréchet distribution.

Example 2.7 (Pareto distribution)**.** *Let $X \sim Pareto(\alpha)$, for some $\alpha > 0$, i.e., $\mathbb{P}(X \geq x) = x^{-\alpha}$. Then, $X \in RV_{-\alpha}$.*

Example 2.8 (Fréchet distribution)**.** *Let $X \sim Fréchet(\sigma, \xi)$, for some $\sigma > 0$ and $\xi > 0$. Then, $X \in RV_{-1/\xi}$.*

Example 2.9 (Generalized Pareto distribution)**.** *Let $X \sim GPD(\sigma, \xi)$, for some $\sigma > 0$ and $\xi > 0$, i.e., $\mathbb{P}(X \geq x) = (1 + \xi x/\sigma)_+^{-1/\xi}$. Then, $X \in RV_{-1/\xi}$.*

A final remark concerning the moments of a regularly varying random variable (see Proposition 1.3.2 in Mikosch (1999)), is that if $X \in RV_{-\alpha}$, then

$$\begin{cases} \mathbb{E}[X^\beta] < +\infty \text{ if } \beta < \alpha, \\ \mathbb{E}[X^\beta] = +\infty \text{ if } \beta > \alpha. \end{cases}$$

There are numerous examples of regularly varying random variables. This simple result can, for example, be used to prove that a random variable is not regularly varying by demonstrating that it either all its moments exist or none of them exist. This property will be utilized in Chapter 5.

2.1.2 Multivariate extremes

Univariate EVT extends to the multivariate case through straightforward generalizations of the MDA and RV assumptions. However, multivariate extremes exhibit much deeper structures. The results and definitions in this section are primarily sourced from Section 5 of Resnick (1987) and Section 6 of Resnick (2007). In what follows, the classic order relation is adapted to $\mathbb{R}^d$: $\mathbf{a} \geq \mathbf{b}$ means for every $j \in \{1, ..., d\}, a_j \geq b_j$ and $\mathbf{a} \not\leqslant \mathbf{b}$ means there exists $j \in \{1, ..., d\}, a_j > b_j$. Univariate operations applied to multivariate objects should be understood componentwise, , *e.g.*, $\mathbf{a}/\mathbf{b} = (a_1/b_1, ..., a_d/b_d)$. Let $\mathbf{X} = (X_1, ..., X_d) \in \mathbb{R}^d$ be a random vector distributed according to F and let $(\mathbf{X}_i)_{i \geq 1}$ be i.i.d. copies of $\mathbf{X}$. Set $\mathbf{M}_n = \bigvee_{i=1}^n \mathbf{X}_i$, where the maximum is taken componentwise.

Definition 2.10 (Multivariate Maximum Domain of Attraction)**.** *Suppose there exist sequences $(\mathbf{b}_n)_{n \geq 1} \in (\mathbb{R}^d)^{\mathbb{N}}$ and $(\mathbf{a}_n)_{n \geq 1} \in (\mathbb{R}^d)^{\mathbb{N}}$, with $\mathbf{a}_n > \mathbf{0}$, and a non-degenerate random vector $\mathbf{Z} \in \mathbb{R}^d$ distributed according to G, so that*

$$\mathcal{L}\left(\frac{\mathbf{M}_n - \mathbf{b}_n}{\mathbf{a}_n}\right) \to \mathcal{L}(\mathbf{Z}) \text{ as } n \to +\infty,$$

or equivalently in terms of distributions,

$$F^n(\mathbf{a}_n \mathbf{x} + \mathbf{b}_n) \underset{n \to +\infty}{\longrightarrow} G(\mathbf{x}). \tag{2.4}$$

In this case, $\mathbf{X}$ *(or* F*) is said to belong to the multivariate maximum domain of attraction of* $\mathbf{Z}$ *(or* G*).*

The limit distributions G are referred to as *Multivariate Extreme Value* distributions. These distributions are more complex than their univariate counterparts: while the marginal distributions of G are univariate extreme value distributions, the joint dependence structure is not predefined, even though it is determined by a limit measure μ. To better understand this measure, similar to the univariate case, the behavior of the upper points of the distribution must be examined using the theory of RV. The transition between the two perspectives is assured by the analogue of Equation (2.2): convergence in Equation (2.4) is equivalent to

$$\lim_{n\to+\infty} n\mathbb{P}\left(\frac{\mathbf{X}-\mathbf{b}_n}{\mathbf{a}_n} \not\leqslant \mathbf{x}\right) = -\log(G(\mathbf{x})).$$

The multivariate RV theory is mandatory to study right-tail distribution. Let $E := \mathbb{R}^d \setminus \{\mathbf{0}\}$ be the punctured Euclidean space. Denote by $\|\cdot\|$ a norm on $\mathbb{R}^d$; by $\mathbb{B}(0,r) := \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq r\}$ the ball of radius $r > 0$ and by $\mathbb{S} := \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| = 1\}$ the unit sphere. A set $A \subset E$ is said to be bounded away zero if there exists $\varepsilon > 0$ such that $A \cap \mathbb{B}(0,\varepsilon) = \emptyset$.

Definition 2.11 (Multivariate regularly varying random variable)**.** *A random vector* $\mathbf{X} \in \mathbb{R}^d$ *is regularly varying with index* $\alpha \geq 0$*, written* $\mathbf{X} \in RV_{-\alpha}(\mathbb{R}^d)$*, if there exist a regularly varying function* b *with index* α *and a nonzero Borel measure* μ *on* E*, finite on all Borel subsets of* E *bounded away from zero, so that*

$$b(t)\mathbb{P}(\mathbf{X} \in tA) \underset{t\to+\infty}{\longrightarrow} \mu(A), \tag{2.5}$$

for any Borel set $A \subset E$ *bounded away from zero and such that* $\mu(\partial A) = 0$*.*

The latter convergence is referred to as vague convergence in $[-\infty,+\infty]^d \setminus \{\mathbf{0}\}$ (see Resnick (2007), Section 3.3 and in particular, Theorem 3.2 for a Portmanteau Theorem), or equivalently as $\mathbb{M}_0$-convergence in E exposed in the next section (see Hult and Lindskog (2006b); Lindskog et al. (2014)). The limit measure μ is called the *exponent measure*. In particular, This measure is homogeneous of degree $-\alpha$, *i.e.*, $\mu(rA) = r^{-\alpha}\mu(A)$. This property decomposes the structure of μ into a radial part and a angular part. Define the *angular measure* (or the *spectral measure*) Φ, which is a finite positive measure on $\mathbb{S}$, given by

$$\Phi(B) = \mu(\{\mathbf{x} \in E : \|\mathbf{x}\| \geq 1, \mathbf{x}/\|\mathbf{x}\| \in B\}), \tag{2.6}$$

for $B \in \mathcal{B}(\mathbb{S})$. The angular measure Φ fully characterizes the dependence structure in extremes. The homogeneity of μ entails the decomposition

$$\mu(\{\mathbf{x} \in \mathbb{R}^d, \mathbf{x}/\|\mathbf{x}\| \in B, \|\mathbf{x}\| \geq r\}) = r^{-\alpha}\Phi(B),$$

for $r \geq 1$. Equivalent statements of multivariate RV are then given in terms of this polar decomposition. Denote by $\boldsymbol{\theta}(\mathbf{x}) := \mathbf{x}/\|\mathbf{x}\|$ the angle of any $\mathbf{x} \in \mathbb{R}^d$.

Theorem 2.12 (Regularly varying random variable)**.** *Let* $\alpha > 0$*. The following statements are equivalent.*

1. $\mathbf{X} \in RV_{-\alpha}(\mathbb{R}^d)$ *with exponent measure* μ*;*

2. *There exists a measure* Φ *on* $\mathbb{S}$ *such that*

$$\frac{\mathbb{P}(\boldsymbol{\theta}(\mathbf{X}) \in B, \|\mathbf{X}\| \geq tr)}{\mathbb{P}(\|\mathbf{X}\| \geq t)} \underset{t\to+\infty}{\longrightarrow} cr^{-\alpha}\Phi(B), \tag{2.7}$$

for all $r > 0$ *and* $B \in \mathcal{B}(\mathbb{S})$ *such that* $\Phi(\partial B) = 0$, *with* $c = \Phi(\mathbb{S})^{-1}$;

3. $\|\mathbf{X}\| \in RV_{-\alpha}$ *and there exists a measure* Φ *on* $\mathbb{S}$ *such that*

$$\mathbb{P}(\boldsymbol{\theta}(\mathbf{X}) \in B \mid \|\mathbf{X}\| \geq t) \underset{t\to+\infty}{\longrightarrow} c\Phi(B),$$

for all $B \in \mathcal{B}(\mathbb{S})$ *such that* $\Phi(\partial B) = 0$, *with* $c = \Phi(\mathbb{S})^{-1}$.

In particular, the measure Φ *is same in* 2. *and* 3., *that is the angular measure associated with* μ, *given in Equation* (2.6).

Observe that the limit measure in Equation (2.7) defines a probability measure on $\mathbb{R}_{\geq 1} \times \mathbb{S}$. Then, if $\mathbf{X} \in RV_{-\alpha}(\mathbb{R}^d)$, there exists a limit random variable $\mathbf{X}_\infty \in \mathbb{R}^d$ such that

$$\mathcal{L}(t^{-1}\mathbf{X} \mid \|\mathbf{X}\| \geq t) \underset{t\to+\infty}{\longrightarrow} \mathcal{L}(\mathbf{X}_\infty). \tag{2.8}$$

Following Equation (2.7), one can decompose the limit random variable $\mathbf{X}_\infty = \|\mathbf{X}_\infty\| \times \boldsymbol{\theta}(\mathbf{X}_\infty)$ with $\|\mathbf{X}_\infty\| \in Pareto(\alpha)$ and $\boldsymbol{\theta}(\mathbf{X}_\infty) \in \mathbb{S}$ a.s., where $\|\mathbf{X}_\infty\|$ and $\boldsymbol{\theta}(\mathbf{X}_\infty)$ are independent.

To simplify the study of the multivariate extremes, without loss of generality, it is convenient to work with standardized marginals. One possible transformation is given by the Pareto (or Fréchet) marginal transformation given by

$$\mathbf{V}(\mathbf{X}) = \left(\frac{1}{1 - F_1(X_1)}, \ldots, \frac{1}{1 - F_d(X_d)}\right), \tag{2.9}$$

where F_j is j-th marginal distribution of $\mathbf{X}$, for $1 \leq j \leq d$. This marginal transformation is particularly interesting to work with (see Proposition 5.10 in Resnick (1987)) since the RV of $\mathbf{X}$ implies the RV of $\mathbf{V}(\mathbf{X})$ with scaling function $b(t) = t$, and an exponent measure that is homogeneous of degree -1. Furthermore, all the marginals of $\mathbf{V}(\mathbf{X})$ are on the same scale which is a necessary property for defining the regression framework in Section 7. Specifically, this transformation set all marginal distributions to unit Pareto distributions. Because the distribution is unknown in practice, an empirical marginal transformation is usually performed that is

$$\hat{\mathbf{V}}(\mathbf{x}) = \left(\frac{1}{1 - \frac{n}{n+1}\hat{F}_1(x_1)}, \ldots, \frac{1}{1 - \frac{n}{n+1}\hat{F}_d(x_d)}\right), \tag{2.10}$$

with $\hat{F}_j(x_j) = (1/n)\sum_{i=1}^n \mathbb{1}\{X_{ij} \leq x_j\}$, for $1 \leq j \leq d$ (see Remark 7.3 for more details).

Another approach involves standardizing the margins to the exponential scale, which is particularly suited for a specific type of limit distribution known as the *Multivariate Generalized Pareto* (MGP) distribution (see Rootzén and Tajvidi (2006); Rootzén et al. (2018); Kiriliouk et al. (2019)). To enjoy the advantages of transforming marginal distributions to an exponential scale, it is essential to first introduce the MGP distribution.

Denote by $\boldsymbol{\eta}$ the vector of lower-end point the corresponding GEV distribution. From the classic MDA assumption (2.10), one can deduce that

$$\mathcal{L}\left(\frac{\mathbf{X}-\mathbf{b}_n}{\mathbf{a}_n} \vee \boldsymbol{\eta} \mid \mathbf{X} \not\leqslant \mathbf{b}_n\right) \to \mathbf{W}, \tag{2.11}$$

as $n \to +\infty$, where $\mathbf{W}$ follows a MGP distribution with cdf H. The limit distributions G and H are *associated*, that is

$$H(\mathbf{x}) = \frac{\log G(\mathbf{x} \wedge \mathbf{0}) - \log G(\mathbf{x})}{\log G(\mathbf{0})}.$$

The positive part of the marginal distributions of the vector $\mathbf{W}$ are univariate GP distributions,

$$\mathbb{P}(W_j \geq x \mid W_j \geq 0) = 1 - H_{0,\sigma_j,\xi_j}(x) = \left(1 + \frac{\xi_j}{\sigma_j}x\right)_+^{-1/\xi_j} \tag{2.12}$$

where σ_j and ξ_j are the GP parameters of G_j and H_{0,σ_j,ξ_j} is a GP cdf (2.3) for $1 \leq j \leq d$. Regarding marginal scaling, for the purpose of multivariate modeling instead of using a Pareto scaling, it is more advantageous to set each marginal's scale and shape parameters to one and zero, respectively. This involves applying the marginal transformation to the exponential scale:

$$e_{\sigma,\xi}(x) = -\log(1 - H_{0,\sigma,\xi}(x)) = \frac{1}{\xi}\log\left(\left(1 + \frac{\xi}{\sigma}x\right)_+\right), \tag{2.13}$$

to each margin with their respective GP parameters. Note that this transformation, when applied to each margin, results in a unit exponential distribution for their positive part, *i.e.*,

$$\mathbb{P}\left(e_{\sigma_j,\xi_j}(W_j) \geq x \mid e_{\sigma_j,\xi_j}(W_j) \geq 0\right) = \exp(-x),$$

for $1 \leq j \leq d$.

Similarly to the classic limit distributions with RV and a multiplicative polar structure, MGP distributions also admit a convenient structure. More precisely, Theorem 7 in Rootzén et al. (2018) states that, a MGP random variable $\tilde{\mathbf{W}}$ with margins on the exponential scale decomposes as

$$\tilde{\mathbf{W}} = E + \mathbf{T} - \max(\mathbf{T}), \tag{2.14}$$

where $E \in \mathbb{R}$ is a unit exponential random variable and $\mathbf{T} \in \mathbb{R}^d$ is a random vector, independent of E.

Remark 2.13 (Parallel structures)**.** *The representations of a MGP vector in Equation* (2.13) *and of a limit regularly varying vector in Equation* (2.14) *are not contradictory; rather, they describe the same object using different choices of marginal distributions. By applying the logarithm function in the exponential scale transformation, the classic multiplicative polar structure is converted into an additive form. Specifically, the common unit exponential random variable in Equation* (2.14) *corresponds to the Pareto random variable* $\|\mathbf{X}_\infty\|$ *in Equation* (2.8)*. The random vector* $\mathbf{T} - \max(\mathbf{T})$*, often termed a spectral vector, is the analogue to the angular random variable* $\boldsymbol{\theta}(\mathbf{X}_\infty)$*, which is sometimes also referred to as a spectral vector, and thus governs the joint dependence structure of the limit variable.*

A deep investigation of the properties of the density of a MGP vector are performed in Rootzén et al. (2018). These properties are extensively used in the procedure of Kiriliouk et al. (2019), which is applied in Chapter 9. The following theorems present the guidelines for constructing MGP densities.

Theorem 2.14 (Theorem 12 in Rootzén et al. (2018))**.** *Let* $\tilde{\mathbf{W}}$ *and* $\mathbf{T}$ *be as in* (2.14)*. Suppose that* $\mathbf{T}$ *admits a density* f_T*, then* $\tilde{\mathbf{W}}$ *also admits a density* h_T *given by*

$$h_T(\mathbf{x}) = \mathbb{1}\{\max(\mathbf{x}) > 0\} \exp(-\max(\mathbf{x})) \int_0^{+\infty} \frac{f_T(\mathbf{x} + \log t)}{t} dt. \tag{2.15}$$

Theorem 2.15 (Theorem 13 in Rootzén et al. (2018))**.** *Let* $\mathbf{U}$ *be random vector in* $\mathbb{R}^d$ *such that* $0 < \mathbb{E}[\exp(U_j)] < +\infty$*, for* $1 \leq j \leq d$*. Suppose that* $\mathbf{U}$ *admits a probability density function* f_U*, then the density* h_U *given by*

$$h_U(\mathbf{x}) = \frac{\mathbb{1}\{\max(\mathbf{x}) > 0\}}{\mathbb{E}[\exp(\max(\mathbf{U}))]} \int_0^{\infty} f_U(\mathbf{x} + \log t) dt, \tag{2.16}$$

is a density associated with a standard MGP distribution.

Equipped with these two theorems, densities for MGP vectors $\tilde{\mathbf{W}}$ can be constructed from ordinary densities. Additionally, by Equation (2.13), densities for general MGP vectors $\mathbf{W}$ can be deduced as

$$h_T^{\boldsymbol{\sigma},\boldsymbol{\xi}}(\mathbf{x}) = h_T(e_{\boldsymbol{\sigma},\boldsymbol{\xi}}(\mathbf{x})) \text{ and } h_U^{\boldsymbol{\sigma},\boldsymbol{\xi}}(\mathbf{x}) = h_U(e_{\boldsymbol{\sigma},\boldsymbol{\xi}}(\mathbf{x})). \tag{2.17}$$

Remark 2.16 (Extended Generalized Pareto distribution)**.** *A recurrent challenge in EVT is to determine an appropriate threshold above which univariate extreme distributions are suitably modeled by a GP distribution. To circumvent this difficult threshold selection, distributions suitable for the entire range of data (not just the extremes) that mimic the behavior of a GP distribution in the tails have been proposed. In cases where the lower tail of the observed distribution behaves as a power law, Naveau et al. (2016) propose the Extended Generalized Pareto (EGP) distribution. The simplest one (used in Part IV of the present thesis) has a cdf given by*

$$F_{\sigma,\xi,\kappa}(x) = \left(1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}\right)^{\kappa},$$

which corresponds to the third family of distributions in Papastathopoulos and Tawn (2013). The parameter κ *controls the lower tail of the distribution, as* $F_{\sigma,\xi,\kappa}(x) \approx cst \times x^{\kappa}$*, as* $x \to 0$*. Notably, data thresholded as in Equation* (2.11) *are more likely to be well-fitted by an EGP distribution than by a classic GP distribution, since only the positive parts of the MGP margins follow a GP distribution. Figure 2.4 illustrates their forms for fixed parameter* $\sigma = 1$*.*

2.2 Infinite-dimensional Extremes

2.2.1 Regular variation in complete separable metric spaces

We recall here the main features of RV in metric spaces, a framework originally introduced by Hult and Lindskog (2006b) as a generalization of the Euclidean case documented in, *e.g.*, Resnick (1987); Bingham et al. (1989); Meerschaert (1984). This framework may be viewed as an adaptation of the 'weak-hash' convergence of boundedly finite measures; one may refer to Section A2.6 in Daley et al. (2003) for further details.

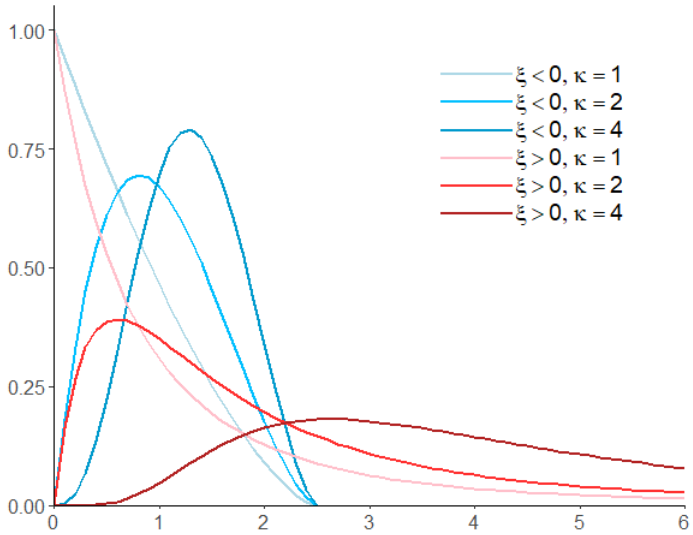


Figure 2.4: EGP probability density functions for $\sigma = 1$.

Let (M, d) be a complete separable metric space, endowed with a multiplication by non-negative real numbers $t > 0$, such that the mapping $(t, x) \in \mathbb{R}_+ \times M \mapsto tx$ is continuous, $1x = x$ and $t_1(t_2 x) = (t_1 t_2)x$. One must assume the existence of an *origin* $0_M \in M$, such that $0x = 0_M$ for all $x \in M$. In Part II, we shall take $M = \mathbb{H}$, a separable, real Hilbert space. Let $M_0 = M \setminus \{0_M\}$. For any subset $A \subset M$, and $t > 0$, we write $tA = \{tx : x \in A\}$. Denote by $\mathcal{C}_0(M)$ the set of bounded and continuous real-valued functions on M_0 which vanish in some neighborhood of 0_M and let $\mathbb{M}_0$ be the class of Borel measures on M_0, which are finite on each Borel subset of M_0 bounded away from 0_M. Then the sequence ν_n *converges to* ν *in* $\mathbb{M}_0$, and we write $\nu_n \xrightarrow{\mathbb{M}_0} \mu$, if $\int f \, d\nu_n \to \int f \, d\nu$ for any $f \in \mathcal{C}_0(M)$. As for the vague (or weak) convergence, there exist versions of the Portmanteau theorem and the mapping theorem for the $\mathbb{M}_0$-convergence (Theorems 2.4 and 2.5 in Hult and Lindskog (2006b)). In particular, to mirror Definition 2.11, it is convenient to note that $\nu_n \xrightarrow{\mathbb{M}_0} \nu$ is equivalent to $\lim_{n \to +\infty} \nu_n(A) = \nu(A)$, for all $A \in \mathcal{B}(M)$ bounded away from zero with $\nu(\partial A) = 0$. A measure ν in $\mathbb{M}_0$ is *regularly varying* if there exist a nonzero measure μ in $\mathbb{M}_0$ and a regularly varying function b such that

$$b(t)\nu(t\,\cdot) \xrightarrow{\mathbb{M}_0} \mu(\cdot), \text{ as } t \to +\infty. \tag{2.18}$$

It follows from (2.18) that the limit measure is necessarily homogeneous, for all $t > 0$ and Borel subset A of M_0, $\mu(tA) = t^{-\alpha}\mu(A)$ for some $\alpha > 0$. Then we say that ν is regularly varying with index $-\alpha$ in M and we write $\nu \in \mathrm{RV}_{-\alpha}(M)$. Hence, RV of $\mathbb{M}_0$-measures extends to random elements valued in M (that is, a Borel measurable map from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to M). Recall incidentally that the extension of the concept of RV to (multivariate) probability measures was originally introduced in Meerschaert (1984).

Definition 2.17 (Regularly varying random element)**.** *A random element $X \in M$ is regularly varying with index $\alpha \geq 0$, written $X \in RV_{-\alpha}(M)$, if there exist a regularly varying function b with index α and a nonzero measure μ in $\mathbb{M}_0$, so that*

$$b(t)\mathbb{P}(X \in tA) \underset{t \to +\infty}{\longrightarrow} \mu(A),$$

for any $A \in \mathcal{B}(M)$ bounded away from zero with $\mu(\partial A)$.

A convenient characterization of regular variation of a random element X is obtained through a polar decomposition. Let $r(x) = d(x, 0_M)$ for $x \in M$. For simplicity, and because it is true in the Hilbert space framework that is our main concern, we focus on the case where the distance to 0_M is homogeneous, although this assumption can be relaxed, as in Segers et al. (2017). Notice that in $\mathbb{H}$, $r(x) = \|x\|$. Introduce a pseudo-angular variable, $\Theta = \theta(X)$ where for $x \in M_0$, $\theta(x) = r(x)^{-1}x$ and let $R = r(X)$. Denote by $\mathbb{S}$ the unit sphere in E relative to r, $\mathbb{S} = \{x \in M : r(x) = 1\}$, equipped with the trace Borel σ-field $\mathcal{B}(\mathbb{S})$. The map $T : \mathbb{M}_0 \to \mathbb{R}_+^* \times \mathbb{S} : x \mapsto (r(x), \theta(x))$ is the polar decomposition. A key quantity throughout this work is the conditional distribution of the angle given that $R > t$ for which we introduce the notation

$$P_{\Theta,t}(\cdot) = \mathbb{P}(\Theta \in \cdot \mid R > t).$$

Several equivalent characterizations of regular variation of X have been proposed in Segers et al. (2017) in terms of the pair of random variable (R, Θ) where $R = r(X)$, thus extending classical characterizations in the multivariate setting, see Resnick (2007). In particular the next statement shall prove to be useful in the subsequent analysis and is the exact analogue of the multivariate characterization 3 in Theorem 2.12 of the previous section.

Proposition 2.18 (Proposition 3.1 in Segers et al. (2017))**.** *A random element X in M is regularly varying with index $\alpha > 0$ if and only if conditions (i) and (ii) below are simultaneously satisfied:*

(i) The radial variable R is regularly varying in $\mathbb{R}$ with index α;

(ii) There exists a probability distribution $P_{\Theta,\infty}$ on the sphere $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ such that $P_{\Theta,t} \xrightarrow{w} P_{\Theta,\infty}$ as $t \to +\infty$.

Remark 2.19. *An interesting fact to note is that, before the clear formalization of RV of measures in metric spaces, authors in Kuelbs and Mandrekar (1974) provided an example of a regularly varying family of measures in a separable Hilbert space. This involves an $\mathbb{H}$-valued random variable that belongs to the domain of attraction of non-Gaussian stable measures. Analogous to the concept of MDA previously presented, a random variable X is said to belong to the domain of attraction of Z, if there exist $(b_n)_{n\geq 1}$ and $(a_n)_{n\geq 1}$, with $a_n > 0$, such that*

$$\mathcal{L}\left(\frac{\sum_{i=1}^n X_i}{a_n} - b_n\right) \to \mathcal{L}(Z),$$

as $n \to +\infty$, where $X_1, ..., X_n$ are i.i.d. copies of X.

A α-stable distribution S with location parameter β and spectral measure Γ, with $0 < \alpha < 2$, is defined by its characteristic functional:

$$\hat{\mu}_S(x) := \mathbb{E}[\exp(i\langle S, x\rangle)] = \exp\left(i\langle x, \beta\rangle - \int_{\mathbb{S}} |\langle x, s\rangle|^\alpha \Gamma(ds) + iC(\alpha, x)\right),$$

with

$$C(\alpha, x) = \begin{cases} \tan\left(\frac{\pi a}{2}\right) \int_{\mathbb{S}(\mathbb{H})} \langle x, s\rangle |\langle x, s\rangle|^{\alpha-1} \Gamma(ds) & \text{if } \alpha \neq 1 \\ \frac{2}{\pi} \int_{\mathbb{S}(\mathbb{H})} \langle x, s\rangle \log|\langle x, s\rangle| \Gamma(ds) & \text{if } \alpha = 1 \end{cases}$$

Then S is regularly varying with index α in $\mathbb{H}$ and $\mathbb{P}(S/\|S\| \in \cdot \mid \|S\| > t) \xrightarrow{w} \Phi(\cdot)$. The spectral measure Γ and the angular measure Φ are linked through the relation: $\Phi(\cdot) = \frac{\Gamma(\cdot)}{\Gamma(\mathbb{S}(\mathbb{H}))}$.

The proof of this result, which is exactly to prove the two points of Proposition 2.18, can be found in Kuelbs and Mandrekar (1974), where it is derived by combining Lemma 4.1 and Theorem 4.11. Notice that the tail index is strictly lower than 2; if $\alpha = 2$, the process would be Gaussian and therefore not regularly varying.

2.2.2 Extremes of $\mathcal{C}[0,1]$-processes

A particular attention has been devoted to extremes in $\mathcal{C}[0,1]$ over the years. Refer to Part III of De Haan and Ferreira (2006) for more details. Section 5.3 compares $\mathcal{C}[0,1]$-RV to $L^2[0,1]$-RV through results from Dombry and Ribatet (2015). Let $X = (X(s))_{s\in[0,1]}$ a stochastic processes with continuous sample-paths and let $(X_i)_{i\geq 1}$ be i.i.d. copies of X. Set $M_n(s) := \bigvee_{i=1}^n X_i(s)$.

Definition 2.20 (Continuous Maximum Domain of Attraction)**.** *Suppose there exist sequences of continuous real-valued functions $(b_n)_{n\geq 1}$ and $(a_n)_{n\geq 1}$, with $a_n(s) > 0$ for all $s \in [0,1]$, and a non-degenerate stochastic process $Z \in \mathcal{C}[0,1]$, so that*

$$\mathcal{L}\left(\frac{M_n - b_n}{a_n}\right) \to \mathcal{L}(Z) \text{ as } n \to +\infty, \tag{2.19}$$

In this case, X is said to belong to the maximum domain of attraction of Z.

For all $s \in [0,1]$, the random variable $Z(s)$ is a univariate extreme value distribution: one can choose $(b_n)_{n\geq 1}$ and $(a_n)_{n\geq 1}$ such that

$$\mathbb{P}(Z(s) \leq x) = \exp(-(1 + \xi(s)x)^{-1/\xi(s)}),$$

with ξ is a continuous function. Similarly to the finite-dimensional case, there exist equivalent formulations of the continuous MDA in terms of convergence of continuous stochastic processes exceeding a threshold. Specifically, Equation (2.19) implies (and is equivalent with additional mild assumptions, see Theorem 3.1 and Theorem 3.2 in Ferreira and de Haan (2014)) that

$$\mathcal{L}\left(\frac{X - b_n}{a_n} \middle| \sup_{s\in[0,1]} \left\{\frac{X(s) - b_n(s)}{a_n(s)}\right\}\right) \to \mathcal{L}(H),$$

as $n \to +\infty$, where H is a *generalized Pareto process*. The process H can be shown to have GP marginal distributions (2.3). Once again, similar to the finite-dimensional case, these statements can be connected to the RV assumption in $\mathcal{C}[0,1]$ (Theorem 9.5.1 in De Haan and Ferreira (2006)).

A key difference from the finite-dimensional scenario lies in the notion of measure convergence, which is characterized by tightness conditions and convergences of the finite-dimensional projections. Ensuring the tightness of a sequence of measures in infinite dimensions is complex and requires careful investigation to identify appropriate conditions. In this context, Hult and Lindskog (2006b) proposes a characterization of RV in $\mathcal{C}[0,1]$ (Theorem 4.4 in Hult and Lindskog (2006b)) using the usual tightness criterion in this space (see Chapter 2 in Billingsley (2013)).

Despite the differences in the concepts of convergence, the RV in $\mathcal{C}[0,1]$ is defined similarly to the multivariate RV : the statements of Theorem 2.12 hold by replacing $\mathbb{R}^d$ by $\mathcal{C}[0,1]$ and by considering the supremum norm (see Hult and Lindskog (2006b); Meinguet and Segers (2010)). In particular, if $X \in RV_{-\alpha}(\mathcal{C}[0,1])$, there exists an angular measure Φ on $\mathbb{S}$ such that Equation (2.7) holds. A key difference with the finite-dimensional case is that the norm measuring the exceedances is restricted to the supremum norm. However, Dombry and Ribatet (2015) shows that the excess function does not necessarily have to be the supremum norm, as shown in the following theorem.

Theorem 2.21 (Theorem 3 in Dombry and Ribatet (2015))**.** *Suppose $X \in RV_{-\alpha}(\mathcal{C}[0,1])$ with angular measure Φ. Let $\ell : \mathcal{C} \to [0,+\infty[$ be an homogeneous function, continuous at the origin and not vanishing Φ-almost everywhere, then there exists a random process $X_{\infty,\ell} \in \mathcal{C}[0,1]$ such that*

$$\mathcal{L}\Big(t^{-1}X \mid \ell(X) > t\Big) \to \mathcal{L}(X_{\infty,\ell}),$$

as $t \to +\infty$.

In addition, they also provide a relationship between the limit extreme distributions arising under RV conditions w.r.t. the supremum norm or the function ℓ (Proposition 2 in Dombry and Ribatet (2015)).

Chapter 3

Theory for Functional Data Analysis

Contents

Functional Data Analysis (FDA) is a statistical framework designed to analyze data where each observation is a function, typically a curve or a surface. These measurements depend on a continuous variable such as time or space. Figure 3.1 displays a functional dataset: the observations represent the evolution of temperature over a week at Orly Airport for the months of July, August, and September from 2020 to 2022. With the increasing availability of functional data, enabled by the improvement of sensors providing ever finer measurements, FDA has become a highly studied field in statistics: ignoring the continuous structure of the data can lead to a significant loss of information. FDA thus offers new perspectives for various applications, from IoT to spectrometry, including predictive maintenance of sophisticated systems (see the two reviews Gertheiss et al. (2023) and Li et al. (2022)). For comprehensive expositions of FDA theory and applications, readers are invited to consult Ramsay and Silverman (2005) and Horváth and Kokoszka (2012). The material in this chapter is mainly drawn from Hsing and Eubank (2015), which provides a self-contained introduction to the mathematical foundations of functional data analysis.

FDA aims to generalize multivariate concepts to the infinite-dimensional case. Hilbert spaces provide a natural generalization of finite-dimensional normed vector spaces, particularly through the convenient extension of the notion of a basis, which allows for practical representations. In addition to Hilbert spaces, the study of linear operators, analogous to matrices in multivariate settings, is essential for building and understanding the theory behind FDA. The results and definitions covered below are crucial for understanding Part II. This chapter is organized as follows.

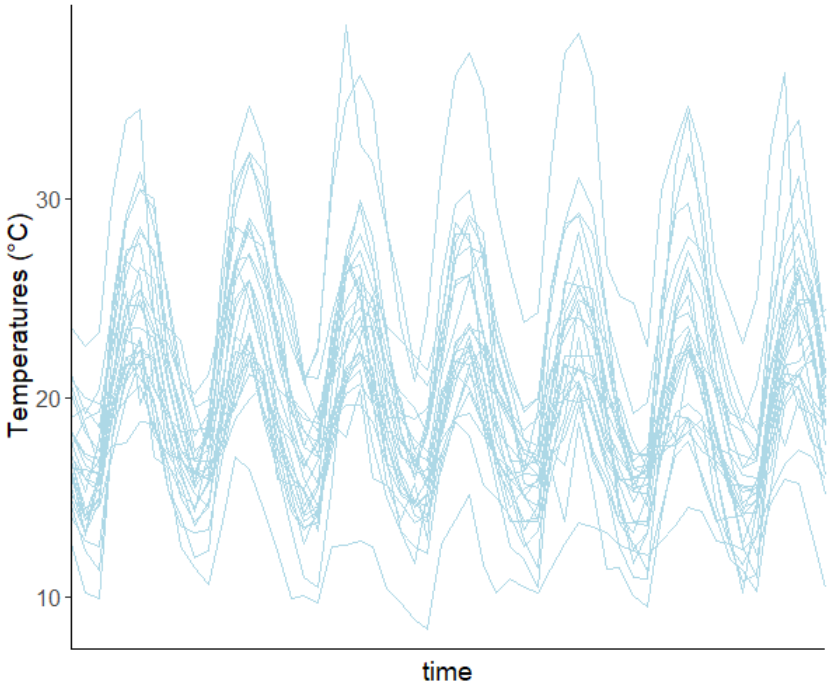


Figure 3.1: Functional dataset comprising temperatures over a week at Orly Airport during July, August, and September from 2020 to 2022.

Section 3.1 introduces basic notions of operator theory, starting with general concepts and facts related to Hilbert spaces and compact operators in Section 3.1.1 and Section 3.1.2. A deeper exposition of a specific class of linear operators, namely *Hilbert-Schmidt* operators, is provided in Section 3.1.3. Chapter 5 aims to propose characterizations of RV adapted to Hilbert spaces, which strongly relies on the concept of weak convergence (see Section 2). The notions of random elements and weak convergence in Hilbert spaces are presented in Sections 3.2 and 3.2.2, respectively. The last section of this chapter discusses two widely used eigendecompositions in particular spaces: the principal components decomposition and the Karhunen-Loève expansion, with a focus on estimation results for the eigenelements involved in principal components analysis in Section 3.3.2.

3.1 Operators on Hilbert Spaces

3.1.1 Basics on Hilbert spaces

The results and definitions provided in this section are primarily drawn from Section 2 of Hsing and Eubank (2015). To begin with, defining a real Hilbert space the concepts of an *inner product* and completeness within a real vector space $\mathbb{V}$. An *inner product* on $\mathbb{V}$ is defined as a function $\langle\cdot,\cdot\rangle$ on $\mathbb{V}\times\mathbb{V}$ satisfying $\langle av_1+v_2,v\rangle = a\langle v_1,v\rangle+\langle v_2,v\rangle$, $\langle v_1,v_2\rangle = \langle v_2,v_1\rangle$ and $\langle v,v\rangle \geq 0$, and $\langle v,v\rangle > 0$ if $v\neq 0$, for $v,v_1,v_2\in\mathbb{V}$ and $a\in\mathbb{R}$. An inner product induces a norm $\|\cdot\|$ such as $\|v\| = \langle v,v\rangle^{1/2}$ for $v\in\mathbb{V}$ satisfying the Cauchy-Schwartz inequality $|\langle v_1,v_2\rangle| \leq \|v_1\|\|v_2\|$, for $v_1,v_2\in\mathbb{V}$. A normed vector space $(\mathbb{V},\|\cdot\|)$ is said to be *complete* if every Cauchy sequence is convergent, *i.e.*, if every sequence $(x_n)_{n\geq 1}$ in $\mathbb{V}$ such that $\sup_{n,m\geq N}\|x_n-x_m\|\to 0$ as $N\to+\infty$ is convergent. Therefore, a Hilbert space can be precisely defined as a complete vector space equipped with an

inner product.

An important advantage of Hilbert spaces over Banach spaces, related to the inner product, is that orthogonality of two elements can be easily defined: $v_1, v_2 \in \mathbb{V}$ are said *orthogonal* if $\langle v_1, v_2 \rangle = 0$. There exist numerous examples of Hilbert spaces.

Example 3.1. *The vector space $\mathbb{R}^d$ equipped with the usual scalar product*

$$\langle x, y \rangle = \sum_{j=1}^{d} x_j y_j,$$

is a Hilbert space.

Example 3.2. *The vector space ℓ^2 of square summable real-valued sequences equipped with the scalar product*

$$\langle (u_n)_{n\geq 1}, (v_n)_{n\geq 1} \rangle = \sum_{n=1}^{+\infty} u_n v_n,$$

is a Hilbert space.

Example 3.3. *The vector space $L^2[0,1]$ of square integrable real-valued function $f : [0,1] \to \mathbb{R}$ equipped with the scalar product*

$$\langle f, g \rangle = \int_0^1 f(x) g(x) dx,$$

is a Hilbert space.

We denote by $\mathbb{H}$ a real Hilbert space equipped with an inner product $\langle \cdot, \cdot \rangle$ and its derived norm $\|\cdot\|$. To offer a comfortable framework, the notion of basis from finite-dimensional vector spaces extends to Hilbert spaces. A sequence $(e_i)_{i\geq 1}$ is called a basis of $\mathbb{H}$ if it spans $\mathbb{H}$, that is every $h \in \mathbb{H}$ decomposes as

$$h = \sum_{i=1}^{+\infty} \langle h, e_i \rangle e_i.$$

The basis is said orthonormal if its elements are of unit norm and pairwise orthogonal. The Hilbert space $\mathbb{H}$ is said to be *separable* if it admits an orthonormal basis. An example of separable Hilbert space is once again given by $L^2[0,1]$.

Example 3.4 (Separable Hilbert space)**.** *The Hilbert space $L^2[0,1]$ equipped with the inner product $\langle f, g \rangle = \int_0^1 f(x) g(x) dx$ is separable. An example of an orthonormal basis of $L^2[0,1]$ is given by the well-known Fourier basis*

$$\{1\} \cup (\sqrt{2}\cos(2\pi n x))_{n\geq 1} \cup (\sqrt{2}\sin(2\pi n x))_{n\geq 1}.$$

3.1.2 Nonnegative compact operators

The concepts and results related to a particular class of linear operators, namely compact operators, are introduced in this section, following the lines of Sections 3 and 4 in Hsing and Eubank (2015). These operators are essential for defining Hilbert-Schmidt operators, central in the framework and findings in Part II. Before discussing

Hilbert-Schmidt operators, we address general linear operators. A linear operator of $\mathbb{H}$ (*i.e.*, from $\mathbb{H}$ to $\mathbb{H}$) is a linear mapping $T : \mathbb{H} \to \mathbb{H}$. In the following, all the operators will be assumed bounded (or equivalently, continuous (Theorem 3.1.2 Hsing and Eubank (2015)): a linear operator T is bounded if its *operator norm* is finite

$$\|T\|_{op} := \sup_{h \in \mathbb{H}} \frac{\langle Th, h \rangle}{\|h\|^2} < +\infty. \tag{3.1}$$

While linear operators can be defined from one Hilbert space $\mathbb{H}_1$ to another $\mathbb{H}_2$, we will focus on the specific case relevant to our purpose: linear operators from and to the same Hilbert space $\mathbb{H}$, which is assumed to be separable hereafter.

Remark 3.5. *To highlight the importance of linear operators, it should be noted that these entities are to Hilbert spaces what matrices are to finite-dimensional vector spaces. In fact, real matrices of size $d_1 \times d_2$ are example of linear operators from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$.*

For the sake of brevity, the study is restricted to nonnegative compact operator. A linear operator T on $\mathbb{H}$ is said to be *compact* if for every bounded sequence $(h_n)_{n \geq 1}$ in $\mathbb{H}$, the sequence $(Th_n)_{n \geq 1}$ contains a convergent subsequence in $\mathbb{H}$.

The set of nonzero eigenvalues of a compact operator form a countable set, where an eigenvalue λ associated with eigenvector φ (or eigenfunction in the case of function spaces such as $L^2[0,1]$) of the operator T satisfies $T\varphi = \lambda\varphi$. A compact operator is said to be *nonnegative* if all its eigenvalues are nonnegative. These eigenvalues can be expressed as the min-max values of optimization problems involving the operator T, where the solutions to these problems are the eigenvectors. This important result is known as the Courant-Fischer theorem.

Theorem 3.6 (Theorem 4.2.7 in Hsing and Eubank (2015))**.** *Let T be a nonnegative, compact operator on $\mathbb{H}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Then,*

$$\lambda_k = \max_{h_1, \ldots, h_k \in \mathbb{H}} \min_{v \in \operatorname{span}\{h_1, \ldots, h_k\}} \frac{\langle Th, h \rangle}{\|h\|^2}$$

and

$$\lambda_k = \min_{h_1, \ldots, h_{k-1} \in \mathbb{H}} \max_{h \in \operatorname{span}(h_1, \ldots, h_{k-1})^{\perp}} \frac{\langle Th, h \rangle}{\|h\|^2},$$

where the maximum and minimum are attained when h is the eigenvector e_k that corresponds to λ_k.

By examining the Courant-Fischer Theorem and the definition of the operator norm (3.1), it becomes evident that the operator norm of a compact nonnegative operator is given by its largest eigenvalue. The cornerstone of the theory of compact operators is found in the following result. If in addition to nonnegative compactness, an operator T is assumed *self-adjoint*, that is $\langle Th_1, h_2 \rangle = \langle h_1, Th_2 \rangle$ for all $h_1, h_2 \in \mathbb{H}$, a eigendecomposition of this operator exists.

Theorem 3.7 (Theorem 4.2.4 in Hsing and Eubank (2015))**.** *Let T be a nonnegative compact, self-adjoint operator on $\mathbb{H}$. The set of nonzero eigenvalues for T is either finite or consists of a sequence which tends to zero. Each nonzero eigenvalue has finite multiplicity and eigenvectors corresponding to different eigenvalues are orthogonal. Let $\lambda_1 \geq \lambda_2 \geq \ldots$ be*

the ordered eigenvalues and let $\varphi_1, \varphi_2, \ldots$ *be the corresponding orthonormal eigenvectors. Then, for all* $h \in \mathbb{H}$

$$Th = \sum_{i=1}^{+\infty} \lambda_i \langle h, \varphi_i \rangle \varphi_i.$$

3.1.3 Hilbert-Schmidt operators

The results and definitions of the presentation are taken from Section 4 in Hsing and Eubank (2015) and Section VIII in Gohberg et al. (2013). We present results from the theory of Hilbert-Schmidt (HS) operators, which are a specific class of compact linear operators (Theorem 4.4.3 in Hsing and Eubank (2015)).

Definition 3.8 (Hilbert-Schmidt operator)**.** *Let* $(e_i)_{i\geq 1}$ *be a orthonormal basis of* $\mathbb{H}$*. A Hilbert-Schmidt operator on* $\mathbb{H}$ *is defined as a linear operator* T *on* $\mathbb{H}$ *satisfying*

$$\sum_{i=1}^{+\infty} \|Te_i\|^2 < +\infty.$$

The family of Hilbert-Schmidt operators on $\mathbb{H}$ *is denoted by* $HS(\mathbb{H})$*.*

Note that Definition 3.8 does not depend on the choice of norm (Theorem 4.4.1 in Hsing and Eubank (2015)). A remarkable property of the set of HS operators $HS(\mathbb{H})$ is that it is itself a Hilbert space equipped with the inner product $\langle T_1, T_2 \rangle_{HS(\mathbb{H})} := \sum_{i=1}^{+\infty} \langle T_1 e_i, T_2 e_i \rangle$ and the norm $\|T\|^2_{HS(\mathbb{H})} = \sum_{i=1}^{+\infty} \|Te_i\|^2$. In addition, if $\mathbb{H}$ is separable with orthonormal basis $(e_i)_{i\geq 1}$ then, $HS(\mathbb{H})$ is also separable with orthonormal basis $(e_i \otimes e_j)_{i,j\geq 1}$, where for all $h_1, h_2 \in \mathbb{H}$, $h_1 \otimes h_2$ is a rank one operator defined by $h_1 \otimes h_2(h) = \langle h_1, h \rangle h_2$, for $h \in \mathbb{H}$. It can be shown that $h_1 \otimes h_2$ are HS operators with HS norm $\|h_1 \otimes h_2\|_{HS(\mathbb{H})} = \|h_1\|\|h_2\|$. If $T \in HS(\mathbb{H})$ is in addition self-adjoint and nonnegative, then it can be established that

$$\|T\|_{HS(\mathbb{H})} = \left(\sum_{i=1}^{+\infty} \lambda_i^2 \right)^{1/2},$$

where $(\lambda_i)_{i\geq 1}$ are the eigenvalues of T. Notice that the Hilbert-Schmidt norm a self-adjoint nonnegative operator T dominates its operator norm, *i.e.*, $\|T\|_{op} \leq \|T\|_{HS(\mathbb{H})}$.

The class of integral operators in $L^2[0,1]$ is an important subset of Hilbert-Schmidt operators, as only this type of operator will be considered in Chapter 6.

Example 3.9 (Integral operator)**.** *Let* $k : [0,1]\times[0,1] \to \mathbb{R}$ *be a kernel function in* $L^2([0,1]\times[0,1])$*. Define the operator on* $K : L^2[0,1] \to L^2[0,1]$ *given for all* $f \in L^2[0,1]$ *by*

$$Kf(t) = \int_0^1 k(s,t) f(s) ds,$$

for all $t \in [0,1]$*. Then, K is a HS operator on* $L^2[0,1]$ *with* $\|K\|_{HS(L^2[0,1])} = \|k\|_{L^2([0,1]\times[0,1])}$*. In addition, if k is symmetric, i.e.,* $k(s,t) = k(t,s)$*, then K is self-adjoint and if k is nonnegative, i.e.,* $\forall (a_1, \ldots, a_n) \in \mathbb{R}^n, (t_1, \ldots, t_n) \in [0,1]^n, \sum_{i,j=1}^n a_i a_j k(t_i, t_j) \geq 0$*, then K is nonnegative.*

In this manuscript, integral operators are presented as examples of Hilbert-Schmidt operators. However, this class of operators has been widely studied independently in the literature (see the numerous sections dedicated to them in Gohberg et al. (2013)).

An important result of HS operator theory is a particular case of the Eckart-Young theorem (Theorem 4.4.7 in Hsing and Eubank (2015)), which is the first step towards the principal components analysis presented in Section 3.3: the best approximation of a self-adjoint nonnegative compact operator T with eigensystem $(\lambda_i, e_i)_{i\geq 1}$ by a sum of rank one operators is given by a sum of the first $\lambda_i e_i \otimes e_i$'s, that is

$$\left\| T - \sum_{i=1}^{n} f_i \otimes g_i \right\|_{HS(\mathbb{H})} \geq \left\| T - \sum_{i=1}^{n} \lambda_i e_i \otimes e_i \right\|_{HS(\mathbb{H})}, \tag{3.2}$$

with $f_1, ..., f_n, g_1, ..., g_n \in \mathbb{H}$.

Remark 3.10 (Trace-class operator)**.** *Trace-class operators form an important subset of Hilbert-Schmidt operators. In particular a self-adjoint nonnegative HS operator T with eigenvalues $(\lambda_i)_{i\geq 1}$ is called trace-class if*

$$\|T\|_{tr} = \sum_{i=1}^{+\infty} \lambda_i < +\infty.$$

In this case, it is easy to see that trace-class operators are HS operators since $\|T\|_{HS(\mathbb{H})} \leq \|T\|_{tr}$ by convexity of the square function $x \mapsto x^2$. If k is continuous and nonnegative in Example 3.9 then the integral operator K is trace-class with

$$\|K\|_{tr} = \int_0^1 k(t,t)dt.$$

Finally, a corollary of the Courant-Fischer theorem (Theorem 3.6) is given by Weyl's inequality. This essential inequality in perturbation theory of Hilbert-Schmidt operators controls the maximum deviation between the eigenvalues of two nonnegative compact operators by the HS norm of their difference, which will be useful to our purpose in Chapter 6.

Theorem 3.11 (Theorem 4.2.8 in Hsing and Eubank (2015))**.** *Let $T, \tilde{T}$ be nonnegative compact operators with eigenvalue sequences $(\lambda_i)_{i\geq 1}$ and $(\tilde{\lambda}_i)_{i\geq 1}$, respectively. Then*

$$\sup_{i\geq 1} |\lambda_i - \tilde{\lambda}_i| \leq \|T - \tilde{T}\|_{HS(\mathbb{H})}.$$

3.2 Probability Theory in Hilbert Spaces

Most of the background gathered in this section may be found with detailed proofs, references and discussions in Sections 7 and 8 of the monograph Hsing and Eubank (2015), which provides a self-contained introduction to mathematical foundations of functional data analysis. Other helpful resources regarding probability and measure theory in Banach spaces and Bochner integrals include Vakhania et al. (2012) or Mikusiński (1978).

3.2.1 Random elements in a Hilbert space

Consider a real separable Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ and denote by $\|\cdot\|$ the associated norm. Let $(e_i)_{i\geq 1}$ be any orthonormal basis of $\mathbb{H}$. Since a separable Hilbert space is a particular instance of a Polish space it follows from basic measure theory in (see, *e.g.*, Vakhania

et al. (2012), Theorem 1.2) that the Borel σ-field $\mathcal{B}(\mathbb{H})$ is generated by the family of mappings $\{h^* : x \mapsto \langle x, h\rangle,\ h \in \mathbb{H}\}$, or in other words, by the class of cylinders

$$\mathcal{C} = \{(h^*)^{-1}(B),\ h \in \mathbb{H},\ B \in \mathcal{B}(\mathbb{R})\}.$$

In addition, since the countable family $(e_i^*)_{i\geq 1}$ separates points in $\mathbb{H}$, it also generates the Borel σ-field, see Proposition 1.4 and its corollary in Vakhania et al. (2012). In other words, if we denote by π_N the projection from $\mathbb{H}$ to $\mathbb{R}^N$ onto the first $N \geq 1$ basis vectors, $\pi_N(x) = (\langle x, e_1\rangle, \ldots, \langle x, e_N\rangle)$, the family of cylinder sets

$$\tilde{\mathcal{C}} = \left\{\pi_N^{-1}(A_1 \times \cdots \times A_N), A_j \in \mathcal{B}(\mathbb{R}), j \leq N, N \geq 1\right\}$$

also generates $\mathcal{B}(\mathbb{H})$. We call $\mathbb{H}$-*valued random element* (or variable) any Borel-measurable mapping X from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to $\mathbb{H}$. The distribution and measurability of X are entirely characterized by the measurabilities and the distributions of these univariate projections, as assessed by the following result (see also Lemma 1.8.3. in van der Vaart and Wellner (1996)).

Theorem 3.12 (Theorem 7.1.2 in Hsing and Eubank (2015))**.** *Let X be a mapping from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into $(\mathbb{H}, \mathcal{B}(\mathbb{H}))$. Then,*

1. *X is measurable if $\langle X, h\rangle$ is measurable for all $h \in \mathbb{H}$ and*
2. *if X is measurable, its distribution is uniquely determined by the (marginal) distributions of $\langle X, h\rangle$ over $h \in \mathbb{H}$.*

Since the family $\tilde{C}$ of cylinder sets is a π-system generating $\mathcal{B}(\mathbb{H})$, it follows that the distributions of all finite-dimensional projections $(\pi_N(X))_{N\geq 1}$ onto a specified basis determine the distribution of X.

Integrability conditions for random elements in $\mathbb{H}$ are understood here in the Bochner sense. Similarly to the construction of the Lebesgue integral, the first step is to define *simple functions*. A function $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{H}$ is called simple if there exist $A_1, ..., A_n \in \mathcal{A}$ and $h_1, ..., h_n \in \mathbb{H}$ such that for all $w \in \Omega, X(w) = \sum_{i=1}^n \mathbb{1}\{w \in A_i\}h_i$, for some $n \geq 1$. Then, such a simple function is said to be *Bochner-integrable* if $\mathbb{P}(A_i) < +\infty$ for $1 \leq i \leq n$, and if so, its *Bochner integral* if defined as $\int_\Omega X d\mathbb{P} = \sum_{i=1}^n \mathbb{P}(A_i)h_i$. Consequently, a random element $X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{H}$ is Bochner integrable if there exists a sequence of simple functions $(X_n)_{n\geq 1}$ such that $\lim_{n\to+\infty} \int_\Omega \|X - X_n\| d\mathbb{P} = 0$ (see Section 2.6 in Hsing and Eubank (2015) for further details) and the Bochner integral of X is defined by $\int_\Omega X d\mathbb{P} = \lim_{n\to+\infty} \int_\Omega X_n d\mathbb{P}$. In our specific case of a separable Hilbert space $\mathbb{H}$, a random element X is Bochner integral if $\mathbb{E}[\|X\|] < \infty$ (Theorem 2.6.5 in Hsing and Eubank (2015)). Then, if $\mathbb{E}[\|X\|] < \infty$, the expectation of X is defined as the Bochner integral $\mathbb{E}[X] = \int_\Omega X d\mathbb{P}$.

A key property of the classic expectation is linearity, and it is also satisfied by the expectation defined in the Bochner sense. Namely if T is a bounded, linear operator from $\mathbb{H}_1$ to $\mathbb{H}_2$, two Hilbert spaces, and if X is a Bochner-integrable random element in $\mathbb{H}_1$ then $T(X)$ is also Bochner-integrable in $\mathbb{H}_2$ and $T(\mathbb{E}[X]) = \mathbb{E}[T(X)]$, see Theorem 3.1.7 in Hsing and Eubank (2015). Many other properties of the classic expectation of a real-valued random variables are preserved, *e.g.*, the dominated convergence theorem. In particular, a version of Jensen's inequality can be formulated for $\mathbb{H}$-valued random variables, see, *e.g.*, pp. 42-43 in Ledoux and Talagrand (1991).

In addition, if $\mathbb{E}[\|X\|^2] < +\infty$, the (centered) *covariance operator* of X can be defined as

$$C_c := \mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])],$$

where the expectations are understood in the Bochner sense. If so, X is said of *second-order*. The covariance operator is a self-adjoint trace-class operator (Theorem 7.2.5 in Hsing and Eubank (2015)), and in particular, a HS operator. This operator satisfies a Hilbert extension of the König-Huygens formula, given by $C_c = \mathbb{E}[X \otimes X] - \mathbb{E}[X] \otimes \mathbb{E}[X]$.

3.2.2 Weak convergence

As our main concern in Chapter 5 is to characterize regular variation in Hilbert spaces in terms of weak convergence of appropriately rescaled variables, some basic facts regarding weak convergence in Hilbert spaces are recalled. Most of the material of this section can be found in Chapter 1.8 of van der Vaart and Wellner (1996) and Chapter 7 of Hsing and Eubank (2015).

By definition a sequence $(X_n)_{n\geq 1}$ of $\mathbb{H}$-valued random variables *weakly converges* (or *converges in distribution*) to a $\mathbb{H}$-valued random variable X, and we write $X_n \xrightarrow{w} X$ (or equivalently, $\mu_n \xrightarrow{w} \mu$ if μ_n denotes the probability distribution of X_n and μ, that of X) if for every bounded, continuous function $f : \mathbb{H} \to \mathbb{R}$, we have $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$. This abstract definition may be difficult to handle for verifying weak convergence in specific examples. However, weak convergence in $\mathbb{H}$ may equivalently be characterized *via* weak convergence of one-dimensional projections and an asymptotic tightness condition, as described next. Notice that, because $\mathbb{H}$ is separable and complete, the Prokhorov theorem applies, *i.e.*, uniform tightness and relative compactness of a family of probability measures are equivalent. Recall that a sequence of probability measures $(\mu_n)_{n\geq 1}$ is uniformly tight if for every $\varepsilon > 0$, there exists a compact set $K \subset \mathbb{H}$ such that $\inf_{n\geq 1} \mu_n(K) \geq 1 - \varepsilon$. Notice that, because $\mathbb{H}$ is separable and complete, any single random element valued in $\mathbb{H}$ is tight (as a consequence of Ulam's Theorem, see Theorem 3.1 in Vakhania et al. (2012)).

Remark 3.13 (On measurability and tightness)**.** *Before proceeding any further, in order to clear out any potential confusion, we emphasize that measurability of the considered maps $X_n : \Omega \to \mathbb{H}$ is not required in van der Vaart and Wellner (1996), while it is assumed in the present work, in which we follow common practice in functional data analysis focusing on Hilbert-valued observations (as, e.g., in Hsing and Eubank (2015)). Notice also that the notion of tightness employed in van der Vaart and Wellner (1996) as a criterion for relative compactness of a family of random variables $(X_n)_{n\geq 1}$, is asymptotic tightness, that is: for all $\varepsilon > 0$, there exists a compact subset K of $\mathbb{H}$, such that for every $\delta > 0$, $\liminf_n \mathbb{P}(X_n \in K^\delta) > 1 - \varepsilon$. Here, K^δ denotes the δ-enlargement of the compact set K, that is, $\{x \in \mathbb{H} : \inf_{y\in K} \|x - y\| < \delta\}$. This is seemingly at odds with other presentations (Prokhorov (1956); Hsing and Eubank (2015)) where the argument is organized around the standard notion of uniform tightness, recalled above. However in a Polish space such as $\mathbb{H}$, the two notions of tightness (asymptotic or uniform) are equivalent (van der Vaart and Wellner (1996), Problem 1.3.9), so that the presentations in van der Vaart and Wellner (1996) and Hsing and Eubank (2015) are actually closer to each other than they might appear at first glance.*

A convenient criterion which is the main ingredient to ensure tightness (hence relative compactness) of a family of random $\mathbb{H}$-valued random variables is termed *asymptotically finite-dimensionality* in van der Vaart and Wellner (1996) and seems to originate from Prokhorov (1956).

Definition 3.14. *A sequence of $\mathbb{H}$-valued random variables $(X_n)_{n\geq 1}$ is asymptotically finite-dimensional if, given a Hilbert basis $(e_i)_{i\geq 1}$, for all ε, $\delta > 0$, there exists a finite subset $I \subset \mathbb{N}_{\geq 1}$ such that*

$$\limsup_n \mathbb{P}\left(\sum_{i \notin I} \langle X_n, e_i \rangle^2 > \delta \right) < \varepsilon.$$

Asymptotic finite-dimensionality combined with uniform tightness of all univariate projections of the kind $\langle X_n, h\rangle, h \in \mathbb{H}$, is sufficient conditions for uniform tightness of the family of random variables $(X_n)_{n\geq 1}$ (see Hsing and Eubank (2015), Theorem 7.7.4). A sufficient condition for the asymptotic finite-dimension of a sequence of $\mathbb{H}$-valued random variables $(X_n)_{n\geq 1}$ involving solely univariate convergences is given in Tsukuda (2017) as the existence of limit $\mathbb{H}$-valued random variable X such that $\mathbb{E}[\|X_n\|^2] \to \mathbb{E}[\|X\|^2] < +\infty$ and $\mathbb{E}[\langle X_n, e_i\rangle^2] \to \mathbb{E}[\langle X, e_i\rangle^2]$ for all $i \geq 1$. It should be noticed that the above property is independent from the specific choice of a Hilbert basis. Also, since knowledge of the distributions of all univariate projections characterizes the distribution of a random Hilbert-valued variable X, asymptotic finite-dimensionality combined with weak convergence of univariate projections (or of finite dimensional ones on a fixed basis) are sufficient to prove weak convergence of a family of random elements in $\mathbb{H}$, as summarized in the next statement. Recall that $\pi_N(x) = (\langle x, e_1\rangle, \ldots, \langle x, e_N\rangle)$ for $x \in \mathbb{H}$ and $N \geq 1$.

Theorem 3.15 (Characterization of weak convergence in $\mathbb{H}$)**.** *A family of $\mathbb{H}$-valued random variables $(X_t)_{t\geq 0}$ converges in distribution to a random variable X if and only if, for any sequence $(t_n)_{n\geq 1}$ such that $t_n \to +\infty$ as $n \to +\infty$, the sequence of random variables $(X_{t_n})_{n\geq 1}$ is asymptotically finite-dimensional and either one of the two conditions below holds:*

1. *the sequence $(\langle X_{t_n}, h\rangle)_{n\geq 1}$ converges in distribution to $\langle X, h\rangle$ for any $h \in \mathbb{H}$;*
2. *the sequence $(\pi_N(X_{t_n}))_{n\geq 1}$ converges in distribution to $\pi_N(X)$ for all $N \geq 1$.*

Proof. The fact that asymptotic finite-dimensionality together with Condition 1. in the statement imply weak convergence, results from Theorem 1.8.4 in van der Vaart and Wellner (1996), in the case where all mappings are measurable. To see that Condition 1. may be replaced with Condition 2. in order to prove weak convergence, note that asymptotic finite-dimensionality implies uniform tightness in the case of a Hilbert space (see Remark 3.13 above). Hence, weak convergence occurs if every subsequential limits coincide. It is so because the family of cylinder sets $\tilde{\mathcal{C}}$ is a measure-determining class. The result is thus proved since convergence in distribution of sequences $(X_{t_n})_{n\geq 1}$ towards X for any sequence $(t_n)_{n\geq 1}$ such that $t_n \to +\infty$ as $n \to +\infty$, is equivalent to convergence in distribution of $(X_t)_{t\geq 0}$ towards X as $t \to +\infty$. ■

3.3 Principal Component Analysis

First, the necessary definitions and mathematical background underlying principal component decomposition of $\mathbb{H}$-valued random elements are presented in Section 3.3.1. A self-contained exposition of the topic may be found in Hsing and Eubank (2015), Chapter 7. Additionally, a presentation of the Karhunen-Loève decomposition, which is an adapted version of the principal component decomposition for mean-square

continuous processes, is given at the end of Section 3.3.1, along with a discussion relating the two decompositions. Secondly, perturbation theory notions, which are particularly useful for developing results in Chapter 6, are detailed in Section 3.3.2.

3.3.1 Eigendecomposition of a random element in a Hilbert space

In the sequel we interchangeably use the terminology *principal component decomposition* or *principal component analysis* (PCA). Because of its optimality properties in terms of L^2-error when $\mathbb{H} = L^2[0,1]$, functional PCA is widely used for a great variety of statistical purposes in functional data analysis. Standard references on this topic are the monographs Ramsay and Silverman (2005) and Horváth and Kokoszka (2012).

Let X be a $\mathbb{H}$-valued random element and assume that $\mathbb{E}[\|X\|^2] < \infty$. Then one may consider the (non-centered) covariance operator

$$C = \mathbb{E}[X \otimes X].$$

The motivation for considering the non-centered version of the covariance is discussed in Remark 6.12 (see also Cadima and Jolliffe (2009) for a comparison between centered and uncentered PCA). The properties stated at the end of Section 3.2.1 for the centered covariance operator hold true also for the uncentered covariance operator, *i.e.*, C is self-adjoint and $C \in HS(\mathbb{H})$, thus C is compact. Also by linearity of the Bochner integration, for any $(h,g) \in \mathbb{H}^2$, we have:

$$Ch = \mathbb{E}[\langle h, X\rangle X] \text{ and } \langle Ch, g\rangle = \mathbb{E}[\langle h, X\rangle\langle X, g\rangle].$$

A key result in functional PCA is the eigen decomposition of the covariance operator (see Theorem 7.2.6 from Hsing and Eubank (2015) regarding the centered covariance operator, which is also valid for the non-centered one):

$$C = \sum_{i=1}^{+\infty} \lambda_i \varphi_i \otimes \varphi_i, \tag{3.3}$$

where $\lambda_1 \geq \lambda_2 \geq \dots$ are the eigenvalues sorted by decreasing order and the φ_i's are orthonormal eigenvectors. The set of nonzero eigenvalues λ_i is either finite or else form a sequence of nonnegative numbers converging to zero such that $\sum_{i\geq 1} \lambda_i < +\infty$. The nonzero eigenvalues have finite multiplicity. The eigen functions φ_i form a Hilbert basis of $\overline{\mathrm{Im}(C)}$. As it is the case for the centered version of C, the decomposition (3.3) immediately derives from the spectral theorem for compact nonnegative self-adjoint operators (Theorem 3.7).

A useful property of the eigen functions $(\varphi_i)_{i\geq 1}$ is that they allow perfect signal reconstruction through the well-known *principal component decomposition*.

Theorem 3.16 (Theorem 7.2.7 in Hsing and Eubank (2015))**.** *Let X be a $\mathbb{H}$-valued random element, with $\mathbb{E}[\|X\|^2] < +\infty$. Suppose that X's covariance operator C has the eigendecomposition (3.3). Then, with probability one,*

$$X = \sum_{i=1}^{+\infty} \langle X, \varphi_i\rangle \varphi_i. \tag{3.4}$$

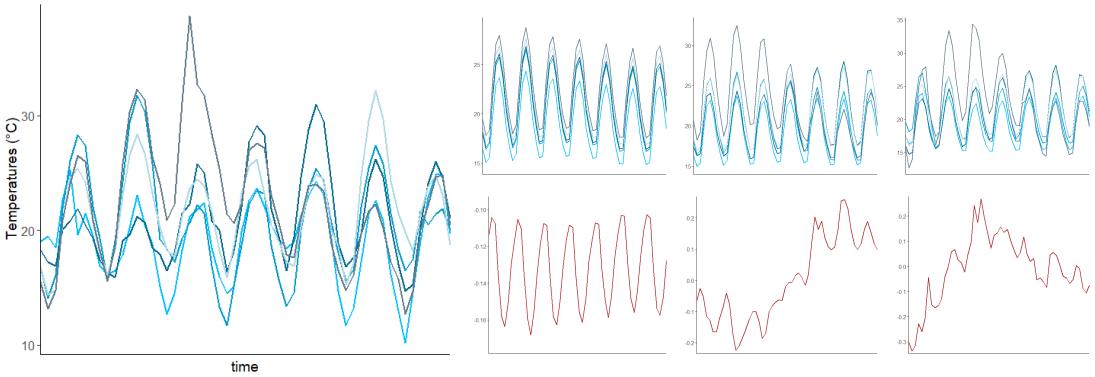


Figure 3.2: Left: Five observations from the Orly temperatures dataset. Right: The five observations decomposed into the first principal component (top) and the first three principal components (bottom).

The *scores* $Z_i = \langle X, \varphi_i \rangle$ satisfy $\mathbb{E}[Z_i^2] = \lambda_i$ and $\mathbb{E}[Z_i Z_j] = 0$ for all $i \neq j$, so that the expansion (3.4) is called *bi-orthogonal*. For all $N \geq 1$, the truncated expansion $\sum_{i \leq N} \langle X, \varphi_i \rangle \varphi_i$ is *optimal* in the sense that it minimizes the integrated mean-squared error

$$\mathbb{E}\left[\left\| X - \sum_{i=1}^{N} \langle X, u_i \rangle u_i \right\|^2\right]$$

over all orthonormal collections $(u_1, \ldots, u_N)$ of $\mathbb{H}$. This is the counterpart of the Eckart-Young theorem for nonnegative self-adjoint Hilbert–Schmidt operator (3.2). The tail behavior of the (summable) eigenvalue sequence $(\lambda_i)_{i \geq 1}$ describes the optimal N-term approximation error, insofar as

$$\sum_{i > N} \lambda_i = \mathbb{E}\left[\left\| X - \sum_{i=1}^{N} \langle X, \varphi_i \rangle \varphi_i \right\|^2\right].$$

Figure 3.2 illustrates PCA on five observations from the Orly temperature dataset and shows the reconstruction of the signals using three principal components.

Another result similar to PCA holds for stochastic processes $X = (X_t)_{t \in K}$ indexed by a compact space K, say $t = [0, 1]$ for simplicity, under a specific continuity assumption. A stochastic process X is said *mean-square continuous* if $\lim_{n \to +\infty} \mathbb{E}\left[(X(t_n) - X(t))^2\right] = 0$ for $t_n \to t$ as $n \to +\infty$. Such mean-square continuous processes are in particular of second-order. Let $m(t) = \mathbb{E}[X(t)]$ denote the mean function of X and $c(t, s) = Cov(X(t), X(s))$ denote the covariance kernel function of X, one can show that X is mean-square continuous iff m and c are continuous (Theorem 7.3.2 in Hsing and Eubank (2015)). By Mercer's theorem (Theorem 4.6.5 in Hsing and Eubank (2015)), the covariance function of X admits the decomposition (3.3)

$$c(s, t) = \sum_{i=1}^{+\infty} \lambda_i \varphi_i(s) \varphi_i(t), \tag{3.5}$$

with the (λ_i, φ_i)'s are the eigenvalues and the continuous eigenfunctions of the associated covariance operator and where the convergence is uniform. In this context, we state the Kosambi-Karhunen-Loève Theorem and one can consider the well-known Karhunen-Loève expansion of a stochastic process X.

Theorem 3.17 (Theorem 7.3.5 in Hsing and Eubank (2015))**.** *Let X be a mean-square continuous stochastic process with mean zero. Suppose that its covariance function has the eigendecomposition* (3.5)*. Then,*

$$\lim_{n\to+\infty} \sup_{t\in[0,1]} \mathbb{E}\left[\left(X(t) - \sum_{i=1}^{n} \langle X, \varphi_i\rangle \varphi_i(t)\right)^2\right] = 0.$$

The *functional PCA* framework is closely related to the celebrated *Karhunen-Loève expansion* in the case where $\mathbb{H} = L^2[0,1]$, however both terms refer to subtly different frameworks, which deserves an explanation. The former framework (which is the one preferred in Part II) relies on a $\mathbb{H}$-valued random element X, with standard results concerning convergence of the expansions of X and its covariance operator in the Hilbert norm and Hilbert-Schmidt norm, respectively, recalled in Section 3.2. Then X's trajectories are in fact equivalence classes of square-integrable functions and the specific value $X_t(\omega)$ of a realization $X(\omega)$ at $t \in [0,1]$ is only defined almost everywhere. In contrast, the latter (Karhunen-Loève) framework relies on a second order stochastic process $X = (X_t)_{t\in[0,1]}$, that is, a collection of random variables, which is continuous is quadratic mean with respect to the index t. Then one must impose additional joint measurability conditions of the mapping $(\omega, t) \mapsto X_t(\omega)$ in order to ensure that the process X is indeed a $\mathbb{H}$-valued random element. In such a case the mean and the covariance operators defined both ways coincide. Also, the Karhunen-Loève Theorem (Loève (1978)) ensures convergence in quadratic mean of the expansion of X_t, uniformly over $t \in [0,1]$. In order to avoid another layer of technicality, and because our main interest indeed lies in the eigenspaces of covariance operators rather than in pointwise reconstruction of the functions, we adopt hereafter the view where X is a $\mathbb{H}$-valued random element, although additional joint measurability assumptions may be imposed in order to fit into the Karhunen-Loève framework.

3.3.2 Perturbation theory related to PCA

Let X be a $\mathbb{H}$-valued random element and assume that $\mathbb{E}\left[\|X\|^2\right] < +\infty$. Let C be its covariance operator with decomposition as in Equation (3.3). Let $X_1, ..., X_n$ i.i.d. copies of X. Define the empirical (centered) covariance operator

$$\hat{C} := \frac{1}{n} \sum_{i=1}^{n} X_i \otimes X_i. \tag{3.6}$$

Like the true covariance operator, the empirical covariance operator $\hat{C}$ is a nonnegative self-adjoint Hilbert-Schmidt operator. Consider its eigenelements $(\hat{\lambda}_i, \hat{\varphi}_i)_{i\geq 1}$. Perturbation theory for PCA addresses how closely C and its empirical counterpart $\hat{C}$ align, particularly examining the proximity of the eigenelements $(\lambda_i, \varphi_i)_{i\geq 1}$ to $(\hat{\lambda}_i, \hat{\varphi}_i)_{i\geq 1}$. For these comparisons to be meaningful, the eigenelements of C must be identifiable, which occurs if the eigenvalues in question are nonzero and have multiplicity one. Significant insights under these conditions, along with the additional assumption of finite fourth moments, are introduced in Dauxois et al. (1982).

Theorem 3.18 (Theorem 2.7 in Horváth and Kokoszka (2012))**.** *Let X be a $\mathbb{H}$-valued random element and assume that $\mathbb{E}\left[\|X\|^4\right] < +\infty$. Suppose that its covariance operator C has the eigendecomposition* (3.3) *so that*

$$\lambda_1 > \lambda_2 > ... > \lambda_p > 0.$$

Let $X_1, ..., X_n$ i.i.d. copies of X, with empirical covariance operator $\hat{C}$ (3.6), with order eigenelements $(\hat{\lambda}_i, \hat{\varphi}_i)_{i\geq 1}$. Then, for each $1 \leq j \leq p$,

$$\limsup_n n\mathbb{E}\left[\|\varphi_j - \hat{c}_j\hat{\varphi}_j\|^2\right] < +\infty, \quad \limsup_n n\mathbb{E}\left[|\lambda_j - \hat{\lambda}_j|^2\right] < +\infty,$$

with $\hat{c}_j = sign(\langle\hat{\varphi}_j, \varphi_j\rangle)$.

Beyond optimal rates of convergence, the asymptotic normality of $\sqrt{n}(\lambda_j - \hat{\lambda}_j)$ and $\sqrt{n}(\varphi_j - \hat{c}_j\hat{\varphi}_j)$ hold (see Sections 2.1 and 2.2 in Dauxois et al. (1982)). More generally, one can consider cases where the eigenvalues have multiplicity greater than one (though still finite, by Theorem 3.7). To infer well-defined quantities, it remains necessary to impose conditions such as $\lambda_p > \lambda_{p+1}$ for some $p \geq 1$. In such scenarios, rather than examining the difference between an eigenfunction and its empirical counterpart individually, one must analyze the difference between the projector onto the p-dimensional eigenspace and its empirical counterpart, with respect to the Hilbert-Schmidt norm, for instance. In this context, Zwald and Blanchard (2005) prove an important result.

Theorem 3.19 (Modified Theorem 3 in Zwald and Blanchard (2005))**.** *Let $A \in HS(\mathbb{H})$ be a self-adjoint nonnegative Hilbert-Schmidt operator with eigenvalues $\lambda_1 > \lambda_2 > \ldots$. Assume that $\lambda_p > \lambda_{p+1}$ for some $p \geq 1$. Let $\delta_p = (\lambda_p - \lambda_{p+1})/2$. Let $B \in HS(\mathbb{H})$ be another self-adjoint nonnegative operator such that $\|B\| < \delta_p/2$ and $(A + B)$ is still a nonnegative operator. Let $\Pi_p(A)$ (resp. $\Pi_p(A + B)$) denote the orthogonal projector onto the p first eigenspaces of A (resp. $(A + B)$). Then,*

$$\|\Pi_p(A) - \Pi_p(A + B)\| \leq \frac{\|B\|}{\delta_p}.$$

Under appropriate assumptions, this result applies to $A = C$ and $B = \hat{C} - C$ (or to a thresholded covariance operator and its difference with a limit covariance operator, see Corollary 6.3), to obtain $\|\Pi_p(C) - \Pi_p(\hat{C})\| \leq \|\hat{C} - C\|/\delta_p$. Regarding this inequality and the Weyl's inequality (Theorem 3.11), it should be noted that the HS norm of the difference between two HS operators fully controls their difference in eigenstructure.

Chapter 4

Statistical Learning

Contents

Statistical learning is a field within statistics and machine learning that focuses on developing methods for making predictions and uncovering patterns from data. It involves building and evaluating models that learn from data, for a given task. Key techniques include regression, classification, or clustering, all aimed at minimizing risks measuring errors of a model. These methods are widely used in applications ranging from meteorological forecasting and medical diagnosis to image recognition and natural language processing. For a precise exposition of statistical learning theory, the reader may be interested in several resources: for a global overview, see Hastie et al. (2009) or James et al. (2013); for classification, see Devroye et al. (2013); for regression, see Györfi et al. (2002); and for concentration inequalities, see Boucheron et al. (2013).

Statistical learning usually focuses on the bulk of the data distribution to obtain models with good average performance. However, some statistical tasks, such as anomaly detection or risk monitoring, which are particularly useful in fields like climatology, finance, and insurance, require a deeper investigation into the behavior of non-normal data, particularly extreme values. Over the last decade, several studies at the intersection of Extreme Value Theory (EVT) and statistical learning have explored dimension reduction (see Engelke and Ivanovs (2021) for a review), anomaly detection and clustering (Goix et al. (2017); Chiapino et al. (2020)), classification (Jalalzai et al. (2018); Clémençon et al. (2023)), cross-validation (Aghbalou et al. (2023)), and principal component analysis (Drees and Sabourin (2021)), among others.

The chapter is structured as follows. Section 4.1 introduces the fundamental task of statistical learning, specifically the prediction problem, through the empirical risk minimization framework. Section 4.2 discusses the use of concentration inequalities to provide statistical guarantees for empirical procedures, extensively used throughout this thesis. In Section 4.3, more advanced concentration results are presented, particularly those related to the theory of Vapnik-Chervonenkis classes. Finally, we focus in Section 4.4 on specific concentration inequalities tailored for extreme values.

4.1 Empirical risk minimization

This section follows the framework of the course Arlot (2018), where proofs and additional discussions (in French) can be found regarding the results presented here.

Consider an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ represent a pair of random variables with an unknown distribution P. The objective is to construct a function $g : \mathcal{X} \to \mathcal{Y}$ which predicts Y from X. To measure the accuracy of predictions, consider a *cost* function $c : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is used, which decreases as the prediction quality improves. The goal is to achieve a prediction that minimizes this cost on average, leading to the definition of the *risk* of a predictive function g

$$R(g) := \mathbb{E}[c(g(X), Y)].$$

The aim is to minimize the risk across all measurable functions. The minimum risk, denoted as $R^* := \inf_g R(g)$, is referred to as the *Bayes risk*. If an optimal predictive function exists, it is denoted as $g^* \in \arg\min_g R(g)$.

We will further examine two classical settings: binary classification with $0-1$-cost and regression with quadratic cost.

Example 4.1 (Binary classification with $0-1$-cost)**.** *Let $\mathcal{Y} = \{0, 1\}$ be the output space and consider the $0-1$-cost function defined as $c : (y, y') \in \mathcal{Y}^2 \mapsto \mathbb{1}\{y \neq y'\} \in \{0, 1\}$. Then the associated risk is defined as $R(g) = \mathbb{P}(g(X) \neq Y)$ and its minimizer, namely the Bayes classifier, is given by $g^*(X) = \mathbb{1}\{\mathbb{E}[Y \mid X] > 1/2\}$ a.s., with $\mathbb{E}[Y \mid X]$ the conditional expectation of Y given X.*

Example 4.2 (Regression with quadratic cost)**.** *Consider the quadratic cost function defined as $c : (y, y') \in \mathcal{Y}^2 \mapsto (y - y')^2 \in \mathbb{R}$. Then the associated risk is defined as $R(g) = \mathbb{E}[(g(X) - Y)^2]$ and its minimizer is given by $g^*(X) = \mathbb{E}[Y \mid X]$ a.s..*

In practice, since the distribution P is unknown, the risk of a predictive function cannot be directly determined. Instead, we observe independent pairs $(X_1, Y_1), ..., (X_n, Y_n)$ identically distributed as (X, Y). The purpose of supervised learning is to infer a predictive function from these observations to accurately predict an output variable Y_{n+1} given a new input observation X_{n+1}. To achieve this, one must minimize the empirical counterpart of the risk

$$\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} c(g(X_i), Y_i).$$

It is convenient to choose a *model*, *i.e.*, a class of function $\mathcal{G}$ over which the minimization of the empirical risk is performed

$$\hat{g}_n \in \arg\min_{g \in \mathcal{G}} \hat{R}_n(g).$$

Example 4.3 (Linear regression)**.** *Assume that $\mathcal{X} \subset \mathbb{R}^d$ and consider the quadratic cost. Consider the linear model, as the class of function $\mathcal{G}_{lin} = \{f : \mathbb{R}^d \to \mathbb{R}, \exists \boldsymbol{\beta} \in \mathbb{R}^d, f(\mathbf{x}) = \langle \boldsymbol{\beta}, \mathbf{x} \rangle\}$. If the matrix $\mathbb{X}$ with rows $\mathbb{X}_i = (\mathbf{X}_i)^T$ for every $1 \leq i \leq n$ is of full rank, then the minimizer of $\hat{R}_n$ over $\mathcal{G}$ is given by*

$$\hat{g}_n(x) = \langle \hat{\boldsymbol{\beta}}, \mathbf{x} \rangle,$$

where $\hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$ *and with* $\mathbf{Y} = (Y_1, ..., Y_n)^T$. *The coefficient* $\hat{\boldsymbol{\beta}}$ *is the well-known Ordinary Least Square (OLS) estimator.*

Choosing a model simplifies the computation of predictive functions but introduces a model bias, referred to as the *approximation error*, defined as $\inf_{g\in\mathcal{G}} R(g) - R^*$. Combined with the estimation error $R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g)$, the *excess of risk* of $\hat{g}_n$ is expressed as

$$R(\hat{g}_n) - R^* = R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) + \inf_{g\in\mathcal{G}} R(g) - R^*.$$

While the approximation error generally cannot be controlled, the estimation error can often be managed using the inequality (Proposition 7 in Arlot (2018))

$$R(\hat{g}_n) - \inf_{g\in\mathcal{G}} R(g) \leq 2\sup_{g\in\mathcal{G}} |\hat{R}_n(g) - R(g)|. \tag{4.1}$$

Obtaining explicit bounds for these quantities lies at the core of concentration inequality theory, which is discussed in the next sections.

4.2 Non-asymptotic analysis

The primary objective of concentration theory is to bound the probability that a function deviates from its expectation. This theory is crucial for obtaining statistical guarantees in ERM, as the right term of inequality (4.1) can be rewritten as

$$|\hat{R}_n(g) - R(g)| = \left|\frac{1}{n}\sum_{i=1}^{n} c_g(X_i, Y_i) - \mathbb{E}[c_g(X, Y)]\right|, \tag{4.2}$$

with $c_g(x, y) = c(g(x), y)$. It is clear from this formulation that concentration theory is encompassed within the theory of empirical processes, since the risk deviation in Equation (4.2) is minimized when the empirical measure $\hat{P}_n = \sum_{i=1}^n \delta_{(X_i,Y_i)}$ closely approximates the true measure P. Several classic inequalities in concentration theory are discussed in Boucheron et al. (2013). Among these, Hoeffding's inequality holds for bounding the deviation of bounded random variables.

Theorem 4.4 (Theorem 2.8 in Boucheron et al. (2013))**.** *Let* $Z_1, ..., Z_n$ *be independent real-valued random variables such that* Z_i *takes its values in* $[a_i, b_i]$ *a.s. for all* $1 \leq i \leq n$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i]) \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

for $\varepsilon > 0$.

Hoeffding's inequality involves a coarse bound of the variance of $\sum_{i=1}^n Z_i$ by $\sum_{i=1}^n (b_i - a_i)^2/4$. In contrast, Bernstein's inequality, or a direct corollary of it, offers a more refined bound for the same quantities through a finer control of the variance term. For the sake of simplicity, we present Bernstein's inequality only for the case of bounded random variables, although a weaker condition on the boundedness of the moments could also be assumed (see Theorem 2.10 in Boucheron et al. (2013)).

Theorem 4.5 (Corollary 2.11 in Boucheron et al. (2013)). *Let $Z_1, ..., Z_n$ be independent real-valued random variables. Assume that there exists positive numbers b and v such that $Z_i \leq b$ a.s. for all $1 \leq i \leq n$ and $\sum_{i=1}^n \mathbb{E}[Z_i^2] \leq v$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \geq \varepsilon \right) \leq \exp\left(- \frac{\varepsilon^2}{2(v + b\varepsilon/3)} \right),$$

for $\varepsilon > 0$.

A final, non-standard inequality will be used to derive statistical guarantees in Chapter 8 and will be referenced hereafter. This inequality can be regarded as a Bernstein-type inequality since its upper bound is identical to that in Theorem 4.5. The difference lies in the finer assumptions, notably the absence of independence between the random variables.

Here and throughout we adopt the shorthand notation $z_{i:j} = z_i, \ldots, z_j$ for $i \leq j$.

Lemma 4.6 (Theorem 3.8 in McDiarmid (1998)). *Let $Z_{1:n}$ with Z_i taking their values in a set $\mathcal{Z}$ and let f be a real-valued function defined on $\mathcal{Z}^n$. Let $X = f(Z_{1:n})$. Consider the positive deviation functions, defined for $1 \leq i \leq n$ and for $z_{1:i} \in \mathcal{Z}^i$*

$$g_i(x_{1:i}) = \mathbb{E}[X | Z_{1:i} = z_{1:i}] - \mathbb{E}[X | Z_{1:i-1} = z_{1:i-1}].$$

Denote by b the maximum deviation

$$b = \max_{1 \leq i \leq n} \sup_{z_{1:i} \in \mathcal{Z}^i} g_i(z_{1:i}).$$

Let v be the supremum of the sum of conditional variances,

$$v := \sup_{(z_1, \ldots, z_n) \in \mathcal{Z}^n} \sum_{i=1}^n \sigma_i^2(f(z_1, \ldots, z_n)),$$

where $\sigma_i^2(f(z_{1:n})) := \mathbb{V}\mathrm{ar}[g_i(z_{1:i-1}, Z_i)]$. If b and v are both finite, then

$$\mathbb{P}\Big(f(X) - \mathbb{E}[f(X)] \geq \varepsilon \Big) \leq \exp\left(\frac{-\varepsilon^2}{2(v + b\varepsilon/3)} \right),$$

for $\varepsilon \geq 0$.

Observe that, the classic Bernstein's inequality derives from Lemma 4.6 by considering independent $Z_1, .., Z_n$ and $f(z_{1:n}) = \sum_{i=1}^n z_i$.

4.3 A Vapnik-Chervonenkis inequality

As noted in the previous section, a critical quantity to control is the estimation error, specifically to bound the term $|\hat{R}_n(g) - R(g)|$ over a class of functions $\mathcal{G}$. This function class must be sufficiently complex to yield a good predictive function (thereby reducing approximation error) but not overly broad to prevent overfitting (thereby controlling the estimation term), aiming for a concentration bound of $\sup_{g \in \mathcal{G}} |\hat{R}_n(g) - R(g)|$. A suitable class with these properties is described by the well-known *Vapnik-Chervonenkis* classes. We give two equivalent definitions for different classes: one for a class of

subsets and one for a class real-valued functions (which is the one useful in Part III). To facilitate the exposition of the following results, we stick to the first definition, which is the most common one, except for Proposition 4.11 which is stated in terms of a class of regression functions, as it will be utilized in this form in Part III.

As mentioned earlier, following the remark below Equation (4.2), it is customary to consider the VC-class of Borel subsets.

Definition 4.7 (Vapnik-Chervonenkis dimension of a class of Borel subsets)**.** *Let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$. Define the shatter coefficient of size $n \geq 1$*

$$S_{\mathcal{A}}(n) := \sup_{x_1,...,x_n \in \mathcal{X}} |\{\{x_1,..,x_n\} \cap A, A \in \mathcal{A}\}|.$$

The class $\mathcal{A}$ is called a VC-class if

$$V_{\mathcal{A}} := \sup\{n \geq 1, S_{\mathcal{A}}(n) = 2^n\} < +\infty.$$

If so, $V_{\mathcal{A}}$ is called the VC-dimension of $\mathcal{A}$.

Following the lines of Section 3.6 in Vapnik (1999), the definition of a VC-class of Borel subsets extends to classes of real-valued functions by considering the family of subgraphs.

Definition 4.8 (Vapnik-Chervonenkis dimension of a class of real functions)**.** *Let $\mathcal{Y} \subset \mathbb{R}$. Let $\mathcal{G}$ be a class of real functions $g : \mathcal{X} \to \mathcal{Y}$. Define the family of subgraphs of $\mathcal{G}$ by*

$$I_{\mathcal{G}} = \Big\{\{(x, y) \in \mathcal{X} \times \mathcal{Y}, y \leq g(x)\}, g \in \mathcal{G}\Big\}.$$

The class $\mathcal{G}$ is called VC-class if $I_{\mathcal{G}}$ is a VC-class, in the sense of a class of Borel subsets. If so, the VC-dimension of $\mathcal{G}$ is the VC-dimension of $I_{\mathcal{G}}$.

A basic, but useful example of a VC-class is given by the class of half-spaces of $\mathbb{R}^d$

Example 4.9 (VC-class)**.** *The class of half-spaces $\mathcal{H}$ of $\mathbb{R}^d$, defined as subsets of the form $\{(x_1,...,x_d) \in \mathbb{R}^d, \beta_0 + \beta_1 x_1 + ... + \beta_d x_d \geq b, (\beta_0, \beta_1, ..., \beta_d) \in \mathbb{R}^{d+1}, b \in \mathbb{R}\}$, is a VC-class of VC-dimension $d + 1$. First, it is easy to construct half-spaces such that $S_{\mathcal{H}}(d) = 2^{d+1}$ (left graph of Figure 4.1). Second, in the case of $d + 2$ points, either all points lie on the boundary of their convex hull (middle graph of Figure 4.1), in which case two "opposite" points cannot be separated from the others, or at least one point lies in the interior of the convex hull (right graph of Figure 4.1) and cannot be separated from the others. These scenarios are illustrated for $d = 2$ in Figure 4.1.*

Similarly, it can be shown mutatis mutandis the class of linear regression functions (in the sense of Definition 4.8) are VC-classes with VC-dimension $d + 1$.

Once again, classes of functions can be replaced by classes of subsets since bounding the deviation term $\sup_{g \in \mathcal{G}} |\hat{R}_n(g) - R(g)|$ is similar to bounding the deviation term $\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|$, where ν_n is an empirical version of the measure ν. In the remainder of the section, we denote by ν the distribution of X on $\mathcal{X}$ and we set $\nu_n = (1/n) \sum_{1 \leq i \leq n} \delta_{X_i}$ the associated empirical measure. We can now state the well-known *Vapnik-Chervonenkis inequality* for class of subsets.

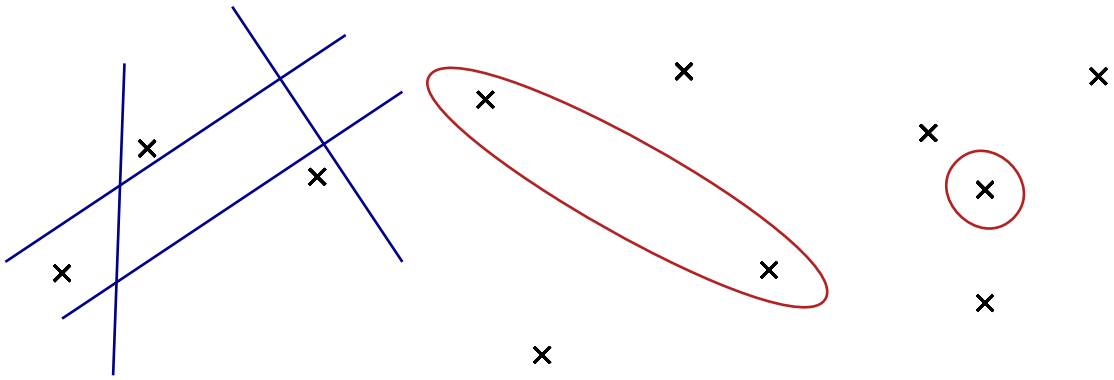


Figure 4.1: Illustration VC-dimension of half-spaces in $\mathbb{R}^2$.

Theorem 4.10 (Theorem 2 in Vapnik and Chervonenkis (2015)). *Let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$ and $\delta > 0$. Then, with probability at least $1-\delta$*

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \leq 2\sqrt{\frac{2}{n}\Big(\log S_{\mathcal{A}}(2n) + \log(4/\delta)\Big)}.$$

This inequality demonstrates its effectiveness when $\mathcal{A}$ is a VC-class of function, so that, with probability at least $1-\delta$,

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \leq 2\sqrt{\frac{2}{n}\Big(V_{\mathcal{A}} \log(2n+1) + \log(4/\delta)\Big)},$$

with $V_{\mathcal{A}}$ the VC-dimension of$\mathcal{A}$. This result follows from the well-known Sauer's lemma (Lemma 1 in Arlot (2018)). This inequality enables thus to bound in expectation the maximal deviation over a possible infinite class. A refinement of this inequality removes the $\log(2n+1)$ term in the inequality (see, *e.g.*, Theorem 3.4 in Boucheron et al. (2005)).

This theorem (as well as Theorem 4.12) relies on a symmetrization argument (see Section 3.7.1 in Arlot (2018)). In the case of class of regression functions, this argument involves bounding the expectation of the maximal deviation as follows

$$\mathbb{E}\Big[\sup_{g\in\mathcal{G}} |\hat{R}_n(g) - R(g)|\Big] \leq 2\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left\{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i c(g(X_i), Y_i)\right\}\right],$$

with $\varepsilon_1, ..., \varepsilon_n$ are i.i.d. Rademacher random variables, independent of $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. The sum on the right side of the inequality is called a *Rademacher average*, which plays a crucial role in the theory of concentration. The following inequality concerning a specific Rademacher average will be a key component in the proof of Theorem 8.4.

Proposition 4.11 (Proposition 2.1 in Giné and Guillou (2001)). *Let $Z_{1:n}$ be i.i.d. random variable of distribution P. Let $\varepsilon_{1:n}$ independent Rademacher random variables, independent of $Z_{1:n}$. Let $\mathcal{G}$ be a measurable uniformly bounded VC class of functions with VC-dimension $V_{\mathcal{G}}$. Let $\sigma^2 \geq \sup_{g\in\mathcal{G}} \mathbb{E}_P[g]$ and $U \geq \sup_{g\in\mathcal{G}} \|g\|_\infty$ be such that $0 < \sigma \leq U$. Then*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} \left|\sum_{i=1}^{n} \varepsilon_i g(Z_i)\right|\right] \leq C\left(V_{\mathcal{G}} U \log(U/\sigma) + \sigma\sqrt{V_{\mathcal{G}} n \log(U/\sigma)}\right),$$

where C is a universal constant.

4.4 Concentration inequalities for rare events

Statistical learning in extreme value theory has only recently become an active area of research. For a more detailed discussion, refer to Section 1.2 in the introduction chapter. This section focuses on concentration inequalities for extremes, central to the results in Sections 6.3 of Part II and 8.2 of Part III. This section draws on findings from Lhaut et al. (2022).

The classic VC inequality is not well-suited for rare events, defined as events $A \in \mathcal{A}$ with $\nu(A) \ll 1$. This is because the inequality does not differentiate between normal and rare events, even though the maximal deviation term $\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|$ is intuitively smaller for rare events. Therefore, an alternative VC-type inequality is presented, which is particularly useful in extreme settings.

Theorem 4.12 (Theorem 2.1 in Anthony and Shawe-Taylor (1993))**.** *Let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$. Then, with probability at least $1-\delta$*

$$\sup_{A\in\mathcal{A}} \frac{\nu_n(A) - \nu(A)}{\sqrt{\nu(A)}} \leq 2\sqrt{\frac{1}{n}\left(\log S_{\mathcal{A}}(2n) + \log(4/\delta)\right)}.$$

For a class $\mathcal{A}$ of rare subsets, that is for every $A \in \mathcal{A}$, $\nu(A) \leq p$ where p is a "small" probability, Theorem 4.12 reformulates as

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \leq 2\sqrt{\frac{p}{n}\left(\log S_{\mathcal{A}}(2n) + \log(4/\delta)\right)},$$

which is a notable improvement over the classic VC-inequality. However, considering that the effective sample size is proportional to np, it is reasonable to achieve a bound involving $\mathcal{S}_{\mathcal{A}}(2np)$ instead of $\mathcal{S}_{\mathcal{A}}(2n)$. This is addressed in the inequalities from Lhaut et al. (2022), with the two principal results presented in the following theorem.

Theorem 4.13 (Theorem 3.1 and Corollary 4.4 in Lhaut et al. (2022))**.** *Let $\mathcal{A}$ be a class of subsets of $\mathcal{X}$, such that $\nu(A) \leq p$ for every $A \in \mathcal{A}$ and $0 < \delta < 1$.*

1. *If $np \geq 4\log(4/\delta)$, then with probability at least $1-\delta$,*

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \leq \frac{2}{3n}\log(4/\delta) + \sqrt{\frac{p}{n}}\left(\sqrt{2\log(4/\delta)} + 2\sqrt{\log(8/\delta) + \log S_{\mathcal{A}}(4np)} + 1\right).$$

2. *If $\mathcal{A}$ is a VC-class, then with probability at least $1-\delta$,*

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \leq \frac{2}{3n}\log(1/\delta) + \sqrt{\frac{2p}{n}}\left(\sqrt{2\log(1/\delta)} + \sqrt{\log(2) + V_{\mathcal{A}}\log(2np+1)} + \frac{\sqrt{2}}{2}\right).$$

Additional concentration inequalities for extremes, along with detailed discussions can be find in Lhaut et al. (2022).

Remark 4.14 (Rare events?)**.** *In accordance with the subsequent sections, a "small" p could, for instance, be defined as $p = 1 - k_n/n$, where $k_n \to +\infty$ and $k_n/n \to 0$ as $n \to +\infty$. Here, n represents the sample size, and k_n denotes the number of extremes in the sample. Thus, in this context, a rare event corresponds to an observation exceeding the k_n-th order statistic of the sample.*

Part II

Functional Extremes

Introduction

The surge in data availability of functional nature has spurred the development of various applications related, *e.g.*, to IoT, spectrometry, predictive maintenance of sophisticated systems (energy networks, aircraft fleets, ...), see, *e.g.*, the review Wang et al. (2016) or Gertheiss et al. (2023), Li et al. (2022) and references therein. This opens new perspectives for Extreme Value Analysis in applications where extremes play a significant role, such as generation of synthetic extreme examples, anomaly detection, or environmental risk assessment.

The main purpose of this part is to develop a general probabilistic and statistical framework for the analysis of extremes of regularly varying random functions in the space $L^2[0,1]$, the Hilbert space of square-integrable, real-valued functions over $[0,1]$, with immediate possible generalization to other compact domains, *e.g.*, spatial ones. A major feature of the proposed framework is the possibility to project the observations onto a finite-dimensional functional space, *via* a modification of the standard functional Principal Component Analysis (PCA) which is suitable for heavy-tailed observations, for which second (or first) moments may not exist.

Recent years have seen a growing interest in the field of Extreme Value Theory (EVT) towards high dimensional problems. On one hand, a particularly active line of research concerns unsupervised dimension reduction for which a variety of of methods have been proposed over the past few years (refer to Section 1.2.2 for more details about dimension reduction suited for extremes). On the other hand, functional approaches in EVT have a long history and are still the subject of recent development in spatial statistics, see, *e.g.*, the recent review Huser and Wadsworth (2022). For statistical applications, typically for spatial extremes, strong parametric assumptions must be made to make up for the infinite-dimensional nature of the problem. Dimension reduction is then limited to choosing a parametric model of appropriate complexity and it is not clear how to leverage dimension reduction tools recently developed for multivariate extremes in this setting. The vast majority of existing works in functional extremes consider the continuous case (see Section 1.2.1 for more details about functional extremes), where it is in particular not clear how to perform dimension reduction on these extremes.

Hereafter, we place ourselves in the Peaks-over-Threshold (PoT) framework: the focus is on the limit distribution of rescaled observations, conditioned upon the event that their norm exceeds a threshold, as this threshold tends to infinity. With continuous stochastic processes, an extreme observation is declared so whenever its supremum norm is large, *i.e.*, above a high quantile. The limiting process arising in this context is a Generalized Pareto process. In the functional PoT framework, the definition of an extreme event depends on the choice of a norm which may be of crucial importance in applications. As an example, in air quality monitoring, it may be more relevant to characterize extreme

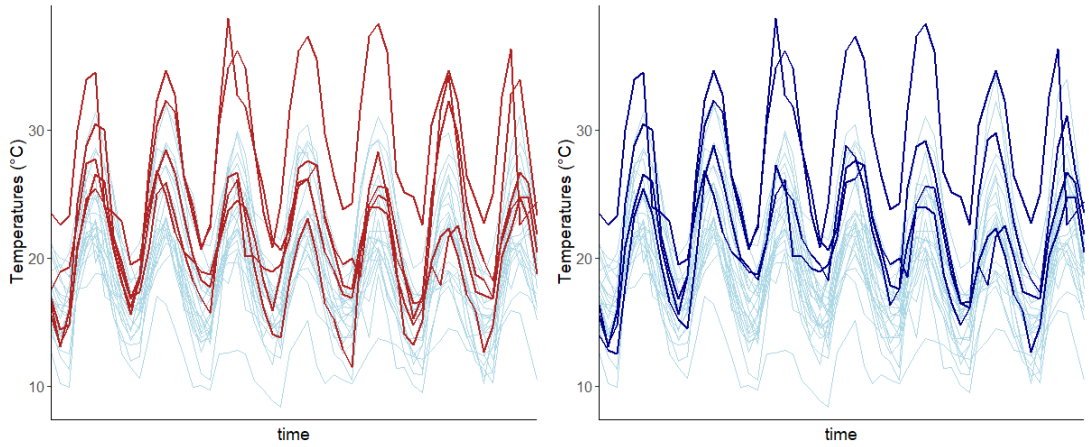


Figure 4.2: Functional extremes on the Orly temperatures dataset w.r.t. the 2-norm (left) and w.r.t. the supremum norm (right).

concentration of pollutants through an integrated criterion over a full 24-hours period, rather than through the maximum hourly record. Figure 4.2 illustrates the different extremes according to the 2-norm and the supremum norm on the Orly temperatures dataset described in Chapter 3. This line of thoughts is the main motivation behind the work of Dombry and Ribatet (2015), where alternative definitions of extreme events are considered by means of a homogeneous cost functional, which gives rise to r-Pareto processes. However the observations are still assumed to be continuous stochastic processes and the framework is not better suited for dimension reduction than those developed in the previously cited works. A standard hypothesis underlying the PoT approach is Regular Variation (RV), which, roughly, may be seen as an assumption of approximate radial homogeneity regarding the distribution of the random object X under study, conditionally on an excess of the norm $\|X\|$ of this object above a high radial threshold. An excellent account of RV of multivariate random vectors is given in the monographs Resnick (1987, 2007). In Hult and Lindskog (2006b) RV is extended to measures on arbitrary complete, separable metric spaces and involves $\mathbb{M}_0$-convergence of measures associated to the distribution of rescaled random objects. One characterization of RV in this context is *via* weak convergence of the pseudo angle $\Theta = \|X\|^{-1}X$ and RV of the (real-valued) norm $\|X\|$. Namely, the law of Θ given that $\|X\| > t$, $\mathcal{L}(\Theta \,|\, \|X\| > t)$, which we denote by $P_{\Theta,t}$, must converge weakly as $t \to +\infty$, towards a limit probability distribution $P_{\Theta,\infty}$ on the unit sphere (see, *e.g.*, Segers et al. (2017); Davis and Mikosch (2008)). In the present work we place ourselves in the general RV context defined through $\mathbb{M}_0$-convergence in Hult and Lindskog (2006b), and we focus our analysis on random functions valued in the Hilbert space $L^2[0,1]$, which has received far less attention (at least in EVT) than the spaces of continuous, semi-continuous or *càdlàg* functions. One main advantage of the proposed framework, in addition to allowing for rough function paths, is to pave the way for dimension reduction of the observations *via* functional PCA of the *angular* component Θ. In this respect the dimension reduction strategy that we propose may be seen as an extension of Drees and Sabourin (2021), who worked in the finite-dimensional setting and derived finite sample guarantees regarding the eigenspaces of the empirical covariance operator for Θ. However their techniques of proof cannot be leveraged in the present context because they crucially rely on the compactness of the unit sphere in $\mathbb{R}^d$, while the unit sphere in an infinite-dimensional Hilbert space is not compact.

Several questions arise. First, when dealing with functional observations, the choice of the norm (thus of a functional space) is not indifferent, since not all norms are equivalent. In particular, there is no reason why RV in one functional space (say, $\mathcal{C}[0,1]$) would be equivalent to RV in a larger space such as $L^2[0,1]$. Also a recurrent issue in the context of weak convergence of stochastic processes is to verify tightness conditions in addition to weak convergence of finite-dimensional projections, in order to ensure weak convergence of the process as a whole. The case of Hilbert valued random variables makes no exception (see, *e.g.*, Chapter 1.8 in van der Vaart and Wellner (1996)). A natural question to ask is then: 'What concrete conditions regarding the angular and radial components $(\Theta, \|X\|)$ in a POT framework, which may be verified in practice on specific generative examples or even on real data, are sufficient in order to ensure tightness and thus RV?'. Regarding the PCA of the angular distribution, the natural extension of the finite-dimensional covariance matrix of extreme angles $C_{t,\mathbb{R}^d} = \mathbb{E}[\Theta\Theta^\top \,|\, \|X\| > t]$ in Drees and Sabourin (2021) where $X \in \mathbb{R}^d$, is the covariance operator $C_t = \mathbb{E}[\Theta \otimes \Theta \,|\, \|X\| > t]$ when $X \in L^2[0,1]$, see Chapter 3 for minimal background regarding probability in Hilbert spaces and covariance operators. One may wonder whether the eigenspaces of C_t indeed converge as $t \to +\infty$ to those of $C_\infty = \mathbb{E}[\Theta_\infty \otimes \Theta_\infty]$, where $\Theta_\infty \sim P_{\Theta,\infty}$, under the RV conditions alone, and whether the results of Drees and Sabourin (2021) regarding concentration of the empirical eigenspaces indeed extend to the infinite-dimensional Hilbert space setting. Although existing studies in the literature have explored functional PCA in $L^2[0,1]$ within the context of EVT, none have addressed the aforementioned questions. For further details on these existing studies, see Section 1.2.1.

Our contribution is twofold. In Chapter 5, we provide a comprehensive description of the notion of RV in a separable Hilbert space which fits into the framework of Hult and Lindskog (2006b). In Chapter 6, we make a first step towards bridging the gap between dimension reduction approaches an functional extremes by considering the functional PCA of the angular variable Θ. Certain technical details are deferred to the Appendices 5.A and 6.A.

Chapter 5

Regular Variation in Hilbert Spaces

Contents

We start Chapter 5 by presenting basic examples of regularly varying random element in a Hilbert space $\mathbb{H}$ relying on weak convergence characterizations provided in van der Vaart and Wellner (1996) (see Proposition 3.15), as sums of simple Hilbert random elements, in Section 5.1. In Section 5.2, we formulate specific characterizations involving finite-dimensional projections and moments of the angular variable Θ, along with several examples and counter-examples illustrating our statements. Section 5.3 discusses the relationships between Regular Variation (RV) in $\mathcal{C}[0,1]$ and in $L^2[0,1]$. The chapter concludes with some perspectives of this work.

5.1 Regularly Varying Random Elements in $\mathbb{H}$

As a warm up we discuss a classic example in EVT, a multivariate multiplicative model within the framework of the multiplicative Breiman's lemma (Basrak et al. (2002b), Proposition A.1). In this setting we show in Proposition 5.1 below that regular variation holds. The proof relies on existing general characterizations such as Equation (2.18). This example will serve as a basis for our simulated data example in Section 6.4.

Proposition 5.1. *Let $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$ be regularly varying with index $\alpha > 0$ and limit measure μ, and let $\mathbb{A} = (A_1, \ldots, A_d)$ be a random vector of $\mathbb{H}$-valued variables A_i, independent of Z, such that $\mathbb{E}\left[\left(\sum_{j=1}^d \|A_j\|_{\mathbb{H}}^2\right)^{\gamma/2}\right] < +\infty$ for some $\gamma > \alpha$. Then,*

$$X = \sum_{j=1}^{d} Z_j A_j$$

is regularly varying in $\mathbb{H}$ with limit measure $\tilde{\mu}(\cdot) = \mathbb{E}\left[\mu\left\{x \in \mathbb{R}^d : \sum_{j=1}^d A_j x \in (\cdot)\right\}\right]$.

Proof. In their Proposition A.1, Basrak et al. (2002b) consider the case where $A_j \in \mathbb{R}^q$ and $\mathbb{A} = (A_1, \dots, A_d)$ is a $q \times d$ matrix. In the proof, they use the operator norm for $\mathbb{A}$, but because all norms are equivalent in that case, their argument remains valid with the finite-dimensional Hilbert-Schmidt norm. In this finite-dimensional context, $\|\mathbb{A}\|$ is equal to $(\sum_{j=1}^{d} \|A_j\|_2^2)^{1/2}$, where $\|\cdot\|_2$ is the Euclidean norm. An inspection of the arguments in their proof shows that they also apply to the case where $A_j \in \mathbb{H}$, up to replacing $\|A_j\|_2$ with $\|A_j\|_{\mathbb{H}}$ and $\|\mathbb{A}\|$ with $(\sum_{j=1}^{d} \|A_j\|_{\mathbb{H}}^2)^{1/2}$. In particular, Pratt's lemma is applicable because Fatou's Lemma is valid for nonnegative Hilbert space-valued functions. ∎

The latter example can be adapted to random sums of regularly varying $\mathbb{H}$-valued random elements, under additional assumptions, by using a conditioning argument.

Proposition 5.2. *Let $Z = (Z_i)_{i\geq 1}$ be a sequence of i.i.d. real-valued regularly varying with index $\alpha > 0$ and with limit measure μ, let $A = (A_i)_{i\geq 1}$ be a sequence of i.i.d. $\mathbb{H}$-valued random variables, independent of Z, such that $\mathbb{E}\Big[(\sum_{j=1}^{\infty} \|A_j\|_{\mathbb{H}}^2)^{\gamma/2}\Big] < +\infty$ for some $\gamma > \alpha$, and let D be an integer-valued nonnegative random variable, independent of A and Z, such that $\mathbb{E}[D] < +\infty$ and $\mathbb{P}(D > x) = o(\mathbb{P}(Z_1 > x))$. Then,*

$$X = \sum_{j=1}^{D} Z_j A_j$$

is regularly varying in $\mathbb{H}$ with limit measure $\mathbb{E}[D]\tilde{\mu}(\cdot)$, where $\tilde{\mu}(\cdot) = \mathbb{E}[\mu\{x \in \mathbb{R} : A_1 x \in \cdot\}]$.

Proof. Let $B \in \mathcal{B}(\mathbb{H})$ be a continuity-set of $\tilde{\mu}$ bounded away from zero, set $b_n = \mathbb{P}(\|Z_1 A_1\| \geq n)^{-1}$ a regularly varying normalizing sequence, and let $k_0 \in \mathbb{N}$ be a non-zero integer.

$$\begin{aligned} b_n \mathbb{P}\Big(n^{-1} \sum_{i=1}^{D} Z_i A_i \in B\Big) &= \sum_{d=1}^{\infty} \mathbb{P}(D = d) b_n \mathbb{P}\Big(n^{-1} \sum_{i=1}^{d} Z_i A_i \in B\Big) \\ &= \Big(\sum_{d=1}^{k_0} + \sum_{d=k_0+1}^{\infty}\Big) \mathbb{P}(D = d) b_n \mathbb{P}\Big(n^{-1} \sum_{i=1}^{d} Z_i A_i \in B\Big) = I_1 + I_2. \end{aligned}$$

The first term is controlled by Proposition 5.1,

$$I_1 \underset{n\to+\infty}{\longrightarrow} \sum_{d=1}^{k_0} \mathbb{P}(D = d) \mathbb{E}\Big[\sum_{i=1}^{d} \mu(\{x \in \mathbb{R} : A_i x \in B\})\Big] = \tilde{\mu}(B) \sum_{d=1}^{k_0} d\,\mathbb{P}\Big(D = d\Big) \underset{k_0\to+\infty}{\to} \tilde{\mu}(B)\mathbb{E}[D].$$

Turning to the second term, since B is bounded away zero, there exists $\varepsilon > 0$ such that $B \subset B^c(0, \varepsilon)$ and then, notice that

$$I_2 \leq \sum_{d=k_0+1}^{\infty} \mathbb{P}(D = d) \frac{\mathbb{P}\Big(\|\sum_{i=1}^{d} Z_i A_i\| \geq n\varepsilon\Big)}{\mathbb{P}(\|Z_1 A_1\| \geq n)} \leq \sum_{d=k_0+1}^{\infty} \mathbb{P}(D = d) \frac{\mathbb{P}\Big(\sum_{i=1}^{d} \|Z_i A_i\| \geq n\varepsilon\Big)}{\mathbb{P}(\|Z_1 A_1\| \geq n)}.$$

The arguments of Faÿ et al. (2006) in their Proposition 4.1 apply and ensure that

$$\limsup_{n\to\infty} \sum_{d=k_0+1}^{\infty} \mathbb{P}(D = d) \frac{\mathbb{P}\Big(\sum_{i=1}^{d} \|Z_i A_i\| \geq n\varepsilon\Big)}{\mathbb{P}(\|Z_1 A_1\| \geq n)} \underset{=}{^{k_0\to+\infty}} o(1),$$

which concludes the proof. ■

The remainder of this section aims at providing some insight on specific properties of regular variation in $\mathbb{H}$, as compared with regular variation in general separable metric spaces as introduced by Hult and Lindskog (2006b) or, at the other end of the spectrum, regular variation in a Euclidean space. On the one hand, we focus on possible finite-dimensional characterizations of regular variation in $\mathbb{H}$, with a view towards statistical applications in which abstract convergence conditions in an infinite-dimensional space cannot be verified on real data, while finite-dimensional conditions may serve as a basis for statistical tests. Although we do not go as far as proposing such rigorous statistical procedures, we do suggest in the experimental section some convergence diagnostics relying on the results gathered in this section. On the other hand we discuss the relationships existing between regular variation in $\mathcal{C}[0,1]$ and regular variation in $\mathbb{H} = L^2[0,1]$.

5.2 Finite-dimensional Characterizations

regularly varying random elements in $\mathbb{H}$ have been present in the literature for a long time, due to strong connections between regular variation and domains of attraction of stable laws in general and in separable Hilbert spaces in particular. As an example, Kuelbs and Mendrekar Kuelbs and Mandrekar (1974) show (through their Lemma 4.1 and their Theorem 4.11) that a random element in $\mathbb{H}$ which is in the domain of attraction of a stable law with index $0 < \alpha < 2$ is regularly varying. However this connection does not yield any finite-dimensional characterization which are our main focus here.

We first recall Proposition 2.1 from Kim and Kokoszka (2022), making a first connection between RV in $\mathbb{H}$ and regular variation of finite-dimensional projections. Let $(e_i)_{i\geq 1}$ be a complete orthonormal system in $\mathbb{H}$. For $\mathcal{I} = (i_1, \dots i_N)$ a finite set of indices with cardinality $N \geq 1$, denote by $\pi_{\mathcal{I}}$ the 'coordinate projection' on the associated finite family, $\pi_{\mathcal{I}}(x) = (\langle x, e_{i_1}\rangle, \dots, \langle x, e_{i_N}\rangle), x \in \mathbb{H}$. In particular we denote by $\pi_N : \mathbb{H} \to \mathbb{R}^N$ the projection onto the N first elements of the basis $(e_i)_{i\geq 1}$.

Proposition 5.3 (regular variation in $\mathbb{H}$ implies multivariate regular variation of finite-dimensional projections)**.** *Let X be a random element of $\mathbb{H}$ that is regularly varying with index $\alpha > 0$. Then, for all finite index set $\mathcal{I}$ of size $N \geq 1$, the multivariate random variable $\pi_{\mathcal{I}} X$ is also regularly varying in $\mathbb{R}^N$.*

One natural question to ask is whether the converse of Proposition 5.3 is true. We answer in the negative in Proposition 5.4 below.

Proposition 5.4 (Multivariate regular variation of finite-dimensional projections does not imply regular variation in $\mathbb{H}$)**.** *The converse of Proposition 5.4 is not true in general. In particular there exists a random element X in $\mathbb{H}$ which is not regularly varying, while, for any $\alpha > 0$,*

1. *for all $N \geq 1$, $\pi_N X$ is regularly varying in $\mathbb{R}^N$ with same index α ;*
2. *the norm of X is regularly varying in $\mathbb{R}$ with index α.*

Sketch of Proof. We construct a random element X in $\mathbb{H}$ in such a way that the probability mass of its angular component Θ, given the radial component R, escapes at infinity as R grows. Here, *at infinity* must be understood as $\text{span}(e_i, i \geq M)$ as $M \to +\infty$. Namely, let $X := R\Theta$ with radial component $R = \|X\| \sim Pareto(\alpha)$ on $[1, +\infty[$ (*i.e.*, $\forall t \geq 1, \mathbb{P}\Big(R_0 \geq t\Big) = t^{-\alpha}$) and define the conditional distribution of Θ given R as the mixture of Dirac masses:

$$\mathcal{L}(\Theta|R) = \frac{1}{\sum_{l=1}^{\lfloor R \rfloor} 1/l} \sum_{i=1}^{\lfloor R \rfloor} \frac{1}{i} \delta_{e_i}.$$

In other words, for $i \leq R$, we have $\Theta = e_i$ with probability proportional to $1/i$. The remainder of the proof, deferred to Appendix 5.A, consists in verifying that (i) all finite-dimensional projections of X are regularly varying; (ii) asymptotic finite-dimensionality (see Definition 3.14) of the family of conditional distributions $P_{\Theta,t}$ does not hold, hence it may not converge to any limit distribution, so that Condition (ii) from Proposition 2.18 does not hold and X may not be regularly varying.

■

The counter-example above suggests that the missing assumption to obtain regular variation in $\mathbb{H}$ is some relative compactness criterion. This is partly confirmed in the next example where the angular variables Θ_t is again a mixture model supported by the e_i's but where the probability mass for the conditional distribution of Θ given $\|X\|$ concentrates around finite-dimensional spaces. The proof, postponed to the Appendix, proceeds by verifying both conditions from Proposition 2.18.

Proposition 5.5. *Let $R \sim Pareto(\alpha)$ on $[1, \infty[$ and define Θ through its conditional distribution given $R = r$, for $r \geq 1$,*

$$\mathcal{L}(\Theta|R = r) = \frac{1}{\sum_{l=1}^{\lfloor r \rfloor} 1/l^2} \sum_{i=1}^{\lfloor r \rfloor} \frac{1}{i^2} \delta_{e_i}.$$

In words, $\Theta \in \{e_1, e_2, ...\}$ and $\forall r \geq 1, \forall 1 \leq j \leq r$, we have $\mathbb{P}(\Theta = e_j \mid R = r) = \frac{1/j^2}{\sum_{l=1}^{\lfloor r \rfloor} 1/l^2}$.

Then, the random element $X = R\Theta$ is regularly varying in $\mathbb{H}$ with index α with limit angular random variable Θ_∞ given by

$$\mathbb{P}(\Theta_\infty = e_j) = \frac{6}{(\pi j)^2},$$

for $j \geq 1$.

The next proposition confirms the intuition given by the preceding examples that asymptotic finite-dimensionality is a necessary additional assumption to regular variation of finite-dimensional projections and of the norm.

Proposition 5.6. *Let X be a $\mathbb{H}$-valued random element. The two conditions below are equivalent.*

1. *X is regularly varying in $\mathbb{H}$ with index $\alpha > 0$, limit measure μ and positive normalizing sequence $(b_n)_{n\geq 1}$, i.e., $\mu_n = b_n \mathbb{P}(X \in n \cdot) \xrightarrow{\mathbb{M}_0} \mu(\cdot)$.*
2. *The family of measures $(\mu_n)_{n\geq 1}$ is relatively compact w.r.t. the $\mathbb{M}_0(\mathbb{H})$-topology, and for all $N \geq 1$, $\pi_N X$ is regularly varying in $\mathbb{R}^N$ with index $\alpha > 0$, limit measure μ_N and positive normalizing sequence $(b_n)_{n\geq 1}$.*

In particular, both the index α and the normalizing sequence $(b_n)_{n\geq 1}$ are the same in assertions 1. and 2. and, for all $N \geq 1$, $\mu_N = \mu \circ \pi_N$.

Proof. 1. $\Rightarrow$ **2.** If X is regularly varying as in the statement 1., then $(\mu_n)_{n\geq 1}$ converges in the $\mathbb{M}_0(\mathbb{H})$ topology and the family is of course relatively compact. Also fix $N \geq 1$ and notice that π_N is a continuous mapping from $(\mathbb{H}, \|\cdot\|)$ to $\mathbb{R}^N$ endowed with the Euclidean norm. The continuous mapping theorem in $\mathbb{M}_0$ (see Hult and Lindskog (2006b), Theorem 2.5) ensures that $\mu_n \circ \pi_N^{-1} \xrightarrow{\mathbb{M}_0} \mu \circ \pi_N^{-1}$ in $\mathbb{R}^N$.

2. $\Rightarrow$ **1.** If μ_n is relatively compact, the sequence μ_n converges in $\mathbb{M}_0(\mathbb{H})$ if and only if any two subsequential limits μ^1, μ^2 coincide. However it follows form the previous implication that in such a case, the finite-dimensional projections of μ^1 and μ^2 coincide, namely $\mu^1 \circ \pi_N^{-1} = \mu^2 \circ \pi_N^{-1} = \mu \circ \pi_N^{-1}$, for all $N \geq 1$. Consider the family of cylinder sets of $\mathbb{H}$ with measurable base, $\mathcal{C} = \{\pi_N^{-1}(A), A \in \mathcal{B}(\mathbb{R}^N), N \geq 1\}$. On $\mathcal{C}$ the measures μ, μ^1 and μ^2 coincide. The cylinder sets family $\mathcal{C}$ is a π-system which generates the Borel σ-field, because it is associated to the family of bounded linear functional $(e_i^*)_{i\geq 1}$ which separates points. Hence, μ, μ^1 and μ^2 coincide on every Borelian set and the proof is complete.

■

Remark 5.7 (Cramér-Wold device for regular variation.)**.** *One may naturally wonder whether a Cramér-Wold device could hold for RV. More precisely, a natural question to ask is whether the condition 'For some fixed $\alpha > 0$, for all $h \in \mathbb{H}$, the random variable $\langle h, X\rangle$ is regularly varying in $\mathbb{R}$ with index α,' would be sufficient to prove RV in $\mathbb{H}$ of the random element X. The answer is no, at least unless additional assumptions are made regarding α. Indeed, this question has been investigated already in the finite-dimensional setting by Basrak et al. (2002a); Hult and Lindskog (2006a). It is shown in Basrak et al. (2002a) that for X valued in $\mathbb{R}^d$, if $\alpha > 0$ is non-integer, then the latter condition indeed implies RV in $\mathbb{R}^d$. Conversely, Hult and Lindskog (2006a) have shown through a counter-example, that such an implication does not hold true when α is an integer, unless additional assumptions are made. As a consequence it cannot hold in a general Hilbert space $\mathbb{H}$ either. Whether or not the technical argument leading to the positive results of Basrak et al. (2002a) may be extended to the Hilbert space setting is left to further research, as this question is not central to our application to functional PCA in Section 6.*

The line of thought of Proposition 5.6 may be pursued further by characterizing the property of relative compactness of a family $(\nu_n)_{n\geq 1} \in \mathbb{M}_0(\mathbb{H})$ through asymptotic finite-dimensionality (see Definition 3.14), following the lines of the proof of Theorem 4.3 in Hult and Lindskog (2006b), relying in particular on Theorem 2.6 of the cited reference. However it is also possible to rely on known characterizations of relative compactness for probability measures, coupled with the polar characterization of regular variation (Proposition 2.18). We propose in this spirit the following simple characterization solely based on weak convergence of univariate and finite-dimensional projections, together with RV of the norm, without additional requirements regarding asymptotic finite-dimensionality. Recall that the distribution $\mathcal{L}(\Theta \mid \|X\| \geq t)$ is denoted $P_{\Theta,t}$.

Theorem 5.8. *Let X be a random element in $\mathbb{H}$ and let Θ_t be a random element in $\mathbb{H}$ distributed on the sphere $\mathbb{S}$ according to the conditional angular distribution $P_{\Theta,t}$. Let $P_{\Theta,\infty}$ denote a probability measure on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ and let Θ_∞ be a random element distributed*

according to $P_{\Theta,\infty}$. The following statements are equivalent.

1. *X is regularly varying with index α with limit angular measure $P_{\Theta,\infty}$, so that $P_{\Theta,t} \xrightarrow{w} P_{\Theta,\infty}$.*
2. *$\|X\|$ is regularly varying in $\mathbb{R}$ with index α, and*
$$\forall h \in \mathbb{H}, \langle \Theta_t, h\rangle \xrightarrow{w} \langle \Theta_\infty, h\rangle \quad \text{as } t \to +\infty.$$
3. *$\|X\|$ is regularly varying in $\mathbb{R}$ with index α, and*
$$\forall N \geq 1, \pi_N(\Theta_t) \xrightarrow{w} \pi_N(\Theta_\infty) \quad \text{as } t \to +\infty.$$

Proof. The fact that 1 implies 2 and 3 is a direct consequence of the polar characterization of regular variation (Proposition 2.18) and of the continuous mapping theorem applied to the bounded linear mappings h^*, $h \in \mathbb{H}$ and π_N, $N \geq 1$.

For the reverse implications (3 $\Rightarrow$ 1) and (2 $\Rightarrow$ 1), in view of Proposition 2.18, we only need to verify that for any sequence $t_n > 0$ such that $t_n \to +\infty$, $\Theta_{t_n} \xrightarrow{w} \Theta_\infty$ in $\mathbb{H}$. From Theorem 3.15, if either Condition 2 or Condition 3 holds true, then it will be so if and only if the family $(P_{\Theta,t_n})_{n\geq 1}$ is asymptotically finite-dimensional.

We use the fact, stated and proved in Tsukuda (2017), that if $(Z_n)_{n\geq 1}$ and Z are $\mathbb{H}$-valued random elements such that, as $n \to +\infty$,
$$\mathbb{E}\Big[\|Z_n\|^2\Big] \to \mathbb{E}\Big[\|Z\|^2\Big], \tag{5.1}$$
and, for all $j \geq 1$,
$$\mathbb{E}\Big[\langle Z_n, e_j\rangle^2\Big] \to \mathbb{E}\Big[\langle Z, e_j\rangle^2\Big], \tag{5.2}$$
then the sequence $(Z_n)_{n\geq 1}$ is asymptotically finite-dimensional.

With $Z_n = \Theta_{t_n}$ and $Z = \Theta_\infty$, Condition (5.1) above is immediately satisfied since $\|\Theta_{t_n}\| = \|\Theta_\infty\| = 1$ almost surely. For the same reason $\mathbb{E}\Big[\langle \Theta_{t_n}, e_j\rangle^2\Big] = \mathbb{E}\Big[\varphi(\langle \Theta_{t_n}, e_j\rangle)\Big]$, where φ is the bounded, continuous function $\varphi(z) = \min(z^2, 1)$. Thus, weak convergence of the projections $\langle \Theta_{t_n}, e_j\rangle$ (Condition 2 or 3 from the statement) together with the continuous mapping theorem imply (5.2), which concludes the proof. ∎

5.3 Regular Variation in $L^2[0,1]$ *vs* Regular Variation in $\mathcal{C}[0,1]$

Turning to the case where $\mathbb{H} = L^2[0,1]$, we discuss the relationships between the notions of RV in $L^2[0,1]$ and in $\mathcal{C}[0,1]$, the space of continuous functions on $[0,1]$. Indeed, any continuous stochastic process $(X_t, t \in [0,1])$ is also a random element in $\mathbb{H} = L^2[0,1]$, as proved in Hsing and Eubank (2015), Theorem 7.4.1, or 7.4.2. It is thus legitimate to ask whether regular variation with respect to one norm implies regular variation for the other norm for such stochastic processes.

Proposition 5.9. *Let X be a continuous process over $[0,1]$. Assume that $X \in RV_{-\alpha}(\mathcal{C}[0,1])$, with $\mathcal{L}(X/\|X\|_\infty | \|X\|_\infty > t) \to \mathcal{L}(\Theta_{\infty,\infty})$, as $t \to +\infty$, where $\Theta_{\infty,\infty}$ is the angular limit process w.r.t. the sup-norm $\|\cdot\|_\infty$. Then, $X \in RV_{-\alpha}(L^2[0,1])$ and the angular limit process $\Theta_{\infty,2}$ (w.r.t. the L^2 norm $\|\cdot\|$) has distribution given by*
$$\mathbb{P}(\Theta_{\infty,2} \in B) = \frac{\mathbb{E}\Big[\|\Theta_{\infty,\infty}\|^\alpha \mathbb{1}\{\Theta_{\infty,\infty}/\|\Theta_{\infty,\infty}\| \in B\}\Big]}{\mathbb{E}\Big[\|\Theta_{\infty,\infty}\|^\alpha\Big]}, \tag{5.3}$$

where $B \in \mathcal{B}(\{x \in L^2[0,1] : \|x\| = 1\})$.

Proof. Since $\|\cdot\|$ is homogeneous and continuous w.r.t. $\|\cdot\|_\infty$ in $\mathcal{C}[0,1]$, Theorem 3 in Dombry and Ribatet (2015) applies (upon choosing $\ell(X) = \|X\|$ with the notations of the cited reference), which yields regular variation of X in $L^2[0,1]$, together with the expression given in (5.3) for the angular measure associated with the L^2 norm $\|\cdot\|$. ■

One may wonder whether the converse is also true, *i.e.*, if $X \in \mathcal{C}[0,1] \cap RV_{-\alpha}(L^2[0,1])$, is it necessarily the case that $X \in RV_{-\alpha}(\mathcal{C}[0,1])$? A counter-example is given in the next proposition.

Proposition 5.10. *The converse statement of Proposition 5.9 is not true in general. There exists a sample-continuous stochastic process indexed by* $t \in [0,1]$ *which is regularly varying in* $L^2[0,1]$ *but not in* $\mathcal{C}[0,1]$.

Proof. We construct a 'spiked' continuous process with controlled L^2-norm, while its sup-norm is super-heavy tailed. Let Z be drawn from a Pareto distribution with parameter $\alpha_Z > 0$, *i.e.*, $\mathbb{P}(Z \geq t) = t^{-\alpha_Z}$ for $t \geq 1$, and define the sample-continuous stochastic process

$$Y(t) = \left(1 - \frac{t}{3Z^2 \exp(-2Z)}\right) \exp(Z) \mathbb{1}\{[0, 3Z^2 \exp(-2Z)[\}.$$

Straightforward computations yield $\|Y\|_\infty = \exp(Z)$ and $\|Y\|_2 = Z$. Let ρ be another independent Pareto-distributed variable with index $0 < \alpha_\rho < \alpha_Z$. Finally, define $X = \rho Y$. Then X is a sample-continuous stochastic process over $[0,1]$. We have $\|X\|_\infty = \rho \exp(Z)$, which is clearly not regularly varying because (see, *e.g.*, Mikosch (1999), Proposition 1.3.2) $\mathbb{E}[\|X\|_\infty^\delta] = +\infty$ for all $\delta > 0$. Thus, X is not regularly varying in $(\mathcal{C}[0,1], \|\cdot\|_\infty)$. On the other hand, the pair (ρ, Y) satisfies the assumptions of Proposition 5.1 with $d = 1$. Hence, $X = \rho Y$ is regularly varying in $\mathbb{H} = L^2[0,1]$.

■

Proposition 5.10 and Proposition 5.9 together show that the framework of $L^2[0,1]$-regular variation encompasses a wider classes of continuous processes than standard $\mathcal{C}[0,1]$ regular variation. This opens a road towards applications of EVT in situations where the relevant definition of an extreme event has to be understood in terms of 'energy' of the (continuous) trajectory, as measured by the L^2-norm, rather than in terms of sup-norm.

5.4 Conclusion

In this chapter, we have conducted an extensive exploration of the concept of RV within Hilbert spaces and established finite-dimensional characterizations. Numerous examples to illustrate these characterizations and counterexamples to underscore the necessity of the various assumptions are provided. A comparison between the RV in $L^2[0,1]$ and the more extensively studied RV in $\mathcal{C}[0,1]$ is also presented. These results can be useful to develop applications at the edge of EVT and signal processing, such as the Principal Component Analysis framework for functional extremes introduced in the next chapter.

5.A Proofs

Proof of Proposition 5.4. Consider as in the sketch of the proof the random element $X := R\Theta$ valued in $\mathbb{H}$, with radial component $R = \|X\| \sim Pareto(\alpha)$ on $[1, +\infty[$, *i.e.*, $\forall t \geq 1, \mathbb{P}\Big(R \geq t\Big) = t^{-\alpha}$ and with angular component $\Theta = X/\|X\|$ defined through its conditional distribution

$$\mathcal{L}(\Theta|R) = \frac{1}{\sum_{k=1}^{\lfloor R \rfloor} 1/k} \sum_{j=1}^{\lfloor R \rfloor} \frac{1}{j} \delta_{e_j}.$$

Notice that $r \mapsto \sum_{i=1}^{\lfloor r \rfloor} \frac{1}{i}$ is slowly varying since we have $\sum_{i=1}^{\lfloor r \rfloor} 1/i \sim \log r$ as $r \to +\infty$. We now check that X satisfies the properties listed in the statement. That $\|X\| = R \in RV_{-\alpha}$ (Condition 1) is obvious. Fix $N \geq 1$, recall that $\pi_N(X)$ is the $\mathbb{R}^N$-valued projection onto the first N elements of the basis $(e_i)_{i\geq 1}$ and denote by $\Theta_N = \pi_N(X)/\|\pi_N(X)\|$ the associated angular component in $\mathbb{R}^N$. Denote also by $R_N = \|\Pi_N(X)\|_{\mathbb{R}^N}$ the radial components of $\pi_N(X)$ in $\mathbb{R}^N$. First, we show that $R_N \in RV_{-\alpha}$. Observe that for all $t \geq N$,

$$\begin{aligned}
\mathbb{P}(R_N \geq t) &= \mathbb{P}(R_N \geq t, R \geq t) \\
&= \mathbb{P}(\Theta \in \{e_1, ..., e_N\}, R \geq t) = \mathbb{E}[\mathbb{1}\{R \geq t\}\mathbb{P}\{\Theta \in \{e_1, ..., e_N\} \mid R\}] \\
&= \mathbb{E}\left[\mathbb{1}\{R \geq t\} \frac{\sum_{i=1}^{N} 1/i}{\sum_{l=1}^{\lfloor R \rfloor} 1/l}\right] = \sum_{i=1}^{N} \frac{1}{i} \int_t^{\infty} \frac{\alpha r^{-(\alpha+1)}}{\sum_{l=1}^{\lfloor r \rfloor} 1/l} \mathrm{d}r.
\end{aligned}$$

Since $r \mapsto \alpha r^{-(\alpha+1)}/(\sum_{l=1}^{\lfloor r \rfloor} 1/l) \in RV_{-(\alpha+1)}$, we have $R_N \in RV_{-\alpha}$ by virtue of Karamata's theorem, see Theorem 2.5.

We next prove that $\mathcal{L}(\Theta_N|R_N \geq t)$ weakly converges as $t \to +\infty$. First, since for all t, the measure $\mathbb{P}(\Theta_N \in \cdot \mid R_n > t)$ is supported by the finite set $\{e_1, ..., e_N\}$. It is sufficient to show convergence of each $\mathbb{P}(\Theta_N = e_j \mid R_N \geq t)$ towards some $p_j \geq 0$ for all j, with $\sum_{j\leq N} p_j = 1$. For fixed $j \leq N$ and for $t \geq N$,

$$\begin{aligned}
\mathbb{P}(\Theta_N = e_j \mid R_N \geq t) &= \frac{\mathbb{P}(\Theta = e_j, R_N \geq t)}{\mathbb{P}(R_N \geq t)} = \frac{\mathbb{P}(\Theta = e_j, R \geq t)}{\mathbb{P}(R_N \geq t)} \\
&= \frac{\mathbb{P}(\Theta = e_j, R \geq t)}{\mathbb{P}(\Theta \in \{e_1, \ldots, e_N\}, R \geq t)} = \frac{\mathbb{E}\left[\mathbb{1}\{R \geq t\} \frac{1/j}{\sum_{l=1}^{\lfloor R \rfloor} 1/l}\right]}{\mathbb{E}\left[\mathbb{1}\{R \geq t\} \frac{\sum_{i=1}^{N} 1/i}{\sum_{l=1}^{\lfloor R \rfloor} 1/l}\right]} = \frac{1/j}{\sum_{l=1}^{N} 1/l}.
\end{aligned}$$

Hence, when t is large enough, $\mathcal{L}(\Theta_N|R_N \geq t) = \sum_{i=1}^{N} (1/i)\delta_{e_i} / \sum_{l=1}^{N} (1/l)$. We have shown that for all $N \geq 1$, $\pi_N(X)$ is regularly varying in $\mathbb{R}^N$ with tail index $-\alpha$.

We now show that $X \notin \mathrm{RV}(\mathbb{H})$. Since $\|X\| \in \mathrm{RV}_{-\alpha}(\mathbb{R})$, according to Proposition 2.18, we have to prove that $\mathcal{L}(\Theta \mid R \geq t)$ does not converge when t tends to infinity. From Theorem 1.8.4. in van der Vaart and Wellner (1996), it is enough to show that the sequence of measures $P_{\Theta,n} = \mathbb{P}(\Theta \in \cdot \mid R > n)$ is not asymptotically finite-dimensional. *i.e.*, that

$$\exists \delta, \varepsilon > 0, \forall d \in \mathbb{N}^*, \limsup_n \mathbb{P}\left(\sum_{i>d} \langle \Theta, e_i \rangle^2 \geq \delta \mid R \geq n\right) \geq \varepsilon.$$

Let $\delta > 0, \varepsilon \in]0,1[$ and $n > d$.

$$\mathbb{P}\Big(\sum_{i>d}\langle\Theta, e_i\rangle^2 \geq \delta \mid R \geq n\Big) = \frac{\mathbb{P}(\Theta \notin \{e_1,\dots,e_d\}, R \geq n)}{\mathbb{P}(R \geq n)}$$
$$= \frac{\mathbb{E}\Big[\mathbb{1}\{R \geq n\}\mathbb{P}(\Theta \notin \{e_1,\dots,e_d\} \mid R)\Big]}{\mathbb{P}(R \geq n)} = \frac{\mathbb{E}\Big[\mathbb{1}\{R \geq n\}\Big(1 - \frac{\sum_{i=1}^{d} 1/i}{\sum_{l=1}^{\lfloor R \rfloor} 1/l}\Big)\Big]}{\mathbb{P}\Big(R \geq n\Big)}$$
$$= \mathbb{E}\Big[1 - \frac{\sum_{i=1}^{d} 1/i}{\sum_{l=1}^{\lfloor R \rfloor} 1/l} \mid R \geq n\Big] \geq 1 - \frac{\sum_{i=1}^{d} 1/i}{\sum_{l=1}^{n} 1/l} \xrightarrow[n\to+\infty]{} 1 \geq \varepsilon. \quad (5.4)$$

Hence, the asymptotic finite-dimensional condition does not hold and X is not regularly varying in $\mathbb{H}$. ■

Proof of the claim in Proposition 5.5. We show that X is regularly varying in $\mathbb{H}$. Following the lines of the proof of Proposition 5.4, it is enough to verify that $P_{\Theta,t} \xrightarrow{w} \Theta_\infty$. Since the common support of $P_{\Theta,t}$ and $P_{\Theta,\infty}$ is discrete we only need to show that $\mathbb{P}(\Theta = e_j \mid R > t) \to 6/(\pi j)^2$ for fixed $j \geq 1$.

For such j, following the steps leading to (5.4), we find

$$\mathbb{P}(\Theta = e_j \mid R \geq t) = \frac{\mathbb{E}\Big[\mathbb{1}\{R \geq t\}\mathbb{P}(\Theta = e_j \mid R)\Big]}{\mathbb{P}(R \geq t)} = \frac{\mathbb{E}\Big[\mathbb{1}\{R \geq t\} j^{-2} / \sum_{l=1}^{\lfloor R \rfloor} l^{-2}\Big]}{\mathbb{P}(R \geq t)}$$
$$= j^{-2}\mathbb{E}\Big[1/\sum_{l=1}^{\lfloor R \rfloor} l^{-2} \,\Big|\, R \geq t\Big] = j^{-2}\int_t^\infty \Big(\sum_{l=1}^{\lfloor r \rfloor} l^{-2}\Big)^{-1} \frac{\alpha t^\alpha}{r^{\alpha+1}}\,\mathrm{d}r.$$

The integrand in the latter display is a regularly varying function of r with exponent $-\alpha - 1$, and an application of Karamata's theorem yields that $\mathbb{P}(\Theta = e_j \mid R \geq t) \to 6/(\pi j^2)$, as $t \to +\infty$. ■

Chapter 6

Principal Component Analysis for Functional Extremes

Contents

In Section 6.1, we introduce the key elements of our analysis, such as the pre-asymptotic covariance operator C_t, the limit covariance operator C_∞, and a distance ρ to measure the similarity between the eigenspaces of the involved covariance operators. Section 6.2 gathers results about the convergence the eigenstructure of C_t towards the eigenstructure of C_∞. In the situation where $n \geq 1$ independent realizations of the random function X are observed, in Section 6.3, we additionally provide statistical guarantees regarding empirical estimation of the pre-asymptotic covariance operator associated in the form of concentration inequalities regarding the Hilbert-Schmidt (HS) norm of the estimation error. These bounds, combined with Regular Variation (RV) of the observed random function X and the results from the preceding section ensure in particular the consistency of the empirical estimation procedure. In Section 6.4 we present experimental results involving real and simulated data illustrating the relevance of the proposed dimension reduction framework.

6.1 Characteristics of the Problem

This section revisits the problem of Principal Component Analysis (PCA) for the specific purpose of Extreme Value Analysis. Motivated by dimension reduction purposes, our goal is to construct a finite-dimensional representation of extreme functions. In other words our primary purpose is to learn a finite-dimensional subspace V of $\mathbb{H} = L^2[0,1]$ such that the orthogonal projections of extreme functions onto V are optimal in terms of angular reconstruction error. Throughout this section we place ourselves in the

setting of regular variation introduced in Section 5 and consider a regularly varying random element X in $\mathbb{H}$ as in Theorem 5.8, with the same notations. Our focus is thus on building a low-dimensional representation of the angular distribution of extremes $P_{\Theta,\infty}$ introduced in the introduction of Part II. We consider the eigendecomposition of the associated covariance operator

$$C_\infty = \mathbb{E}[\Theta_\infty \otimes \Theta_\infty] = \sum_{j\geq 1} \lambda_\infty^j \varphi_j^\infty \otimes \varphi_j^\infty,$$

where $\Theta_\infty \sim P_{\Theta,\infty}$, and the φ_j^∞'s and λ_∞^j's are eigenfunctions and eigenvalues of C_∞ following the notations of Section 3.3. If $P_{\Theta,\infty}$ is sufficiently concentrated around a finite-dimensional subspace of moderate dimension p, a reasonable approximation of $P_{\Theta,\infty}$ is provided by its image measure *via* the projection onto $V_\infty^p = \text{span}(\varphi_j^\infty, j \leq p)$. Independently from such sparsity assumptions, the space V_∞^p minimizes the reconstruction error (3.3.1) of the orthogonal projection relative to Θ_∞. It is also the unique minimizer as long as $\lambda_\infty^p > \lambda_\infty^{p+1}$, as discussed in the background Section 3.3.

Our main results bring finite sample guarantees regarding an empirical version of V_∞^p constructed using the $k \ll n$ observations with largest norm. In this respect our work may be seen as an extension of Drees and Sabourin (2021), where finite-dimensional observations $X \in \mathbb{R}^d$ are considered, to an infinite-dimensional ambient space. However, our proof techniques are fundamentally different from those involved in the aforementioned reference. Indeed their analysis relies on Empirical Risk Minimization arguments relative to the reconstruction risk at infinity, $R_\infty(V) = \lim_{t\to+\infty} \mathbb{E}[\|\Theta - \Pi_V \Theta\|^2 \mid R > t]$, where Π_V denotes the orthogonal projection onto V. The main ingredients of their analysis are (i) the fact that V_∞^p minimizes the risk at infinity (ii) compactness of the unit sphere (or of any bounded, closed subset of $\mathbb{R}^d$). In the present setting such compactness properties do not hold, and we follow an entirely different path, as we investigate the convergence of an empirical version of C_∞ in the Hilbert-Schmidt norm, and then rely on perturbation theory for covariance operators in order to control the deviations of its eigenspaces. We thus consider the pre-asymptotic covariance operator

$$C_t = \mathbb{E}[\Theta \otimes \Theta \mid R > t] = \mathbb{E}[\Theta_t \otimes \Theta_t]. \tag{6.1}$$

In the sequel, the discrepancy between finite-dimensional linear subspaces of $\mathbb{H}$ is measured in terms of the HS norm of the difference between orthogonal projections, namely we define a distance ρ between finite-dimensional subspaces V, W of $\mathbb{H}$, by

$$\rho(V, W) = \|\Pi_V - \Pi_W\|_{HS(\mathbb{H})}.$$

Incidentally, it should be noticed that Drees and Sabourin (2021) denote by ρ the operator norm of the difference between the projections and that their results regarding convergence of eigenspaces are stated relative to the operator norm. Since the HS norm dominates the operator norm, our results are indeed stronger in nature than those of the cited references as claimed in the Introduction, even in the finite-dimensional case.

The problem is now fully outlined, allowing us to refine the initial plan provided at the beginning of the chapter by specifying the nature of the forthcoming results. We show in Section 6.2 that the first p eigenfunctions of the pre-asymptotic operator C_t generate a vector space V_t^p converging to V_∞^p whenever $\lambda_\infty^p > \lambda_\infty^{p+1}$. Second, we establish in Section 6.3 the consistency of the empirical subspace $\widehat{V}_t^p$ (the one generated by the

first p eigenfunctions of an empirical version of C_t) and we derive nonasymptotic guarantees for its deviations, based on concentration inequalities regarding the empirical covariance operator.

6.2 The Pre-asymptotic Covariance Operator and its Eigenspaces

Since perturbation theory allows to control the deviations of eigenvectors and eigenvalues of a perturbed covariance operator, a natural first step in our analysis is to ensure that the pre-asymptotic operator C_t introduced in (6.1) may indeed be seen as a perturbed version of the asymptotic operator C_∞, as shown next.

Theorem 6.1 (Convergence of the pre-asymptotic covariance operator)**.** *In the setting of Theorem 5.8, as $t \to +\infty$, the following convergence in the Hilbert-Schmidt norm holds true,*

$$\|C_t - C_\infty\|_{HS(\mathbb{H})} \to 0\,.$$

Proof. Let $(t_n)_{n\geq 1}$ be a nondecreasing sequence of reals converging to infinity. Recall from Theorem 2.18 that regular variation of X implies weak convergence of the sequence Θ_{t_n} towards Θ_∞. Using the fact that the mapping $h \in \mathbb{H} \mapsto h \otimes h \in HS(\mathbb{H})$ is continuous, also $\Theta_{t_n} \otimes \Theta_{t_n}$ converges weakly towards $\Theta_\infty \otimes \Theta_\infty$.

Since the separability of $(\mathbb{H}, \langle\cdot,\cdot\rangle)$ implies the separability of $(HS(\mathbb{H}), \langle\cdot,\cdot\rangle_{HS(\mathbb{H})})$ (see Blanchard et al. (2007), Section 2.1), we may apply the Skorokhod's Representation theorem to the weakly converging sequence $\Theta_{t_n} \otimes \Theta_{t_n}$. Thus there is a probability space $(\Omega', \mathcal{F}, \mathbb{P}')$, and random elements Y_n, $n \geq 1$ and Y_∞ in $HS(\mathbb{H})$ defined on the probability space $(\Omega', \mathcal{F}, \mathbb{P}')$, such that $\Theta_{t_n} \otimes \Theta_{t_n} \overset{d}{=} Y_n$, $\Theta_\infty \otimes \Theta_\infty \overset{d}{=} Y_\infty$ and Y_n converges to Y_∞ almost surely with respect to $\mathbb{P}'$.

A Jensen's type inequality in Hilbert spaces (see pp. 42-43 in Ledoux and Talagrand (1991)) yields $\|C_{t_n} - C_\infty\|_{HS(\mathbb{H})} \leq \mathbb{E}[\|Y_n - Y_\infty\|_{HS(\mathbb{H})}]$. The dominated convergence theorem applied to the vanishing sequence of random variables $\|Y_n - Y_\infty\|_{HS(\mathbb{H})}$ (which are bounded by the constant 2) completes the proof.

■

Remark 6.2. *An alternative way to obtain the weak convergence of $\Theta_t \otimes \Theta_t$, which is key in the proof of Theorem 6.1, is to leverage Proposition 3.2 in Kokoszka et al. (2019), which ensures that the operator $X \otimes X$ is regularly varying in $HS(\mathbb{H})$. Since $\Theta \otimes \Theta$ is indeed the angular component of $X \otimes X$, the result follows by an application of Proposition 2.18.*

The next result concerns the convergence of eigenspaces and is obtained by combining tools from operator perturbation theory with the result from Theorem 6.1. In order to avoid additional technicalities we consider in the next statement an integer p such that $\lambda_\infty^p > \lambda_\infty^{p+1} \geq 0$, that is, a positive the spectral gap. Notice that such a p necessarily exists since $\|C_\infty\|_{HS(\mathbb{H})}^2 = \sum_{j=1}^\infty (\lambda_\infty^j)^2 < \infty$.

Corollary 6.3 (Convergence of pre-asymptotic eigenspaces)**.** *Let $p \in \mathbb{N}^*$ be such that $\lambda_\infty^p > \lambda_\infty^{p+1}$. Then, as t tends to infinity,*

$$\rho(V_t^p, V_\infty^p) \to 0.$$

Proof. According to Theorem 3 in Zwald and Blanchard (2005), for A and B two Hilbert-Schmidt operators on a separable Hilbert space, and an integer p such that the ordered eigenvalues of A satisfy $\lambda^p(A) > \lambda^{p+1}(A)$, if $\|B\|_{HS(\mathbb{H})} < \gamma^p := \frac{\lambda^p(A)-\lambda^{p+1}(A)}{2}$ is such that $A+B$ is still a positive operator, then following inequality holds

$$\rho(V^p, W^p) \leq \frac{\|B\|_{HS(\mathbb{H})}}{\gamma^p},$$

where V^p and W^p are respectively the eigenspaces spanned by the first p eigenvectors of A and $A+B$. From Theorem 6.1, the operators $A = C_\infty$ and $B = C_\infty - C_t$ satisfy the required assumptions stated above for t sufficiently large, and $\|B\|_{HS(\mathbb{H})}$ may be chosen arbitrarily small, which concludes the proof. ■

Remark 6.4 (Convergence of eigenvalues and choice of p)**.** *Even though the eigenvalues of C_∞ are not the main focus of our work, they are involved in the conditions of Corollary 6.3 through the requirement of a positive spectral gap. Of course these eigenvalues are unknown, however Weyl's inequality (see Theorem 3.11) ensures that* $\sup_{j\geq 1} |\lambda_t^j - \lambda_\infty^j| \leq \|C_t - C_\infty\|_{HS(\mathbb{H})}$*. Identification of an integer p for which the eigengap is positive may thus be achieved using consistent estimates of the λ_t^j's for t large enough.*

6.3 Empirical Estimation: Consistency and Concentration Results

We now turn to statistical properties of empirical estimates of C_t and its eigendecomposition based on an independent sample $X_1, ..., X_n$ distributed as X. For simplicity we shall assume in the sequel that the radial variable has no atoms ($\mathbb{P}(R = r) = 0$ for all $r > 0$) in order to avoid ambiguities in the definition of order statistics of the norm. This (mild) assumption could be relaxed at the price of additional minor technicalities, *e.g.*, by assuming that it holds only above a certain high quantile, and assuming that k/n is small enough.

Let $t_{n,k}$ denote the generalized quantile of the norm of order $1 - k/n$, namely $t_{n,k} = \inf\{t > 0 : \mathbb{P}(\|X\| \leq t) \geq 1 - k/n\}$. Then $t_{n,k}$ is uniquely defined and by the above assumption it holds that $\mathbb{P}(\|X\| \geq t_{n,k}) = k/n$. Denote by $X_{(1)}, \dots X_{(n)}$ the permutation of the sample such that $\|X_{(1)}\| \geq \|X_{(2)}\| \geq ... \geq \|X_{(n)}\|$. Also by the above assumption, there are no ties among the $\|X_i\|$'s and this permutation is again uniquely defined, with probability one. Accordingly, let $\Theta_{(i)}, R_{(i)}$ denote the angular and radial components of $X_{(i)}$. Then $\|X_{(k)}\| = R_{(k)}$ is an empirical version of $t_{n,k}$, which we shall sometimes denote by $\widehat{t_{n,k}}$. Following standard practice in Peaks-Over-Threshold analysis, we consider a fixed number of excesses k above the latter random radial threshold. Even though our main results are of non-asymptotic nature, letting $k, n \to \infty$ with $k/n \to 0$ yields consistency guarantees such as Corollary 6.11 below. Equipped with these notations the pre-asymptotic covariance operator at threshold $t_{n,k}$ is

$$C_{t_{n,k}} := \mathbb{E}[\Theta_{t_{n,k}} \otimes \Theta_{t_{n,k}}] = \frac{n}{k}\mathbb{E}[\Theta \otimes \Theta \mathbb{1}\{R \geq t_{n,k}\}],$$

and its empirical counterpart is given by

$$\widehat{C_k} := \frac{1}{k}\sum_{i=1}^{k} \Theta_{(i)} \otimes \Theta_{(i)}.$$

Remark 6.5 (Choice of k)**.** *Choosing the number k of observations considered as extreme is key in practice and corresponds to a difficult and recurrent topic in EVT. A wide variety of methods have been proposed in univariate problems (Caeiro and Gomes (2016); Scarrott and MacDonald (2012)), some rules of thumb exist in multivariate settings, based on visual inspection of angular histograms (Coles and Tawn (1994)) or stability under rescaling of the radial distribution (Stărică (1999)) with little theoretical foundations. We leave this question outside the scope of the paper. However, visual diagnostics are proposed in our numerical study based on Hill plots and convergence checking based on the finite-dimensional characterizations of RV stated in Theorem 5.8.*

Our analysis of the statistical error $\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})}$ involves the intermediate pseudo empirical covariance

$$\overline{C}_t := \frac{1}{\mathbb{P}(\|X_1\| \geq t)} \frac{1}{n} \sum_{i=1}^{n} \Theta_i \otimes \Theta_i \mathbb{1}\{R_i \geq t\}.$$

evaluated at $t = t_{n,k}$. Since $t_{n,k}$ is unknown, $\overline{C}_{t_{n,k}} = k^{-1} \sum_{i=1}^{n} \Theta_i \otimes \Theta_i \mathbb{1}\{R_i \geq t_{n,k}\}$ is not observable, although its deviation from $\widehat{C}_k$ may be controlled by the classical Bernstein inequality (Proposition 6.7). Our point of departure is the following decomposition of the statistical error,

$$\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})} \leq \|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})} + \|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})}. \tag{6.2}$$

We analyze separately the two terms in the right-hand side of (6.2) in the next two propositions.

Proposition 6.6. *Let $\delta \in (0,1)$. With probability larger than $1 - \delta/2$, we have*

$$\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})} \leq \frac{1 + 4\sqrt{\log(2/\delta)}}{\sqrt{k}} + \frac{8\log(2/\delta)}{3k}.$$

Sketch of Proof. A Bernstein-type concentration inequality from McDiarmid (1998) which is applicable to arbitrary functions of n variables with controlled conditional variance and conditional range (Theorem 3.8 of the reference, recalled in Lemma 4.6 from Chapter 4) ensures that

$$\mathbb{P}\Big(\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})} - \mathbb{E}\Big[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\Big] \geq \varepsilon\Big) \leq \exp\left(\frac{-k\varepsilon^2}{4(1+\varepsilon/3)}\right).$$

In order to control the expected deviation $\mathbb{E}\Big[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\Big]$ in the left-hand side, we first use the bound

$$\mathbb{E}\Big[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\Big] \leq \mathbb{E}\Big[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|^2_{HS(\mathbb{H})}\Big]^{1/2}$$

and then the fact that, if $A_1, \ldots, A_n$ are independent centered $\mathbb{H}$-valued random elements, $\mathbb{E}\Big[\|\sum_{i=1}^{n} A_i\|^2\Big] = \sum_{i=1}^{n} \mathbb{E}\Big[\|A_i\|^2\Big]$ (Lemma 6.14 in Appendix 6.A). We apply this result to A_i chosen as the deviation of the operator $\Theta_i \otimes \Theta_i \mathbb{1}\{R_i \geq t_{n,k}\}$ from its expectation, which yields

$$\mathbb{E}\Big[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\Big] \leq 1/\sqrt{k}.$$

Lemma 6.15 finishes the proof, as detailed in Appendix 6.A. ∎

We now turn to the second term $\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})}$ in the error decomposition (6.2).

Proposition 6.7. *Let $\delta \in (0,1)$. With probability larger than $1-\delta/2$, we have*

$$\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})} \leq \sqrt{\frac{8\log(4/\delta)}{k}} + \frac{4\log(4/\delta)}{3k}\,.$$

Proof. Because the $\Theta_i \otimes \Theta_i$'s have HS norm equal to one, we may write

$$\begin{aligned}
\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})} &= \frac{1}{k}\left\| \sum_{i=1}^{n} \Theta_i \otimes \Theta_i (\mathbb{1}\{t_{n,k} \leq R_i\} - \mathbb{1}\{R_{(k)} \leq R_i\}) \right\|_{HS(\mathbb{H})} \\
&\leq \frac{1}{k} \sum_{i=1}^{n} |\, \mathbb{1}\{t_{n,k} \leq R_i\} - \mathbb{1}\{R_{(k)} \leq R_i\}\,| \\
&= \frac{\mathbb{1}\{t_{n,k} \leq R_{(k)}\}}{k} \sum_{i=1}^{n} (\mathbb{1}\{t_{n,k} \leq R_i\} - \mathbb{1}\{R_{(k)} \leq R_i\}) + \cdots \\
&\qquad \frac{\mathbb{1}\{t_{n,k} > R_{(k)}\}}{k} \sum_{i=1}^{n} (\mathbb{1}\{R_{(k)} \leq R_i\} - \mathbb{1}\{t_{n,k} \leq R_i\}).
\end{aligned}$$

Also, since we have assumed that the distribution of R has no atoms we have with probability one, $R_{(1)} > R_{(2)} > \ldots R_{(k)} > R_{(k+1)}$. Thus the number of R_i's such that $R_i \geq R_{(k)}$ is exactly k, so that

$$\begin{aligned}
\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})} &\leq \frac{\mathbb{1}\{t_{n,k} \leq R_{(k)}\}}{k}\left[\left(\sum_{i=1}^{n} \mathbb{1}\{t_{n,k} \leq R_i\}\right) - k\right] + \cdots \\
&\qquad \frac{\mathbb{1}\{t_{n,k} > R_{(k)}\}}{k}\left(k - \sum_{i=1}^{n} \mathbb{1}\{t_{n,k} \leq R_i\}\right) \\
&= \frac{1}{k}\left|k - \sum_{i=1}^{n} \mathbb{1}\{R_i \geq t_{n,k}\}\right|,
\end{aligned}$$

where the last line follows from the fact that on the event $\{t_{n,k} \leq R_{(k)}\}$ it holds that $\sum_{i=1}^{n} \mathbb{1}\{t_{n,k} \leq R_i\} \geq k$, while on the complementary set, the inequality is reversed.

Notice that $\sum_{i=1}^{n} \mathbb{1}\{R_i \geq t_{n,k}\}$ follows a Binomial distribution with parameters $(n, k/n)$. The (classic) Bernstein's inequality as stated, *e.g.*, in McDiarmid (1998), Theorem 2.7, yields

$$\mathbb{P}\Big(\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})} \geq \varepsilon\Big) \leq \mathbb{P}\left(\left|\sum_{i=1}^{n} \mathbb{1}\{R_i \geq t_{n,k}\} - k\right| \geq k\varepsilon\right) \leq 2\exp\left(\frac{-k\varepsilon^2}{2(1+\varepsilon/3)}\right).$$

Solving for ε and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any nonnegative numbers a, b, we obtain the upper bound in the statement. ∎

We are now ready to state a non-asymptotic guarantee regarding the deviations (in the Hilbert-Schmidt norm) of the empirical covariance operator.

Theorem 6.8. *Let $\delta \in (0,1)$. With probability larger than $1-\delta$, we have*

$$\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})} \leq \frac{1 + 4\sqrt{\log(2/\delta)} + \sqrt{8\log(4/\delta)}}{\sqrt{k}} + \frac{8\log(2/\delta) + 4\log(4/\delta)}{3k}.$$

Proof. Observe that the following inclusion between adverse events holds true because of (6.2),

$$\left\{\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})} \geq \varepsilon_1 + \varepsilon_2\right\} \subset \left\{\|\widehat{C}_k - \overline{C}_{t_{n,k}}\|_{HS(\mathbb{H})} \geq \varepsilon_1\right\} \cup \left\{\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})} \geq \varepsilon_2\right\},$$

for all $\varepsilon > 0$. A union bound, Proposition 6.6 and Proposition 6.7 conclude the proof. ■

Remark 6.9 (Tightness of the upper bound, asymptotics)**.** *The bound obtained in Theorem 6.8 constitutes a minimal guarantee regarding covariance estimation of the extremes. By no means do we claim optimality regarding the multiplicative constants, which we have not tried to optimize, as revealed by an inspection fo the proof where the decomposition of the adverse event into two events of same probability may be sub-optimal. However the leading term of the error as $k \to +\infty$ is an explicit, moderate constant and the rate of convergence is $1/\sqrt{k}$, which matches known asymptotic rates in the literature of tail empirical processes in the univariate or multivariate case (see, e.g., Einmahl and Mason (1988) or Aghbalou et al. (2024a), Theorem 3). We leave to further research the question of the asymptotic behavior of $\widehat{C}_k - C_{t_{n,k}}$ as $k, n \to +\infty$, $k/n \to 0$, a problem which could be attacked by means of Lindeberg central limit theorems in Hilbert spaces (Kundu et al. (2000)).*

Combining Theorem 6.1 and Theorem 6.8, the following consistency result is immediate.

Corollary 6.10 (Consistency)**.** *The empirical covariance of extreme angles $\widehat{C}_k$ is consistent, i.e., as n, $k \to +\infty$ such that $k/n \to 0$, we have*

$$\|\widehat{C}_k - C_\infty\|_{HS(\mathbb{H})} \to 0 \text{ in probability.}$$

Theorem 6.8 also provides a control of the deviations of the empirical eigenspaces, with a proof paralleling the one of Corollary 6.3. In the following statement we denote by $\widehat{V}_k^p$ such an eigenspace, that is, the linear space generated by the first p eigenfunctions of $\widehat{C}_k$.

Corollary 6.11 (Deviations of empirical eigenspaces)**.** *Let $p \geq 1$ satisfying the same positive eigengap assumption as in Corollary 6.3, that is $\gamma_\infty^p := (\lambda_\infty^p - \lambda_\infty^{p+1})/2 > 0$. Denote the pre-asymptotic eigengap by*

$$\gamma_t^p = \frac{\lambda_t^p - \lambda_t^{p+1}}{2}.$$

Let n, k be large enough so that $\gamma_{t_{n,k}}^p > 0$ (see Remark 6.4 for the fact that $\gamma_{t_{n,k}}^p \to \gamma_\infty^p > 0$). For $\delta \in (0,1)$, with probability larger than $1-\delta$, we have

$$\rho(\widehat{V}_k^p, V_{t_{n,k}}^p) \leq \frac{B(n,k,\delta)}{\gamma_{t_{n,k}}^p},$$

where $B(n,k,\delta)$ is the upper bound on the deviations of $\widehat{C}_k$ stated in Theorem 6.8. In particular, we have the following consistency result: as n, $k \to +\infty$ s.t. $k/n \to 0$,

$$\rho(\widehat{V}_k^p, V_\infty^p) \to 0 \text{ in probability.}$$

Remark 6.12 (Uncentered *vs* centered covariance operators)**.** *Throughout this article, we only consider uncentered covariance operators, which leads to uncentered PCA. We have chosen to consider uncentered PCA for the sake of notational simplicity mainly. Also the*

mathematical expressions do not involves first moment terms, which shortens the proofs of our main statistical and probabilistic results.

Whereas centered PCA aims at finding directions of highest variability around the mean, uncentered PCA exhibits directions of highest absolute variability (around 0). An in-depth comparison between the two is presented in Cadima and Jolliffe (2009) in the finite-dimensional case.

Considering the feasibility of an extension of our results to centered PCA, notice first that since Θ_t is bounded, the continuous mapping theorem applies and entails that in the setting of Theorem 6.1, $\mathbb{E}[\Theta_t] \to \mathbb{E}[\Theta_\infty]$. Consequently, we have $\Theta_t - \mathbb{E}[\Theta_t] \xrightarrow{w} \Theta_\infty - \mathbb{E}[\Theta_\infty]$, thus the arguments of the proof remain valid when considering centered covariance operators. Concerning extending our concentration results of Section 8.2, It is worth noting that such extensions have been obtained for kernel PCA, outside the extreme value setting, by Blanchard et al. (2007), with 'slow' rates of convergence of order $O(1/\sqrt{n})$. However fast convergence rates of order $O(1/n)$ are obtained in the latter reference in the uncentered case *only, based on localized risk-minimization arguments which are significantly different from our techniques of proof. The authors leave as an open question the possibility to obtain fast rates for centered PCA, as empirical centering induces additional slow rate terms of order $O(1/\sqrt{n})$ in their analysis. Because our statistical results in this paper consist of slow rates only (of order $O(1/\sqrt{k})$) it is reasonable to conjecture that accounting for the error attached to first moment estimation would merely bring additional terms of order $O(1/\sqrt{k})$ in the upper bound, which would not change the nature of our results.*

6.4 Illustrative Numerical Experiments

Two possible applications of PCA for functional extremes are considered here. In both contexts, our goal is to assess the usefulness of the proposed functional PCA method for extremes by comparing it with the closest alternative, namely functional PCA of the full sample (not only extremes). On the one hand, a typical objective is to identify likely profiles of extreme events, by which we mean a finite-dimensional subspace of $\mathbb{H}$ with basis given by the eigenfunctions of C_∞ with the highest eigenvalue. In this context, extreme functional PCA serves as a pattern identification tool for a qualitative interpretation. This line of thought is illustrated in Section 6.4.1 on a toy simulated dataset in the multiplicative model of Proposition 5.1.

On the other hand, functional PCA of extremes may be viewed as a data compression tool allowing to represent functional extremes in a finite-dimensional manner, with optimal reconstruction properties which would not be achieved by standard functional PCA. The relevance of this approach is demonstrated in Section 6.4.2 with an electricity demand dataset which is publicly available on the `CRAN` network. On this occasion we also propose visual diagnostics for functional regular variation according to finite-dimensional characterizations proposed in Section 5.

The electricity demand dataset `sundaydemand` considered in Section 6.4.2 is available in the `R` package `fds`. It contains half-hourly electricity demands on Sundays in Adelaide between 6/7/1997 and 31/3/2007. It is made of $n = 508$ observations X_i, each of them being represented as a vector of size 48, indicating the recorded half-hour demand on day i. Here an 'angle' is in practice the profile of the half-hour records over one day, *i.e.*, the original curve rescaled by its L^2-norm.

In our toy example (Section 6.4.1) we generate a functional regularly varying dataset of same dimension $d = 48$ with larger sample size $n = 10e3$, according to Proposition 5.1. With the notations of the latter example, we choose $Z \in \mathbb{R}^6$ with independent components, with $Z_1 \sim \text{Pareto}(0.5)$, $Z_2 \sim 0.8 * \text{Pareto}(0.5)$, $Z_3 \sim \mathcal{N}(m = 0, \sqrt{\sigma^2} = 20)$, $Z_4 \sim \mathcal{N}(m = 0, \sqrt{\sigma^2} = 0.8 * 20)$, $Z_5 \sim \mathcal{N}(m = 0, \sqrt{\sigma^2} = 0.6 * 20)$, $Z_6 \sim \mathcal{N}(m = 0, \sqrt{\sigma^2} = 0.4 * 20)$, where $\mathcal{N}(m, \sqrt{\sigma^2})$ is the normal distribution with mean m and variance σ^2. The first two components have a heavier tail than the last four, which may be considered at noise above sufficiently high level. The angular measure on the sphere of $\mathbb{R}^4$ is concentrated on the canonical basis vectors (e_1, e_2).

The $L^2[0,1]$ functions A_j's are chosen deterministically for simplicity, namely $A_j(x) = \sin(2\pi\omega_j x)$, $j \in \{1,3,5\}$ and $A_j(x) = \cos(2\pi\omega_j x)$, $j \in \{2,4,5\}$, with $(\omega_1, \ldots, \omega_6) = (2,3,1,4,5,6)$. In this setting the angular measure of extremes in $L^2[0,1]$ is concentrated on a two-dimensional subspace, namely the one generated by (A_1, A_2). In particular, the extreme covariance operator is given by $C_\infty = (A_1 \otimes A_1 + A_2 \otimes A_2)/2$.

From a numerical perspective, all scalar products in $L^2[0,1]$ are approximated in this work by the Euclidean scalar product in $\mathbb{R}^{48}$, which corresponds to a Riemann midpoint rule. For simplicity, and because the choice of the unit scale is also arbitrary, we dispense with standardizing by the half-hour width between records. Several numerical solutions exist to perform the eigendecomposition of the empirical covariance operator. However the considered datasets are moderately high dimensional and because all observations are regularly sampled in time we may use the simplest strategy, which is to perform the eigendecomposition of second moment matrix $\mathbb{X}^\top \mathbb{X} \in \mathbb{R}^{48\times 48}$ where $\mathbb{X}_{i,j}$ is the j^{th} time record on the i^{th} day. In practice we rely on the `svd` function in `R` issuing the singular value decomposition of X based on a LAPACK routine. This boils down to choosing as a basis for $L^2[0,1]$ a family of indicator functions centered at the observation times. Alternative orthonormal families in $L^2[0,1]$ (typically, the Fourier basis or wavelet basis) may be preferred in higher dimensional contexts or with irregularly sampled observations.

6.4.1 Pattern identification of functional extremes

With the synthetic dataset described above, we compare the output of functional PCA applied to extreme angular data, to the one obtained using all possible angles, *i.e.*, we compare the eigendecomposition of $\widehat{C}_k$ with that of $\widehat{C}_n$. First, the number k of observations considered as extreme k must be chosen. In this simulated case, calibration is possible and we use the Hilbert-Schmidt norm of the error $\|\widehat{C}_k - C_\infty\|_{HS(\mathbb{H})}$ as a calibration criterion. For each candidate value $k \in \{100, 200, ..., 2900, 3000\}$, we generate 500 datasets of size $n = 10e3$ in the above described model, resulting in 500 independent replicates of $\widehat{C}_k$. The average Hilbert-Schmidt norm of the error is displayed in Figure 6.1 and suggests choosing $k = 500$, which we do in the remainder of this section. The scree-plot (*i.e.*, the graph of ordered eigenvalues, normalized by their sum) for both operators $\widehat{C}_k$ and $\widehat{C}_n$ is displayed in the upper panel of Figure 6.2. The gap between the first two eigenvalues and the remaining ones is more pronounced with $\widehat{C}_k$ than with $\widehat{C}_n$, indicating that the method we promote is able to uncover a sparsity pattern at extreme levels which would not be as relevant for the bulk behavior. The limit measure of extremes (μ) is indeed concentrated on a two-dimensional subspace, as opposed to the distribution of the full dataset which support has dimension four. In addition, the 'true' extreme angular pattern, which is a superposition of two periodic

signals with frequencies (1,7), is easily recognized by inspecting the shape of the first two eigenfunctions of the extreme covariance $\widehat{C}_k$ (solid lines, first two panels of the second row in Figure 6.2) while these frequencies are perturbed by shorter tailed 'noise' with the full covariance $\widehat{C}_n$ (dotted lines). The discrepancy between extreme and non-extreme eigenfunctions vanishes for the third eigenfunction, which may be considered as 'noise' as far as extremes are concerned.

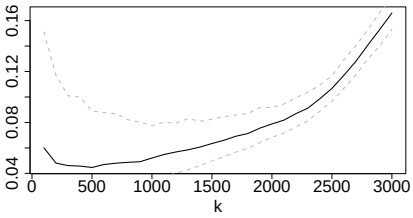


Figure 6.1: Simulated data: Errors $\|\widehat{C}_k - C_\infty\|_{HS(\mathbb{H})}$ as a function of k. Solid line: averaged errors over 500 experiments. Dotted lines: 90% bootstrap confidence interval.

6.4.2 Optimal reconstruction of functional extremes on the electricity demand dataset

Here we investigate the L^2 reconstruction error when projecting new (test) angular observations on the eigenspaces issued from the spectral decomposition of the empirical covariance operator $\widehat{C}_k$. Another important goal of this section is to provide guidelines and graphical diagnostic tools allowing to check whether functional regular variation in L^2 may reasonably be assumed for a given functional dataset. For simplicity, we ignore in this illustrative study any temporal dependence from week to week.

First, regular variation must be checked and an appropriate number k of extreme observations should be selected for estimating C_∞ with $\widehat{C}_k$. A Gaussian QQ-plot (not shown) suggests that the radial quantile is potentially heavy-tailed. In view of Proposition 5.6, 2., one should check RV of the radial variable and weak convergence of univariate projections $\langle \Theta_t, h \rangle$. Regarding the radial variable $R = \|X\|$, we propose to inspect a Hill plot and a Pareto quantile plot (Beirlant et al. (2006), Chapter 2). Visual inspection (Figure 6.3) suggests a stability region for the Hill estimator of $\gamma = 1/\alpha$ (left panel) between $k = 100$ and $k = 200$. We recall that if $\|X\|$ can be assumed to be regularly varying, α is the positive index such that $\mathbb{P}(\|X\| > t) = t^{-\alpha} L(t)$ where L is a slowly varying function. Choosing $k = 150$ corresponds to an empirical quantile level $1 - k/n \approx 0.7$, for which the Pareto quantile plot (right panel) is reasonably linear. For $k = 150$ the estimated regular variation index with the Hill estimator $\hat{\gamma}$ is $\hat{\alpha} = 1/\hat{\gamma} = 12.4$ (0.95 CI: $[10.7 - 14.8]$). The obtained value of $\hat{\alpha}$ is high relative to other settings, which may be considered as only weak evidence of RV. To rule out the hypothesis that $\|X\|$ could be in the domain of attraction of some Extreme Value Distribution with $\gamma \leq 0$ (Weibull and Gumbel domains), we consider a profile likelihood approach in the Generalized Pareto model (see, *e.g.*, Coles et al. (2001)). Negative values of γ fall outside the profile likelihood 95% confidence interval. The likelihood ratio test based on the deviance statistic issues a p-value of 0.01, which is in fact strong evidence for RV, despite the large value of $\hat{\alpha}$.

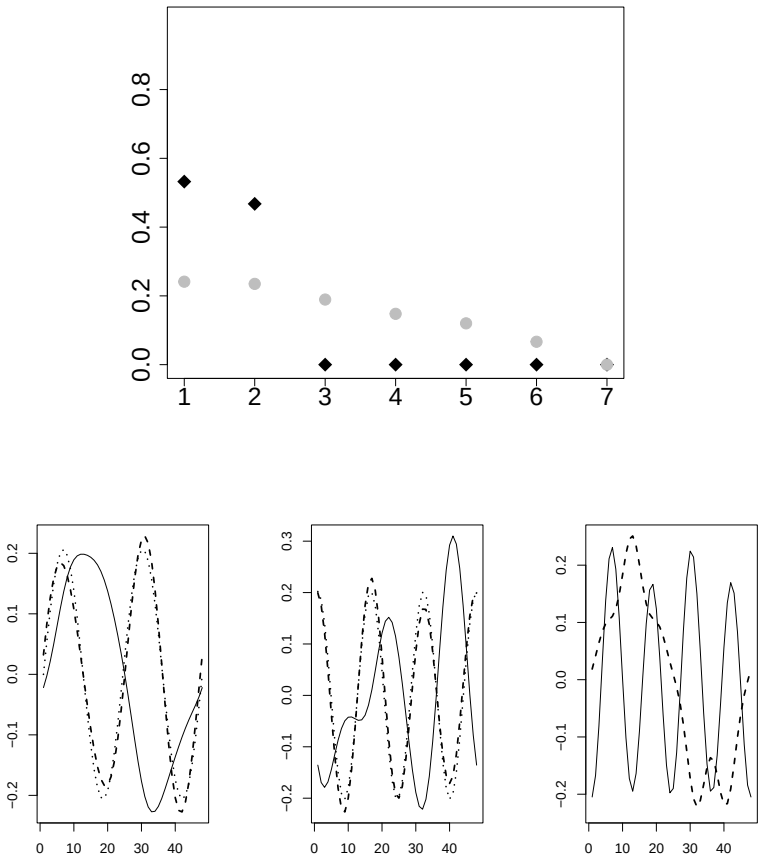


Figure 6.2: Simulated data: Scree plots and first three eigenfunctions. Diamond shaped dots and dashed lines: angular functional PCA of extremes ($\widehat{C}_k$). Round dots and solid lines: angular functional PCA of the full dataset ($\widehat{C}_n$). Dotted lines on the first two plots, bottom left: (normalized) functions A_1, A_2, *i.e.*, support of the angular measure for extremes.

The condition of weak convergence of projections $\langle \Theta_t, h \rangle$ is obviously difficult to check in practice, in particular because it must hold for any h. As a default strategy we propose to check convergence of the (absolute value of) the first moment, namely convergence of $\mathbb{E}|\langle \Theta_t, h \rangle|$ as $t \to +\infty$, for a finite number of 'appropriate' functions $h_j, j \in \{1, \ldots, J\}$. The context of daily records suggests a periodic family, namely we choose $h_j(x) = \sin(2\pi j x)$, for $j \in \{1, 2, 3, 4, 6, 8\}$. Figure 6.4 displays the six plots of the empirical conditional moment $\frac{1}{k}\sum_{i=1}^{k} |\langle \Theta_{(i)}, h_j \rangle|$, as a function of k. The plots confirm the existence of a relative stability region around $k = 150$.

Turning to performance assessment, our interest lies in the mean squared angular reconstruction error of an orthogonal projector π,

$$R(\pi, t) = \mathbb{E}\Big[\|\Theta - \pi(\Theta)\|^2 \mid \|X\| > t \Big].$$

Our goal is to assess the performance of $\hat{\pi}_k$, the orthogonal projector onto the first p-dimensional eigenspace of $\widehat{C}_k$. We fix $p = 2$ throughout. We compare $\hat{\pi}_k$ with natural alternatives, namely the orthogonal projectors onto the principal eigenspaces of the

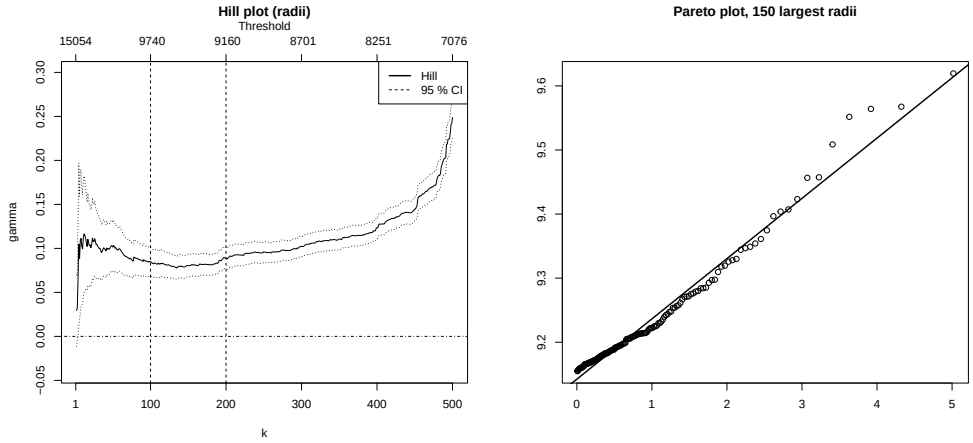


Figure 6.3: Hill plot (left) and Pareto quantile plot (right) for the radial variable of the air quality dataset. Dotted vertical lines on the Hill plot: stability region.

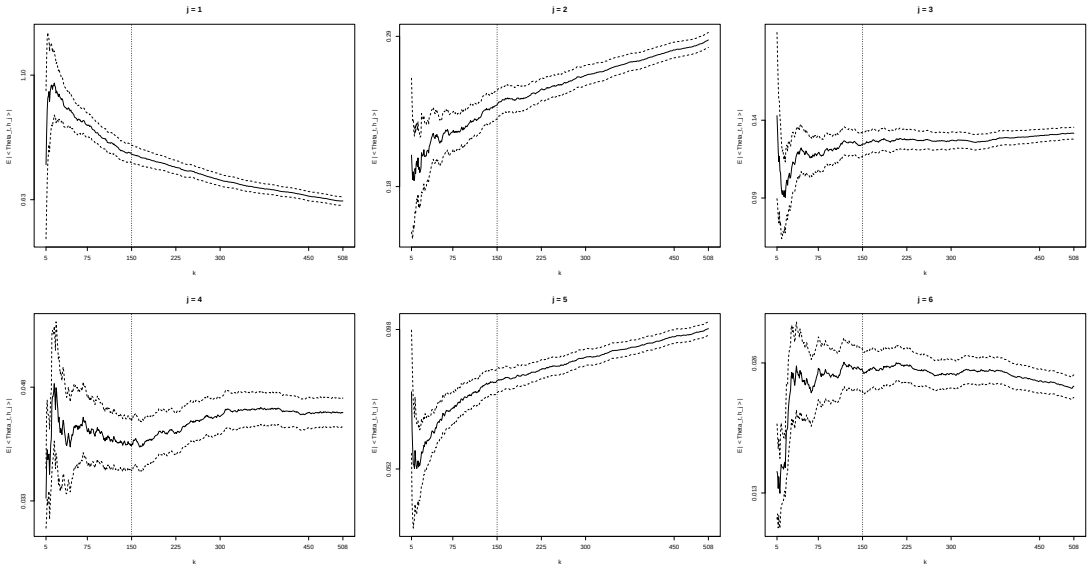


Figure 6.4: Electricity demand data: first moment of $|\langle \Theta, h_j \rangle|$ conditioned upon $R \geq R_{(k)}$, as a function of k, for $h_j(x) = \sin(2\pi j x)$, $x \in [0,1]$.

empirical covariance operator of respectively, all angular data, $\widehat{C}_n = n^{-1} \sum_{i \leq n} \Theta_i \otimes \Theta_i$, and a random subsample of size k, $\widetilde{C}_{\sigma,k} = k^{-1} \sum_{i \leq k} \Theta_{\sigma(i)} \otimes \Theta_{\sigma(i)}$, where σ is a random permutation of $\{1,\ldots,n\}$. We denote respectively by $\hat{\pi}_n$ and $\tilde{\pi}_{\sigma,k}$ the projectors onto the principal eigenspaces of the latter two operators.

We perform two experiments, the results of which are displayed in Table 6.1 and Table 6.2. In the first experiment (Table 6.1) we report cross-validation estimates of the risks $R(\hat{\pi}_k, t_{n,k})$, $R(\hat{\pi}_n, t_{n,k})$ and $\mathbb{E}_\sigma[R(\tilde{\pi}_{\sigma,n}, t_{n,k})]$. Namely the following procedure is repeated $N = 1000$ times. First, a validation index set $\mathcal{V}$ of size 30 is randomly chosen among $\{1,\ldots,k\}$, where we recall $k = 150$. Then, three covariance operators are constructed: $\widehat{C}(k,\mathcal{V})$ is an average of the $\Theta_{(i)} \otimes \Theta_{(i)}$'s over $i \in \{1,\ldots,k\} \setminus \mathcal{V}$; $\widehat{C}(n,\mathcal{V})$ is an average over the full index set $\{1,\ldots,n\} \setminus \mathcal{V}$; and $\widetilde{C}(\sigma,k,\mathcal{V})$ is an average over a random subset of size $k - |\mathcal{V}|$ among $\{1,\ldots,n\} \setminus \mathcal{V}$. Denoting by $\hat{\pi}_k(\mathcal{V}), \hat{\pi}_n(\mathcal{V})$ and $\tilde{\pi}_{\sigma,k}(\mathcal{V})$

the associated projectors, three hold-out risks are obtained,

$$\hat{R}(\pi, \mathcal{V}) = |\mathcal{V}|^{-1} \sum_{i \in \mathcal{V}} \|\Theta_{(i)} - \pi(\Theta_{(i)})\|^2,$$

for $\pi \in \{\hat{\pi}_k(\mathcal{V}), \hat{\pi}_n(\mathcal{V}), \tilde{\pi}_{\sigma,k}(\mathcal{V})\}$.

The cross-validation estimates $\hat{R}_{CV}(\hat{\pi}_k), \hat{R}_{CV}(\hat{\pi}_n), \hat{R}_{CV}(\tilde{\pi}_{\sigma,k})$ reported in Table 6.1 are the averages of the three hold-out risks $\hat{R}(\hat{\pi}_k(\mathcal{V}), \mathcal{V}), \hat{R}(\hat{\pi}_n(\mathcal{V}), \mathcal{V}), \hat{R}(\tilde{\pi}_{\sigma,k}(\mathcal{V}), \mathcal{V})$ over the $N = 1000$ replications resulting in different random choices of $\mathcal{V}$ and σ.

$\hat{R}_{CV}(\hat{\pi}_k)$	$\hat{R}_{CV}(\hat{\pi}_n)$	$\hat{R}_{CV}(\tilde{\pi}_{\sigma,k})$
6.2 (1.6)	8.8 (1.8)	9.1 (2.2)

Table 6.1: Angular reconstruction error ($*10^2$) when projecting on principal eigenspaces of $\widehat{C}_k$ (first column), $\widehat{C}_n$ (second column) and $\tilde{C}_{\sigma,k}$ (third column). The reported numbers are the cross-validation estimates obtained by averaging $N = 1000$ hold-out estimates $\hat{R}(\pi, \mathcal{V})$. The numbers in parentheses are the standard deviation of the hold-out risks over the N replications.

In the second experiment, extrapolation risks are compared, these are risks of the kind $R(\pi, t)$ where t is even larger than the largest observation of the training set. To this end, we consider a single validation set $\mathcal{V} = \{1, \ldots, 50\}$. For this single (extreme) validation set, we report in Table 6.2 the hold-out risks $\hat{R}(\pi, \mathcal{V})$ described in the latter experiment. For simplicity, a single random permutation σ of the remaining indices $\{51, \ldots, n\}$ is considered for the third column. The numbers in parentheses are the estimated standard deviations of the hold-out risks viewed as averages of independent observations.

$\hat{R}(\hat{\pi}_k(\mathcal{V}), \mathcal{V})$	$\hat{R}(\hat{\pi}_n(\mathcal{V}), \mathcal{V})$	$\hat{R}(\tilde{\pi}_{\sigma,k}(\mathcal{V}), \mathcal{V})$
3.8 (0.6)	7.1 (0.9)	6.7 (0.8)

Table 6.2: Extrapolation errors ($*10^3$): hold-out risks with validation set $\mathcal{V}$ chosen as the most extreme fraction of the observations, $X_{(1)}, \ldots, X_{(50)}$. The numbers in parentheses are the estimated standard deviations.

The conclusion is the same for both experiments reported in Table 6.1 and Table 6.2: performing functional PCA on the fraction of the angular data corresponding to the most extreme angles significantly reduces the reconstruction error, despite the reduced size of the training set. Comparison between the second and the third columns of each panel illustrates the negative impact of reducing the training sample size, while comparing the first and the third columns shows the bias reduction achieved by localizing on the tail region. Comparing the first and second columns shows the overall benefit of the proposed approach compared with a standard PCA of all angles. On this particular example the bias-variance trade-off favors our approach.

6.5 Conclusion

In this chapter, we have established non-asymptotic guarantees for a dimension reduction technique applied in extreme regions. These guarantees were derived using concentration inequalities for the ρ-distance between extreme Hilbert eigenspaces

associated with a PCA procedure and its empirical counterpart. While our contribution is essentially of theoretical nature, basic experiments, both on synthetic and real-world datasets, have also been carried out, with promising results. These findings not only validate the viability of our theoretical framework but also open avenues for practical applications. In particular, the present part gathers the theoretical guarantees for potential dimension reduction steps performed in machine learning tasks for extremes, such as the regression task developed in the subsequent part.

6.A Proofs

Proof of Proposition 6.6. Our main tool to derive a concentration bound on $\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}$ is a Bernstein-type inequality, Theorem 3.8 in McDiarmid (1998) which is recalled in Section 4.2, Lemma 4.6. Here and throughout we adopt the shorthand notation $x_{i:j} = x_i, \dots, x_j$ for $i \le j$.

In order to apply this inequality to our purposes we need to write the empirical pre-asymptotic operator (or its surrogate $\overline{C}_{t_{n,k}}$) as a function f_t of the sample $X_{1:n}$. With this in mind, we introduce a thresholded angular functional

$$\begin{array}{rccc} \theta_t : & \mathbb{H} & \longrightarrow & \mathbb{S} \\ & x & \longmapsto & \theta_t(x) = \mathbb{1}\{\|x\| \ge t\}\|x\|^{-1}x\,. \end{array}$$

Observe that with this notation, $\Theta_i \mathbb{1}\{R_i > t\} = \theta_t(X_i)$. Consider now the function

$$\begin{array}{rccc} f_t : & \mathbb{H}^n & \longrightarrow & \mathbb{R} \\ & x_{1:n} & \longmapsto & f_{t_{n,k}}(x_{1:n}) = \frac{1}{k}\|\sum_{i=1}^n (\theta_t(x_i)\otimes\theta_t(x_i) - \mathbb{E}[\theta_t(X)\otimes\theta_t(X)]\|_{HS(\mathbb{H})}\,. \end{array}$$

Notice that $f_{t_{n,k}}(X_{1:n}) = \|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}$ which is the focus of Proposition 6.6.

Lemma 6.13 (Deviations of $f_{t_{n,k}}(X_{1:n})$)**.** *With the above notations, we have*

$$\mathbb{P}\Big(f_{t_{n,k}}(X_{1:n}) - \mathbb{E}\Big[f_{t_{n,k}}(X_{1:n})\Big] \ge \varepsilon\Big) \le \exp\left(\frac{-k\varepsilon^2}{4(1+\frac{\varepsilon}{3})}\right).$$

Proof. We apply Lemma 4.6 to the function $f = f_{t_{n,k}}$. To do so we derive upper bounds on the maximum deviation term b and on the maximum sum of variances σ^2 from the statement. Let $x_{1:n} \in \mathbb{H}^n$. The maximum deviation b is bounded by $2/k$ since by independence among X_i's, with the notations of Lemma 4.6,

$$\begin{aligned} g_i(x_{1:i}) &= \mathbb{E}\Big[f_{t_{n,k}}(x_1,\dots,x_{i-1},x_i,X_{i+1},\dots,X_n) - f_{t_{n,k}}(x_1,\dots,x_{i-1},X_i,X_{i+1},\dots,X_n)\Big] \\ &\le \frac{1}{k}\mathbb{E}\Big[\|\theta_{t_{n,k}}(x_i)\otimes\theta_{t_{n,k}}(x_i) - \theta_{t_{n,k}}(X_i)\otimes\theta_{t_{n,k}}(X_i)\|_{HS(\mathbb{H})}\Big] \\ &\le \frac{1}{k}\Big(\mathbb{1}\{\|x_i\| \ge t_{n,k}\} + \mathbb{P}(\|X\| \ge t_{n,k})\Big) \le \frac{1+k/n}{k} \le \frac{2}{k}, \end{aligned}$$

where the first inequality comes from the triangle inequality $|\|a\| - \|b\|| \le \|a-b\|$, and the second one from the fact that $\|s\otimes s\|_{HS(\mathbb{H})} = 1$ if $\|s\| = 1$.

There remains to bound the variance term. Since for every $1 \le i \le n$, by the tower rule for conditional expectations, $\mathbb{E}\Big[g_i(x_{1:i-1}, X_i)\Big] = 0$, we may write, for Y_i and independent copy of X_i,

$$\begin{aligned} \sigma_i^2(f_{t_{n,k}}(x_1,\dots,x_n)) &= \mathbb{E}\Big[(f_{t_{n,k}}(x_1,\dots,x_{i-1},Y_i,X_{i+1},\dots,X_n) - f_{t_{n,k}}(x_1,\dots,x_{i-1},X_i,X_{i+1},\dots,X_n))^2\Big] \\ &\le \frac{1}{k^2}\mathbb{E}\Big[\|\theta_{t_{n,k}}(Y_i)\otimes\theta_{t_{n,k}}(Y_i) - \theta_{t_{n,k}}(X_i)\otimes\theta_{t_{n,k}}(X_i)\|^2_{HS(\mathbb{H})}\Big] \\ &\le \frac{2}{k^2}\mathbb{E}\Big[\|\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\|^2_{HS(\mathbb{H})}\Big] = \frac{2\mathbb{P}(\|X\| \ge t_{n,k})}{k^2} = \frac{2}{nk}. \end{aligned}$$

Hence, $\hat{v}$ is bounded from above by $2/k$. Injecting the upper bounds on $\hat{v}$ and b in Lemma 4.6 concludes the proof. ∎

The following intermediate lemma proves useful for bounding the expected deviation in the left-hand side of Lemma 6.13.

Lemma 6.14. *Let $A_1, \ldots, A_n$ be independent centered random elements in $\mathbb{H}$. Then*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{n} A_i\right\|^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\|A_i\|^2\right].$$

Proof. The left-hand side equals $\sum_{i=1}^{n} \mathbb{E}\left[\|A_i\|^2\right] + 2\sum_{1\le i<l\le n} \mathbb{E}[\langle A_i, A_l\rangle]$. Since the A_i's are independent with mean 0, for all $1 \le i < l \le n$,

$$0 = \langle \mathbb{E}[A_i], \mathbb{E}[A_l]\rangle = \mathbb{E}\left[\langle A_i, \mathbb{E}[A_l]\rangle\right] = \mathbb{E}\left[\langle A_i, \mathbb{E}[A_l|A_i]\rangle\right] = \mathbb{E}\left[\mathbb{E}[\langle A_i, A_l\rangle|A_i]\right] = \mathbb{E}[\langle A_i, A_l\rangle],$$

which concludes the proof. ∎

We are now ready to obtain a bound on $\mathbb{E}\left[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\right]$.

Lemma 6.15.

$$\mathbb{E}\left[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\right] \le \frac{1}{\sqrt{k}}. \tag{6.3}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})}\right] &= \frac{n}{k}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \theta_{t_{n,k}}(X_i)\otimes\theta_{t_{n,k}}(X_i) - \mathbb{E}\left[\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\right]\right\|_{HS(\mathbb{H})}\right] \\
&\le \frac{n}{k}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n} \theta_{t_{n,k}}(X_i)\otimes\theta_{t_{n,k}}(X_i) - \mathbb{E}\left[\theta_{t_{n,k}}(X_i)\otimes\theta_{t_{n,k}}(X_i)\right]\right\|^2_{HS(\mathbb{H})}\right]^{1/2} \\
&= \frac{n}{k}\frac{1}{\sqrt{n}}\mathbb{E}\left[\left\|\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X) - \mathbb{E}\left[\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\right]\right\|^2_{HS(\mathbb{H})}\right]^{1/2} \\
&= \frac{\sqrt{n}}{k}\left(\mathbb{E}\left[\|\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\|^2_{HS(\mathbb{H})}\right] - \left\|\mathbb{E}\left[\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\right]\right\|^2_{HS(\mathbb{H})}\right)^{1/2} \\
&\le \frac{\sqrt{n}}{k}\mathbb{E}\left[\|\theta_{t_{n,k}}(X)\otimes\theta_{t_{n,k}}(X)\|^2_{HS(\mathbb{H})}\right]^{1/2} \le \frac{\sqrt{n}}{k}\mathbb{P}(\|X\| \ge t_{n,k})^{1/2} = \frac{1}{\sqrt{k}},
\end{aligned}$$

where the second identity derives from Lemma 6.14, The last inequality follows from $\|\theta(x)\otimes\theta(x)\|_{HS(\mathbb{H})} = 1$. ∎

Combining Lemma 6.13 and Lemma 6.15, together with the definition of $f_{t_{n,k}}$, we obtain that with probability at least $1 - \gamma/2$,

$$\|\overline{C}_{t_{n,k}} - C_{t_{n,k}}\|_{HS(\mathbb{H})} \le \frac{1}{\sqrt{k}} + \frac{4}{3k}\log(2/\gamma) + 4\left(\left(\frac{\log(2/\gamma)}{3k}\right)^2 + \frac{\log(2/\gamma)}{k}\right)^{1/2}.$$

Simplifying the above display with $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ yields the statement of Proposition 6.6.

Part III

On Regression in Extreme Regions

Introduction

Regression is a predictive problem of crucial importance in statistical learning, covering a wide variety of applications. In the standard setup, $(\mathbf{X}, Y)$ is a pair of random variables defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution P, where the target Y is a square integrable real-valued random variable (the output) and the predictor (or covariable) $\mathbf{X}$ is a random vector with marginal distribution ρ taking its values in some measurable space $\mathcal{X}$ modeling some input information hopefully useful to predict Y. The predictive learning problem consists in building, from a training dataset $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ composed of $n \geq 1$ independent copies of $(\mathbf{X}, Y)$, a mapping $f : \mathcal{X} \to \mathbb{R}$ in order to compute a 'good' prediction $f(\mathbf{X})$ for Y, with the quadratic risk

$$R_P(f) = \mathbb{E}\Big[(Y - f(\mathbf{X}))^2\Big] \tag{6.4}$$

as close as possible to that of $f^*(X) = \mathbb{E}[Y \mid \mathbf{X}]$, which obviously minimizes (6.4) over the space $L_2(\rho)$ of square integrable functions of $\mathbf{X}$: $R_P^* := \min_{f \in L_2(\rho)} R_P(f) = R_P(f^*)$. A natural strategy consists in solving the Empirical Risk Minimization problem (ERM in abbreviated form) $\min_{f \in \mathcal{F}} R_{\hat{P}_n}(f)$, where $\mathcal{F} \subset L_2\Big(\rho\Big)$ is a closed and convex class of functions sufficiently rich to include a reasonable approximant of f^* and $\hat{P}_n$ is an empirical version of P based on $\mathcal{D}_n$.

The performance of predictive functions $\hat{f}$ obtained by *least square regression*, has been extensively investigated in the statistical learning literature Györfi et al. (2002); Massart (2007). Under the assumption that the tails of the random pairs $(f(\mathbf{X}), Y)$ are subgaussian and appropriate complexity conditions are satisfied by the class $\mathcal{F}$, confidence upper bounds for the excess of quadratic risk $R_P(\hat{f}) - R_P^* = \mathbb{E}[(Y - \hat{f}(\mathbf{X}))^2 \mid \mathcal{D}_n] - R_P^*$ have been established in Lecué and Mendelson (2013) by means of concentration inequalities for empirical processes Boucheron et al. (2013).

Here we consider the problem of building prediction functions which would be reliable in a 'crisis scenario' where the covariates vector takes unusually large values and thus belongs to regions where few or even no such large examples have been observed in the past. Notice incidentally that it could be thus viewed as a specific, never tackled yet, *few shot learning problem*, see, *e.g.*, Wang et al. (2020). We place ourselves in a finite-dimensional setting, $\mathcal{X} \subset \mathbb{R}^d$. The distribution of X is not subgaussian and in particular its support is unbounded. Covariates are considered as extreme when their norm $\|X\|$ exceeds some (asymptotically) large threshold $t > 0$. The choice of the norm is unimportant in theory, and is typically determined by the application context.

The threshold t depends on the observations, since 'large' should be naturally understood as large relative to the vast majority of data observed. Hence, extreme observations are rare by nature and severely underrepresented in the training dataset with overwhelming probability. Consequently, the impact of prediction errors in extreme

regions of the input space on the global regression error of $\hat{f}$ is generally negligible. Indeed, the law of total probability yields

$$R_P(f) = \mathbb{P}(\|\mathbf{X}\| \geq t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big] + \mathbb{P}(\|\mathbf{X}\| < t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| < t\Big]. \quad (6.5)$$

The above decomposition involves a conditional error term relative to excesses of $\|\mathbf{X}\|$ above t, which we term *conditional quadratic risk* (or simply *conditional risk*)

$$R_t(f) := \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big].$$

It is the purpose of the subsequent analysis to construct a predictive function $\hat{f}$ that (approximately) minimizes $R_t(f)$ for all $t > t_0$, with t_0 being a large threshold. It is important to note that an approximate minimizer of R_t might not be suitable for minimizing $R_{t'}$ when $t' > t$. To ensure robust extrapolation performance for our learned function, we focus on obtaining a prediction function, $\hat{f}$, that minimizes the *asymptotic conditional quadratic risk* defined as

$$R_\infty(f) := \limsup_{t \to +\infty} R_t(f) = \limsup_{t \to +\infty} \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big]. \quad (6.6)$$

It is immediate to see that any function that coincides with the regression function $f^*(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$ on the region $\{\mathbf{x} \in \mathcal{X}, \|\mathbf{x}\| \geq t\}$ minimizes the risk functional R_t, for all $t > 0$, and thus also R_∞. In other words $R^*_\infty := \inf_f R_\infty(f) = R_\infty(f^*)$. However, even though f^* provides a straightforward theoretical solution, f^* is of course unknown.

In view of Equation (6.5) it is evident that an estimate $\hat{f}$ of f^* produced by an ERM strategy with good overall empirical performances, may not necessarily enjoy good performances when restricted to extreme regions. Put another way, there is no guarantee that the conditional risk $R_t(\hat{f})$ (or $R_\infty(\hat{f})$) would be small. However, accurate prediction in extreme regions turns out to be crucial in certain practical (safety) applications, in environmental sciences, dietary risk analysis or finance/insurance for instance.

To summarize, the *Regression Problem on Extremes* refers here to the the task of constructing a prediction function $\hat{f}$ based on $\mathcal{D}_n$ which approximately minimizes R_∞. Notice that our choice of the squared error is motivated by simplicity and for illustrative purpose, extensions to other losses may be achieved at the price of additional (minor) technicalities.

In order to develop a specific ERM framework relative to R_∞ with provable guarantees, regularity assumptions are required regarding the tail behavior of the pair $(\mathbf{X}, Y)$, with respect to the first component. Multivariate Regular Variation (RV) hypotheses are very flexible in the sense that they correspond to a large nonparametric class of heavy-tailed distributions. These assumptions, or slightly weaker ones such as *Maximum Domain of Attraction* conditions are at the heart of Extreme Value Analysis (EVA) (see, *e.g.*, the monographs Beirlant et al. (2006); De Haan and Ferreira (2006); Resnick (1987)). They are frequently used in applications where the impact of extreme observations should be enhanced, or not neglected at the minimum.

In the past few decades, numerous papers have combined Extreme Value Theory (EVT) with statistical learning techniques, covering areas such as clustering, dimension reduction, and anomaly detection (see Section 1.2). One of the primary objectives of this paper is to establish sufficient and reasonable conditions for extending the

results of Jalalzai et al. (2018), which develop a framework for binary classification in extreme regions (see Section 1.2 for more details), to a broader context encompassing statistical regression with a continuous target and an appropriate real-valued loss. It must be noted that the above risk functionals defined above are the same for the classification replacing the quadratic loss by the $0-1$-loss. The continuous nature of the target in the regression problem considered here requires fundamentally different assumptions and proof techniques compared with the binary classification setting. In particular one natural generalization of the assumptions made in the cited reference would be to assume RV of the conditional distributions $\mathcal{L}(\mathbf{X}|Y=y)$, almost everywhere. This somewhat intricate generalization leads to measure theoretic complications and is difficult to verify in practice and also on theoretical examples. We propose to bypass this issue by requiring instead a joint form of RV of the pair $(\mathbf{X}, Y)$, see our Assumption 7.2. We show that this condition is satisfied in various examples worked out in Section 7.2. Another major improvement of the present work upon Jalalzai et al. (2018), with implications for applications related to climate extremes, is to offer a novel perspective upon extreme value prediction within regularly varying random vectors, see Example 7.10.

It should also be pointed out that the problem of regression in extreme regions can be assimilated to a specific *transfer learning* problem, see, *e.g.*, Pan and Yang (2010). Indeed, the objective pursued is to learn a regression function that is nearly optimal in the target (limit) extremal domain, based on source training data in a pre-asymptotic regime. Unlike pre-existing transfer learning and domain adaptation approaches, the methodology we develop does not rely on inverse probability weighting Clémençon et al. (2016), estimating/learning propensity score functions Bertail et al. (2021) or the use of Markov kernels Pfister and Bühlmann (2024), but exploits a multivariate RV assumption to estimate the target loss with guarantees.

This part is divided into two chapters. Chapter 7 details the algorithmic approach we propose for regression on extremes and elaborates on the probability framework considered for regression in extreme regions. Chapter 8 presents the probabilistic and statistical results that justify the algorithmic procedure. The soundness of the proposed approach is demonstrated through various numerical experiments. Certain technical details are deferred to the Appendices 7.A and 8.A.

Chapter 7

A Regular Variation Framework for Regression on Extremes

Contents

In this chapter, we propose a probabilistic framework in which regression on extremes may be addressed, together with a dedicated algorithmic approach in Section 7.1, the latter being analyzed next in the subsequent chapter. The foundational assumption of this part, namely Regular Variation w.r.t. the first component, is detailed Section 7.2. Section 7.3 introduces the limit regression function central in the analysis of Chapter 8, along with a necessary technical condition and scenarios where these conditions are met. Section 7.4 demonstrates the reliability of the developed framework through various concrete examples based on classical Extreme Value Theory (EVT) assumptions. Finally, Section 7.5 connects the proposed Regular Variation (RV) assumption with the framework developed in Lindskog et al. (2014). The chapter concludes with final remarks.

Here and throughout, $(\mathbf{X}, Y)$ is a pair of random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with distribution P, where Y is real-valued with marginal distribution G and $\mathbf{X} = (X_1, \ldots, X_d)$ takes its values in $\mathbb{R}^d$, $d \geq 1$. We sometimes denote by $\mathcal{L}(Z)$ the distribution of a random variable Z. Recall from the Introduction section that $\|\cdot\|$ is any norm on $\mathbb{R}^d$. We denote by $\mathbb{S}$ the unit sphere for this norm and by $\mathbb{B} := \{\mathbf{x} \in \mathbb{R}^d, \|\mathbf{x}\| \leq 1\}$ the unit ball. Let $E = \mathbb{R}^d \setminus \{\mathbf{0}\}$ be the punctured Euclidean space. For any measurable subset A of $\mathbb{R}^d$ we denote by $\mathcal{B}(A)$ the Borel σ-algebra on A. The boundary and the closure of A are respectively denoted by ∂A and $\bar{A}$, and we set $tA = \{t\mathbf{x} : \mathbf{x} \in A\}$ for all $t \in \mathbb{R}$. By $\mathbb{1}\{\mathcal{E}\}$ is meant the indicator function of any event $\mathcal{E}$ and the integer part of any $u \in \mathbb{R}$ is denoted by $\lfloor u \rfloor$. For any $\mathbf{x} \in E$, we denote by $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ the angular component of $\mathbf{x}$ for conciseness.

7.1 ROXANE Algorithm

In order to help the reader understand the general workflow of the part, we begin by introducing the algorithm ROXANE (Regression On eXtreme ANglEs) that we promote to solve the Regression Problem on Extremes stated in the Introduction, formulated as the minimization of the risk functional R_∞ defined in (6.6). The remainder of this work aims at developing a framework that fully justifies Algorithm 7.1 below.

Algorithm 7.1 Regression On eXtreme ANglEs (ROXANE)

INPUT: Training dataset $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ with $(\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$; class $\mathcal{H}$ of predictive functions $h : \mathbb{S} \to \mathbb{R}$; number $k \leq n$ of 'extreme' observations among training data.
Truncation: Sort the training data by decreasing order of magnitude of the input information $\|\mathbf{X}_{(1)}\| \geq \ldots \geq \|\mathbf{X}_{(n)}\|$ and form a set of k *extreme training observations*

$$\Big\{(\mathbf{X}_{(1)}, Y_{(1)}), \ldots, (\mathbf{X}_{(k)}, Y_{(k)})\Big\}.$$

Empirical quadratic risk minimization: based on the extreme training dataset, solve the optimization problem

$$\min_{h \in \mathcal{H}} \frac{1}{k} \sum_{i=1}^{k} \Big(Y_{(i)} - h\big(\boldsymbol{\theta}(\mathbf{X}_{(i)})\big) \Big)^2, \tag{7.1}$$

where $\boldsymbol{\theta}(x) = \mathbf{x}/\|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$.
OUTPUT: Solution $\hat{h}$ to problem (7.1) and predictive function $\widehat{f}(\mathbf{x}) = (\hat{h} \circ \boldsymbol{\theta})(\mathbf{x})$ to be used for predictions of Y based on new examples $\mathbf{X}$ such that $\|\mathbf{X}\| \geq \|\mathbf{X}_{(k)}\|$.

Notice that the ROXANE algorithm can be implemented with any optimization heuristic solving the quadratic risk minimization problem (7.1), refer to, *e.g.*, Györfi et al. (2002). The study of dedicated numerical techniques is beyond the scope of the present paper.

A key feature of the ROXANE Algorithm is that its training step involves the *angular* component of extremes solely. It returns a prediction function $\widehat{f}$ which only depends on the angular component $\boldsymbol{\theta}(\mathbf{X})$ of a new input $\mathbf{X}$. This apparently arbitrary choice turns out to be fully justified under RV assumptions, which are introduced and discussed in the following subsections. To wit, the main theoretical advantage of considering angular prediction function is to ensure the convergence of the conditional risk R_t, as $t \to +\infty$. In practice, rescaling all extremes (in the training set and in new examples) onto a bounded set allows a drastic increase in the density of available training examples and a clear extrapolation method beyond the envelope of observed examples.

Based on the background on multivariate RV recalled in Section 2.1.2, we introduce a modified version of the standard framework (*regular variation with respect to the first component*) in Section 7.2 which is suitable for the regression problem considered here, in the sense that the ROXANE Algorithm turns out to enjoy probabilistic and statistical guarantees in this context. We thoroughly discuss the relevance of our assumptions by working out several sufficient conditions and examples. We state our main probabilistic results in Section 8.1, establishing connections between different risks and their corresponding minimizers, thus bringing a first (probabilistic) justification regarding

the angular nature of the prediction function in Algorithm 7.1. Statistical guarantees are deferred to Section 8.2.

7.2 Regular Variation with respect to the First Component

We now describe rigorously the framework we consider for regression in extreme regions, which may be seen as a natural, 'one-component' extension of standard multivariate RV assumptions recalled in Section 2.1.2.

For simplicity, we suppose that Y is bounded through this paper. This assumption can be naturally relaxed at the price of additional technicalities (*i.e.*, tail decay hypotheses).

Assumption 7.1. *The random variable Y is bounded: there exists $M \in (0, +\infty)$ such that with probability one, $Y \in I = [-M, M]$.*

The following hypothesis concerns the asymptotics, as $t \to +\infty$, of the conditional distribution of the pair $(\mathbf{X}, Y)$ given that $\|\mathbf{X}\| > t$. It may be viewed as one-component extension of (2.5).

Assumption 7.2. *There exists a nonzero Borel measure μ on $\mathbb{O} = E \times I$, which is finite on sets bounded away from $\mathbb{C} = \{\mathbf{0}\} \times I$, and a regularly varying function $b(t)$ with index $\alpha > 0$ such that*

$$\lim_{t \to +\infty} b(t) \mathbb{P}\Big(t^{-1}\mathbf{X} \in A, Y \in C\Big) = \mu(A \times C), \tag{7.2}$$

for all $A \in \mathcal{B}(E)$ bounded away from zero and $C \in \mathcal{B}(I)$ such that $\mu(\partial(A \times C)) = 0$.

Assumption 7.2 could be understood as a multivariate extension of the *One-Compo-nent Regular Variation* framework developed in Hitz and Evans (2016). It should be noticed that Assumption 7.2 fits into the framework of RV in $\mathbb{M}_{\mathbb{O}}$ developed in Lindskog et al. (2014) as an extension of Hult and Lindskog (2006b), where $\mathbb{O} = E \times I = (\mathbb{R}^d \times I) \setminus (\{\mathbf{0}\} \times I)$ and where the scalar multiplication is defined as $\lambda(\mathbf{x}, y) = (\lambda \mathbf{x}, y)$. More details regarding the connections between Assumption 7.2 and Lindskog et al. (2014) are provided in Section 7.5.

Remark 7.3 (Pre-Processing)**.** *Because the goal of this paper is to explain main ideas to tackle the problem of regression on extremes, the input are assumed to be regularly varying with same marginal index while in practice, this condition may be satisfied only after some marginal standardization. This is a recurrent theme in multivariate extreme value theory. For binary-valued Y, in the classification setting, Clémençon et al. (2023) consider a marginal standardization based on ranks, following Einmahl and Segers (2009); Einmahl et al. (2001). They prove an upper bound on the statistical error term induced by this transformation which is of the same order of magnitude as the error when marginal distributions are known, a simplified case considered in Jalalzai et al. (2018). In our experiments with real data, this pre-processing step is not necessary. We leave this technical question outside the scope of this paper.*

In the sequel we refer to the limit measure μ as the *joint limit measure* of $(\mathbf{X}, Y)$. Under Assumption 7.2, $\mathbf{X}$'s marginal distribution is regularly varying with *marginal limit measure*

$$\mu_X(A) = \lim_{t \to +\infty} b(t) \mathbb{P}(\mathbf{X} \in tA) = \lim_{t \to +\infty} b(t) \mathbb{P}(\mathbf{X} \in tA, Y \in I) = \mu(A \times I),$$

with $A \in \mathcal{B}(E)$ bounded away from zero and such that $\mu(\partial(A \times I)) = 0$. We also naturally introduce the *joint angular measure* of $(\mathbf{X}, Y)$ denoted by Φ, which is a finite measure on $\mathbb{S} \times I$ given by

$$\Phi(B \times C) = \mu\{(\mathbf{x}, y) \in E \times I : \|\mathbf{x}\| \geq 1, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C\}. \tag{7.3}$$

With this notation, under Assumption 7.2 it holds that

$$\frac{\mathbb{P}(\boldsymbol{\theta}(\mathbf{X}) \in B,\, Y \in C, \|\mathbf{X}\| \geq tr)}{\mathbb{P}(\|\mathbf{X}\| \geq t)} \underset{t \to +\infty}{\longrightarrow} c\, r^{-\alpha} \Phi(B \times C), \tag{7.4}$$

where $c = \Phi(\mathbb{S} \times I)^{-1} = \mu((E \setminus \mathbb{B}) \times I)^{-1}$, for all $C \in \mathcal{B}(I)$, $B \in \mathcal{B}(\mathbb{S})$, such that $\Phi(\partial(B \times A)) = 0$ and $r \geq 1$. The latter statement is proved in Appendix 7.A, Theorem 7.11. To lighten the notation, we assume without loss of generality that b is chosen so that $\mu((E \setminus \mathbb{B}) \times I) = 1$ and thus $c = 1$ and Φ is a probability measure on $\mathbb{S} \times I$. In particular, the joint limit measure μ and the joint angular measure Φ are linked through the relation

$$\mu(\{\mathbf{x} \in E :\ \|\mathbf{x}\| \geq r, \boldsymbol{\theta}(\mathbf{x}) \in B\} \times C) = r^{-\alpha} \Phi(B \times C),$$

for all $C \in \mathcal{B}(I), B \in \mathcal{B}(\mathbb{S})$ and $r > 0$. Observe that

$$\lim_{t \to +\infty} \frac{\mathbb{P}(\boldsymbol{\theta}(\mathbf{X}) \in B, Y \in C, \|\mathbf{X}\| \geq t)}{\mathbb{P}(\|\mathbf{X}\| \geq t)} = \Phi(B \times C),$$

for all $B \in \mathcal{B}(\mathbb{S}), C \in \mathcal{B}(I)$, such that $\Phi(\partial(B \times C)) = 0$. In words, Φ is the asymptotic joint probability distribution of $(\theta(\mathbf{X}), Y)$ given that $\|\mathbf{X}\| \geq t$ as $t \to +\infty$. Notice also that $\mathbf{X}$'s angular (probability) measure writes $\Phi_X(B) = \Phi(B \times I)$.

Let P_∞ denote the limit conditional distribution on $E \setminus \mathbb{B} \times I$ of the pair $(\mathbf{X}/t, Y)$ given that $\|\mathbf{X}\| \geq t$, *i.e.*,

$$P_\infty(A \times C) = \lim_{t \to +\infty} \mathbb{P}(\mathbf{X}/t \in A\,, Y \in C \mid \|\mathbf{X}\| \geq t) \tag{7.5}$$

for all $A \in \mathcal{B}(E \setminus \mathbb{B})$ and $C \in \mathcal{B}(I)$ such that $\mu(\partial(A \times C)) = 0$, and let $(\mathbf{X}_\infty, Y_\infty)$ denote a random pair with distribution P_∞. It follows immediately from (7.4) and from our choice $c = 1$, that P_∞ indeed exists and is determined by (Φ, α), namely

$$\begin{aligned} &P_\infty((\mathbf{x}, y) : \|\mathbf{x}\| > r, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C) \\ &\qquad = \lim_{t \to +\infty} \mathbb{P}(\|\mathbf{X}\|/t \geq r, \boldsymbol{\theta}(\mathbf{X}) \in B, Y \in C \mid \|\mathbf{X}\| \geq t) = r^{-\alpha} \Phi(B \times C), \end{aligned}$$

where B, C, r are as in Equation (7.4). In other words, if T denotes the pseudo-polar transformation with respect to the first component $T(\mathbf{x}, y) = (\|\mathbf{x}\|, \boldsymbol{\theta}(\mathbf{x}), y)$ on $E \setminus \mathbb{B} \times I$, and if ν_α is the Pareto measure $\nu_\alpha([r, \infty)) = r^{-\alpha}$, then the following tensor product decomposition holds true in polar coordinates,

$$P_\infty \circ T^{-1} = \nu_\alpha \otimes \Phi.$$

Observe that, under Assumptions 7.1 and 7.2, the random variable Y_∞ is almost-surely bounded in amplitude by $M < +\infty$.

7.3 The Extreme Bayes Regression Function

Equipped with these notations, it is natural to consider the squared error loss of a prediction function f, under the distribution P_∞. We call this key quantity the *extreme quadratic risk*, denoted by R_{P_∞}, defined as

$$R_{P_\infty}(f) := \mathbb{E}\Big[(Y_\infty - f(\mathbf{X}_\infty))^2\Big],$$

for $f \in \mathcal{F}$ a class of real-valued bounded Borel-measurable functions defined on $E \setminus \mathbb{B}$. As will become clear in the subsequent analysis, although our objective R_∞ and the extreme risk R_{P_∞} are two different functionals, they turn out to be connected through their minimizers under an additional technical assumption stated below. In the sequel we let $f^*_{P_\infty}$ denote the minimizer of R_{P_∞} among all measurable functions. Standard arguments from statistical learning theory show immediately that $f^*_{P_\infty}$ is defined (up to a negligible set) by a conditional expectation, $f^*_{P_\infty}(X_\infty) = \mathbb{E}[Y_\infty \mid \mathbf{X}_\infty]$.

Remark 7.4 (Heavy-tailed input vs heavy-tailed output)**.** *Attention should be paid to the fact that the heavy-tail assumption is here on the distribution of the input/explanatory random variable X, in contrast to other works devoted to regression such as Brownlees et al. (2015), Mendelson (2017) or Lugosi and Mendelson (2019) where it is the loss/response that is supposedly heavy-tailed. In the EVT literature, similarly, the vast majority of existing works in a regression context are concerned with extreme values of the target, in particular for extreme quantiles regression (El Methni et al. (2012); Daouia et al. (2013); Chavez-Demoulin et al. (2014); Daouia et al. (2023))*

Assumption 7.5. *The extreme regression function $f^*_{P_\infty}$ is continuous on $\mathbb{R}^d \setminus \{\mathbf{0}\}$ and as t tends to infinity,*

$$\mathbb{E}\Big[|f^*(\mathbf{X}) - f^*_{P_\infty}(\mathbf{X})| \,|\, \|\mathbf{X}\| \geq t\Big] \to 0.$$

The next proposition highlights the weakness of Assumption 7.5, as long as Assumptions 7.1 and 7.2 are satisfied.

Proposition 7.6 (Sufficient conditions for Assumption 7.5)**.** *Let $(\mathbf{X}, Y)$ satisfy Assumptions 7.1 and 7.2. Then Assumption 7.5 also holds if one of the three conditions (i), (ii), (iii) below holds*

(i) The regression function f^* is continuous on $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \geq 1\}$ and as $t \to +\infty$,

$$\sup_{\|\mathbf{x}\| \geq t} |f^*(\mathbf{x}) - f^*_{P_\infty}(\mathbf{x})| \to 0; \tag{7.6}$$

(ii) The conditional distributions of Y given $\mathbf{X} = \mathbf{x}$ (resp. Y_∞ given $\mathbf{X}_\infty = \mathbf{x}$) admit densities $p_{Y|\mathbf{x}}(y)$ (resp. $p^\infty_{Y|\mathbf{x}}(y)$) w.r.t. the Lebesgue measure on I, for all $\mathbf{x} \neq \mathbf{0}$. In addition for all $y \in I$, the mapping $\mathbf{x} \mapsto p_{Y|\mathbf{x}}(y)$ (resp. $\mathbf{x} \mapsto p_{Y|\mathbf{x}}(y)$) is continuous, and $\sup_{\|\mathbf{x}\| \geq 1, y \in I} p_{Y|\mathbf{x}}(y) < +\infty$. Finally the following uniform convergence holds true,

$$\sup_{\|\mathbf{x}\| \geq t, y \in I} |p_{Y|\mathbf{x}}(y) - p^\infty_{Y|\mathbf{x}}(y)| \underset{t \to +\infty}{\longrightarrow} 0; \tag{7.7}$$

(iii) The random pair $(\mathbf{X}, Y)$ (resp. $(\mathbf{X}_\infty, Y_\infty)$) has a continuous density p (resp. q) w.r.t. the Lebesgue measure, and the densities converge uniformly, in the sense that

$$\sup_{(\boldsymbol{\omega}, y) \in \mathbb{S} \times I} \Big| b(t) t^d p(t\boldsymbol{\omega}, y) - q(\boldsymbol{\omega}, y) \Big| \underset{t \to +\infty}{\longrightarrow} 0, \tag{7.8}$$

where $b(t) = \mathbb{P}(\|\mathbf{X}\| \geq t)^{-1}$. *In addition,* q *is uniformly lower bounded on the unit sphere by a positive constant,*

$$\inf_{\omega \in \mathbb{S}, y \in I} q(\omega, y) > 0. \tag{7.9}$$

It should be noticed that Condition (iii) in Proposition 7.6 is a 'one-component variant' of standard assumptions regarding RVs of densities (De Haan and Resnick (1987); Cai et al. (2011)), further discussed in Example 7.10 below. We refer to Appendix 7.A for a proof of Proposition 7.6.

7.4 Examples of Valid Regression Models

We now work out several examples of regression settings in which our Assumptions 7.1, 7.2 and 7.5 are satisfied.

Proposition 7.7 (Noise model with heavy-tailed random design)**.** *Suppose that* $\mathbf{X}$ *is a regularly varying random vector in* $\mathbb{R}^d$*, independent from a real-valued random variable* ε *modeling some noise and consider a target*

$$Y = g(\mathbf{X}, \varepsilon),$$

where $g : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ *is a bounded, continuous mapping. Assume also that there exists a function* $g_\theta : \mathbb{S} \times \mathbb{R} \to \mathbb{R}$ *such that, for all* $z \in \mathbb{R}$

$$\sup_{\|\mathbf{x}\| \geq t} |g(\mathbf{x}, z) - g_\theta(\mathbf{x}/\|\mathbf{x}\|, z)| \to 0, \tag{7.10}$$

as $t \to +\infty$*. Then, the random pair* $(\mathbf{X}, Y)$ *fulfills Assumptions 7.1, 7.2 and 7.5. In particular, the limit distribution* P_∞ *in Equation (7.5) is given by*

$$P_\infty = \mathcal{L}(\mathbf{X}_\infty, g_\theta(\mathbf{X}_\infty/\|\mathbf{X}_\infty\|, \varepsilon)),$$

where $\mathbf{X}_\infty$ *follows the limit distribution*

$$Q_\infty = \lim_{t \to +\infty} \mathcal{L}(t^{-1}\mathbf{X} \mid \|\mathbf{X}\| \geq t).$$

The proof of the claim made in Proposition 7.7 is deferred to Appendix 7.A. Concrete examples arise within the broader context of this generic example, such as the additive noise model $Y = \tilde{g}(\mathbf{X}) + \varepsilon$ and the multiplicative noise model $Y = \varepsilon \tilde{g}(\mathbf{X})$. In both cases, Condition (7.10) holds true whenever $\tilde{g}$ satisfies the similar condition

$$\sup_{\|\mathbf{x}\| \geq t} |\tilde{g}(\mathbf{x}) - \tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{x}))| \to 0,$$

for some angular function $\tilde{g}_\theta$, with minor additional regularity assumptions.

Corollary 7.8 (Additive noise model with heavy-tailed random design)**.** *Consider the additive noise model*

$$Y = \tilde{g}(\mathbf{X}) + \varepsilon,$$

where $\mathbf{X}$ *is a regularly varying random vector in* $\mathbb{R}^d$ *such that*

$$\mathcal{L}\left(t^{-1}\mathbf{X} \mid \|\mathbf{X}\| \geq t\right) \to \mathcal{L}(\mathbf{X}_\infty),$$

as $t \to +\infty$, ε is a bounded real-valued random variable defined on the same probability space independent from $\mathbf{X}$ and $\tilde{g}_\theta$ is a bounded, continuous function on $\mathbb{R}^d$ which converges uniformly to some angular mapping $\tilde{g}_\theta : \mathbb{S} \to \mathbb{R}$, in the sense that

$$\sup_{\|\mathbf{x}\| \geq t} |\tilde{g}(\mathbf{x}) - \tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{x}))| \to 0 \text{ as } t \to +\infty.$$

Then, the random pair $(\mathbf{X}, Y)$ satisfies the requirements of Proposition 7.7 with $M = \sup_{\mathbf{x}\in\mathbb{R}^d} |\tilde{g}(\mathbf{x})| + \|\varepsilon\|_\infty$. The limit distribution P_∞ in Equation (7.5) is

$$P_\infty = \mathcal{L}\Big(\mathbf{X}_\infty, \tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{X}_\infty)) + \varepsilon\Big).$$

Corollary 7.9 (Multiplicative noise model with heavy-tailed random design)**.** *Consider the multiplicative noise model*

$$Y = \varepsilon \tilde{g}(\mathbf{X}),$$

where $(\mathbf{X}, \varepsilon)$ and $\tilde{g}$ are as in Corollary 7.8. Then, the random pair $(\mathbf{X}, Y)$ satisfies the requirements of Proposition 7.7 with $M = \sup_{\mathbf{x}\in\mathbb{R}^d} |\tilde{g}(\mathbf{x})| \times \|\varepsilon\|_\infty$ and the limit distribution P_∞ in (7.5) is given by $P_\infty = \mathcal{L}(\mathbf{X}_\infty, \varepsilon\tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{X}_\infty)))$, where $\tilde{g}_\theta$ and $\mathbf{X}_\infty$ are as in Corollary 7.8.

The next example establishes a strong connection between the considered regression setting and typical situations considered in Extreme Value Analysis where the goal is to predict the occurrence and/or the intensity of unusually large events. The technical proofs of the main claims are gathered in Appendix 7.A.

Proposition 7.10 (Predicting a missing component in a regularly varying random vector))**.** *In this example we show that our assumptions are met when considering a random vector $\tilde{\mathbf{X}}$ with a regularly varying* density*, where the target Y is one missing component from the vector, or more precisely a normalized version of that missing component. The normalization allows to satisfy our boundedness constraint Assumption 7.1. We believe this example could be particularly useful in applications, for imputation of missing data with heavy tails.*

Let $\tilde{\mathbf{X}} \in \mathbb{R}^{d+1}$ have continuous density p, and $b(t) = \mathbb{P}(\|\tilde{\mathbf{X}}\| \geq t)^{-1}$, where $\|\cdot\|$ is the L^p norm on $\mathbb{R}^{d+1}$ for some $p \in [1, +\infty)$. Assume that b is regularly varying with index α for some $\alpha > 0$, and that there exists a positive function q on $\mathbb{R}^{d+1}$ such that for all $\tilde{\mathbf{x}} \neq \mathbf{0}$,

$$t^{d+1} b(t) p(t\tilde{\mathbf{x}}) - q(\tilde{\mathbf{x}}) \underset{t\to+\infty}{\longrightarrow} 0. \tag{7.11}$$

Assume in addition that the convergence is uniform on the sphere,

$$\sup_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} |t^{d+1} b(t) p(t\boldsymbol{\omega}) - q(\boldsymbol{\omega})| \underset{t\to+\infty}{\longrightarrow} 0, \tag{7.12}$$

where $\mathbb{S}_{d+1}$ denotes the unit sphere of $\mathbb{R}^{d+1}$. This assumption is used in De Haan and Resnick (1987); Cai et al. (2011). It is shown in these references that (7.11) and (7.12) imply that $\tilde{\mathbf{X}}$ is regularly varying with index α. More precisely with $\mu(A) = \int_A q(\tilde{\mathbf{x}})\,\mathrm{d}\tilde{\mathbf{x}}$ for any measurable set $A \subset E$, we have $b(t)\mathbb{P}(\tilde{\mathbf{X}}/t \in \cdot) \to \mu(\cdot)$ in the sense of vague convergence. Necessarily q is homogeneous of order $-\alpha - d - 1$. Also the continuity of p implies that of q. Assume finally that

$$\min_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega}) > 0.$$

Another useful feature of this setting is that, if (7.11) and (7.12) hold, then also

$$\sup_{\|\tilde{\mathbf{x}}\|\geq 1} |p(t\tilde{\mathbf{x}})t^{d+1}b(t) - q(\tilde{\mathbf{x}})| \underset{t\to+\infty}{\longrightarrow} 0. \tag{7.13}$$

Let $\mathbf{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$ *and* $Y = \tilde{X}_{d+1}/\|\tilde{\mathbf{X}}\|$. *The norm* $\|\mathbf{x}\|$ *also denotes the* L^p*-norm in* $\mathbb{R}^d$ *when it is clear from the context that* $\mathbf{x} \in \mathbb{R}^d$. *Then*

(i) The pair $(\mathbf{X}, Y)$ *satisfies Assumptions 7.1, 7.2 and 7.5;*

(ii) The limit pair $(\mathbf{X}_\infty, Y_\infty)$ *for* $(\mathbf{X}, Y)$ *defined in (7.5) has distribution*

$$\mathcal{L}\left(\left(\tilde{\mathbf{X}}_{\infty,1:d}, \frac{\tilde{X}_{\infty,d+1}}{\|\tilde{\mathbf{X}}_\infty\|}\right) | \, \|\tilde{\mathbf{X}}_{\infty,1:d}\| \geq 1\right),$$

where $\tilde{\mathbf{X}}_{\infty,1:d}$ *denotes the* d*-dimensional vector* $(\tilde{X}_{\infty,1}, \ldots, \tilde{X}_{\infty,d})$.

It is important to observe that predicting Y *allows to predict* $\tilde{X}_{d+1}$, *as*

$$Y = \frac{\tilde{X}_{d+1}}{\|\tilde{\mathbf{X}}\|_p} \iff \tilde{X}_{d+1} = \frac{Y\|\mathbf{X}\|_p}{(1-|Y|^p)^{1/p}}. \tag{7.14}$$

In our experiments with real data we consider this prediction example on a financial dataset.

As will be shown in the forthcoming sections, Assumptions 7.1, 7.2 and 7.5 provide sufficient regularity and stability conditions allowing to justify the *angular* ERM approach taken in Algorithm 7.1.

7.5 Regular Variation w.r.t. the First Component: Parallel with Lindskog et al. (2014)

This section makes explicit the connection between Assumption 7.2 and the RV framework on a metric space developed in Lindskog et al. (2014). We also provide alternative formulations of Assumption 7.2. Following whenever possible the notations of Lindskog et al. (2014), let $\mathcal{Z} = \mathbb{R}^d \times I$ where we recall $I = [-M, M]$ (in Lindskog et al. (2014) the ambient space $\mathcal{Z}$ is denoted by $\mathbb{S}$ which interferes with our notation for the unit sphere). The ambient space $\mathcal{Z}$ is endowed with the Euclidean product metric,

$$d((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)) = \sqrt{\|\mathbf{x}_1 - \mathbf{x}_2\|^2 + (y_1 - y_2)^2},$$

so that $(\mathcal{Z}, d)$ is a complete separable metric space. Define a scalar 'multiplication' on $\mathcal{Z}$ as $\lambda.(\mathbf{x}, y) = (\lambda\mathbf{x}, y)$, $\lambda > 0$, which is continuous and satisfies the associativity property $\lambda_1.(\lambda_2.\mathbf{z}) = (\lambda_1\lambda_2).\mathbf{z}$, and $1.\mathbf{z} = \mathbf{z}$. This scalar multiplication induces a scaling operation on sets, $\lambda A = \{\lambda.\mathbf{z}, \mathbf{z} \in A\}$ for $A \subset \mathcal{Z}$. Consider the set $\mathbb{C} = \{\mathbf{0}\} \times I \subset \mathcal{Z}$. Then $\mathbb{C}$ is a closed set which is preserved by the above scaling operation, *i.e.*, it is a closed cone. For $\mathbf{z} = (\mathbf{x}, y)$ we have $d(\mathbf{z}, \mathbb{C}) = \|\mathbf{x}\|$, whence $d(\mathbf{x}, \mathbb{C}) < d(\lambda\mathbf{x}, \mathbb{C})$ for $\lambda > 1$. Thus Assumptions A1, A2, A3 in Lindskog et al. (2014), Section 3, are satisfied. Let $\mathbb{O} = \mathcal{Z} \setminus \mathbb{C}$ and introduce $\mathbb{C}^r = \{\mathbf{z} \in \mathbb{O} : d(\mathbf{z}, \mathbb{C}) > r\}, r \geq 0$. In Lindskog et al. (2014), the class of Borel measures on $\mathbb{O}$ whose restriction to $\mathcal{Z} \setminus \mathbb{C}^r$ is finite for any $r > 0$ is denoted by $\mathbb{M}_\mathbb{O}$. Then convergence of a sequence of measures $\mu_n \in \mathbb{M}_\mathbb{O}$ towards $\mu \in \mathbb{M}_\mathbb{O}$ is defined as convergence of

functional evaluations $\mu_n(f) \to \mu(f)$ for $f \in \mathcal{C}_{\mathbb{O}}$, the class of continuous functions on $\mathcal{Z}$ which vanish on a neighborhood of $\mathbb{C}$, *i.e.*, whose support is a subset of $\mathbb{C}^r$ for some $r > 0$. A measure $\nu \in \mathbb{M}_{\mathbb{O}}$ is called *regularly varying* with limit measure $\mu \in \mathbb{M}_{\mathbb{O}}$ and scaling sequence $b_n \in \mathbb{R}$, if b_n is increasing, regularly varying in $\mathbb{R}$ and if the sequence of measures $b_n \nu(n\cdot)$ converges in $\mathbb{M}_{\mathbb{O}}$ towards μ (see Definitions 3.1, 3.2 in Lindskog et al. (2014)). From the Portmanteau Theorem 2.1 in Lindskog et al. (2014) and the series of equivalences in Theorem 3.1 of the same reference, our Assumption 7.2 is equivalent to assuming that the distribution P of the random pair $(\mathbf{X}, Y)$ is regularly varying in $\mathbb{M}_{\mathbb{O}}$ with scaling sequence b_n and limit measure μ, with the notations of Section 7.2.

Theorem 7.11. *Let $\mathbb{O}, \mathbb{C}$ be defined as above the statement, let $\mu \in \mathbb{M}_{\mathbb{O}}$ be a nonzero measure and let $b(t)$ be a regularly varying function on $\mathbb{R}^+$ with index $\alpha > 0$. Let $(\mathbf{X}, Y) \sim P$ be a random pair valued in $\mathbb{R}^d \times I$. The following assertions are equivalent.*

(i) The random pair $(\mathbf{X}, Y)$ satisfies Assumption 7.2 from the main paper with limit measure μ and normalizing function b.

(ii) For any bounded and continuous function $h : \mathbb{O} \to \mathbb{R}$ that vanishes in a neighborhood of $\mathbb{C}$, i.e., whose support is included in $\mathbb{C}^r$ for some $r > 0$,

$$\lim_{t \to +\infty} b(t) \mathbb{E}\left[h(t^{-1}\mathbf{X}, Y)\right] = \int_{\mathbb{O}} h \, \mathrm{d}\mu.$$

(iii) There exists a finite measure Φ on $\mathbb{S} \times I$ such that

$$\frac{\mathbb{P}(\boldsymbol{\theta}(\mathbf{X}) \in B, Y \in A, \|\mathbf{X}\| \geq tr)}{\mathbb{P}(\|\mathbf{X}\| \geq t)} \underset{t \to +\infty}{\longrightarrow} cr^{-\alpha} \Phi(B \times A)$$

for all $r > 0$ and $A \in \mathcal{B}(I)$, $B \in \mathcal{B}(\mathbb{S})$ such that $\Phi(\partial(B \times A)) = 0$, with $c = \Phi(\mathbb{S} \times I)^{-1}$.

Proof. *(i)* $\Leftrightarrow$ *(ii)*. Condition *(ii)* in the statement is precisely Definition 3.2 of RV in $\mathbb{M}_{\mathbb{O}}$ of Lindskog et al. (2014), regarding the measure P restricted to $\mathbb{O}$. The equivalence with our Assumption 7.2 is a direct application of the Portmanteau Theorem 2.1 in Lindskog et al. (2014).

(iii) $\Leftrightarrow$ *(ii)*. We generalize the argument of Lindskog et al. (2014), Example 3.4 and we verify that we fit into the context of Example 3.5 of the same reference. The argument in Example 3.5 (see also Example 3.4) in Lindskog et al. (2014) relies on a continuous mapping argument (Theorem 2.3 in the same reference). Introduce the 'polar coordinate transform' $T(\mathbf{x}, y) = (\|\mathbf{x}\|, \boldsymbol{\theta}(\mathbf{x}), y)$, for $(\mathbf{x}, y) \in \mathbb{O}$, where we recall $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$. Then T is a homeomorphism from $\mathbb{O}$ onto $\mathbb{O}' = (\mathbb{R}_+ \setminus \{0\}) \times \mathbb{S} \times I = \mathcal{Z}' \setminus \mathbb{C}'$ with $\mathcal{Z}' = \mathbb{R}_+ \times \mathbb{S} \times I$, $\mathbb{C}' = \{0\} \times \mathbb{S} \times I$. The space $\mathcal{Z}'$ is endowed with a continuous scalar multiplication $\lambda.(r, \boldsymbol{\omega}, y) = (\lambda r, \boldsymbol{\omega}, y)$ for $\lambda \geq 0$, which is compatible with the mapping T in the sense that $\lambda.T(\mathbf{z}) = T(\lambda.\mathbf{z})$. The scalar multiplication on $\mathcal{Z}'$ satisfies the same associativity and monotonicity properties as the one on $\mathcal{Z}$. The mapping T has the property that if $A' \subset \mathbb{O}'$ is bounded away from $\mathbb{C}'$ then also $T^{-1}(A') \subset \mathbb{O}$ is bounded away from $\mathbb{C}$. The conditions of Example 3.5 in Lindskog et al. (2014) are thus satisfied, so that regular variation of the joint distribution P (restricted to $\mathbb{O}$) in $\mathbb{M}_{\mathbb{O}}$ is equivalent to RV of the image measure $T_\star P$ (restricted to $\mathbb{O}'$), with limit measure $\mu' = T_\star \mu$, and with the same scaling function $b(t)$. In other words Condition *(ii)* is equivalent to

the fact that for any measurable sets $B \subset \mathbb{S}, C \in I$ such that $\mu(\partial(\mathcal{C}_B \times C)) = 0$, where $\mathcal{C}_B = \{t\boldsymbol{\omega}, t \geq 1, \boldsymbol{\omega} \in B\}$, we have

$$\begin{aligned} b(t)\mathbb{P}(\|\mathbf{X}\| > tr, \boldsymbol{\theta}(\mathbf{X}) \in B, Y \in C) &\underset{t\to+\infty}{\longrightarrow} \mu(\{(\mathbf{x}, y) : \|\mathbf{x}\| \geq r, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C\}) \\ &= \mu(r.\{(\mathbf{x}, y) : \|\mathbf{x}\| \geq 1, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C\}) \\ &= r^{-\alpha}\mu(\{(\mathbf{x}, y) : \|\mathbf{x}\| \geq 1, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C\}), \end{aligned}$$

where the last identity follows from the homogeneity of μ (Theorem 3.1 in Lindskog et al. (2014)). Define the angular measure Φ on $\mathbb{S} \times I$ as in (7.3) from the main paper, $\Phi(B \times C) = \mu(\{(\mathbf{x}, y) \in \mathbb{O} : \|\mathbf{x}\| \geq 1, \boldsymbol{\theta}(\mathbf{x}) \in B, y \in C\})$. Then Φ is a finite measure and the latter display writes equivalently

$$b(t)\mathbb{P}(\|\mathbf{X}\| > tr, \boldsymbol{\theta}(\mathbf{X}) \in B, Y \in C) \underset{t\to+\infty}{\longrightarrow} r^{-\alpha}\Phi(B \times C), \tag{7.15}$$

for all measurable sets $B \subset \mathbb{S}, C \in I$ such that $\Phi(\partial(B \times C)) = 0$. If (7.15) holds then also, taking $B = \mathbb{S}, C = I, r = 1$ we have

$$b(t)\mathbb{P}(\|\mathbf{X}\| > t) \underset{t\to+\infty}{\longrightarrow} \Phi(\mathbb{S} \times I),$$

and taking the ratio of (7.15) with the latter displays yields Condition *(iii)* of the statement. Conversely if *(iii)* holds, then letting $b(t) = \Phi(\mathbb{S} \times I)/\mathbb{P}(\|\mathbf{X}\| > t)$, we obtain (7.15), which is equivalent to Condition *(ii)*. ■

7.6 Conclusion

In this chapter, we introduce an algorithmic procedure named ROXANE, designed to handle regression tasks in extreme regions. To support the soundness of this algorithm, we develop a framework for extreme problems where extremality is measured with respect to a specific component. We propose the novel assumption of regular variation with respect to the first component and extend this to establish the classical properties of regular variation under this hypothesis. Finally, we present typical regression scenarios where these working assumptions are satisfied. These regression situations are the subject of experimental studies in the next section, where probabilistic and statistical guarantees regarding the ROXANE algorithm are proved.

7.A Proofs

Proofs for Section 7.3

Proof of Proposition 7.6. We show that if Assumptions 7.1 and 7.2 both hold true, then each condition (i), (ii), or (iii) of the statement imply Assumption 7.5. In fact we show that (iii)⇒ (ii) ⇒ (i) ⇒ Assumption 7.5.

Condition (i) ⇒ Assumption 7.5. The continuity of $f^*_{P_\infty}$ follows from the continuity of f^* and the uniform convergence (7.6). Also, the convergence in Assumption 7.5 is a direct consequence of convergence (7.6).

Condition (ii) ⇒ Condition (i). For $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| \geq t \geq 1$, we have

$$\begin{aligned} |f^*(\mathbf{x}) - f^*_{P_\infty}(\mathbf{x})| &= \left| \int_{y\in I} y p_{Y|\mathbf{x}}(y) dy - \int_{y\in I} y p^\infty_{Y|\mathbf{x}}(y) dy \right| \\ &\leq M^2 \sup_{\|\mathbf{x}\|\geq t, y\in I} |p_{Y|\mathbf{x}}(y) - p^\infty_{Y|\mathbf{x}}(y)|. \end{aligned}$$

Thus, uniform convergence in (7.6) follows from (7.7). The continuity of f^* is ensured by an application of the dominated convergence theorem to the parametric integral $f^*(\mathbf{x}) = \int_I y p_{Y|\mathbf{x}}(y)\, \mathrm{d}y$, using the fact that for all $y \in I$, $\mathbf{x} \mapsto p_{Y|\mathbf{x}}(y)$ is continuous and that $\sup_{\|\mathbf{x}\|\geq 1, y\in I} p_{Y|\mathbf{x}}(y) < +\infty$.

Condition (iii) ⇒ Condition (ii). We first show that uniform convergence (7.7) holds true. Notice first that the density q of μ is necessarily homogeneous in its first component, $q(t\mathbf{x}, y) = t^{-\alpha-d} q(\mathbf{x}, y)$ for $\mathbf{x} \neq \mathbf{0}$ (This follows from the homogeneity of μ and a change of variable in the first component when integrating over a region $tA \times B$ where $A \subset \mathbb{R}^d \setminus \{0\}$ and $B \subset I$). Thus for $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| \geq 1$ and $y \in I$, we have

$$p_{Y|\mathbf{x}}(y) = \frac{p(\mathbf{x}, y)}{p_X(\mathbf{x})} \quad \text{and} \quad p^\infty_{Y|\mathbf{x}}(y) = \frac{q(\mathbf{x}, y)}{q_X(\mathbf{x})} = \frac{q(\mathbf{x}/\|\mathbf{x}\|, y)}{q_X(\mathbf{x}/\|\mathbf{x}\|)},$$

where we denote by p_X (resp. q_X) the marginal density of X (resp. X_∞) given by $p_X(\mathbf{x}) = \int_I p(\mathbf{x}, y) dy$ (resp. $q_X(\mathbf{x}) = \int_I q(\mathbf{x}, y) dy$). Then, for $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}, y \in I$, introducing the function $h(t) = t^d b(t)$, the left-hand side in Equation (7.7) writes as

$$\begin{aligned} \left| \frac{p(\mathbf{x}, y)}{p_X(\mathbf{x})} - \frac{q(\mathbf{x}/\|\mathbf{x}\|, y)}{q_X(\mathbf{x}/\|\mathbf{x}\|)} \right| &= \left| \frac{h(\|\mathbf{x}\|) p(\mathbf{x}, y)}{h(\|\mathbf{x}\|) p_X(\mathbf{x})} - \frac{q(\mathbf{x}/\|\mathbf{x}\|, y)}{q_X(\mathbf{x}/\|\mathbf{x}\|)} \right| \\ &\leq \underbrace{h(\|\mathbf{x}\|) p(\mathbf{x}, y) \left| \frac{1}{h(\|\mathbf{x}\|) p_X(\mathbf{x})} - \frac{1}{q_X(\mathbf{x}/\|\mathbf{x}\|)} \right|}_{A(\mathbf{x},y)} + \dots \\ &\quad \underbrace{\frac{\left| h(\|\mathbf{x}\|) p(\mathbf{x}, y) - q(\mathbf{x}/\|\mathbf{x}\|, y) \right|}{q_X(\mathbf{x}/\|\mathbf{x}\|)}}_{B(\mathbf{x},y)}. \end{aligned} \tag{7.16}$$

Regarding the numerator of the term $B(\mathbf{x}, y)$ above, notice that for $\|\mathbf{x}\| \geq t$,

$$\begin{aligned} |h(\|\mathbf{x}\|) p(\mathbf{x}, y) - q(\mathbf{x}/\|\mathbf{x}\|, y)| &= |h(t(\|\mathbf{x}\|/t)) p(t(\|\mathbf{x}\|/t)(\mathbf{x}/\|\mathbf{x}\|), y) - q(\mathbf{x}/\|\mathbf{x}\|, y)| \\ &\leq \sup_{s\geq t, (\boldsymbol{\omega}, y)\in \mathbb{S}\times I} |h(s) p(s\boldsymbol{\omega}, y) - q(\boldsymbol{\omega}, y)| \to 0, \end{aligned}$$

as t tends to infinity, by uniform convergence (7.8). This, together with the lower bound (7.9) on q, implies that as $t \to +\infty$,

$$\sup_{\|\mathbf{x}\|>t, y\in I} B(\mathbf{x}, y) \to 0.$$

Turning to the term $A(\mathbf{x}, y)$ in (7.16), observe first that

$$A(\mathbf{x}, y) = h(\|\mathbf{x}\|)p(\mathbf{x}, y)\left|\frac{h(\|\mathbf{x}\|)p_X(\mathbf{x}) - q_X(\mathbf{x}/\|\mathbf{x}\|)}{h(\|\mathbf{x}\|)p_X(\mathbf{x})q_X(\mathbf{x}/\|\mathbf{x}\|)}\right|.$$

Notice that for $\|\mathbf{x}\| > t$,

$$\begin{aligned}|h(\|\mathbf{x}\|)p_X(\mathbf{x}) - q_X(\mathbf{x}/\|\mathbf{x}\|)| &= \left|\int_I (h(\|\mathbf{x}\|)p(\mathbf{x}, y) - q(\mathbf{x}/\|\mathbf{x}\|, y))dy\right| \\ &\leq 2M \sup_{s\geq t, (\boldsymbol{\omega}, y)\in \mathbb{S}\times I} |h(s)p(s\boldsymbol{\omega}, y) - q(\boldsymbol{\omega}, y)| := U(t), \qquad (7.17)\end{aligned}$$

where the upper bound $U(t)$ vanishes as $t \to +\infty$ because of (7.8). Now, for $\|\mathbf{x}\| > t$ and $y \in I$,

$$A(\mathbf{x}, y) \leq \frac{\sup_{\|\mathbf{x}\|\geq t, y\in I} h(\|\mathbf{x}\|)p(\mathbf{x}, y)}{\inf_{\|\mathbf{x}\|>t} h(\|\mathbf{x}\|)p_X(\mathbf{x}) \inf_{\boldsymbol{\omega}\in\mathbb{S}} q_X(\boldsymbol{\omega})} U(t).$$

Regarding the numerator of the above display, recall that the density function q is continuous on the compact set $\mathbb{S}$, whence it is upper bounded. Because of uniform convergence (7.8), it is also true that $\sup_{\|\mathbf{x}\|\geq t, y\in I} h(\|\mathbf{x}\|)p(\mathbf{x}, y)$ is upper bounded by a finite constant for t large enough. In addition, our lower bound assumption (7.9) on q together with uniform convergence (7.17) show that the denominator is ultimately (as $t \to +\infty$) lower bounded by a positive constant. Summarizing, we have shown that $\sup_{\|\mathbf{x}\|>t, y\in\mathbb{S}} A(\mathbf{x}, y) \to 0$ as $t \to +\infty$, finishing the proof of (7.7).

It remains to prove that for all $y \in I$, $\mathbf{x} \mapsto p(\mathbf{x}, y)/p_X(\mathbf{x})$ is continuous and that $p(\mathbf{x}, y)/p_X(\mathbf{x})$ is uniformly bounded. For all $y \in I$, the continuity of $\mathbf{x} \mapsto p(\mathbf{x}, y)/p_X(\mathbf{x})$ follows from the continuity of p. Notice again that for $\mathbf{x} \in \mathbb{R}^d$ and $y \in I$

$$\frac{p(\mathbf{x}, y)}{p_X(\mathbf{x})} = \frac{h(\|\mathbf{x}\|)p(\mathbf{x}, y)}{h(\|\mathbf{x}\|)p_X(\mathbf{x})}.$$

The numerator uniformly converges to q, which is uniformly bounded. The denominator uniformly converges to q_X, which is uniformly lower bounded by Equation (7.9). Then $\sup_{\|\mathbf{x}\|\geq 1, y\in I}(p(\mathbf{x}, y)/p_X(\mathbf{x}))$ is finite, which concludes the proof. ∎

Proofs for Section 7.4

In this section, we show that a generic heavy-tailed regression model (Example 7.7) satisfies the requirements of our assumptions. Subsequently, we establish that two widely used models, the additive and multiplicative noise models, constitute particular instances of that generic model.

Proof of Proposition 7.7. Assumption 7.1 is fulfilled with $M = \sup_{\mathbf{x},z\in\mathbb{R}^d\times\mathbb{R}} |g(\mathbf{x},z)|$. Regarding Assumption 7.2 and the limit distribution, we consider a bounded and Lipschitz function $l : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$. For all $t > 0$, writing $\boldsymbol{\Theta} = \|\mathbf{X}\|^{-1}\mathbf{X}$, we have

$$\begin{aligned}\mathbb{E}\Big[l(t^{-1}\mathbf{X}, Y) \mid \|\mathbf{X}\| \geq t\Big] &= \mathbb{E}\Big[l(t^{-1}\mathbf{X}, g(\mathbf{X},\varepsilon)) \mid \|\mathbf{X}\| \geq t\Big] \\ &= \mathbb{E}\Big[l(t^{-1}\mathbf{X}, g_\theta(\boldsymbol{\Theta},\varepsilon)) \mid \|\mathbf{X}\| \geq t\Big] + \dots \\ &\quad \mathbb{E}\Big[l(t^{-1}\mathbf{X}, g(\mathbf{X},\varepsilon)) - l(t^{-1}\mathbf{X}, g_\theta(\boldsymbol{\Theta},\varepsilon)) \mid \|\mathbf{X}\| \geq t\Big].\end{aligned}$$

Since ε is independent from $\mathbf{X}$, writing $\boldsymbol{\Theta}_\infty = \|\mathbf{X}_\infty\|^{-1}\mathbf{X}_\infty$, the RV of $\mathbf{X}$ and continuity of l and g_θ imply that

$$\mathbb{E}\Big[l(t^{-1}\mathbf{X}, g_\theta(\boldsymbol{\Theta},\varepsilon)) \mid \|\mathbf{X}\| \geq t\Big] \to \mathbb{E}[l(\mathbf{X}_\infty, g_\theta(\boldsymbol{\Theta}_\infty,\varepsilon))]. \tag{7.18}$$

Because l is Lipschitz continuous (for some Lipschitz constant C) and $\mathbf{X}$ and ε are independent, we have

$$\begin{aligned}&\Big|\mathbb{E}\Big[l(t^{-1}\mathbf{X}, g(\mathbf{X},\varepsilon)) - l(t^{-1}\mathbf{X}, g_\theta(\boldsymbol{\Theta},\varepsilon))\Big|\|\mathbf{X}\| \geq t\Big]\Big| \\ &\leq C\mathbb{E}\Big[|g(\mathbf{X},\varepsilon) - g_\theta(\boldsymbol{\Theta},\varepsilon)| \mid \|\mathbf{X}\| \geq t\Big] \\ &\leq C\mathbb{E}\Big[\sup_{\|\mathbf{x}\|\geq t} |g(\mathbf{x},\varepsilon) - g_\theta(\boldsymbol{\theta}(\mathbf{x}),\varepsilon)|\Big].\end{aligned}$$

The right-hand side tends to zero as $t \to +\infty$, from the dominated convergence theorem which applies because $\sup_{\|\mathbf{x}\|\geq t} |g(\mathbf{x},\varepsilon) - g_\theta(\mathbf{x}/\|\mathbf{x}\|,\varepsilon)| \leq M$ and because of our model assumption (7.10). Thus Assumption 7.2 is satisfied and $P_\infty = \mathcal{L}(\mathbf{X}_\infty, g_\theta(\boldsymbol{\Theta}_\infty,\varepsilon))$.

We now show that Assumption 7.5 also holds true by proving the stronger condition (i) from Proposition 7.6. For $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| \geq t$, we have by independence of $\mathbf{X}$ and ε,

$$\begin{aligned}|f^*(\mathbf{x}) - f^*_{P_\infty}(\boldsymbol{\theta}(\mathbf{x}))| &= \Big|\mathbb{E}[g(\mathbf{x},\varepsilon)] - \mathbb{E}[g_\theta(\boldsymbol{\theta}(\mathbf{x}),\varepsilon)]\Big| \\ &\leq \mathbb{E}\Big[\sup_{\|\mathbf{x}\|\geq t}\Big|g(\mathbf{x},\varepsilon) - g_\theta(\boldsymbol{\theta}(\mathbf{x}),\varepsilon)\Big|\Big],\end{aligned}$$

which entails as in (7.18) that $\sup_{\|\mathbf{x}\|\geq t} |f^*(\mathbf{x}) - f^*_{P_\infty}(\mathbf{x}/\|\mathbf{x}\|)| \to 0$, as $t \to +\infty$. Since g is assumed continuous and bounded, f^* is continuous. Thus, the sufficient condition (i) from Proposition 7.6 is satisfied, which shows that Assumption 7.5 holds true. ∎

Proof of Corollary 7.8. Because ε is almost surely bounded, there exists $m_\varepsilon \in \mathbb{R}_+$ a nonnegative real-number such that $\varepsilon \overset{a.s.}{\in} [-m_\varepsilon, +m_\varepsilon]$. Consider the mapping $g : (\mathbf{x}, z) \in \mathbb{R}^d \times [-m_\varepsilon, +m_\varepsilon] \mapsto g(\mathbf{x}) + z$ and $g_\theta : (\boldsymbol{\omega}, z) \in \mathbb{S} \times [-m_\varepsilon, +m_\varepsilon] \mapsto \tilde{g}_\theta(\boldsymbol{\omega}) + z$. The function g is continuous and bounded by $M = \sup_{\mathbf{x}\in\mathbb{R}^d} |\tilde{g}(\mathbf{x})| + m_\varepsilon$ and the pair (g, g_θ) satisfies Equation (7.10). Indeed for all $z \in [-m_\varepsilon, +m_\varepsilon]$,

$$\sup_{\|\mathbf{x}\|\geq t} |g(\mathbf{x},z) - g_\theta(\boldsymbol{\theta}(\mathbf{x}),z)| = \sup_{\|\mathbf{x}\|\geq t} |\tilde{g}(\mathbf{x}) - \tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{x}))| \to 0,$$

as $t \to +\infty$, which concludes the proof. ∎

Proof of Corollary 7.9. Consider the mapping $g(\mathbf{x}, z) = z\tilde{g}(\mathbf{x})$ and $g_\theta(\boldsymbol{\omega}, z) = z\tilde{g}_\theta(\boldsymbol{\omega})$. Let m_ε be as in the proof of Proposition 7.8. On the domain $\mathbb{R}^d \times [-m_\varepsilon, m_\varepsilon]$, the function g is continuous and bounded by $M = m_\varepsilon \sup_{\mathbf{x}\in\mathbb{R}^d} |\tilde{g}(\mathbf{x})|$. The pair (g, g_θ) satisfies (7.10) since for all $z \in [-m_\varepsilon, +m_\varepsilon]$

$$\sup_{\|\mathbf{x}\|\geq t} |g(\mathbf{x}, z) - g_\theta(\mathbf{x}/\|\mathbf{x}\|, z)| \leq m_\varepsilon \sup_{\|\mathbf{x}\|\geq t} |\tilde{g}(\mathbf{x}) - \tilde{g}_\theta(\boldsymbol{\theta}(\mathbf{x}))| \xrightarrow[t\to\infty]{} 0,$$

which concludes the proof. ∎

Proof of Proposition 7.10. Let $\tilde{E} = \mathbb{R}^{d+1} \setminus \{0_{\mathbb{R}^{d+1}}\}$, $E = \mathbb{R}^d \setminus \{0_{\mathbb{R}^d}\}$. and for simplicity denote both by $\mathbb{B}_d$ the d-dimensional unit ball and its image by the canonical embedding $\mathbb{R}^d \to \mathbb{R}^{d+1}$, *i.e.*, $\mathbb{B}_d = \{\tilde{\mathbf{x}} \in \mathbb{R}^{d+1} : \|(\tilde{x}_1, \ldots, \tilde{x}_d)\| \leq 1, \tilde{x}_{d+1} \in \mathbb{R}\}$. For $\tilde{\mathbf{x}} \in \mathbb{R}^{d+1}$ we denote by $\mathbf{x}$ the first d coordinates of $\tilde{\mathbf{x}}$, $\mathbf{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)$. Denote by φ the continuous mapping sending $\tilde{\mathbf{X}}$ to $(\mathbf{X}, Y)$, *i.e.*,

$$\begin{aligned} \varphi : \quad & E \times \mathbb{R} \to E \times (-1, 1) \\ & \tilde{\mathbf{x}} = (\mathbf{x}, z) \mapsto (\mathbf{x}, y) = (\mathbf{x}, z/\|(\mathbf{x}, z)\|). \end{aligned}$$

Equipped with these notations, we may proceed with the proof.

(a) Assumption 7.1 is trivially satisfied because $|Y| \leq 1$.

(b) We now show that Assumption 7.2 holds with limit pair $(\mathbf{X}_\infty, Y_\infty)$ as in the second part of the statement. Equipped with the notations introduced above, the pair defined in the statement may be written as $(\mathbf{X}_\infty, Y_\infty) = \varphi(\tilde{\mathbf{X}}_\infty)$, where $\tilde{\mathbf{X}}_\infty$ is well defined by RV of the full vector $\tilde{\mathbf{X}}$. We need to show that for any bounded, continuous function g,

$$\mathbb{E}\Big[g(\mathbf{X}/t, Y) \,|\, \|\tilde{\mathbf{X}}\| \geq t\Big] \to \mathbb{E}\Big[g \circ \varphi(\tilde{\mathbf{X}}_\infty) \,|\, \|\tilde{\mathbf{X}}_{\infty,1:d}\| \geq 1\Big].$$

However $(\mathbf{X}/t, Y) = \varphi(\tilde{\mathbf{X}}/t)$ and $\|\mathbf{X}\| \geq t \Rightarrow \|\tilde{\mathbf{X}}\| \geq t$. Thus

$$\begin{aligned} &\mathbb{E}[g(\mathbf{X}/t, Y) \,|\, \|\mathbf{X}\| \geq t] \\ &\quad = \frac{\mathbb{E}\Big[g \circ \varphi(\tilde{\mathbf{X}}/t) \mathbb{1}\{\|\mathbf{X}/t\| \geq 1\} \mathbb{1}\{\|\tilde{\mathbf{X}}/t\| \geq 1\}\Big]}{\mathbb{P}(\|\tilde{\mathbf{X}}/t\| \geq 1)} \frac{\mathbb{P}(\|\tilde{\mathbf{X}}/t\| \geq 1)}{\mathbb{P}(\|\mathbf{X}/t\| \geq 1)} \\ &\quad = \mathbb{E}\Big[g \circ \varphi(\tilde{\mathbf{X}}/t) \mathbb{1}\{\|\mathbf{X}/t\| \geq 1\} \,|\, \|\tilde{\mathbf{X}}\| \geq t\Big] \frac{\mathbb{P}(\|\tilde{\mathbf{X}}/t\| \geq 1)}{\mathbb{P}(\|\mathbf{X}/t\| \geq 1)} \\ &\quad \to \mathbb{E}\Big[g \circ \varphi(\tilde{\mathbf{X}}_\infty) \mathbb{1}\{\|\tilde{\mathbf{X}}_{\infty,1:d}\| \geq 1\}\Big] \frac{1}{\mathbb{P}(\|\tilde{\mathbf{X}}_{\infty,1:d}\| \geq 1)}, \end{aligned}$$

where the convergence of the first term in the latter expression is obtained by approaching the (discontinuous) function $\mathbb{1}\{\|\mathbf{x}\| \geq 1\}$ by continuous ones and using the fact that the boundary of $\mathbb{B}_d$ in $\mathbb{R}^{d+1}$ is not a cone, whence it cannot carry any positive μ-mass (a standard feature of radially homogeneous measures).

(c) We now prove that Assumption 7.5 holds true by proving the stronger condition (7.6) which rephrase in our setting as

$$\sup_{\|\mathbf{x}\|=1} |f^*(t\mathbf{x}) - f^*_{P_\infty}(t\mathbf{x})| \xrightarrow[t\to+\infty]{} 0. \tag{7.19}$$

Indeed if (7.19) holds, then $\sup_{s\geq t} \sup_{\|\mathbf{x}\|=1} |f^*(s\mathbf{x}) - f^*_{P_\infty}(s\mathbf{x})| \xrightarrow[t\to+\infty]{} 0$, so that

$$\begin{aligned} \sup_{\|\mathbf{x}\|\geq t} |f^*(\mathbf{x}) - f^*_{P_\infty}(\mathbf{x})| &= \sup_{\|\mathbf{x}\|\geq 1} |f^*(t\mathbf{x}) - f^*_{P_\infty}(t\mathbf{x})| \\ &= \sup_{s\geq t} \sup_{\|\mathbf{x}\|=1} |f^*(s\mathbf{x}) - f^*_{P_\infty}(s\mathbf{x})| \xrightarrow[t\to+\infty]{} 0. \end{aligned}$$

Notice first that for $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| \geq 1$, $f^*(\mathbf{x})$ and $f^*_{P_\infty}(\mathbf{x})$ may be written in terms of integrals

$$f^*(\mathbf{x}) = \int_{z\in\mathbb{R}} \frac{z}{\|(\mathbf{x},z)\|} \frac{p(\mathbf{x},z)}{p(\mathbf{x})}\,\mathrm{d}z,$$

where for simplicity we denote by $p(\mathbf{x})$ the marginal density of the first d components of $\tilde{\mathbf{X}}$ at $\mathbf{x}$, and also $p(\mathbf{x},z)$ the joint density at $\tilde{\mathbf{x}} = (\mathbf{x},z)$.

In the present setting, $f^*_{P_\infty}$ is defined as $f^*_{P_\infty}(X_\infty) = \mathbb{E}[Y_\infty \mid \mathbf{X}_\infty]$. Introduce a random vector $\tilde{\mathbf{Z}} = (\tilde{Z}_1,\dots,\tilde{Z}_{d+1})$ distributed as $\mathcal{L}(\tilde{\mathbf{X}}_\infty \,|\, \|\tilde{X}_{\infty,1:d}\| \geq 1)$. Then $\tilde{\mathbf{Z}}$ has density $Cq(\mathbf{x},z)$ on $\mathbb{B}_d^c \times \mathbb{R}$, and marginal density for its first d components, $Cq(\mathbf{x}) := \int_{\mathbb{R}} Cq(\mathbf{x},z)\,\mathrm{d}z$. With these notations we have $(\mathbf{X}_\infty, Y_\infty) \overset{d}{=} (\tilde{Z}_{1:d}, \tilde{Z}_{d+1}/\|\tilde{Z}_{1:d}\|)$, whence $f^*_{P_\infty}(\tilde{Z}_{1:d}) = \mathbb{E}\Big[\tilde{Z}_{d+1}/\|\tilde{\mathbf{Z}}\| \,|\, \tilde{Z}_{1:d}\Big]$ almost surely. We obtain, for $\|\mathbf{x}\| \geq 1$,

$$f^*_{P_\infty}(\mathbf{x}) = \int_{\mathbb{R}} \frac{z}{\|(\mathbf{x},z)\|} \frac{Cq(\mathbf{x},z)}{Cq(\mathbf{x})}\,\mathrm{d}z = \int_{\mathbb{R}} \frac{z}{\|(\mathbf{x},z)\|} \frac{q(\mathbf{x},z)}{q(\mathbf{x})}\,\mathrm{d}z.$$

Combining the latter two displays we obtain

$$|f^*(\mathbf{x}) - f^*_{P_\infty}(\mathbf{x})| \leq \int_{z\in\mathbb{R}} \left| \frac{p(\mathbf{x},z)}{p(\mathbf{x})} - \frac{q(\mathbf{x},z)}{q(\mathbf{x})} \right| \mathrm{d}z. \tag{7.20}$$

Introduce as in Lemma 7.12 the function $h(t) = t^{d+1}/\mathbb{P}\big(\|\tilde{\mathbf{X}}\| \geq t\big)$. For $\|\mathbf{x}\| = 1$, by a change of variable $r = z/t$ in (7.20), we obtain

$$\begin{aligned} |f^*(t\mathbf{x}) - f^*_{P_\infty}(t\mathbf{x})| &\leq \int_{r\in\mathbb{R}} \left| \frac{p(t\mathbf{x},tr)}{p(t\mathbf{x})} - \frac{q(t\mathbf{x},tr)}{q(t\mathbf{x})} \right| t\,\mathrm{d}r \\ &= \int_{r\in\mathbb{R}} \left| \frac{h(t)p(t\mathbf{x},tr)}{t^{-1}h(t)p(t\mathbf{x})} - \frac{q(\mathbf{x},r)}{q(\mathbf{x})} \right| \mathrm{d}r, \end{aligned}$$

since by homogeneity of q, it holds that $q(t\mathbf{x},tr) = t^{-d-1-\alpha}q(\mathbf{x},r)$ while $q(t\mathbf{x}) = t^{-d-\alpha}q(x)$. Thus

$$\sup_{\|\mathbf{x}\|=1} |f^*(t\mathbf{x}) - f^*_{P_\infty}(t\mathbf{x})| \leq \int_{r\in\mathbb{R}} \underbrace{\sup_{\|\mathbf{x}\|=1} \left| \frac{h(t)p(t\mathbf{x},tr)}{t^{-1}h(t)p(t\mathbf{x})} - \frac{q(\mathbf{x},r)}{q(\mathbf{x})} \right|}_{J(t,r)} \mathrm{d}r. \tag{7.21}$$

We have the following controls over the quantities in the latter integrand:

1. $q(\mathbf{x})$ is lower bounded by a positive constant (Lemma 7.13)
2. $\sup_{\|\mathbf{x}\|=1} |h(t)t^{-1}p(t\mathbf{x}) - q(\mathbf{x})| \underset{t\to+\infty}{\longrightarrow} 0$ (Lemma 7.12),
3. For all fixed r, because of (7.13), and since $\|(\mathbf{x},r)\| \geq \|\mathbf{x}\|$,

$$\begin{aligned} \sup_{\|\mathbf{x}\|=1} |h(t)p(t\mathbf{x},tr) - q(\mathbf{x},r)| &\leq \sup_{\|\tilde{\mathbf{u}}\|\geq 1} |h(t)p(t\tilde{\mathbf{u}}) - q(\tilde{\mathbf{u}})| \\ &\underset{t\to+\infty}{\longrightarrow} 0. \end{aligned}$$

Thus, combining 1., 2., 3. above, for fixed r, the integrand $J(t,r)$ in (7.21) converges to 0 as $t \to +\infty$. In order to apply the dominated convergence theorem, we verify that $J(t,r)$ is upper bounded by an integrable function of r. The argument is somewhat similar to the one in the proof of Lemma 7.12. We decompose the integrand as

$$\begin{aligned} J(t,r) &\leq \underbrace{\sup_{\|\mathbf{x}\|=1} \frac{h(t)}{h(t\|(\mathbf{x},r)\|)}}_{A(t,r)} \underbrace{\sup_{\|\mathbf{x}\|=1} \frac{h(t\|(\mathbf{x},r)\|)p\Big(t\|(\mathbf{x},r)\|\boldsymbol{\theta}(\mathbf{x},r)\Big)}{t^{-1}h(t)p(t\mathbf{x})}}_{B(t,r)} + \dots \\ &\dots \underbrace{\sup_{\|\mathbf{x}\|=1} \frac{q(\mathbf{x},r)}{q(\mathbf{x})}}_{C(t,r)} \\ &= A(t,r)B(t,r) + C(t,r). \end{aligned}$$

From the proof of Lemma 7.12 (see Equation (7.22)) we have that for $t \geq t_0$ large enough, and for all $r \in \mathbb{R}$,

$$A(t,r) \leq 2\|(\mathbf{x},r)\|^{-d-\alpha/2-1} \leq 2(1+r^p)^{\frac{-d-\alpha/2-1}{p}},$$

an integrable function of r.

The numerator and the denominator in the definition of $B(t,r)$ converge as $t \to +\infty$, uniformly over $\|\mathbf{x}\| \geq 1$ and $r \in \mathbb{R}$, respectively to $q(\mathbf{x},r)$ and $q(\mathbf{x})$. The latter quantity is lower bounded (Lemma 7.13) and $q(\mathbf{x},r)$ is uniformly bounded for $\|\mathbf{x}\| = 1$ (by homogeneity). Thus, for some constant $C > 0$, for all $t \geq t_1$ with some large enough $t_1 \geq t_0$, we have

$$B(t,r) \leq C.$$

By homogeneity of q and Lemma 7.13 again, we have

$$\begin{aligned} C(t,r) &\leq \sup_{\|\mathbf{x}\|=1} \|(\mathbf{x},r)\|^{-\alpha-d-1} \frac{\max_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega})}{c} \\ &= (1+r^p)^{\frac{-\alpha-d-1}{p}} \frac{\max_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega})}{c}, \end{aligned}$$

which is an integrable function of r.

Combining the bounds regarding $A(t,r), B(t,r), C(t,r)$, we have shown that $A(t,r)\times B(t,r) + C(t,r)$ is upper bounded by an integrable function of r. The proof of the condition (7.6) is complete. It remains to show that $f^*_{P_\infty}$ is continuous on $\|\mathbf{x}\| \geq 1$. Recall that for $\mathbf{x} \in \mathbb{R}^d \setminus \{0_{\mathbb{R}^d}\}$,

$$f^*_{P_\infty}(\mathbf{x}) = \frac{1}{q(\mathbf{x})} \int_{\mathbb{R}} \frac{z}{\|(\mathbf{x},z)\|} q(\mathbf{x},z) dz.$$

The continuity of p implies that of q by Equation (7.12). By homogeneity of q, we have

$$\begin{aligned} \frac{z}{\|(\mathbf{x},z)\|} q(\mathbf{x},z) \leq q(\mathbf{x},z) &= \|(\mathbf{x},z)\|^{-d-\alpha-1} q(\boldsymbol{\theta}(\mathbf{x},z)) \\ &\leq (1+z^p)^{\frac{-d-\alpha-1}{p}} \max_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega}). \end{aligned}$$

Since $z \mapsto (1+z^p)^{\frac{-d-\alpha-1}{p}}$ is integrable over $\mathbb{R}$, the dominated convergence theorem for continuity applies twice and entails that $\mathbf{x} \mapsto \int_{\mathbb{R}} \frac{z}{\|(\mathbf{x},z)\|} q(\mathbf{x},z)dz$ and $\mathbf{x} \mapsto \frac{1}{q(\mathbf{x})}$ are continuous and then $f^*_{P_\infty}$ is continuous. The proof is complete. ■

Lemma 7.12 (Uniform Convergence of marginals of p.)**.** *Under the assumptions of Example 7.10, we have*

$$\sup_{\|\mathbf{x}\|=1} \left| \int_{\mathbb{R}} t^{-1} h(t) p(t\mathbf{x}, z)\, \mathrm{d}z - q(\mathbf{x}) \right| \underset{t\to+\infty}{\longrightarrow} 0, \quad \textit{where}$$
$$q(\mathbf{x}) = \int_z q(\mathbf{x}, z)\, \mathrm{d}z, \quad \textit{and } h(t) = t^{d+1}/\mathbb{P}(\|\tilde{\mathbf{X}}\| \geq t).$$

Proof. We adapt the arguments of the proof of Theorem 2.1 of De Haan and Resnick (1987) to our context. With the notation h from our statement, our uniform convergence assumption (7.12) becomes

$$\sup_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} |h(t)p(t\boldsymbol{\omega}) - q(\boldsymbol{\omega})| \underset{t\to+\infty}{\longrightarrow} 0.$$

Now

$$\int_{\mathbb{R}} t^{-1} h(t) p(t\mathbf{x}, z)\, \mathrm{d}z = \int_{\mathbb{R}} h(t) p(t\mathbf{x}, tr)\, \mathrm{d}r,$$

so that

$$\sup_{\|\mathbf{x}\|=1} \left| \int_{\mathbb{R}} t^{-1} h(t) p(t\mathbf{x}, z)\, \mathrm{d}z - q(\mathbf{x}) \right| \leq \int_{\mathbb{R}} \sup_{\|\mathbf{x}\|=1} \left| h(t)p(t\mathbf{x}, tr) - q(\mathbf{x}, r) \right| \mathrm{d}r.$$

For fixed $r \in \mathbb{R}$, because $\|(\mathbf{x}, r)\| \geq \|\mathbf{x}\| \geq 1$, the integrand in the right-hand side is less than

$$\sup_{\|\tilde{\mathbf{u}}\|\geq 1} \left| h(t)p(t\tilde{\mathbf{u}}) - q(\tilde{\mathbf{u}}) \right|.$$

The latter display tends to zero as $t \to +\infty$ because of (7.13). To conclude, we need to upper bound the integrand by an integrable function of r, in order to apply dominated convergence. We thus write

$$\begin{aligned}
&\sup_{\|\mathbf{x}\|=1} \left| h(t)p(t\mathbf{x}, tr) - q(\mathbf{x}, r) \right| \\
&\leq \sup_{\|\mathbf{x}\|=1} h(t)p(t\mathbf{x}, tr) + \sup_{\|\mathbf{x}\|=1} q(\mathbf{x}, r). \\
&= \underbrace{\sup_{\|\mathbf{x}\|=1} \frac{h(t)}{h(t\,\|(\mathbf{x},r)\|)}}_{A(t,r)} \underbrace{\sup_{\|\mathbf{x}\|=1} h(t\,\|(\mathbf{x},r)\|)p(t\,\|(\mathbf{x},r)\|\,\boldsymbol{\theta}(\mathbf{x},r))}_{B(t,r)} + \underbrace{\sup_{\|\mathbf{x}\|=1} q(\mathbf{x},r)}_{C(t,r)},
\end{aligned}$$

where $\boldsymbol{\theta}(\mathbf{x}, r) \in \mathbb{S}_{d+1}$.

- The function h is regularly varying with positive index $d+1+\alpha$. By Karamata representation (Proposition 0.5 of Resnick (1987)), for t large enough (say $t \geq t_0$), for any $s \geq 1$, we have
$$\frac{h(t)}{h(ts)} \leq 2s^{-d-\frac{\alpha}{2}+1}.$$
Thus for $t \geq t_0$, for all $r \in \mathbb{R}$,
$$A(t,r) \leq 2\|(\mathbf{x},r)\|^{-d-\alpha/2-1} \leq 2(1+r^p)^{\frac{-d-\alpha/2-1}{p}}, \tag{7.22}$$
which is an integrable function of r for any $d \geq 1, \alpha > 0$.
- because $\|(\mathbf{x},r)\| \geq \|\mathbf{x}\| \geq 1$ we have for all $t \geq t_0$ large enough, uniformly over $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$ and $r \in \mathbb{R}$,
$$\left|h(t\,\|(\mathbf{x},r)\|)p(t\,\|(\mathbf{x},r)\|\,\boldsymbol{\theta}(x,r)) - q(\theta(\mathbf{x},r))\right| \leq 1,$$
thus for $t \geq t_0$, for all r,
$$B(t,r) \leq \sup_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega}) + 1,$$
which is a finite constant.
- We may also upper bound $C(t,r)$ by an integrable function of r, since by homogeneity of q,
$$\begin{aligned} C(t,r) &= \sup_{\|\mathbf{x}\|=1} \|(\mathbf{x},r)\|^{-d-\alpha-1} q\Big(\boldsymbol{\theta}(\mathbf{x},r)\Big) \\ &\leq \max_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} (q(\boldsymbol{\omega}))(1+r^p)^{\frac{-d-\alpha-1}{p}} \end{aligned}$$
which is integrable for $d \geq 1$ and $\alpha > 0$.

As a consequence of the above three points, the quantity $A(t,r)B(t,r) + C(t,r)$ is upper bounded by an integrable function of r. The result follows by dominated convergence. ■

Lemma 7.13 (Upper and lower bounds for the marginals of q)**.** *Under the conditions of Example 7.10, there exists positive constants $c, C > 0$ such that for all $x \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = 1$,*
$$c \leq \int q(\mathbf{x},z)\,\mathrm{d}z \leq C.$$

Proof. For $\mathbf{x} \in \mathbb{R}^d$ such that $\|\mathbf{x}\| = 1$, and $z \in \mathbb{R}$ we have
$$q(\mathbf{x},z) = (1+z^p)^{\frac{-\alpha-d-1}{p}} q(\boldsymbol{\theta}(\mathbf{x},z)).$$
The results follows with $c = (\min_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega})) \int (1+z^p)^{\frac{-\alpha-d-1}{p}}\,\mathrm{d}z$ and
$$C = (\max_{\boldsymbol{\omega}\in\mathbb{S}_{d+1}} q(\boldsymbol{\omega})) \int (1+z^p)^{\frac{-\alpha-d-1}{p}}\,\mathrm{d}z.$$
■

In this section, we show that a generic heavy-tailed regression model (Example 7.7) satisfies the requirements of our assumptions. Subsequently, we establish that two widely used models, the additive and multiplicative noise models, constitute particular instances of that generic model.

Chapter 8

Regression on Extremes

Contents

In Section 8.1, we show that a predictive rule using the angular information only *i.e.*, of the form $f(\mathbf{X}) = h(\mathbf{X}/\|\mathbf{X}\|)$, where h is a real-valued function defined the hypersphere $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\}$ reaches the best possible performances w.r.t. the asymptotic risk. Subsequently, we study the performance of a predictive rule based on a training sample $\{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ composed of $n \geq 1$ independent copies of the pair $(\mathbf{X}, Y)$. Non-asymptotic bounds for the excess of asymptotic risk of such an empirical (preasymptotic) risk minimizer are established in Section 8.2, demonstrating its near optimality. Beyond these theoretical guarantees, the performance of empirical risk minimization on extreme covariates is supported by various numerical experiments, on real and simulated datasets, displayed in Section 8.3. Some concluding remarks are collected in Section 8.4.

For the sake of clarity, we recall the main objects introduced in Chapter 7 to which the results in this chapter apply. Under Assumption 7.2, there exists two random variables $(\mathbf{X}_\infty, Y_\infty) \in E \setminus \mathbb{B} \times I$ with distribution P_∞ such that $\mathcal{L}(t^{-1}\mathbf{X}, Y \mid \|\mathbf{X}\| \geq t) \to \mathcal{L}(\mathbf{X}_\infty, Y_\infty)$, as $t \to +\infty$. The extreme risk is defined as $R_{P_\infty}(f) = \mathbb{E}\left[(f(\mathbf{X}_\infty) - Y_\infty)^2\right]$ and the asymptotic risk is defined as $R_\infty(f) = \limsup_{t \to +\infty} \mathbb{E}\left[(f(\mathbf{X}) - Y)^2 \mid \|\mathbf{X}\| \geq t\right]$

8.1 Structural Analysis of Minimizers: Conditional, Asymptotic and Extreme Risks

The main purposes of this subsection are to show that under the assumptions previously listed in Chapter 7, (i) the extreme quadratic risk R_{P_∞} is minimized by angular prediction functions, that is functions depending on the input through the angle only ;

(*ii*) Although R_∞ and R_{P_∞} are different risk functionals, they are connected through their respective minimizers and minimum values.

The first objective (*i*) above is easily tackled. Indeed, the discussion below Equation (7.5) shows that, under Assumption 7.2, letting $\mathbf{\Theta}_\infty = \boldsymbol{\theta}(\mathbf{X}_\infty)$ denote the angular component of X_∞, the random pair $(\mathbf{\Theta}_\infty, Y_\infty)$ is independent from the norm $\|\mathbf{X}_\infty\|$, and in particular Y_∞ and $\|\mathbf{X}_\infty\|$ are independent. Hence, the only useful piece of information carried by $\mathbf{X}_\infty$ to predict Y_∞ is its angular component $\mathbf{\Theta}_\infty$. As a consequence the Bayes regression function satisfies $f^*_{P_\infty}(\mathbf{X}_\infty) = \mathbb{E}[Y_\infty \mid \mathbf{X}_\infty] = \mathbb{E}[Y_\infty \mid \mathbf{\Theta}_\infty]$ almost-surely. As a consequence we may write $f^*_{P_\infty} = h_\infty \circ \boldsymbol{\theta}$ for some function h_∞ defined on the sphere $\mathbb{S}$. Finally, Assumption 7.5 ensures that h_∞ may be chosen as a continuous function. We summarize the discussion in the following lemma.

Lemma 8.1. *Under Assumptions 7.1, 7.2, 7.5, the extreme risk R_{P_∞} has a minimizer (among all measurable functions) which may be written as $f^*_{P_\infty}(x) = h_\infty \circ \boldsymbol{\theta}(\mathbf{x})$ where $h_\infty : \mathbb{S} \to I$ is a bounded, continuous function.*

The next result brings answers regarding the objective (*ii*) outlined above, by establishing a key connection between the (seemingly) different problems of minimizing R_∞ on the one hand, and minimizing R_{P_∞} on the other hand. The extreme risk R_{P_∞} and the asymptotic risk R_∞ are two different functionals, so that the regression function $f^*_{P_\infty}$ is only defined as a minimizer of the extreme risk R_{P_∞} and not the asymptotic risk R_∞. In the sequel we denote by $R^*_{P_\infty}$ the minimum value of the extreme risk, *i.e.*, $R^*_{P_\infty} := \inf_{f \text{ measurable}} R_{P_\infty}(f) = R_{P_\infty}(f^*_{P_\infty})$.

Theorem 8.2. *Under Assumptions 7.1 and 7.2, we have*

(i) For any angular function of the kind $f(\mathbf{x}) = h \circ \boldsymbol{\theta}(\mathbf{x})$, where h is a continuous function defined on $\mathbb{S}$, the conditional risk converges to the extreme risk, i.e.,

$$R_t(f) \xrightarrow[t \to +\infty]{} R_{P_\infty}(f).$$

Thus for such prediction functions, $R_\infty(f) = \lim_{t \to +\infty} R_t(f) = R_{P_\infty}(f)$.

If in addition Assumption 7.5 is satisfied, then the following assertions hold true.

(ii) As $t \to +\infty$, the minimum value of R_t converges to that of R_{P_∞}, i.e., $R^*_t \xrightarrow[t \to +\infty]{} R^*_{P_\infty}$.

(iii) The minimum values of R_∞ and R_{P_∞} coincide, i.e., $R^*_\infty = R^*_{P_\infty}$.

(iv) The regression function $f^*_{P_\infty}$ minimizes the asymptotic conditional quadratic risk:

$$R^*_\infty = R_\infty(f^*_{P_\infty}).$$

The proof is deferred to Appendix 8.A. Observe that Theorem 8.2 does not assert that $R_t(f)$ converges to $R_{P_\infty}(f)$ for all f, but the convergence holds true for angular predictors $f = h \circ \boldsymbol{\theta}$ (Property *(i)* in the statement). Property *(iv)* discloses that the solution $f^*_{P_\infty}$ of the extreme risk minimization problem, which is of angular type, is also a minimizer of the asymptotic conditional quadratic risk R_∞ (and that the minima coincide). Because $f^*_{P_\infty} = h_\infty \circ \boldsymbol{\theta}$ is of angular type, we thus obtain, under Assumptions 7.1, 7.2 and 7.5,

$$\inf_{f \,\text{measurable}} R_\infty(f) = \inf_{h \,\text{measurable}} R_\infty(h \circ \boldsymbol{\theta}). \tag{8.1}$$

In other words, the search for minimizers of R_∞ may indeed be restricted to angular prediction functions, which provides a first heuristic justification for the ROXANE algorithm. However in order to provide rigorous guarantees for the predictive performance of minimizers of the empirical criterion (7.1) computed by means of the ROXANE algorithm, further assumptions regarding the class $\mathcal{H}$ of angular predictors are needed. In particular these additional assumptions ensure uniformity of the convergence result *(i)* from Theorem 8.2. This is the focus of the next section.

8.2 Statistical Guarantees

This section provides a non-asymptotic analysis of the approach proposed for regression on extremes. An upper confidence bound for the excess of R_∞-risk of a solution of (7.1) is established, when the class $\mathcal{H}$ over which empirical minimization is performed is of controlled complexity, see Assumption 8.3 below.

The rationale behind the ROXANE algorithm is to find an angular predictive function that nearly minimizes the asymptotic conditional quadratic risk R_∞ (6.6). Our ERM strategy thus consists in solving an empirical version of the non-asymptotic optimization problem

$$\min_{h\in\mathcal{H}} R_t(h\circ\boldsymbol{\theta}).$$

Recall that a heuristic justification for considering angular classifiers is provided by Equation (8.1), which is itself a consequence of Theorem 8.2. The radial threshold t is chosen as a relatively high quantile of the empirical distribution of the radii $\|\mathbf{X}_i\|$. In particular, let $t_{n,k}$ denote the $1-k/n$ quantile of the norm $\|\mathbf{X}\|$, where $k \ll n$ is large enough so that a statistical analysis remains realistic, but small enough so that the distribution of $(\mathbf{X}, Y)$ given that $\|\mathbf{X}\| > t_{n,k}$ is close to the limit P_∞, see (7.5). Then an empirical version of $t_{n,k}$ is $\hat{t}_{n,k} = \|\mathbf{X}_{(k)}\|$, the k^{th} largest order statistic of the norm already introduced in Algorithm 7.1. In practice the number k of retained extreme statistics in a recurrent issue in Extreme Value Analysis, for which no definite theoretical answer exists, but which is a standard bias/variance compromise. In our experiments, following standard practice we choose k by inspection of stability regions in Hill plots. In addition, in a regression setting we consider feature importance summaries relative to the radial variable, see Section 8.3 for details.

Summarizing, the objective minimized in Algorithm 7.1 may be viewed as an empirical version of the conditional risk $R_{t_{n,k}}$ for a predictive mapping of the form $h\circ\boldsymbol{\theta}$. In the sequel we denote by $\hat{R}_k$ this empirical objective

$$\hat{R}_k(f) = \frac{1}{k}\sum_{i=1}^{k}(Y_{(i)} - f(\mathbf{X}_{(i)}))^2. \tag{8.2}$$

We point out that the statistic above is not an average of independent random variables, as it involves extreme order statistics of the norm. Thus investigating its concentration properties is far from straightforward. The minimum is taken over a class $\mathcal{H}$ of continuous bounded functions on $\mathbb{S}$ of controlled complexity but hopefully rich enough to contain a reasonable approximant of h_∞ introduced in Lemma 8.1. The following assumption regarding $\mathcal{H}$ will turn out to be sufficient to obtain a control of the deviations of the empirical risk. In order to avoid measurability issues regarding supremum deviations over the class $\mathcal{H}$, it is assumed throughout that $\mathcal{H}$ is *pointwise*

measurable (see van der Vaart and Wellner (1996), Example 2.3.4), *i.e.*, that there exists a countable family $\mathcal{H}_0 \subset \mathcal{H}$, such that for all $\omega \in \mathbb{S}$ and all $h \in \mathcal{H}$, there is a sequence $(h_i)_{i\geq 1} \in \mathcal{H}_0$ such that $h_i(\boldsymbol{\omega}) \to h(\boldsymbol{\omega})$. This mild condition is satisfied in most practical cases, in particular by parametric classes $\mathcal{H}$, *i.e.*, classes indexed by a finite dimensional parameter $\boldsymbol{\beta} \in \mathbb{R}^p$, which depend continously on the parameter, *i.e.*, such that $\|h_\beta - h_{\beta_n}\|_{\infty,\mathbb{S}} \to 0$ as $\boldsymbol{\beta}_n \to \boldsymbol{\beta}$.

Assumption 8.3. *The pointwise measurable class $\mathcal{H}$ is a family of continuous, real-valued functions defined on $\mathbb{S}$; of VC dimension $V_\mathcal{H} < +\infty$, and uniformly bounded by the same constant as the target Y (see Assumption 7.1), $\forall h \in \mathcal{H}, \forall \boldsymbol{\omega} \in \mathbb{S}$, $|h(\boldsymbol{\omega})| \leq M$.*

Under the complexity hypothesis above, the following result provides an upper confidence bound for the maximal deviations between the conditional quadratic risk $R_{t_{n,k}}$ and its empirical version $\hat{R}_k$, uniformly over the class $\mathcal{H}$.

Notice that a similar result is obtained in Aghbalou et al. (2023) (Lemma A.3) in the more complex setting of cross validation. For the sake of completeness, we provide a detailed proof in Appendix 8.A.

Theorem 8.4. *Suppose that Assumptions 7.1 and 8.3 are satisfied. Let $\delta \in (0,1)$. We have with probability larger than $1-\delta$*

$$\sup_{h\in\mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \leq \frac{8M^2\sqrt{2\log(3/\delta)} + C\sqrt{V_\mathcal{H}}}{\sqrt{k}} + \frac{16M^2\log(3/\delta)/3 + 4M^2 V_\mathcal{H}}{k},$$

where C is a universal constant.

Notice that Theorem 8.4 controls only the statistical deviations between the sub-asymptotic risk $R_{t_{n,k}}$ and its empirical version $\hat{R}_k$. A control of the bias term $R_{t_{n,k}} - R_\infty$ is given next, under appropriate complexity assumptions controlling the complexity of class $\mathcal{H}$. In particular Assumption 8.3 can be traded against a total boundedness assumption (Case 1. in Proposition 8.5 below) which is further discussed below (Remark 8.6). Regarding the second set of assumption (Case 2.), notice that for $t \geq 1$, the conditional distribution $\Phi_{\theta,t} = \mathcal{L}(\boldsymbol{\theta}(\mathbf{X}) | \|\mathbf{X}\| \geq t)$ is absolutely continuous w.r.t. $\Phi_{\theta,1} = \mathcal{L}(\boldsymbol{\theta}(\mathbf{X}) | \|\mathbf{X}\| \geq 1)$. Indeed for any measurable set $A \subset \mathbb{S}$, if $\mathbb{P}(\boldsymbol{\Theta} \in A | \|\mathbf{X}\| \geq 1) = 0$ then also for any $t \geq 1$, $\mathbb{P}(\boldsymbol{\Theta} \in A | \|\mathbf{X}\| \geq t) = 0$. Denote by $\phi_{\theta,t}$ the probability density of the former angular distribution with respect to the latter.

Proposition 8.5. *Suppose that Assumptions 7.1 and 7.2 are satisfied. Let $\mathcal{H}$ be a class of real-valued, continuous functions on $\mathbb{S}$. Assume that one of the two following conditions is satisfied.*

1. *$\mathcal{H}$ is totally bounded in the space $(\mathcal{C}(\mathbb{S}), \|\cdot\|_\infty)$ of continuous functions on $\mathbb{S}$ endowed with the supremum norm, or*

2. *$\mathcal{H}$ fulfills Assumption 8.3 and in addition, suppose that the conditional densities $\phi_{\theta,t}$ introduced above the statement satisfy*

$$\sup_{t\geq 1,\, \boldsymbol{\omega}\in\mathbb{S}} \phi_{\theta,t}(\boldsymbol{\omega}) = D,$$

for some $0 < D < +\infty$.

Then, as t tends to infinity, we have

$$\sup_{h\in\mathcal{H}} \left| R_t(h\circ\boldsymbol{\theta}) - R_\infty(h\circ\boldsymbol{\theta}) \right| \to 0.$$

The proof of Proposition 8.5 is deferred to Appendix 8.A.

Remark 8.6 (Totally bounded family of regression functions)**.** *Relying on a topological assumption on a set of regression functions such as total boundedness (i.e., $\mathcal{H}$ may be covered by finitely many balls of radius ε, for any $\varepsilon > 0$) is rather uncommon in statistical learning. However it turns out that this condition encompasses several standard algorithms. Namely, if $\mathcal{H}$ is a parametric family indexed by a bounded parameter set, i.e., $\mathcal{H} = \{h_\beta, \boldsymbol{\beta} \in B\}$ for some $B \subset \mathbb{R}^d$ of finite diameter, and if h_β is Lipschitz-continuous with respect to $\boldsymbol{\beta}$, i.e., for some $C > 0$, $\|h_\beta - h_\gamma\|_\infty \leq C\|\boldsymbol{\beta} - \boldsymbol{\gamma}\|$ for all $\boldsymbol{\beta}, \boldsymbol{\gamma} \in B$, then $\mathcal{H}$ satisfies Condition 1. from Proposition 8.5. As an example consider set of functions $h_\beta(\boldsymbol{\omega}) = \langle \boldsymbol{\beta}, \boldsymbol{\omega} \rangle$ for $\boldsymbol{\omega} \in \mathbb{S}$ with a bounded parameter set $B = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}\|_q \leq \lambda\}$ for some fixed $\lambda > 0$, where $\|\cdot\|_q$ is the L^q norm on $\mathbb{R}^d$, $q \geq 1$. The case $q = 2$ (resp. $q = 1$) corresponds to a constrained Ridge (resp. Lasso) regression.*

Remark 8.7 (Bounded angular densities)**.** *The second condition in Proposition 8.5 implies that the angular measure $\Phi_{\theta,t}$ for large t may not concentrate around sets that are negligible with respect to the 'bulk' angular measure $\Phi_{\theta,1}$. This excludes situations where the limit angular measure Φ_θ concentrates on lower dimensional subcones of $\mathbb{R}^d$, whereas $\Phi_{\theta,1}$ does not necessarily do so. This concentration phenomenon as $t \to +\infty$ is precisely the framework considered in recent works on unsupervised dimension reduction for extremes where the goal is to uncover sparsity patterns in the limit angular measure Φ_θ which may not be representative of the bulk behavior (Goix et al. (2016, 2017); Meyer and Wintenberger (2021); Chiapino et al. (2019); Drees and Sabourin (2021); Cooley and Thibaud (2019)). How to relax Condition 2. in order to encompass such frameworks even though the family $\mathcal{H}$ does not satisfy Condition 1. is left to future research.*

The corollary below summarizes the main results of Section 8 in the form of an upper confidence bound for the excess of R_∞-risk for any solution $\hat{f}_k$ of the problem

$$\min_{h\in\mathcal{H}} \hat{R}_k(h\circ\boldsymbol{\theta}).$$

Corollary 8.8 (Summary)**.** *Let $\hat{f}_k = \hat{h}_k \circ \boldsymbol{\theta}$ be the prediction function issued by Algorithm 7.1. Let Assumptions 7.1, 7.2, 7.5 and 8.3 be satisfied. Recall h_∞ from Lemma 8.1 and that, from Theorem 8.2, $R_\infty(h_\infty \circ \boldsymbol{\theta}) = \inf_{h \text{ measurable}} R_\infty(h\circ\boldsymbol{\theta}) = R^*_\infty$.*

For any $\delta > 0$, with probability at least $1 - \delta$, the excess R_∞-risk of $\hat{f}_k$ satisfies

$$R_\infty(\hat{f}_k) - R^*_\infty \leq D_k + B_1(t_{n,k}) + B_2(\mathcal{H}), \tag{8.3}$$

where D_k, B_1, B_2 are respectively a deviation term and two bias terms,

$$\begin{cases} D_k = \left(16M^2\sqrt{2\log(3/\delta)} + 2C\sqrt{V_\mathcal{H}}\right)/\sqrt{k} + \ldots \\ \qquad \left(32M^2\log(3/\delta)/3 + 8M^2 V_\mathcal{H}\right)/k & \textit{(deviations)} \\ B_1(t) = 2\sup_{h\in\mathcal{H}} |R_\infty(h\circ\boldsymbol{\theta}) - R_t(h\circ\boldsymbol{\theta})| & \textit{(threshold bias)} \\ B_2(\mathcal{H}) = \inf_{h\in\mathcal{H}} R_\infty(h\circ\boldsymbol{\theta}) \; - \; R_\infty(h_\infty\circ\boldsymbol{\theta}) & \textit{(class bias).} \end{cases}$$

The first bias term $B_1(t_{n,k})$ in the above bound converges to zero as $n \to +\infty$, $k \to +\infty$, $k/n \to 0$ whenever the conditions of Proposition 8.5 are met.

Proof. Assume for simplicity that the infimum of the R_∞-risk over the class $\mathcal{H}$ is reached, *i.e.*, $\exists h_{\mathcal{H}} \in \mathcal{H} : R_\infty(h_{\mathcal{H}} \circ \boldsymbol{\theta}) = \inf\{R_\infty(h \circ \boldsymbol{\theta}), h \in \mathcal{H}\}$ (if this is not the case, consider an ε-minimizer h_ε for arbitrarily small ε, and proceed). Thus

$$\begin{aligned} R_\infty(\hat{f}_k) - R^*_\infty \leq\ & R_\infty(\hat{h}_k \circ \boldsymbol{\theta}) - R_{t_{n,k}}(\hat{h}_k \circ \boldsymbol{\theta}) + R_{t_{n,k}}(\hat{h}_k \circ \boldsymbol{\theta}) - \hat{R}_k(\hat{h}_k \circ \boldsymbol{\theta}) + \dots \\ & \hat{R}_k(\hat{h}_k \circ \boldsymbol{\theta}) - \hat{R}_k(h_{\mathcal{H}} \circ \boldsymbol{\theta}) + \hat{R}_k(h_{\mathcal{H}} \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h_{\mathcal{H}} \circ \boldsymbol{\theta}) + \dots \\ & R_{t_{n,k}}(h_{\mathcal{H}} \circ \boldsymbol{\theta}) - R_\infty(h_{\mathcal{H}} \circ \boldsymbol{\theta}) + R_\infty(h_{\mathcal{H}} \circ \boldsymbol{\theta}) - \inf_{h \text{ measurable}} R_\infty(h \circ \boldsymbol{\theta}) \\ & + \inf_{h \text{ measurable}} R_\infty(h \circ \boldsymbol{\theta}) - \inf_{f \text{ measurable}} R_\infty(f). \end{aligned}$$

Because $\hat{h}_k \circ \boldsymbol{\theta}$ minimizes $\hat{R}_k$ and considering identity (8.1) (which holds because of Assumptions 7.1, 7.2, 7.5), the above decomposition simplifies into

$$\begin{aligned} R_\infty(\hat{f}_k) - R^*_\infty \leq\ & 2 \sup_{h \in \mathcal{H}} |R_\infty - R_{t_{n,k}}|(h \circ \boldsymbol{\theta}) + 2 \sup_{h \in \mathcal{H}} |R_{t_{n,k}} - \hat{R}_k|(h \circ \boldsymbol{\theta}) + \dots \\ & R_\infty(h_{\mathcal{H}} \circ \boldsymbol{\theta}) - \inf_{h \text{ measurable}} R_\infty(h \circ \boldsymbol{\theta}). \end{aligned}$$

The result follows by plugging in the deviation bound from Theorem 8.4. ∎

As it is generally the case in statistics of extremes, two types of bias terms are involved in the upper bound (8.3) of Corollary 8.8. The first bias term $B_1(t)$ results from the substitution of the conditional quadratic risk $R_{t_{n,k}}$ for its asymptotic limit R_∞. While the weak additional assumptions of Proposition 8.5 ensure that this bias term vanishes as $k/n \to 0$, a quantification of its decay rate would require second-order conditions, *e.g.*, by extending the second order regular variation setting of Resnick and de Haan (1996) to our context of joint regular variation.

The second bias term is a model bias, induced by restricting the family of all measurable functions on $\mathbb{S}$ to the class $\mathcal{H}$ of controlled combinatorial complexity. It should be noted that under the conditions of the statement, Identity (8.1) ensures that restricting to angular predictors does not induce any additional bias term compared with considering a standard class for predictors taking the full covariate (including the radius) as input.

Remark 8.9 (Rate of convergence)**.** *To establish the concentration bound stated in Theorem 8.4, we employ general concentration results that are not ideally tailored for a regression context. A more detailed investigation might yield a bound on the stochastic error term of order $O(\log(k)/k)$, as suggested by standard concentration results (refer to Györfi et al. (2002), Section 11). This refined study is left to future work.*

Remark 8.10 (Extensions)**.** *This article presents a rigorous formulation and investigation of the regression problem involving an output variable confronted to a heavy-tailed input variable, a so far unexplored topic in academic research. Subsequently, we anticipate that the straightforward adaptation of the proposed methodology to incorporate regularized risk formulations or diverse cost functions holds the potential for practical utility and improvements. These extensions lie outside the scope of this paper and are deferred to further works.*

Remark 8.11 (Alternative to ERM)**.** *In the case where the output/response variable Y is heavy-tailed (or possibly contaminated by a heavy-tailed noise), robust alternatives to the ERM approach exist and are preferable (see Lugosi and Mendelson (2019)). Extension of these robust alternatives to the present context of heavy-tailed input is beyond the scope of this paper but will be the subject of further research.*

8.3 Numerical Experiments

We now investigate the performance of the approach previously described and theoretically analyzed for regression on extremes from an empirical perspective on several simulated and real datasets. The code used to run our experiments is available at https://bitbucket.org/nathanhuet/extremeregression. The Mean Square Error (MSE) in extreme regions of angular regression functions output by specific implementations of the ROXANE algorithm are compared to those of the classic regression functions, learned in a standard fashion. On this occasion we propose a simple graphical diagnostic procedure allowing to check visually whether the data meet our assumptions, in particular Assumption 7.2 which is central in our work. More precisely we inspect the relative importance of the radial variable $\|\mathbf{X}\|$ for predicting Y above increasing radial thresholds. We consider in Section 8.3.1 simulated data in the additive and multiplicative models which are particular instances of Example 7.7. In Section 8.3.2 we consider a real-life financial dataset which has already been studied in an EVA context Meyer and Wintenberger (2023).

8.3.1 Simulated data

As a first go, we focus on predictive performance of the ROXANE algorithm in terms of MSE, with simulated data respectively from an additive noise model and from a multiplicative noise model with heavy-tailed design, $Y = \tilde{g}_0(\mathbf{X}) + \varepsilon_0$, and $Y = \varepsilon_1 \tilde{g}_1(\mathbf{X})$, where $\|\cdot\| = \|\cdot\|_2$, $\tilde{g}_0(\mathbf{x}) = \boldsymbol{\beta}^T \boldsymbol{\theta}(\mathbf{x})(1+1/\|\mathbf{x}\|)$, and $\tilde{g}_1(\mathbf{x}) = \cos(1/\|\mathbf{x}\|) \sum_{i=1}^{d/2} (\boldsymbol{\theta}(\mathbf{x})_{2i-1} - 1/\|\mathbf{x}\|^2) \times \sin((\boldsymbol{\theta}(\mathbf{x})_{2i} - 1/\|\mathbf{x}\|^2)\pi)$, for $\mathbf{x} \in E$.

Both models satisfy our assumptions (see Corollaries 7.8 and 7.9 in Chapter 7). In the additive model (resp. in the multiplicative model) the design $\mathbf{X}$ is generated according to a multivariate extreme value distribution from the logistic family Stephenson (2003) with dependence parameter $\xi = 1$, which means that extreme observations occur very close to the axes (resp. $\xi = 0.7$, meaning that the angular component of extreme observations is relatively spread-out in the positive orthant of the unit sphere). The input 1-d marginals are standard Pareto with shape parameter $\alpha = 1$ (resp. $\alpha = 3$) and the noise ε_0 is defined as a truncated Gaussian variable on $[-1, 1]$ with zero mean and standard deviation $\sigma_0 = 0.1$, *i.e.*, ε_0 admits the probability density $f_{\varepsilon_0}(x) = \mathbb{1}\{|x| \leq 1\} \exp(-x^2/(2\sigma_0^2)) / \int_{-1}^{1} \exp(-z^2/(2\sigma_0^2)) dz$. For the multiplicative model, ε_1 is again a truncated Gaussian variable on $[0, 2]$ with mean $\mu = 1$ and standard deviation $\sigma_1 = 0.1$, *i.e.*, ε_1 has density $f_{\varepsilon_1}(x) = \mathbb{1}\{0 \leq x \leq 2\} \exp(-(x-\mu)^2/(2\sigma_1^2)) / \int_0^2 \exp(-(z-\mu)^2/(2\sigma_1^2)) dz$.

The simulated data is of dimension $d = 7$ (resp. $d = 14$). For both models, the size of the training dataset is $n_{train} = 10\,000$, and the number of extreme observations retained for training the ROXANE algorithm is set to $k_{train} = 1000$ ($= n_{train}/10$). The size of the test dataset is $n_{test} = 100\,000$ and the $k_{test} = 10\,000$ ($= n_{test}/10$) largest instances are used to evaluate predictive performance on extreme covariates. We consider three different regression algorithms implemented in the *scikit-learn* library Pedregosa et al. (2011) with the default parameters, namely Ordinary Least Squares (OLS), Support Vector Regression (SVR), and Random Forest (RF). Predictive functions are learned using respectively (i) the full training dataset, (ii) a reduced dataset composed of the k_{train} largest observations $\mathbf{X}_{(1)}, \ldots \mathbf{X}_{(k_{train})}$, and (iii) an angular dataset $\boldsymbol{\Theta}_{(1)}, \ldots \boldsymbol{\Theta}_{(k_{train})}$ consisting of the angles of the k_{train} largest observations. These three options correspond respectively to (i) the default strategy (using the full dataset), (ii) a 'reasonable'

Methods/Models	Train on $\mathbf{X}$	Train on $\mathbf{X} \mid \|\mathbf{X}\|$ large	Train on $\boldsymbol{\Theta} \mid \|\mathbf{X}\|$ large
Add.: OLS	$23_{\pm 29}$	$3_{\pm 6}$	$\mathbf{0.003}_{\pm \mathbf{0.001}}$
SVR	$0.13_{\pm 0.01}$	$0.05_{\pm 0.02}$	$\mathbf{0.003}_{\pm \mathbf{0.001}}$
RF	$0.012_{\pm 0.004}$	$0.007_{\pm 0.002}$	$\mathbf{0.004}_{\pm \mathbf{0.001}}$
Mult.: OLS	$0.006_{\pm 0.001}$	$0.003_{\pm 0.001}$	$\mathbf{0.001}_{\pm \mathbf{0.001}}$
SVR	$0.0041_{\pm 0.0002}$	$0.0038_{\pm 0.0004}$	$\mathbf{0.0034}_{\pm \mathbf{0.0003}}$
RF	$0.0020_{\pm 0.0001}$	$0.0013_{\pm 0.0001}$	$\mathbf{0.0004}_{\pm \mathbf{0.0001}}$

Table 8.1: Average MSE (and standard deviation) for regression functions trained using all observations, extreme observations and angles of extreme observations, over 10 independent replications of the dataset generated in the additive and the multiplicative noise models.

naive strategy (training on extreme covariates for the purpose of predicting from extreme covariates), (*iii*) the strategy that we promote in this paper, corresponding to Algorithm 7.1. We evaluate the performance of the outputs using the MSE computed on the test set. Table 8.1 shows the average MSE's when repeating this experiment across $E = 10$ independent replications of the dataset. For the additive model the regression parameter $\boldsymbol{\beta}$ is randomly chosen for each replication, namely each entry of $\boldsymbol{\beta}$ is drawn uniformly at random over the interval $[0, 1]$.

With both models, the approach we promote for regression on extremes clearly outperforms its competitors, no matter the algorithm (*i.e.*, the model bias) considered. This paper being the first to consider regression on extremes (see Remark 8.11 for a description of regression problems of different nature with heavy-tailed data), no other alternative approach is documented in the literature.

Besides prediction performance, we propose to assess the validity of our main modeling assumption (Assumption 7.2) by inspecting the *variable importance* (*a.k.a. feature importance*, see, *e.g.*, Grömping (2015) and the references therein) of the radial variable $\|\mathbf{X}\|$ compared with the angular variables $\boldsymbol{\Theta}_j, 1 \leq j \leq d$, for the purpose of predicting the target Y. Indeed, under Assumption 7.2, the variables Y and $\|\mathbf{X}\|$ are asymptotically independent conditional on $\{\|\mathbf{X}\| > t\}$ as $t \to +\infty$, so that the variable importance of $\|\mathbf{X}\|$, when restricting the training set to regions above increasingly large radial thresholds, should in principle vanish.

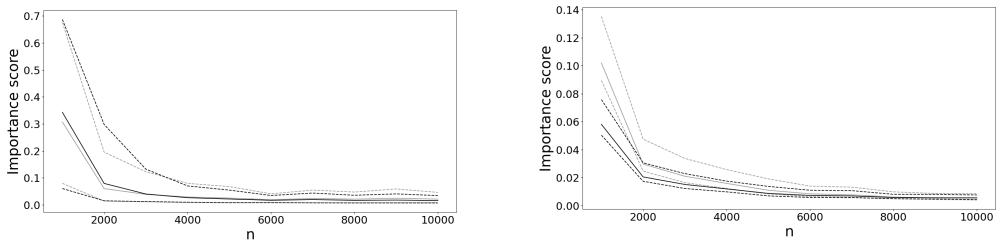
We consider here two widely used measures of feature importance, Gini importance (or Mean Decrease of Impurity, Breiman (2017); Wei et al. (2015)) and Permutation feature importance Breiman (2001); Wei et al. (2015) in the context of Random Forest prediction, as implemented in the *scikit-learn* library. Gini importance measures a mean decrease of impurity in a forest of trees, between parent nodes involving a split on the considered variables, and their child nodes. Gini score is normalized so that the sum of all importance scores across variables equals 1. Permutation importance compares the prediction performance of the original input dataset with the same dataset where the values of the considered variable have been randomly shuffled. A large score indicates a high predictive value of the variable for both measures.

The aim of this second experiment is to illustrate the decrease of the radial feature importance for reduced datasets involving increasingly (relatively) large inputs. To cancel out the perturbation effect of reduced sample sizes, we fix a training size

$k_{imp} = 1000$ and we simulate increasingly large datasets of size

$$n_{imp} \in \{k_{imp}, 2k_{imp}, \dots, 10k_{imp}\}$$

in the additive and multiplicative models described above. Then for $j \in \{1, \dots, 10\}$ the k_{imp} largest observations in terms of $\|\mathbf{X}\|$ among $n_{imp} = jk_{imp}$ are retained, a random forest is fitted with input variables $(\|\mathbf{X}\|, \mathbf{\Theta}_1, \dots, \mathbf{\Theta}_d)$, and the Gini and Permutation scores are computed. Figure 8.1 shows the average scores obtained over 10 independent experiments, together with interquantile ranges, as a function of the full sample size n_{imp}. In both models, the decrease of both scores is obvious. In particular in terms of Gini measure, the relative importance of the radius decreases from 38% to 1% for the additive model and from 6% to < 1% for the multiplicative model.



Additive model Multiplicative model

Figure 8.1: Average permutation and Gini importance measures of the radial variable using the RF algorithm in the additive noise model (left) and the multiplicative noise model (right) over 10 replications, as a function of the total sample size n_{imp} for fixed extreme training size k_{imp}. Solid black line: average Gini importance. Solid grey line: average Permutation importance. Dashed lines: empirical 0.8-interquantile ranges.

8.3.2 Real data

Encouraged by this first agreement between theoretical and numerical results, experiments on real data are conducted. We place ourselves in the setting of Example 7.10 where the target is one particular variable in a multivariate regularly varying random vector. We consider a financial dataset, namely *49 Industry Portfolios [Daily]* from Kenneth R. French - Data Library (https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). A study of extremal clustering properties within this dataset has already been carried out by Meyer and Wintenberger (2023). This dataset comprises daily returns of 49 industry portfolios, within the time span from January 5th, 1970 to October 31st, 2023. Rows containing any NA values are removed, resulting in a dataset of dimension $d = 49$ and size $n = 13577$. Figure 8.2 displays a Hill plot of the radial variable (w.r.t. $\|\cdot\|_2$), with a rather wide stability region, roughly between $k = 500$ and $k = 2000$, which suggests that RV is indeed present, with RV index $\alpha \approx 3.2$. We consider separately the first three variables as output (target) variables, namely *Agric* (*i.e.*, "Agriculture"), *Food* (*i.e.*, "Food Products"), and *Soda* (*i.e.*, "Candy and Soda"). Each choice of a target variable yields a regression problem consisting of predicting the target based on a covariate vector of dimension $d = 10$ composed of the 10 variables $(\tilde{X}_1, \dots, \tilde{X}_d)$ which are the most correlated with the target $\tilde{X}_{d+1}$. Following the workflow of Proposition 7.10, as an intermediate step, Algorithm 7.1 is used to predict $Y = X_{d+1}/\|\tilde{\mathbf{X}}\|$ where $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_{d+1})$. The output $\hat{Y}$ of Algorithm 7.1 is then plugged in the formula $\hat{X}_{d+1} = Y\|\mathbf{X}\|/\sqrt{1 - Y^2}$ where $\mathbf{X} = (X_1, \dots, X_d)$, which yields an

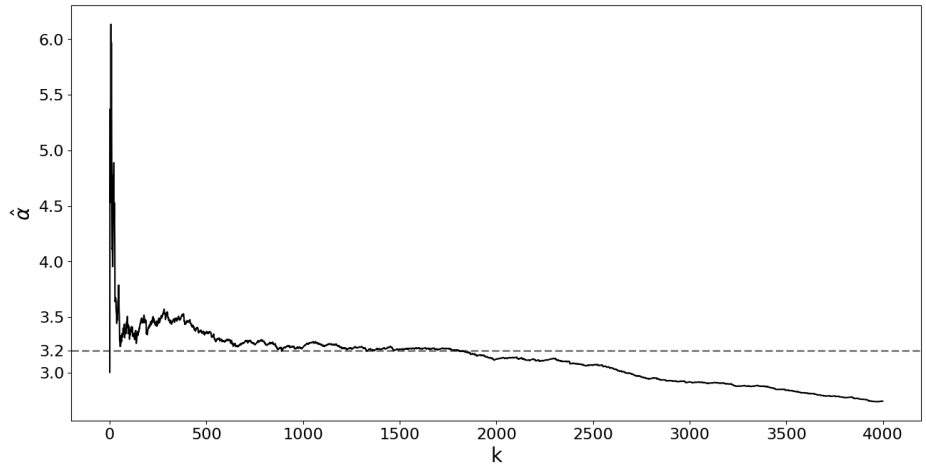


Figure 8.2: Hill plot for the radial variable of the 49 Industry Portfolio Daily dataset: estimation of the extreme value index $\gamma = 1/\alpha$ with the Hill estimator using the k largest order statistics of $\|\mathbf{X}\|$, as a function of k.

estimate $\hat{X}_{d+1}$ for the target variable. The dataset is randomly split into a test set of size $n_{test} = 4073$ (30% of the data), and a train set of size $n_{train} = 9504 = n - n_{test}$. As suggested by the Hill plot (Figure 8.2), the number k_{train} of extreme observations used at the training step is set to $k_{train} = \lfloor n_{train}/5 \rfloor = 1900$. On the other hand, at the testing step, to evaluate the extrapolation performance of our method, we fix k_{test} to a smaller fraction of the test set, $k_{test} = \lfloor n_{test}/10 \rfloor = 407$. In this setting, paralleling our experiments with simulated data, we compare the performance of regression functions learned using the full training dataset, the truncated version composed of the the k_{train} largest observations and the angles of the truncated version. For the sake of realism, we report the MSE regarding prediction of the original target variable $\tilde{X}_{d+1}$, *i.e.*, $(\tilde{X}_{d+1} - \hat{X}_{d+1})^2$, which would be of greater interest in applications than the error in the transformed variable $(Y - \hat{Y})^2$. Notice that our theory provides guarantees for the latter, not the former. The results gathered in Table 8.2 are the average MSE's obtained when repeating 10 times the procedure described above with random splits of the dataset into a train and a test set. These results provide evidence that conditionally on the other (covariate) variables being large, our method ensures, in most cases, better reconstruction of the target variable than the default strategy (first column) and the intermediate strategy (second column). For predicting the *Soda* variable however, the default strategy with OLS obtains the best scores. This suggests that convergence of the conditional distribution of excesses towards its limit as in (2.5) is somewhat slower for the subvector $(\tilde{X}_1, \ldots, \tilde{X}_{d+1})$ where $\tilde{X}_{d+1}$ is *Soda* and $\tilde{X}_1, \ldots, \tilde{X}_d$ are the 10 selected variables based on their correlation with *Soda*. This intuition is confirmed by the graphs of variable importance displayed in Figure 8.3, again paralleling the ones of Figure 8.1 and fully described in Section 8.3.1. In Figure 8.3, for simplicity, the importance scores are computed in a prediction task where the covariate vector includes all the available variables, except from the target (48 of them). Also the target variable for the RF algorithm is the rescaled variable $Y = \tilde{X}_{d+1}/\|\tilde{\mathbf{X}}\|$. Whereas the radial importances decreases monotonically when the target variable in *Agric* and *Food*, the third panel dedicated to the target variable *Soda* displays a local maximum in radial importance

Methods/Models	Train on $\mathbf{X}$	Train on $\mathbf{X}$ | $\|\|\mathbf{X}\|\|$ large	Train on Θ | $\|\|\mathbf{X}\|\|$ large
Agric: OLS	3.30±0.47	3.26±0.47	**3.25±0.44**
SVR	4.76±0.56	3.98±0.51	**3.74±0.50**
RF	3.47±0.47	3.48±0.47	**3.28±0.52**
Food: OLS	0.69±0.087	**0.678±0.082**	0.680±0.085
SVR	1.8±0.4	1.3±0.4	**0.87±0.08**
RF	0.70±0.13	0.72±0.12	**0.63±0.08**
Soda: OLS	**2.35±0.21**	2.37±0.21	2.42±0.21
SVR	4.0±0.5	3.1±0.5	**2.8±0.2**
RF	2.46±0.28	2.46±0.25	**2.34±0.18**

Table 8.2: Average MSE (and standard deviation) for predictive functions learned using all observations, extremes (20%) and angles of the extreme observations with output variables *Agric* over 10 random splits of each dataset.

around $n = 11\,000$. This value corresponds to a ratio $k/n \approx 0.12$ which is near the ratio 1/10 considered for the testing step in our experimental results reported in Table 8.2. This may explain our comparatively poor results for this particular variable. However for all three target variables, overall, both Gini and Permutation importance score decrease significantly, as the ratio k/n decreases. In particular for Gini importance, the relative radial importances are approximately $2\% \approx 1/48$ when $n = k$, which is to be expected when all variables have equal importance. On the other hand when $n = 10k$, all three Gini importances are less than 1%.

8.4 Conclusion

We have provided a sound ERM approach to the generic problem of statistical regression on extreme values. The asymptotic framework we have developed crucially relies on the (novel) notion of *joint regular variation* w.r.t. some multivariate component. When the distribution of the couple $(\mathbf{X}, Y)$ is regularly varying w.r.t. $\mathbf{X}$'s component, the problem can be stated and analyzed in a rigorous manner. We have described the optimal solution and proved that it can be nearly recovered with non-asymptotic guarantees by implementing a variant of the ERM principle, based on the angular information carried by a fraction of the largest observations only. We have also carried out numerical experiments to support the approach promoted, highlighting the necessity of using a dedicated methodology to perform regression on extreme samples with guarantees. Once validated, the ROXANE procedure is used in the next chapter in a detailed applied study to obtain predictions of extreme sea levels at a site given nearby extreme values.

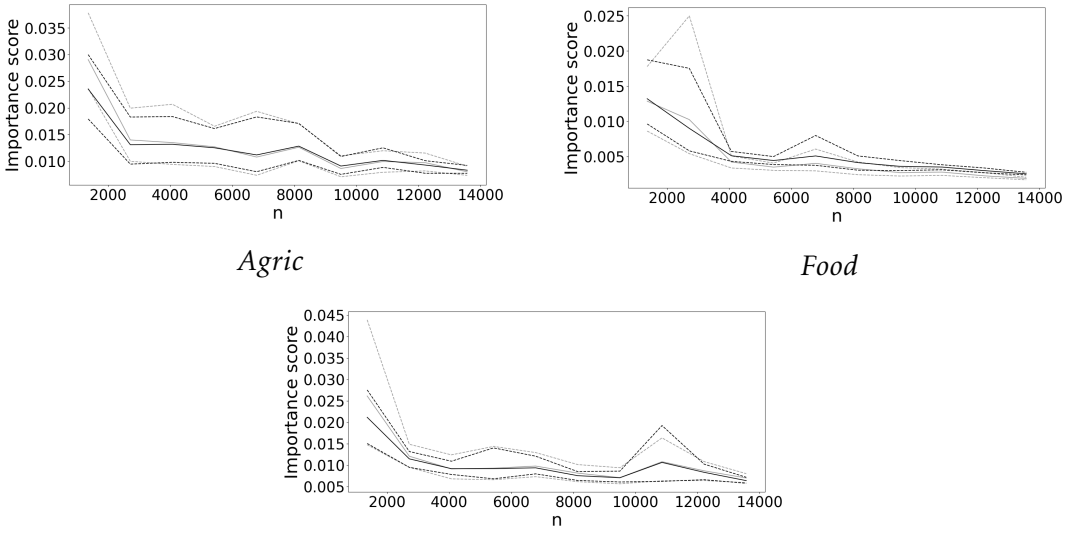


Figure 8.3: Average permutation and Gini importance measures of the radial variable for predicting *Agric* (top left), *Food* (top right) and *Soda* (bottom) variables using the RF over 10 randomly shuffled datasets. At each measurement, 1357 extreme observations are selected from a dataset whose total size increases from 1357 to 13570 with increments of 1357. Solid black line: average Gini importance. Solid grey line: average Permutation importance. Dashed lines: empirical 0.8-interquantile ranges.

8.A Proofs

Proof of Theorem 8.2.

(i) In view of Characterization (iii) from Theorem 7.11 (see also (7.4) in the main paper), Assumption 7.2 implies that the conditional distribution

$$\mathcal{L}(\boldsymbol{\Theta}, Y, \|\mathbf{X}\|/t \mid \|\mathbf{X}\| > t)$$

converges weakly to the distribution of $(\boldsymbol{\Theta}_\infty, Y_\infty, \|\mathbf{X}_\infty\|)$. Now if $f = h \circ \boldsymbol{\theta}$ is a prediction function on $\mathbb{R}^d$, where h is a continuous function defined on $\mathbb{S}$, then by compactness of $\mathbb{S}$ the function $(\boldsymbol{\omega}, y) \mapsto (h(\boldsymbol{\omega}) - y)^2$ is automatically bounded and continuous on the domain $\mathbb{S} \times [-M, M]$. Thus by weak convergence we obtain as $t \to +\infty$,

$$R_t(f) = \mathbb{E}\Big[(h(\boldsymbol{\Theta}) - Y)^2 \,|\, \|\mathbf{X}\| > t\Big] \to \mathbb{E}\Big[(h(\boldsymbol{\Theta}_\infty) - Y_\infty)^2\Big] = R_{P_\infty}(f).$$

(ii) Recall that $R_t^* = R_t(f^*)$ where f^* is the regression function for the pair $(\mathbf{X}, Y)$ and $R_{P_\infty}^* = R_{P_\infty}(f_{P_\infty}^*)$ where $f_{P_\infty}^*$ is the regression function for the pair $(\mathbf{X}_\infty, Y_\infty)$ defined in Lemma 8.1. Now we decompose R_t^* as

$$\begin{aligned} R_t^* &= \mathbb{E}\Big[(Y - f^*(\mathbf{X}))^2 \,|\, \|\mathbf{X}\| \geq t\Big] \\ &= \underbrace{\mathbb{E}\Big[(Y - f_{P_\infty}^*(\mathbf{X}))^2 \,|\, \|\mathbf{X}\| \geq t\Big]}_{A_t} + \underbrace{\mathbb{E}\Big[(f_{P_\infty}^*(\mathbf{X}) - f^*(\mathbf{X}))^2 \,|\, \|\mathbf{X}\| \geq t\Big]}_{B_t} + \dots \\ &\quad \dots \underbrace{2\mathbb{E}\Big[(Y - f_{P_\infty}^*(\mathbf{X}))(f_{P_\infty}^*(\mathbf{X}) - f^*(\mathbf{X})) \,|\, \|\mathbf{X}\| \geq t\Big]}_{C_t}. \end{aligned}$$

The first term A_t is simply $R_t(f_{P_\infty}^*)$. From Lemma 8.1, $f_{P_\infty}^*$ is an angular function, thus Property (i) of the statement implies that $A_t \to R_{P_\infty}(f_{P_\infty}^*)$, which is $R_{P_\infty}^*$.

We now show that the second and third terms B_t, C_t vanish. We use that, as a consequence of Assumption 7.1, $\forall \mathbf{x} \in \mathbb{R}^d$, $|f_{P_\infty}^*(\mathbf{x})| \leq M$ and, $|f^*(\mathbf{x})| \leq M$. Thus

$$B_t \leq 4M^2 \mathbb{E}\Big[|f_{P_\infty}^*(\mathbf{X}) - f^*(\mathbf{X})| \,|\, \|\mathbf{X}\| \geq t\Big].$$

Assumption 7.5 ensures that the latter display converges to 0 as $t \to \infty$. Similarly, using Assumptions 7.1 and 7.5 again, we obtain

$$|C_t| \leq 4M^2 \mathbb{E}\Big[|f_{P_\infty}^*(\mathbf{X}) - f^*(\mathbf{X})| \,|\, \|\mathbf{X}\| \geq t\Big] \underset{t \to +\infty}{\longrightarrow} 0.$$

We have proved that $R_t^* \underset{t \to +\infty}{\longrightarrow} R_{P_\infty}^*$.

(iii) Recall from the introduction that $R_\infty^* = R_\infty(f^*) = \limsup_t R_t(f^*)$. Because of (ii), in fact $R_t(f^*)$ converges to $R_{P_\infty}^*$. Thus

$$\limsup_t R_t(f^*) = \lim_t R_t(f^*) = R_{P_\infty}^*,$$

and the result follows.

(iv) From Property (iii) of the statement, we have $R^*_\infty = R_{P_\infty}(f^*_{P_\infty})$. Now, Property (i) of the statement and the angular nature of $f^*_{P_\infty}$ (Lemma 8.1) imply that $R_{P_\infty}(f^*_{P_\infty}) = R_\infty(f^*_{P_\infty})$. ■

Proof of Theorem 8.4. The key ingredient of the proof of Theorem 8.4 is a Bernstein-type inequality due to McDiarmid (1998) which is recalled in Section 4.2, Lemma 4.6.

Introduce an intermediate risk functional

$$\tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) = \frac{1}{k} \sum_{i=1}^{n} (h(\boldsymbol{\theta}(\mathbb{X}_i)) - Y_i)^2 \mathbb{1}\{\|\mathbb{X}_i\| \geq t_{n,k}\},$$

and notice that $\mathbb{E}[\tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta})] = R_{t_{n,k}}(h \circ \boldsymbol{\theta})$. Our proof is based on the following risk decomposition,

$$\begin{aligned} \sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \leq & \sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| + \dots \\ & \sup_{h \in \mathcal{H}} \left| \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right|. \end{aligned} \tag{8.4}$$

Regarding the first term on the right-hand side of Inequality (8.4),

$$\begin{aligned} & \sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \\ & \quad = \sup_{h \in \mathcal{H}} \frac{1}{k} \left| \sum_{i=1}^{n} \left(h \circ \boldsymbol{\theta}(\mathbb{X}_i) - Y_i \right)^2 \left(\mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\} - \mathbb{1}\{\|\mathbf{X}_i\| \geq \|\mathbf{X}_{(k)}\|\} \right) \right| \\ & \quad \leq \frac{4M^2}{k} \sum_{i=1}^{n} \left| \mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\} - \mathbb{1}\{\|\mathbf{X}_i\| \geq \|\mathbf{X}_{(k)}\|\} \right|. \end{aligned}$$

The number of nonzero terms inside the sum in the above display is the number of indices i such that ' $\|\mathbf{X}_i\| < \|\mathbf{X}_{(k)}\|$ and $\|\mathbf{X}_i\| \geq t_{n,k}$', or the other way around. In other words

$$\begin{aligned} & \left\{ \left| \mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\} - \mathbb{1}\{\|\mathbf{X}_i\| \geq \|\mathbf{X}_{(k)}\|\} \right| \neq 0 \right\} \subset \\ & \qquad \left(\left\{ t_{n,k} \leq \|\mathbf{X}_i\| < \|\mathbf{X}_{(k)}\| \right\} \cup \left\{ \|\mathbf{X}_{(k)}\| \leq \|\mathbf{X}_i\| < t_{n,k} \right\} \right). \end{aligned}$$

Considering separately the cases where $\|\mathbf{X}_{(k)}\| \leq t_{n,k}$ and $\|\mathbf{X}_{(k)}\| > t_{n,k}$ we obtain

$$\sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \leq \frac{4M^2}{k} \left| \sum_{i=1}^{n} \mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\} - k \right|.$$

Notice that $\sum_{i=1}^{n} \mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\}$ follows a Binomial distribution with parameters $(n, \frac{k}{n})$. The (classical) Bernstein inequality as stated, *e.g.*, in McDiarmid (1998), Theorem 2.7, yields

$$\begin{aligned} \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \geq \varepsilon \right) & \leq \mathbb{P}\left(\left| \sum_{i=1}^{n} \mathbb{1}\{\|\mathbf{X}_i\| \geq t_{n,k}\} - k \right| \geq k\varepsilon/(4M^2) \right) \\ & \leq 2 \exp\left(\frac{-k\varepsilon^2}{32M^4 + 8M^2\varepsilon/3} \right). \end{aligned} \tag{8.5}$$

We now turn to the second term of Inequality (8.4), and we apply Lemma 4.6 to the function

$$f((\mathbb{x}_1,y_1),\ldots,(\mathbb{x}_n,y_n)) = \sup_{h\in\mathcal{H}}\left|\frac{1}{k}\sum_{i=1}^{n}\left(h\circ\boldsymbol{\theta}(\mathbb{x}_i)-y_i\right)^2\mathbb{1}\{\|\mathbf{x}_i\|\geq t_{n,k}\}-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|,$$

so that $f((\mathbf{X}_1,Y_1),\ldots,(\mathbf{X}_n,Y_n)) = \sup_{h\in\mathcal{H}}\left|\tilde{R}_{t_{n,k}}(h\circ\boldsymbol{\theta})-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|$. With the notations of Lemma 4.6, the maximum of the positive deviations and the maximum sum of variances satisfy respectively $b\leq 4M^2/k$ and $\hat{v}\leq 16M^4/k$. Thus

$$\begin{aligned}\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left|\tilde{R}_{t_{n,k}}(h\circ\boldsymbol{\theta})-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|-\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\tilde{R}_{t_{n,k}}(h\circ\boldsymbol{\theta})-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|\right]\geq\varepsilon\right)\\ \leq\exp\left(\frac{-k\varepsilon^2}{32M^4+8M^2\varepsilon/3}\right).\end{aligned}\tag{8.6}$$

The last step consists in bounding from above the expected deviations in the above display, that is

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\tilde{R}_{t_{n,k}}(h\circ\boldsymbol{\theta})-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|\right].$$

Let $\varepsilon_1,\ldots,\varepsilon_n$ be n independent, $\{0,1\}$-valued Rademacher random variables and introduce the Rademacher average

$$\mathcal{R}_k^{\varepsilon} = \sup_{h\in\mathcal{H}}\frac{1}{k}\left|\sum_{i=1}^{n}\varepsilon_i(h\circ\boldsymbol{\theta}(\mathbf{X}_i)-Y_i)^2\mathbb{1}\{\|\mathbf{X}_i\|\geq t_{n,k}\}\right|.$$

Following a standard symmetrization argument as, *e.g.*, in the proof of Lemma 13 in Goix et al. (2015), we obtain

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\tilde{R}_{t_{n,k}}(h\circ\boldsymbol{\theta})-R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right|\right]\leq 2\mathbb{E}\left[\mathcal{R}_k^{\varepsilon}\right].\tag{8.7}$$

Let $(\mathbf{X}_1^k,Y_1^k),\ldots,(\mathbf{X}_n^k,Y_n^k)$ be independent replicates, also independent from the $\mathbf{X}_i,Y_i$'s, such that $\mathcal{L}(\mathbf{X}_i^k,Y_i^k)=\mathcal{L}\big((\mathbf{X},Y)\,|\,\|\mathbf{X}\|\geq t_{n,k}\big)$. By Lemma 2.1 of Lhaut et al. (2022), we have

$$\sum_{i=1}^{n}\varepsilon_i(h\circ\boldsymbol{\theta}(\mathbf{X}_i)-Y_i)^2\mathbb{1}\{\|\mathbf{X}_i\|\geq t_{n,k}\}\overset{d}{=}\sum_{i=1}^{\mathcal{K}}\varepsilon_i(h\circ\boldsymbol{\theta}(\mathbf{X}_i^k)-Y_i^k)^2,$$

where $\mathcal{K}\sim Bin(n,k/n)$ is independent from the $\varepsilon_i,\mathbf{X}_i,Y_i$'s. Then, write

$$\mathbb{E}\left[\mathcal{R}_k^{\varepsilon}\right]=\frac{1}{k}\mathbb{E}\left[\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{\mathcal{K}}\varepsilon_i(h\circ\boldsymbol{\theta}(\mathbf{X}_i^k)-Y_i^k)^2\right|\,\middle|\,\mathcal{K}\right]\right].\tag{8.8}$$

We first control the conditional expectation in the above display for any fixed value $\mathcal{K}=m\leq n$. For this purpose, we apply Proposition 2.1 of Giné and Guillou (2001) to the class of functions $\mathcal{G}=\{g(\mathbf{x},y)=(h\circ\boldsymbol{\theta}(\mathbf{x})-y)^2, h\in\mathcal{H}\}$.

Notice first that for $g_i(\mathbf{x}, y) = (h_i \circ \boldsymbol{\theta}(\mathbf{x}) - y)^2, i = 1, 2$ and Q any probability measure on $\mathbb{R}^d \times [-M, M]$ we have

$$\begin{aligned}\|g_1 - g_2\|_{L^2(Q)} &= \sqrt{\mathbb{E}_Q[(h_1 \circ \boldsymbol{\theta}(\mathbf{X}) - h_2 \circ \boldsymbol{\theta}(\mathbf{X}))(h_1 \circ \boldsymbol{\theta}(\mathbf{X}) + h_2 \circ \boldsymbol{\theta}(\mathbf{X}) - 2Y)]^2} \\ &\leq 4M \|h_1 - h_2\|_{L^2(Q_X \circ \boldsymbol{\theta}^{-1})},\end{aligned}$$

where Q_X is the marginal distribution of Q regarding the first component $X \in \mathbb{R}^d$. Thus the covering number $\mathcal{N}(\mathcal{G}, L_2(Q), \tau)$ for the class $\mathcal{G}$, relative to any $L_2(Q)$ radius τ is always less than than $\mathcal{N}(\mathcal{H}, L_2(\tilde{Q}), \tau/(4M))$ for the class $\mathcal{H}$, where $\tilde{Q} = Q_X \circ \theta^{-1}$. Now the class $\mathcal{H}$ has envelope function $H = M \mathbb{1}_{\mathbb{S}}(\cdot)$ and has VC-dimension $V_{\mathcal{H}} < +\infty$, thus Theorem 2.6.7 in van der Vaart and Wellner (1996) yields a control of its covering number,

$$\mathcal{N}(\mathcal{H}, L_2(\tilde{Q}), \tau M) \leq (A/\tau)^{2V_{\mathcal{H}}}$$

for some universal constant $A > 0$ not depending on $\tilde{Q}$ nor $\mathcal{H}$. We obtain

$$\mathcal{N}(\mathcal{G}, L_2(Q), \tau) \leq (4AM^2/\tau)^{2V_{\mathcal{H}}}.$$

Now $\mathcal{G}$ has envelope function $G = 4M^2 \mathbb{1}_{\mathbb{R}^d \times \mathbb{S}}$. The previous display writes equivalently

$$\mathcal{N}(\mathcal{G}, L_2(Q), \tau \|G\|_{L^2(Q)}) \leq (A/\tau)^{2V_{\mathcal{H}}}. \tag{8.9}$$

Inequality (8.9) is precisely the first step of the proof of Proposition 2.1 in Giné and Guillou (2001) (see Inequality 2.2 in the cited references), so that their upper bound on the Rademacher process applies with VC constant $v = 2V_{\mathcal{H}}$. The upper bound of their statement involves $\sigma^2 = \sup_g \mathbb{E}g^2 \leq 16M^4$ and $U = \sup_g \|g\|_\infty \leq 4M^2$, thus we may take $\sigma = U = 4M^2$. We obtain

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left| \sum_{i=1}^{m} \varepsilon_i (h \circ \boldsymbol{\theta}(\mathbf{X}_i^k) - Y_i^k)^2 \right| \right] \leq C'(4M^2 V_{\mathcal{H}} + \sqrt{m V_{\mathcal{H}}}),$$

for some other universal constant C'. Injecting the latter control into (8.8) yields, using the concavity of the squared root function and $\mathbb{E}[\mathcal{K}] = k$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_k^\varepsilon\right] &\leq \frac{1}{k} C'(4M^2 V_{\mathcal{H}} + \mathbb{E}[\sqrt{\mathcal{K}}] \sqrt{V_{\mathcal{H}}}) \\ &\leq \frac{1}{k} C'(4M^2 V_{\mathcal{H}} + \sqrt{k} \sqrt{V_{\mathcal{H}}}).\end{aligned} \tag{8.10}$$

Combining (8.7) and (8.10) we obtain

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left| \tilde{R}_{t_{n,k}}(h \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \right] \leq 2\mathbb{E}\left[\mathcal{R}_k^\varepsilon\right] \leq C\left(\frac{4M^2 V_{\mathcal{H}}}{k} + \sqrt{\frac{V_{\mathcal{H}}}{k}}\right), \tag{8.11}$$

with $C = 2C'$. Finally, combining Equations (8.5), (8.6) and (8.11) yields

$$\begin{aligned}&\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| \hat{R}_k(h \circ \boldsymbol{\theta}) - R_{t_{n,k}}(h \circ \boldsymbol{\theta}) \right| \geq \varepsilon + C\left(\frac{4M^2 V_{\mathcal{H}}}{k} + \sqrt{\frac{V_{\mathcal{H}}}{k}}\right)\right) \\ &\qquad \leq 3 \exp\left(\frac{-k\varepsilon^2}{16(8M^4 + M^2\varepsilon/3)}\right),\end{aligned}$$

which concludes the proof after solving for $3\exp\left(\frac{-k\varepsilon^2}{16(8M^4+M^2\varepsilon/3)}\right) = \delta$. ∎

Proof of Proposition 8.5.

1. For $t \geq 1$ and $h \in \mathcal{H}$, write $r_t(h) = R_t(h \circ \boldsymbol{\theta})$. For all $h_1, h_2 \in \mathcal{H}$, and $t \geq 1$, we have

$$\begin{aligned}
&|r_t(h_1) - r_t(h_2)| = |R_t(h_1 \circ \boldsymbol{\theta}) - R_t(h_2 \circ \boldsymbol{\theta})| \\
&= \left| \mathbb{E}\left[h_1(\mathbf{X})^2 - h_2(\mathbf{X})^2 + 2Y(h_1(\mathbf{X}) - h_2(\mathbf{X})) \mid \|\mathbf{X}\| \geq t \right] \right| \\
&\leq \mathbb{E}\Big[|(h_1(\mathbf{X}) + h_2(\mathbf{X}))(h_1(\mathbf{X}) - h_2(\mathbf{X}))| \mid \|\mathbf{X}\| \geq t \Big] + \cdots \\
&\quad \cdots 2\mathbb{E}\Big[|Y(h_1(\mathbf{X}) - h_2(\mathbf{X}))| \mid \|\mathbf{X}\| \geq t \Big] \\
&\leq 4M \|h_1 - h_2\|_\infty, \qquad (8.12)
\end{aligned}$$

   where we have used Assumption 7.1 to obtain the last inequality. Similarly,

$$\begin{aligned}
&R_{P_\infty}(h_1 \circ \boldsymbol{\theta}) - R_{P_\infty}(h_2 \circ \boldsymbol{\theta}) \\
&\leq \mathbb{E}|(h_1(\boldsymbol{\Theta}_\infty) + h_2(\boldsymbol{\Theta}_\infty))(h_1(\boldsymbol{\Theta}_\infty) - h_2(\boldsymbol{\Theta}_\infty))| + \ldots \\
&\quad \ldots 2\mathbb{E}|Y_\infty(h_1(\boldsymbol{\Theta}_\infty) - h_2(\boldsymbol{\Theta}_\infty))| \\
&\leq 4M \|h_1 - h_2\|_\infty, \qquad (8.13)
\end{aligned}$$

   Let $\varepsilon > 0$. By total boundedness $\exists h_1, \ldots, h_L \in \mathcal{H}$ such that $\cup_{i=1,\ldots,L} B(h_i, \varepsilon) \supset \mathcal{H}$. Here $B(h, \varepsilon)$ denotes the ball of radius ε in $(\mathcal{C}(\mathbb{S}), \|\cdot\|)$. Now because of Assumption 7.2 (see Theorem 8.2, (i)) we have $r_t(h_i) \to R_{P_\infty}(h_i \circ \boldsymbol{\theta})$ as $t \to \infty$, for all fixed i. Thus there exists some $T > 0$ such that for all $i \in \{1, \ldots, L\}$ $|r_t(h_i) - R_{P_\infty}(h_i \circ \boldsymbol{\theta})| \leq \varepsilon$. Now for any $h \in \mathcal{H}$ and $t \geq T$, using (8.12) and (8.13) there exists $i \leq L$ such that

$$\max(|r_t(h) - r_t(h_i)|, |R_{P_\infty}(h \circ \boldsymbol{\theta}) - R_{P_\infty}(h_i \circ \boldsymbol{\theta})| \leq 4M\varepsilon,$$

   so that

$$\begin{aligned}
&|r_t(h) - R_{P_\infty}(h \circ \boldsymbol{\theta})| \\
&\leq |r_t(h) - r_t(h_i)| + |r_t(h_i) - R_{P_\infty}(h_i \circ \boldsymbol{\theta})| + [R_{P_\infty}(h_i \circ \boldsymbol{\theta}) - R_{P_\infty}(h \circ \boldsymbol{\theta})| \\
&\leq 8M\varepsilon + \varepsilon.
\end{aligned}$$

   Because $R_{P_\infty}(h \circ \boldsymbol{\theta}) = R_\infty(h \circ \boldsymbol{\theta})$ (Theorem 8.2-(i)), the proof is complete.

2. The VC-class property of $\mathcal{H}$ (Assumption 8.3) ensures that for any probability measure Q on $\mathbb{S}$, and any $\varepsilon > 0$, the covering number $\mathcal{N}(\varepsilon, \mathcal{H}, L_1(Q))$ is finite (see, *e.g.*, van der Vaart and Wellner (1996), Section 2.6.2). Our first step is to build such a probability measure Q which dominates both the $\Phi_{\theta,t}$'s and Φ_θ, in such a way that $\mathbb{E}\Big[|h_1 - h_2|(\boldsymbol{\Theta}) \,|\, \|\mathbf{X}\| > t\Big]$ and $\mathbb{E}\Big[|h_1 - h_2|(\boldsymbol{\Theta}_\infty)\Big]$ are both controlled by $\int_{\mathbb{S}} |h_1 - h_2| \mathrm{d}Q = \|h_1 - h_2\|_{L_1(Q)}$. Let $Q = \frac{1}{2}(\Phi_{\theta,1} + \Phi_\theta)$. Then Φ_θ is absolutely continuous with respect to Q, and so is each $\phi_t, t \geq 1$, in view of the discussion above the statement in the main paper.

   In addition we have $\sup_{\boldsymbol{\omega} \in \mathbb{S}} |\mathrm{d}\Phi_\theta / \mathrm{d}Q(\boldsymbol{\omega})| \leq 2$ and from Condition 2. also

$$\sup_{\boldsymbol{\omega} \in \mathbb{S}, t \geq 1} |\mathrm{d}\Phi_{\theta,t} / \mathrm{d}Q(\boldsymbol{\omega})| \leq 2D.$$

   For any h_1, h_2 in $\mathcal{H}$, following the argument leading to (8.12) we obtain

$$\begin{aligned}
\left| r_t(h_1) - r_t(h_2) \right| &\leq 4M \int_{\mathbb{S}} \left| h_1 - h_2 \right| \mathrm{d}\Phi_t \\
&\leq 8MD \int_{\mathbb{S}} \left| h_1 - h_2 \right| \mathrm{d}Q = 8MD \|h_1 - h_2\|_{L_1(Q)}.
\end{aligned}$$

Also,

$$
\begin{aligned}
&|R_{P_\infty}(h_1 \circ \boldsymbol{\theta}) - R_{P_\infty}(h_2 \circ \boldsymbol{\theta})| \\
&\leq \mathbb{E}|(h_1 g(\boldsymbol{\Theta}_\infty) + h_2(\boldsymbol{\Theta}_\infty))(h_1(\boldsymbol{\Theta}_\infty) - h_2(\boldsymbol{\Theta}_\infty))| + \dots \\
&\quad \dots 2\mathbb{E}|Y_\infty(h_1(\boldsymbol{\Theta}_\infty) - h_2(\boldsymbol{\Theta}_\infty))| \\
&\leq 4M\mathbb{E}\left[|h_1 - h_2|(\boldsymbol{\Theta}_\infty)\right] \\
&\leq 8M\|h_1 - h_2\|_{L_1(Q)}.
\end{aligned}
$$

Let $\varepsilon > 0$. Since the covering number of the class $\mathcal{H}$ for the $L_1(Q)$-norm is finite, for some $L \leq \mathcal{N}(\varepsilon, \mathcal{H}, L_1(Q))$, there exists $h_1, \dots, h_L \in \mathcal{H}$ such that each $h \in \mathcal{H}$ is at $L_1(Q)$-distance at most ε from one of the h_i's. The rest of the proof follows the same lines as the argument following (8.13), up to replacing the infinity norm with the $L_1(Q)$-norm on $\mathcal{H}$. ■

Part IV

Application: Extreme Sea Levels

Introduction

With the rise of sea levels due to global warming, it becomes increasingly important to model and predict extreme sea levels that can lead to flooding, such as the North Sea flood of 1953 (Figure 1.1, McRobie et al. (2005)). This study of extremes is particularly valuable for computing precise estimates of extreme return periods, which correspond to the average time between extreme events (Coles et al. (2001)). To infer these return periods, researchers may conduct direct studies on sea level modeling (Lennon et al. (1963); Suthons (1963)). However, a more common approach in the literature involves using the decomposition of sea levels into a deterministic tidal component and a stochastic surge component, then focusing on the study of extreme surges (Pugh and Vassie (1978, 1980); Tawn (1992)). Indirect methods based on this decomposition often face the challenge of modeling the tide-surge dependence structure (Idier et al. (2012)). To bypass this step, skew surges — the difference between the maximum observed sea level during a tide and the maximum predicted sea level for the same tide — are considered, as they are shown to be independent of the tidal component (Williams et al. (2016)). Additional details about modeling extreme sea levels are discussed in Section 1.2.3 and the references therein.

Figure 8.4: Photos taken during a visit of the SHOM, Brest, March 2024. Left: former Port-Tudy tide gauge. Right: current Brest tide house, with tide staff on the wall.

To avoid modeling the tide-surge interaction, this study focuses solely on sea levels and skew surges, and aims to capture the dependence of these quantities across different tide gauges. Modeling the multivariate extremal dependence structure is an important part of extreme value literature, particularly for spatial extremes (see Huser and Wadsworth (2022)). We propose to learn the extreme spatial dependence structure among observations from various stations over their common time range. This learned model is then utilized to reconstruct sea levels and skew surges at a site based on extreme values at nearby sites. Our primary objective is to provide practitioners in the field with statistical learning tools based on the latest advancements in the area of multivariate extremes. Moreover, this study serves as an opportunity to implement the ROXANE method (Part III) to solve a real-world problem and to compare its performance on this example with a method that is more akin to traditional parametric statistics than to nonparametric statistical learning.

In the first method, consistent with traditional extreme value analysis for sea levels, we fit an appropriate extreme distribution to the data. The sea level and skew surge data clearly exhibit asymptotic dependence. Additionally, an observation is declared extreme if at least one of its input components is extreme, since a single large sea level or skew surge can cause flooding at the recorded station, regardless of conditions at the other stations. Therefore, we model these extremes using a multivariate generalized Pareto distribution, which is particularly suited to this type of data (see Theorem 2.1 in Rootzén and Tajvidi (2006)). Specifically, we follow the parametric procedure of Kiriliouk et al. (2019) to fit a density to the data at low computational cost.

In the second approach, to avoid restrictive assumptions about the distribution of the observations, we use a predictive method based on a regression algorithm. Developing statistical learning methods in extreme settings is an important research subject nowadays (see Section 1.2). We use the regression procedure proposed in Part III (Huet et al. (2023)), which is designed for extreme value predictive problems where the extremality is measured w.r.t. covariates, which are the values at the long-term stations. Our goal is to learn a predictive function over the common time range of the data that predicts values at the time-limited stations based on extreme values at the output stations with a large number of records.

Chapter 9

Modeling and Prediction of Extreme Sea Levels

Contents

In this chapter, we present the results of two prediction procedures for extreme sea levels and skew surges. In the first approach, the multivariate data are modeled using a Multivariate Generalized Pareto (MGP) distribution, which is discussed at the end of Section 2.1.2. In a second time, we choose to explore the method proposed in Part III, namely the ROXANE algorithm. Specifically, we work within the framework of Proposition 7.10 "Predicting a missing component in a regularly varying random vector". It is important to note that, following the lines of Section 2.1.2, both models are valid in the same extreme framework, *i.e.*, under the classic multivariate maximum domain of attraction assumption (2.4).

Both procedures require rescaling of the marginal observations to a common scale: unit exponential scales for the MGP procedure and unit Pareto distributions for the ROXANE procedure.

For the MGP modeling, recall from Section 2.1.2 that if W is a MGP vector in $\mathbb{R}^d$ with parameters (σ, ξ), then the positive part of the margins of $\mathbf{W}$ are GP distribution

$$\mathbb{P}(W_j \geq x \mid W_j \geq 0) = H_{0,\sigma_j,\xi_j}(x) = \left(1 + \frac{\xi_j}{\sigma_j} x\right)_+^{-1/\xi_j},$$

for $1 \leq j \leq d$. Consequently, an appropriate marginal transformation involves setting $(\sigma_j, \xi_j) = (1, 0)$, for all $1 \leq j \leq d$, which transforms the positive part of the MGP vector into a unit exponential distribution (2.14). The parameters (σ_j, ξ_j) must be estimated, for $1 \leq j \leq d$, which requires modeling the right tail of the observations with a GP distribution.

Conversely, in the ROXANE routine, the entire marginal distribution needs to be modeled, since we want to apply to each margin the Pareto transformation given by

$$p(x) = \frac{1}{1 - F(x)},$$

where F is the cdf of the considered margin (2.9). Marginal observations from multivariate extremes may include non-extreme observations (*e.g.*, a storm affecting one input but not another). These residual non-extreme marginal observations belong to the left tail of the marginal distribution at each station (see, *e.g.*, Figure 9.2). Therefore, the GP distribution is not appropriate to model these data.

In this context, as in Legrand et al. (2023), we use an Extended Generalized Pareto (EGP) distribution as our marginal model to meet the requirement of modeling the whole distribution with GP behavior in the right-tail of the observations. Specifically, we consider the model EGPD3 from Papastathopoulos and Tawn (2013), whose cdf we recall from Section 2.1.2

$$F_{\sigma,\xi,\kappa}(x) = \left(1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}\right)^{\kappa}, \tag{9.1}$$

with $\sigma > 0$, $\xi \in \mathbb{R}$, $\kappa \in \mathbb{R}$ and $x \in [0, +\infty[$ if $\xi \geq 0$ and $x \in [0, -\sigma/\xi]$ otherwise.

The chapter is structured as follows. Section 9.1 introduces the sea level and skew surge datasets. In Section 9.2, we describe the marginal preprocessing and the threshold selection, along with the two prediction schemes detailed in two algorithms. We describe the results of these procedures applied on the sea level and skew surge dataset in Section 9.3. Two additional studies conducted on two other output stations are deferred to Appendix 9.A.

9.1 Sea Level Data

Our study focuses on the Atlantic French coast, utilizing data from tide gauges provided by the sea level observations network RONIM (Réseau d'Observation du Niveau de la Mer) SHOM, managed by the 'Service hydrographique et océanographique de la Marine' (SHOM). We specifically analyze two key variables: maximal observed sea levels and skew surges, defined as the differences between the maximal predicted sea levels and the maximal observed sea levels during a full tide. The dataset consists of hourly validated data, each associated with a timestamp. Our inference targets both sea levels and skew surges, with a particular focus on large sea levels, which are crucial for flood risk monitoring. Notably, large sea levels can be derived using convolution methods between maximal predicted sea levels and skew surges (see Haigh et al. (2010)).

Intuitively, predicting missing observations at a given station ideally requires at least two stations with long-range observations on each side of the target station. Indeed, it seems reasonable to suppose in this setting that most of the extreme values at the

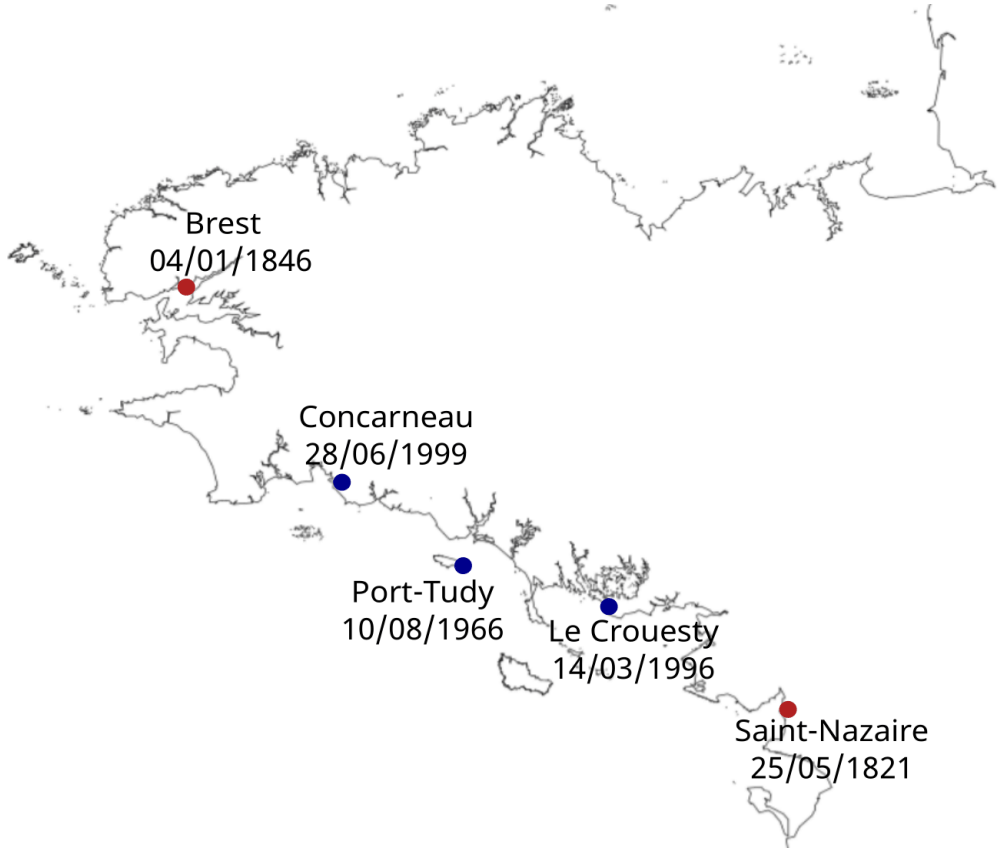


Figure 9.1: Map showing the locations and the temporal depths of each station. Red dots indicate input stations, while blue dots indicate output stations.

output station will be associated with at least one extreme value at a input station, while, if only one input station is considered, some extremes at the output station could be missed. Consequently, our study utilizes Brest and Saint-Nazaire as the two long-range input stations. Our objective is to predict and generate values at three stations with limited historical records, referred to as the output stations: Port-Tudy, Concarneau, and Le Crouesty. Consequently, three separate studies are conducted, each focusing on either maximum sea levels or skew surges. The locations and temporal depths of the five stations are showed in Figure 9.1.

Remark 9.1 (a bit of history)**.** *The dates displayed in Figure 9.1 represent the first available data on* `data.shom.fr`*. Systematic measurements of sea levels at Brest date back to the 17th century. January 4, 1846 marks the deployment of the first tide-gauge in France (and one of the first of the world). The tide gauge in Saint-Nazaire was deployed in 1863; measurements between 1821 and 1863 were not conducted regularly (see* `refmar.shom.fr` *for details).*

In the subsequent analysis, X_B and X_N represent sea levels or skew surges at the Brest and Saint-Nazaire stations, while Y represents sea levels or skew surges at an output station. Whenever necessary, we specify by an upper index sl a quantity related to sea levels and by an upper index ss a quantity related to skew surges.

To evaluate the performance of our method, each of the three datasets is divided into two subsets. The training set, comprising the most recent observations, is used to fit the marginal and joint models, while the test set, consisting of the earliest observations, is

used to assess the performance of the fitted model. The dates for splitting into training and test sets are 31 December 1999 for Port-Tudy, 2010 for Concarneau, and 2014 for Le Crouesty, ensuring that each set contains approximately the same amount of data.

In practice, the marginal observations of the output station are unknown and thus are not used to determine the extremality of an observation. Instead, an observation is considered extreme if at least one of the two input station values exceeds a large threshold. A first pre-selection is performed such that we retain only triplets (X_B, X_N, Y) satisfying

$$X_B \geq q_B^{0.5} \text{ or } X_N \geq q_N^{0.5}, \tag{9.2}$$

where q_B^ρ (resp. q_N^ρ) is the empirical ρ-quantile at Brest (resp. at Saint-Nazaire). This initial thresholding is distinct from the common extreme value theory problem of selecting an extreme threshold above which observations are considered extreme (addressed in Section 9.2.1). This pre-processing step aims solely to reduce the dataset size by considering only relevant observations, thereby avoiding computational burden. Note that the initial thresholding is set low enough to retain all observations of interest in the study (see the left columns of Figures 9.3 and 9.4). In the following discussions, when we mention the "data" we are referring to the data that remains after this first thresholding step, unless stated otherwise. In applying the three algorithms described below, we use only the data remaining after this thresholding step. Specifically, when we mention modeling "the entire marginal distributions" in the previous section, we are referring to modeling the marginal distributions of the retained data.

9.2 Methods

In this section, we outline the full details of the methodologies used in both approaches. Each aims to achieve accurate predictions with respect to a least-square criterion, specifically, to obtain values as close as possible to the expectation of Y given $\mathbf{X}$. In Section 9.2.1, we present the marginal modeling procedure and the method to choose a multivariate threshold above which a GP distribution is deemed an appropriate model for the marginal distributions, central in the MGP procedure. These two steps are common to both methods. In Section 9.2.2, we consider a plug-in method: we seek to model $(\mathbf{X}, Y)$ by fitting a density, selected from multiple candidates, to the data. An estimate of $\mathbb{E}[Y \mid \mathbf{X}]$ is then obtained by averaging generated data according to the associated conditional density given $\mathbf{X}$. In the second approach, detailed in Section 9.2.3, no modeling step is required; instead, the goal is to derive a predictive function $\hat{h}$ by minimizing an empirical risk of the form $\sum(Y_i - h(\mathbf{X}_i))^2$.

9.2.1 Univariate study and threshold selection

As previously discussed, we model the marginal distributions using EGP distributions (9.1). Given that the left-end point of the support of an EGP distribution is zero, it is necessary to relocate our data to set their minimum to zero before performing the marginal fittings. Regarding the selection of thresholds above which each margin is considered plausible for a Generalized Pareto (GP) distribution, it is important to note that the threshold is not necessarily constant and may vary from station to station. In this study, the threshold is determined independently for each margin based on fundamental observations concerning EGP and GP distributions. As stated in Naveau et al. (2016), the EGP distribution behaves similarly to a GP distribution for large

Algorithm 9.1 Marginal modeling and threshold selection.

INPUT: Training dataset $\mathcal{D}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ with $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ input/target pair.
Relocalisation: Shift the data in order to obtain nonnegative entries. Define

$$\mathbf{Z}_{i,m} = \mathbf{Z}_i - \mathbf{m}$$

with $\mathbf{m} = (m_1, ..., m_{d+1})$, with $m_j = \min_{1 \le i \le n} Z_{ij}$, for $1 \le j \le d+1$.
Marginal Fitting: Fit an EGP distribution to the respective margins of the $\mathbf{Z}_{i,m}$'s to obtain triplets of estimated parameters $(\boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\kappa}) := ((\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X), (\sigma_Y, \xi_Y, \kappa_Y)) \in \mathbb{R}^{3d+3}$.

Thresholds selection: Select the largest $c_{X,j} \in supp(F_{\sigma_{X,j}, \xi_{X,j}, \kappa_{X,j}})$ for $1 \le j \le d$ and $c_Y \in supp(F_{\sigma_Y, \xi_Y, \kappa_Y})$ such that

$$\frac{d^2 F_{\sigma_{X,j}, \xi_{X,j}, \kappa_{X,j}}(c_{X,j})}{dx^2} = 0,$$

$$\frac{d^2 F_{\sigma_Y, \xi_Y, \kappa_Y}(c_Y)}{dx^2} = 0$$

OUTPUT: estimated marginal EGP parameters $(\boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\kappa}) = ((\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X), (\sigma_Y, \xi_Y, \kappa_Y)) \in \mathbb{R}^{3d+3}$ and a multivariate threshold $\mathbf{t} = (\mathbf{t}_X, t_Y) = ((m_1 + c_{X,1}, ..., m_d + c_{X,d}), m_{d+1} + c_Y)$.

values. Since the GP density is a strictly convex function for $\xi > -1/2$ (a condition that is always met in this study), the EGP density is also strictly convex for sufficiently large values. Consequently, each marginal threshold is defined as the lowest point above which the fitted EGP density is convex. This point is among the zeros of $\frac{d^2 F_{\sigma,\xi,\kappa}(x)}{dx^2}$: if $\kappa < 2$, it corresponds to the unique zero; if $\kappa \ge 2$ it corresponds to the largest zero.

Remark 9.2 (Selection of thresholds and marginal models)**.** *In this paper, we propose modeling the margins using a specific type of EGP distribution, as it effectively fits our data. However, any marginal model that accurately fits the data can be employed. Thus, alternative marginal modeling approaches can be utilized without altering the multivariate procedures. For example, one could use any EGP family introduced in Naveau et al. (2016), or the conventional approach of modeling with a GP distribution above a preselected threshold and the empirical cdf below the threshold (as in, e.g., Heffernan and Tawn (2004)). Similarly, the threshold selection method can vary. It can be determined using stability plot diagnostics, as discussed in Kiriliouk et al. (2019); Legrand et al. (2023); Huet et al. (2023).*

9.2.2 Multivariate generalized Pareto procedure

In this section, we propose a method to model the extremal dependence of sea levels and skew surges between stations. The goal of this procedure is to predict values at the output stations based on the values at the Brest and Saint-Nazaire stations. To achieve this, we fit a density to the data and obtain predictions by averaging values generated from the corresponding conditional fitted density given the input values. This approach allows us to estimate the conditional expectation $\mathbb{E}[Y \mid \mathbf{X}]$. Additionally, an underlying outcome of this procedure is the ability to sample new observations from the joint fitted density for data augmentation purposes.

Algorithm 9.2 MGP predictive algorithm.

INPUT: Training dataset $\mathcal{D}_n = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$ with $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ input/target pair; fitted GP parameters $(\boldsymbol{\sigma}, \boldsymbol{\xi}) = ((\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X), (\sigma_Y, \xi_Y)) \in \mathbb{R}^{2d+2}$; a multivariate threshold $\mathbf{t} = (\mathbf{t}_X, t_Y)$; $\mathcal{H} = \{\mathcal{H}_1, ..., \mathcal{H}_N\}$ set of N classes of density functions.
Truncation: Form a set of extreme observations with shifted $\mathbf{Z}_{i,t}$'s

$$\mathbf{Z}_{i,t} = \mathbf{Z}_i - \mathbf{t}, \quad \text{for all } i \in I_{ext} := \Big\{ i \in \{1, ..., n\}, \mathbf{X}_i \nleq \mathbf{t}_X \Big\}.$$

Marginal Exponential Transformation: Apply the exponential transformation (9.3) to each margin of the extreme observations

$$\tilde{\mathbf{Z}}_i = (\tilde{\mathbf{X}}_i, \tilde{Y}_i) = e_{\boldsymbol{\sigma}, \boldsymbol{\xi}}(\mathbf{Z}_{i,t}), \quad \text{for all } i \in I_{ext}.$$

Density selection: Fit each density model to the data and select the density $\hat{h} \in \mathcal{H}$ with the smallest AIC.
OUTPUT: Near-optimal density function $\hat{h}$ in $\mathcal{H}$ and a procedure to be used for predictions of Y_{n+1} based on new observations $\mathbf{X}_{n+1}$ such that $\mathbf{X}_{n+1} \nleq \mathbf{t}_X$ so that

- Generate a sample $(\hat{\tilde{Y}}^1_{n+1}, ..., \hat{\tilde{Y}}^L_{n+1})$ *via* rejection sampling from the conditional density
$$\hat{h}_{|\tilde{\mathbf{X}}_{n+1}}(\tilde{y}) := \frac{\hat{h}(\tilde{\mathbf{X}}_{n+1}, \tilde{y})}{\int_{\mathbb{R}} \hat{h}(\tilde{\mathbf{X}}_{n+1}, s) ds}.$$
- Backtransform the sample *via*
$$(\hat{Y}^1_{n+1}, ..., \hat{Y}^L_{n+1}) = (e^{-1}_{\sigma_Y, \xi_Y}(\hat{\tilde{Y}}^1_{n+1}) + t_Y, ..., e^{-1}_{\sigma_Y, \xi_Y}(\hat{\tilde{Y}}^L_{n+1}) + t_Y).$$
- Obtain a prediction of Y_{n+1} by the Monte Carlo average $\hat{Y}_{n+1} = (1/L) \sum_{l=1}^{L} \hat{Y}^l_{n+1}$.

We outline the procedure in Algorithm 9.2, which produces an 'optimal' density that fits the data. This approach shares similarities with the one proposed in Kiriliouk et al. (2019), with some changes dictated by the nature of the considered data, summarized in the next paragraph. Marginal distribution study and threshold selection have to be performed beforehand *via* Algorithm 9.1 to obtain the GP parameters and the multivariate threshold. Note that, because an EGP distribution behaves as a GP distribution in the right tail, we choose as GP parameters, the parameters (σ, ξ) obtained from the fitted EGP in Algorithm 9.1. From Equation (2.13), recall that the marginal transformation to exponential scale is given by

$$e_{\boldsymbol{\sigma}, \boldsymbol{\xi}}(\mathbf{x}) = \Big(-\log(1 - H_{0, \sigma_{X,1}, \xi_{X,1}}(x)), ..., -\log(1 - H_{0, \sigma_{X,d}, \xi_{X,d}}(x)), -\log(1 - H_{0, \sigma_Y, \xi_Y}(x)) \Big), \quad (9.3)$$

where $(\boldsymbol{\sigma}, \boldsymbol{\xi}) = ((\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X), (\sigma_Y, \xi_Y))$ and $H_{0,\sigma,\xi}(x) = (1 + \xi x/\sigma)^{-1/\xi}_+$ is the cdf of a GP distribution with parameters $\mu = 0, \sigma$ and ξ (2.3).

The advantage of this procedure lies in the availability of a wide range of suitable density models to fit standard MGP observations, *i.e.*, observations after applying the transformation $e_{\sigma,\xi}$. Indeed, as outlined at the end of Section 2.1.2, a standard MGP

vector $\tilde{\mathbf{W}}$ decomposes as $\tilde{\mathbf{W}} = E + \mathbf{T} - \max(\mathbf{T})$ (2.14). Hence, for any density $f_T : \mathbb{R}^d \to \mathbb{R}$ of $\mathbf{T}$, Theorem 2.14 allows one to deduce a density h_T for $\tilde{\mathbf{W}}$ given by

$$h_T(\mathbf{x}) = \mathbb{1}\{\max(\mathbf{x}) > 0\} \exp(-\max(\mathbf{x})) \int_0^{+\infty} \frac{f_T(\mathbf{x} + \log t)}{t} dt.$$

In other words, for any density $f_T : \mathbb{R}^d \to \mathbb{R}$, there exists a candidate density to fit the transformed data. Additionally, Theorem 2.15 provides another method to construct densities for standard MGP vectors.

As in Kiriliouk et al. (2019), a censored likelihood approach is employed in Algorithm 9.2 rather than the classical likelihood approach: only triplets with three positive observations are considered when fitting the density. Up to the estimation of a density $\hat{h} \in \mathcal{H}$, our procedure mirrors that in Kiriliouk et al. (2019). However, after this step, no further fitting is conducted in our procedure. In contrast, Kiriliouk et al. (2019) computes an appropriate density in the original scale by simultaneously fitting the parameters of the retained density model and the GP marginal parameters. The latter step is computationally intensive: in our case, it involves fitting between 7 to 11 parameters on over 3000 observations and results in inferior performance compared with the procedure described in Algorithm 9.2.

9.2.3 Angular regression procedure

In this section, we discuss the ROXANE procedure and summarize it in Algorithm 9.3. Recall that the distribution of $\mathbf{Z}$ is modeled by an EGP distribution with cdf $F_{\sigma,\xi,\kappa}$ given in Equation (9.1). As suggested in Section 2.1.2, it is convenient in this settings to work with Pareto marginal distributions. In the general case, the marginal Pareto transformation is given in Equation (2.9). Here, because the marginal distributions are modeled by EGP distributions, this marginal transformation corresponds to

$$p_{\sigma,\xi,\kappa}(x) = \frac{1}{1 - F_{\sigma,\xi,\kappa}(x)}. \tag{9.4}$$

The complex form of the prediction function $\hat{g}$ in Equation (9.6) arises from the various pre-processing steps performed in the procedure. These steps are: shifting the input variable subtracting $\mathbf{m}_X$; applying the marginal Pareto transformation by $p_{\boldsymbol{\sigma}_X,\boldsymbol{\xi}_X,\boldsymbol{\kappa}_X}$; applying the angular input transformation $x \mapsto x/\|x\|_r$; making predictions in angular Pareto scale $\hat{h}$; backtransforming from angular Pareto scale to Pareto scale *via* $y \mapsto y\|x\|_r/(1-y^r)^{1/r}$; and finally, backtransforming from Pareto scale to original scale by $p^{-1}_{\sigma_Y,\xi_Y,\kappa_Y}(\cdot) + m_Y$.

Remark 9.3 (Role of the EGP distribution)**.** *This remark provides a summary of the roles played by Algorithm 9.1 and the EGP distribution in the prediction procedures discussed.*

For the MGP approach, Algorithm 9.1 is utilized to determine a multivariate threshold $\mathbf{t} = (\mathbf{t}_X, t_Y)$ *above which the GP distribution is deemed an appropriate model for the marginal distributions. In other words, the MGP distribution serves as a suitable joint model for observations* $\mathbf{X}$ *conditional on the event* $\mathbf{X} \not\leqslant \mathbf{t}X$*. Additionally, because the EGP distribution behave as a GP distribution in the right tail, the parameters* σ *and* ξ *from the fitted EGP model are also used as the GP parameters for the marginal distributions above the threshold* $\mathbf{t}$*. Thus, each margin of the data is transformed to an exponential scale using*

$$e_{\sigma,\xi}(x) = -\log(1 - H_{0,\sigma,\xi}(x)),$$

Algorithm 9.3 ROXANE regression algorithm.

INPUT: Training dataset $\mathcal{D}_n = \{\mathbf{Z}_1, ..., \mathbf{Z}_n\}$ with $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ input/target pair; fitted EGP parameters $(\boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\kappa}) = ((\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X), (\sigma_Y, \xi_Y, \kappa_Y)) \in \mathbb{R}^{3d+3}$; a multivariate threshold $\mathbf{t} = (\mathbf{t}_X, t_Y)$; a L^r-norm $\|\cdot\|_r$ for $r \in [1, +\infty[$; a class Γ of angular predictive function $\gamma : \mathbb{S}^{d-1} \to [0, 1]$.
Truncation: Form a set of extreme observations with shifted $\mathbf{Z}_i$'s

$$\mathbf{Z}_{i,m} = \mathbf{Z}_i - \mathbf{m},$$

with $\mathbf{m} = (\mathbf{m}_X, m_Y) = \min_{1 \le i \le n}\{\mathbf{Z}_i\}$ and for all $i \in I_{ext} = \left\{i \in \{1, ..., n\}, \mathbf{X}_i \not\preccurlyeq \mathbf{t}_X\right\}$.
Marginal Pareto Transformation: Apply the Pareto transformation (9.4) to each margin of the extreme observations

$$\tilde{\mathbf{Z}}_i = (\tilde{\mathbf{X}}_i, \tilde{Y}_i) = p_{\boldsymbol{\sigma}, \boldsymbol{\xi}, \boldsymbol{\kappa}}(\mathbf{Z}_{i,m}), \quad \text{for all } i \in I_{ext}.$$

Angular rescaling: Form the angular components of the Pareto scale observations,

$$\boldsymbol{\Theta}_{X,i} = \tilde{\mathbf{X}}_i / \|\tilde{\mathbf{X}}_i\|_r,$$

$$\Theta_{Y,i} = \tilde{Y}_i / \|\tilde{\mathbf{Z}}_i\|_r.$$

Empirical quadratic risk minimization: based on the extreme transform dataset, solve the optimization problem

$$\min_{h \in \mathcal{H}} \sum (\Theta_{Y,i} - h(\boldsymbol{\Theta}_{X,i}))^2. \tag{9.5}$$

OUTPUT: Solution $\hat{h}$ to problem (9.5) and a predictive function $\hat{g}$ given by

$$\hat{g} : \mathbf{x} \in \mathbb{R}^d \mapsto$$
$$p^{-1}_{\sigma_Y, \xi_Y, \kappa_Y}\left(\left(\frac{\hat{h}(p_{\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X}(\mathbf{x} - \mathbf{m}_X)/\|p_{\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X}(\mathbf{x} - \mathbf{m}_X)\|_r)\|p_{\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X}(\mathbf{x} - \mathbf{m}_X)\|_r}{1 - \hat{h}(p_{\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X}(\mathbf{x} - \mathbf{m}_X)/\|p_{\boldsymbol{\sigma}_X, \boldsymbol{\xi}_X, \boldsymbol{\kappa}_X}(\mathbf{x} - \mathbf{m}_X)\|_r)^r}\right)^{1/r}\right) + m_Y, \tag{9.6}$$

to be used for predictions of Y_{n+1} based on new observation $\mathbf{X}_{n+1}$ such that $\mathbf{X}_{n+1} \not\preccurlyeq \mathbf{t}_X$.

to set the part above the threshold for each margin to unit exponential distribution and where we read that $H_{0,\sigma,\xi}$ is a GP distribution (2.3). The conditional distribution below the threshold is also transformed using the same transformation, but with no underlying precise distribution. Note that in this procedure, the parameter κ resulting from Algorithm 9.1 is not utilized.

For the ROXANE approach, Algorithm 9.1 is utilized to determine a multivariate threshold $\mathbf{t} = (\mathbf{t}_X, t_Y)$ such that the observations $\mathbf{X}$ satisfying $\mathbf{X} \not\preccurlyeq \mathbf{t}_X$ follow the limit distribution displayed in Section 7.2 - specifically, the independence between the radius and the angle of $\mathbf{X}$ - is valid. Additionally, Algorithm 9.1 is employed to model the entire distribution of the data - referring here to the retained thresholded dataset (9.2) - using an EGP distribution.

This modeling is mandatory for transforming all the data to Pareto margins via

$$p_{\sigma,\xi,\kappa}(x) = \frac{1}{1 - F_{\sigma,\xi,\kappa}(x)},$$

where $F_{\sigma,\xi,\kappa}$ *is the EGP cdf* (9.1).

9.3 Results

The two algorithmic procedures presented in Section 9.2 are applied separately to the sea level and skew surge datasets. Recall that we have pre-selected approximately half of the observations, ensuring that $X_B \geq q_B^{0.5}$ or $X_N \geq q_N^{0.5}$. For sea levels, the training set comprises 8271 observations for sea levels and 9213 observations for skew surges ranging from 10-08-1966 to 31-12-1999 and the test set comprises by 7483 observations for sea levels and 6024 observations for skew surges ranging from 01-01-2000 to 31-12-2023.

Two additional similar studies for Concarneau and Le Crouesty stations as output station are deferred to the appendix.

9.3.1 Marginal fitting and threshold selection

The initial step common to both procedures involves modeling the marginal distributions using the EGP distribution and selecting the threshold as described in Algorithm 9.1. The marginal fitting of the training data is carried out using the R package GAMLSS R. A. Rigby and D. M. Stasinopoulos (2005), employing the EGP family le Carrer (2022). The estimated parameters are presented in Table 9.1. The histogram plots with the corresponding fitted densities are illustrated in Figure 9.2. As outlined in Section 9.2.1, we determine the final threshold, above which observations are considered extreme, as the lowest point above which the fitted EGP density is convex. These thresholds are indicated by the red dotted vertical lines in Figure 9.2. The chosen thresholds range from quantiles of order 0.86 to 0.88, as summarized in Table 9.1. This thresholding results in final training sets composed of 2,436 sea levels and 3,154 skew surges, and test sets consisting of 1,963 sea levels and 1,883 skew surges. We denote by $\hat{\mathbf{t}} := (\hat{t}_B, \hat{t}_N, \hat{t}_Y) = (q_B^{0.88}, q_N^{0.87}, q_T^{0.87})$ the multivariate selected threshold and by $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}_B, \hat{\sigma}_N, \hat{\sigma}_Y)$, $\hat{\boldsymbol{\xi}} = (\hat{\xi}_B, \hat{\xi}_N, \hat{\xi}_Y)$ and $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_B, \hat{\kappa}_N, \hat{\kappa}_Y)$ the estimated EGP parameters. Extreme training and test sets are then formed with observations $\mathbf{X}_i$ such that

$$X_{B,i} \geq t_B \text{ or } X_{N,i} \geq t_N.$$

Figures 9.3 and 9.4 are bivariate scatterplots of observations at each station. These plots illustrate the strong correlation between stations, although this dependence is weaker for the most extreme observations which correspond in Figures 9.3 and 9.4 to the observations $\mathbf{X}_i$ such that $X_B \geq q_B^{0.98}$ or $X_N \geq q_N^{0.98}$.

9.3.2 Joint procedures

The multivariate procedure presented in Sections 9.2.2 and 9.2.3 are now applied to the extreme sea level and skew surge observations. Recall that the selected thresholds and EGP parameters for each margin, computed on the training set are summarized in Table 9.1.

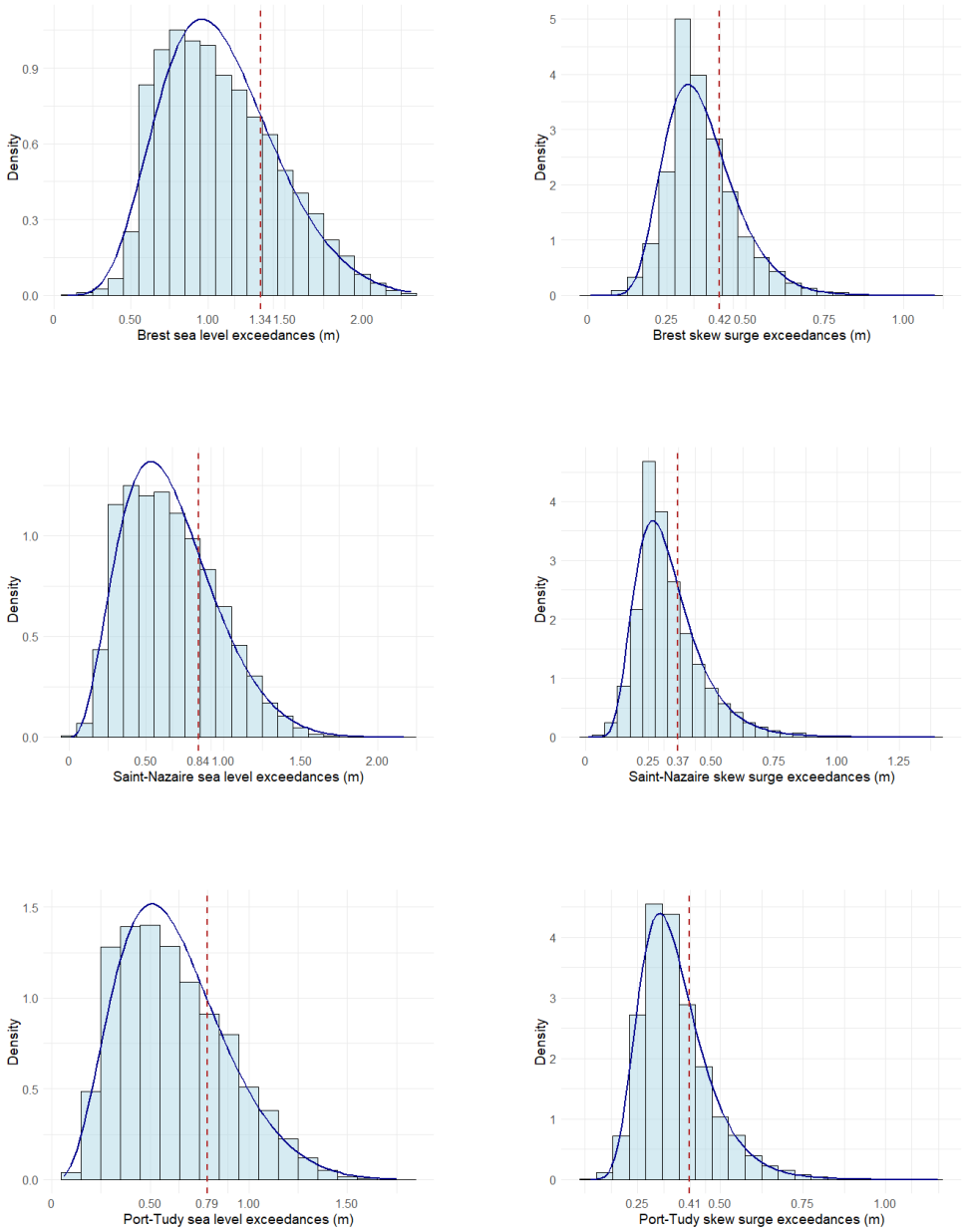


Figure 9.2: Histograms of sea level exceedances (left column) and skew surge exceedances (right column) of the training set at the three stations. The blue curves represent the fitted EGP densities, with parameters specified in Table 9.1. The dotted vertical red lines represent the smallest point above which each fitted density is convex, which correspond to the chosen marginal thresholds.

9.3.2.1 Plug-in method: MGP procedure

We describe first the MGP procedure for sea levels; the analysis for skew surges applies *mutatis mutandis*.

Modeling of the data. As described in Algorithm 9.2, the extreme observations are shifted by subtracting $\hat{\mathbf{t}}$, and then transformed to an exponential scale as follows,

$$\tilde{\mathbf{Z}}_i := (\tilde{X}_{B,i}, \tilde{X}_{N,i}, \tilde{Y}_i) = e_{\hat{\boldsymbol{\sigma}},\hat{\boldsymbol{\xi}}}(\mathbf{Z}_i - \hat{\mathbf{t}}).$$

Parameters/Stations		Brest	Saint-Nazaire	Port-Tudy
SL:	$\hat{\sigma}$	0.52	0.40	0.36
	$\hat{\xi}$	-0.18	-0.18	-0.17
	$\hat{\kappa}$	7.76	3.90	4.44
	t	$q_B^{0.88} \approx 7.11$	$q_N^{0.87} \approx 5.91$	$q_T^{0.87} \approx 5.19$
SS:	$\hat{\sigma}$	0.13	0.10	0.09
	$\hat{\xi}$	-0.092	0.004	-0.010
	$\hat{\kappa}$	15.12	13.05	38.68
	t	$q_B^{0.88} \approx 0.19$	$q_N^{0.86} \approx 0.15$	$q_T^{0.86} \approx 0.12$

Table 9.1: Point estimates of EGP parameters for sea levels and skew surges at the three stations. The chosen thresholds, determined using the procedure described in Algorithm 9.1, are shown in the fourth row in each sub-table.

Following the lines of Equations (2.15) and (2.16), we propose potential densities for $\mathbf{T}$ or $\mathbf{U}$ to deduce a suitable density for $\tilde{\mathbf{Z}}$ *via* Equations (2.15) and (2.16). The proposed density families for $\mathbf{T}$ and $\mathbf{U}$ are those described in Kiriliouk et al. (2019), specifically multivariate distribution with independent components where the marginals distributions are either all reverse exponential distributions or all Gumbel distributions. For detailed formulations of the resulting candidate densities for $\tilde{\mathbf{Z}}$ refer to Section 7 in Kiriliouk et al. (2019). To facilitate generation from these distributions and to avoid numerous approximations, we restrict our study to densities with explicit forms. As in Kiriliouk et al. (2019), we use a censored likelihood criterion to select the density. This involves maximizing the classical product likelihood function using only uncensored observations, where an observation is censored if any of its components are negative. This approach excludes the smallest observations, thereby enhancing performance for the most extreme observations, which are of primary interest. Finally, the retained density f_T for $\mathbf{T}$ is associated with independent Gumbel components with a common dependence parameter $\alpha > 0$ and varying locations parameters $\boldsymbol{\beta} = (\beta_B, \beta_N, \beta_Y) \in \mathbb{R}^3$, *i.e.*,

$$f_T(\mathbf{x}) = \alpha^3 \prod_{j \in \{B,N,Y\}} \exp(-\alpha(x_j - \beta_j)) \exp(-\exp(-\alpha(x_j - \beta_j))),$$

for $\mathbf{x} = (x_B, x_N, x_Y)$. Following Equation (2.16), the corresponding density for $\tilde{\mathbf{Z}}$ is given by

$$h_T(\mathbf{x}) = \mathbb{1}\{\max(\mathbf{x}) > 0\} \alpha^2 \Gamma(3) \exp(-\max(\mathbf{x})) \frac{\prod_{j \in \{B,N,Y\}} \exp(\alpha(x_j - \beta_j))}{\sum_{j \in \{B,N,Y\}} \exp(\alpha(x_j - \beta_j))}, \tag{9.7}$$

for $\mathbf{x} = (x_B, x_N, x_Y)$. The computed parameters for the Gumbel density are summarized in Table 9.2. For identifiability purpose, β_Y is set to zero.

Predictions on the test set. To obtain predictive values for the extreme Port-Tudy observations of the test set, we generate 100 values *via* rejection sampling of the conditional density

$$h^{\boldsymbol{\sigma},\boldsymbol{\xi}}_{T|(\tilde{X}_B,\tilde{X}_N)}(\tilde{y}) = \frac{h_T^{\boldsymbol{\sigma},\boldsymbol{\xi}}(\tilde{X}_B, \tilde{X}_N, \tilde{y})}{\int_{\mathbb{R}} h_T^{\boldsymbol{\sigma},\boldsymbol{\xi}}(\tilde{X}_B, \tilde{X}_N, s) ds}.$$

These generated values in exponential scale are then backtransformed to original scale using the inverse function $e^{-1}_{\hat{\sigma}_Y,\hat{\xi}_Y}(\cdot) + t_Y$. Point estimates are derived as the Monte Carlo averages of these 100 samples. For visual assessment of the goodness-of-fit, the bottom

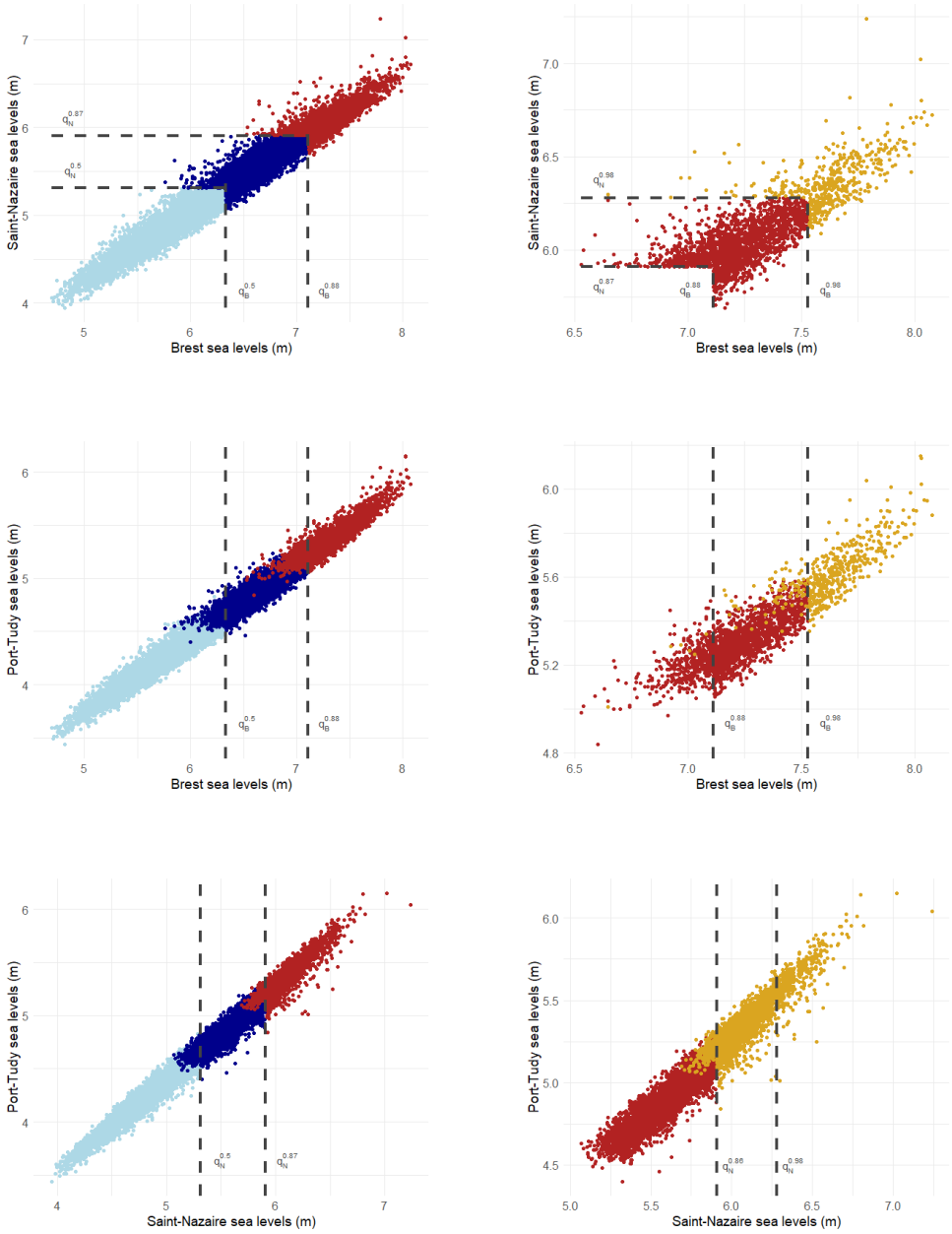


Figure 9.3: Left: Bivariate scatterplots of sea levels in the training set for each station. Light blue points represent the observations that were removed before the analysis. Dark blue and red points represent the remaining observations used for inference, with red points representing the extreme observations above the threshold specified in Table 9.1. Right: Bivariate scatterplots focusing on the extreme sea levels at each station (red points from the left plots). Orange points represent sea level observations where $X_B \geq q_B^{0.98}$ or $X_N \geq q_N^{0.98}$.

row of Figure 9.5 depicts QQ-plots comparing the estimated quantiles of sea levels and skew surges with the observed quantiles. Figure 9.6 shows the predicted curves, along with 0.95-bootstrap confidence intervals computed on the generated sample on the years 1999, 1989 and 1979. The 0.95-coverage probability associated with those confidence intervals is 0.91, which is the proportion of test observations falling within the computed 0.95-confidence intervals.

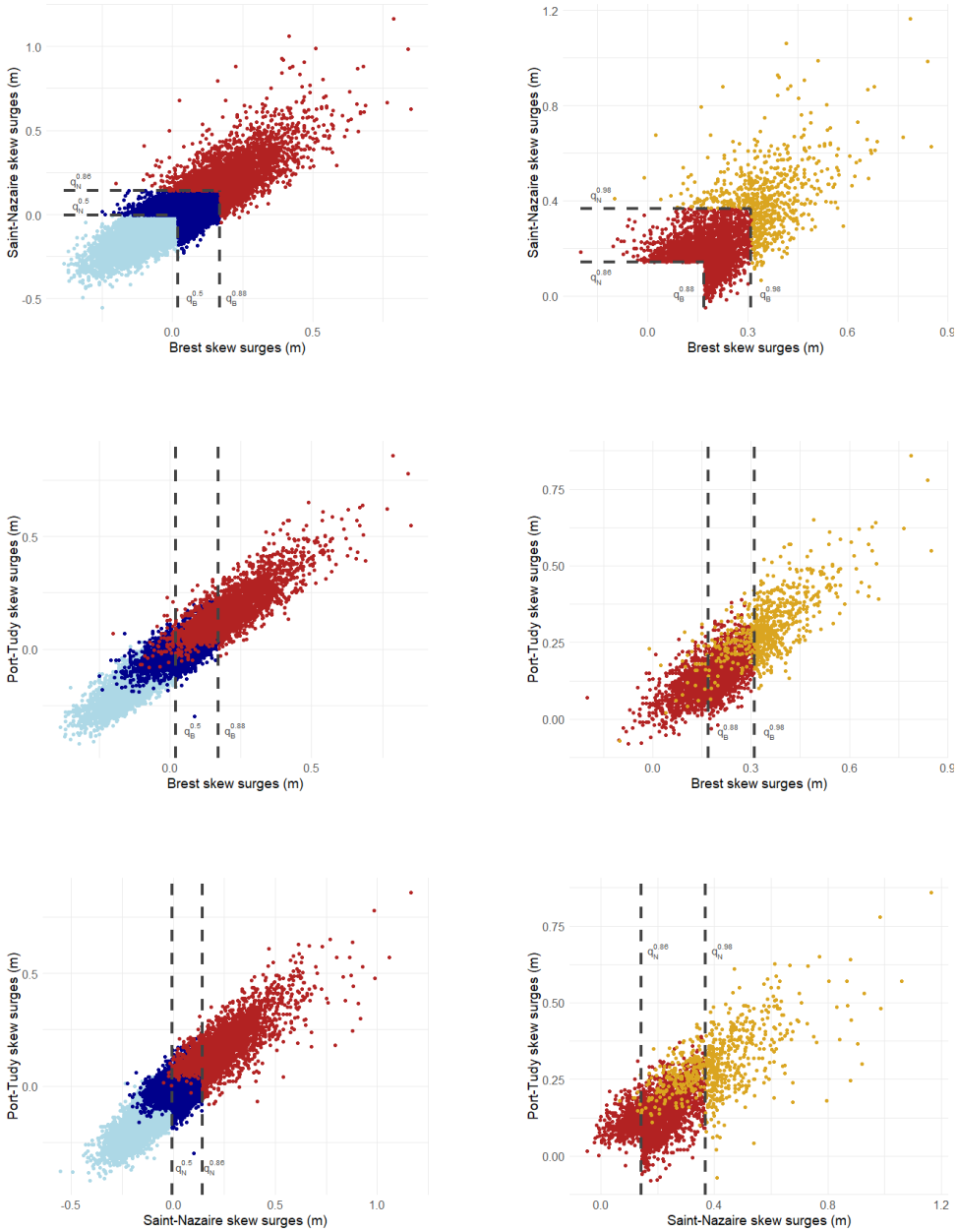


Figure 9.4: Left: Bivariate scatterplots of skew surges in the training set for each station. Light blue points represent the observations that were removed before the analysis. Dark blue and red points represent the remaining observations used for inference, with red points representing the extreme observations above the threshold specified in Table 9.1. Right: Bivariate scatterplots focusing on the extreme skew surges at each station (red points from the left plots). Orange points represent sea level observations where $X_B \geq q_B^{0.98}$ or $X_N \geq q_N^{0.98}$.

9.3.2.2 Regression method: ROXANE procedure

We now describe the ROXANE procedure for the sea levels, the analysis for skew surges applied *mutatis mutandis*. As described in Algorithm 9.3, the extreme observations are shifted by subtracting $\mathbf{m}$, and then transformed to a Pareto scale as follows,

$$\tilde{\mathbf{Z}}_i := (\tilde{X}_{B,i}, \tilde{X}_{N,i}, \tilde{Y}_i) = p_{\boldsymbol{\sigma},\boldsymbol{\xi},\boldsymbol{\kappa}}(\mathbf{Z}_i - \mathbf{m}),$$

Estimated parameters	α	β_B	β_N	β_Y
	6.63	-0.17	-0.06	0.00

Table 9.2: Point estimates of the dependence and location parameters of Gumbel density model given in (9.7).

with $\mathbf{m} = (\mathbf{m}_X, m_Y) = \min_{1\leq i\leq n}\{\mathbf{Z}_i\}$, where the $\mathbf{Z}_i$'s refer to the retained observations after initial thresholding (9.2). We then consider the angular parts of the input and the output variables w.r.t. the L^2-norm that are

$$\mathbf{\Theta}_{X_i} := (\Theta_{B,i}, \Theta_{N,i}) := \frac{(\tilde{X}_{B,i}, \tilde{X}_{N,i})}{\sqrt{\tilde{X}_{B,i}^2 + \tilde{X}_{N,i}^2}},$$

$$\Theta_{Y,i} := \frac{\tilde{Y}_i}{\sqrt{\tilde{X}_{B,i}^2 + \tilde{X}_{N,i}^2 + \tilde{Y}_i^2}}.$$

Two regression algorithms are used to predict the angular output observations $\Theta_{Y,i}$ based on the angular input observations $\mathbf{\Theta}X, i$: Ordinary Least Squares (OLS) from the stats package and Random Forest (RF) from the randomForest package. Following Equation (7.14) of Proposition 7.10, the predicted angular values $\hat{\Theta}_{Y,i}$ on the test set are then backtransformed to the Pareto scale *via*

$$\hat{\tilde{Y}}_i = \sqrt{\frac{\hat{\Theta}_{Y,i}\sqrt{X_{B,i}^2 + X_{N,i}^2}}{1 - \hat{\Theta}_{Y,i}^2}},$$

and then to the original scale *via* the inverse function $p^{-1}_{\sigma_Y,\xi_Y,\kappa_Y}(\cdot) + m_Y$. For visual assessment of the goodness-of-fit, the top and middle rows of Figure 9.5 depict QQ-plots comparing the estimated quantiles of sea levels and skew surges with the observed quantiles for both the OLS and RF algorithms. Figure 9.6 shows the predicted curves, along with 0.95-bootstrap confidence intervals computed on the generated sample on the years 1999, 1989 and 1979.

9.3.3 Discussion and comparison of the methods

In this section, we analyze the results obtained from the various models applied to our data. First, the EGP marginal model demonstrates a satisfactory fit, especially in the right tail of the distribution for data exceeding the selected thresholds, as illustrated in Figure 9.2. Regarding the joint models, all models exhibit reasonably good performance, as evidenced by Figure 9.6. The QQ-plots indicate that the generated distributions closely mimic the behavior of the observed distribution in the extreme regions. However, there is one notable exception: the prediction of the largest skew surge using the ROXANE OLS procedure. This surge, recorded on November 7, 1969, was the highest ever at Brest. Meteorological records suggest that a severe storm impacted northern and western France, heavily affecting Brest while leaving Port-Tudy and Saint-Nazaire relatively unaffected.

Meteorological archives indicate that a storm impacted northern and western France, heavily affecting Brest while leaving Port-Tudy and Saint-Nazaire relatively unaffected. This discrepancy likely led to an overestimation by the model. For the lower tail of

the extreme observations, the QQ-plots reveal poor model performance due to several factors:

- for the MGP procedure, a censored likelihood is used to obtain accurate predictions for the most extreme observations, which results in the neglect of less extreme observations;
- for ROXANE procedure, the estimation of the κ parameter lacks robustness and can be significantly overestimated. While this parameter does not influence the prediction of the largest observations, it has a substantial effect on the smallest ones. This is because the backtransformation $p^{-1}_{\sigma_Y,\xi_Y,\kappa_Y}$ behaves as $x^{1/\kappa}$ near zero. Consequently, a small estimation error in Pareto scale can lead to a large estimation error in original scale. For instance, with $\kappa = 39$ (as for the skew surges), a minor error of 0.01 in Pareto scale results in a substantial error of 0.89 meters in the original scale;
- common to both procedures, an extreme observation might be extreme only at one station, such as Brest or Saint-Nazaire, and not at the other two stations. Hence, the models cannot distinguish these observations from typical observations where values are large at Port-Tudy and another station. This issue could be addressed by including wind-related data into the models, as wind conditions significantly influence skew surges (see Pugh and Woodworth (2014)).

Given the challenges associated with the lower tail of the extreme distributions, it is crucial to evaluate the performance of both models primarily in relation to the largest observations, which aligns with the objectives of EVT. Figure 9.6 clearly illustrates that the most significant misestimations occur for the smallest values. Additionally, the models demonstrate reduced precision for earlier years. One possible reason is the failure to account for the time trend due to global warming (Seneviratne et al. (2021)). Although we assume in our study that this trend is present for all three stations and can be ignored, if the trend behavior differs across stations, it must be considered.

To compare the performance of the two procedures, we use Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics. Table 9.3 contains the errors computed for both the entire test set and the most extreme subset of the test set, consisting of observations such that $Y_i \geq q_Y^{0.5}$, where $q_Y^{0.5}$ is calculated on the test set only. Overall, the MGP procedure performs better, correctly modeling the smallest values despite using a censored likelihood criterion. However, the ROXANE procedure when paired with the OLS algorithm, performs better in the most extreme regions. This is more evident in the studies for Concarneau and Le Crouesty (refer to Tables 9.5 and 9.7 in Appendix 9.A).

Training models/Errors	RMSE(80%)	MAE(80%)	RMSE(90%)	MAE(90%)
SL: ROX RF	0.087	0.065	0.081	0.059
ROX OLS	0.081	**0.059**	**0.077**	**0.055**
MGP	**0.080**	**0.059**	0.079	0.056
SS: ROX RF	0.084	0.062	0.083	0.061
ROX OLS	0.083	0.059	0.079	**0.056**
MGP	**0.078**	**0.057**	**0.077**	**0.056**

Table 9.3: RMSE and MAE of predicted sea levels and skew surges at Port-Tudy station from the ROXANE procedure with RF regression (ROX RF), ROXANE procedure with OLS regression (ROX OLS) and MGP procedure (MGP). Errors are computed for the entire test set (first two columns) and for the subset of the test set comprising the most extreme observations with respect to the Port-Tudy value (last two columns).

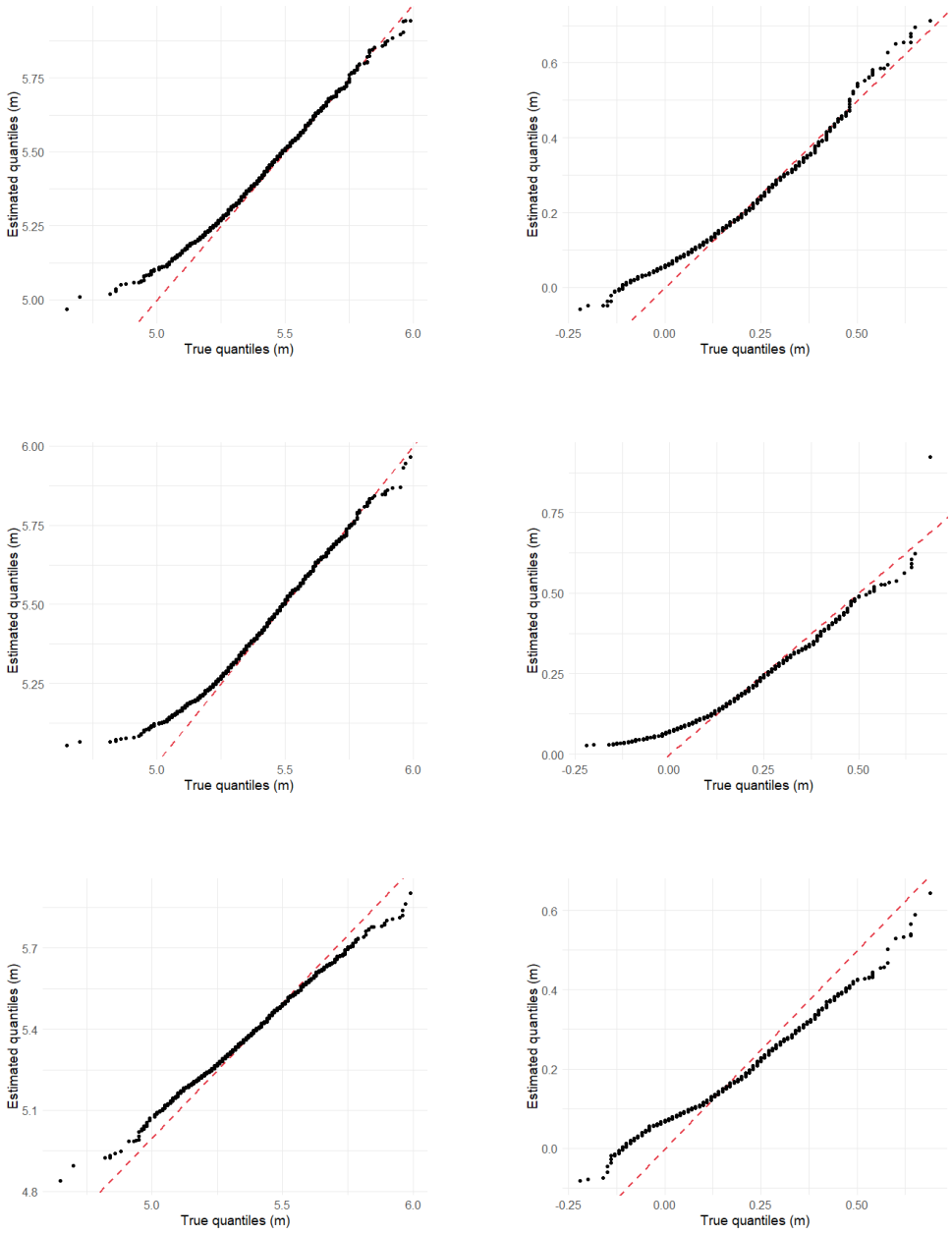


Figure 9.5: QQ-plots comparing observed sea level quantile (left column, x-axis) and skew surge quantiles (right column, x-axis) to estimated quantile (y-axis) from the predictions given by algorithms of Sections 9.2.3 and 9.2.2. The plots show results from the ROXANE procedure with RF regression (top row), ROXANE procedure with OLS regression (middle row), and MGP procedure (bottom row). The red line represents the identity line $x = y$.

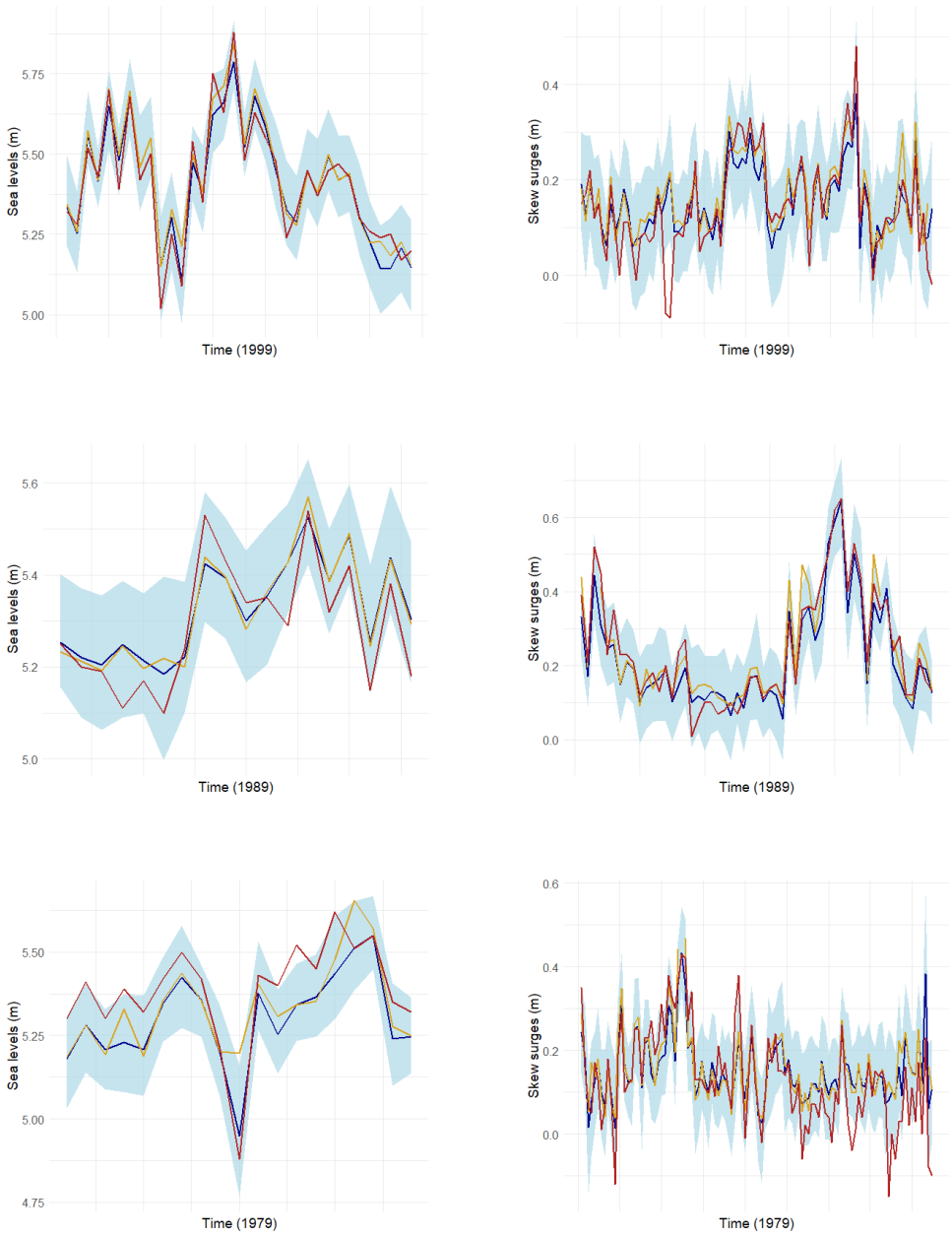


Figure 9.6: Predicted sea levels (left column) and skew surges (right column) at Port-Tudy station for the years 1999 (top row), 1989 (middle row), 1979 (bottom row). Red curves represent the true values on the test set; orange curves represent the predicted values by the ROXANE procedure with OLS algorithm; blue curves represent the predicted values by the MGP procedure with bootstrap 0.95 confidence intervals (light blue).

9.4 Conclusion

An in-depth experimental investigation into extremal joint dependence structure between sea levels and skew surges across various sites along the French Atlantic coast has been conducted. The study demonstrates that an EGP distribution fitted to the marginal observations at each site shows satisfactory performances. We introduce two novel methodologies that have not previously been applied to sea level or skew surge modeling: a procedure for deriving an optimal MGP density and a regression algorithm tailored for extreme value. Both are employed in prediction tasks for missing values at a site, given nearby extreme values, with the underlying goal of reconstructing past extreme values at sites with limited historical records. The methods are extensively compared and discussed, revealing that both approaches yield valid results of significant importance for practitioners, each offering distinct advantages: one provides better point estimates, while the other offers a robust generative model.

9.A Additional Studies at Le Crouesty and Concarneau

Two similar analyses to the one in Section 9.3 are conducted with the sea levels and skew surges at Concarneau and Le Crouesty as output stations.

9.A.1 Concarneau study

Table 9.4 presents the EGP parameters and the selected thresholds provided by Algorithm 9.1. The fitted EGP densities and the selected thresholds are illustrated in Figure 9.7. Figure 9.8 displays the QQ-plots comparing the true quantiles against the quantiles estimated with the OLS ROXANE, RF ROXANE and the MGP procedures. The RMSE and MAE associated with the three procedure are summarized in Table 9.5.

Parameters/Stations		Brest	Saint-Nazaire	Concarneau
SL:	$\hat{\sigma}$	0.52	0.40	0.33
	$\hat{\xi}$	-0.19	-0.19	-0.14
	$\hat{\kappa}$	5.90	3.76	5.29
	t	$q_B^{0.87} \approx 7.14$	$q_N^{0.87} \approx 5.94$	$q_T^{0.87} \approx 5.06$
SS:	$\hat{\sigma}$	0.11	0.10	0.12
	$\hat{\xi}$	-0.08	-0.003	-0.06
	$\hat{\kappa}$	20.78	10.67	15.23
	t	$q_B^{0.88} \approx 0.44$	$q_N^{0.86} \approx 0.35$	$q_T^{0.87} \approx 0.40$

Table 9.4: Point estimates of EGP parameters for sea levels and skew surges at the three stations. The chosen thresholds, determined using the procedure described in Algorithm 9.1, are shown in the fourth row of each sub-table.

Training models/Errors	RMSE(80%)	MAE(80%)	RMSE(90%)	MAE(90%)
SL: ROX RF	0.069	0.053	0.070	0.055
ROX OLS	**0.062**	**0.046**	**0.064**	**0.048**
MGP	0.066	0.050	0.071	0.053
SS: ROX RF	0.072	0.056	0.070	0.056
ROX OLS	**0.063**	**0.049**	**0.063**	**0.051**
MGP	0.064	0.050	0.071	0.058

Table 9.5: RMSE and MAE of predicted sea levels and skew surges at Concarneau station from the ROXANE procedure with RF regression (ROX RF), ROXANE procedure with OLS regression (ROX OLS) and MGP procedure (MGP). Errors are computed for the entire test set (first two columns) and for the subset of the test set comprising the most extreme observations with respect to the Concarneau value (last two columns).

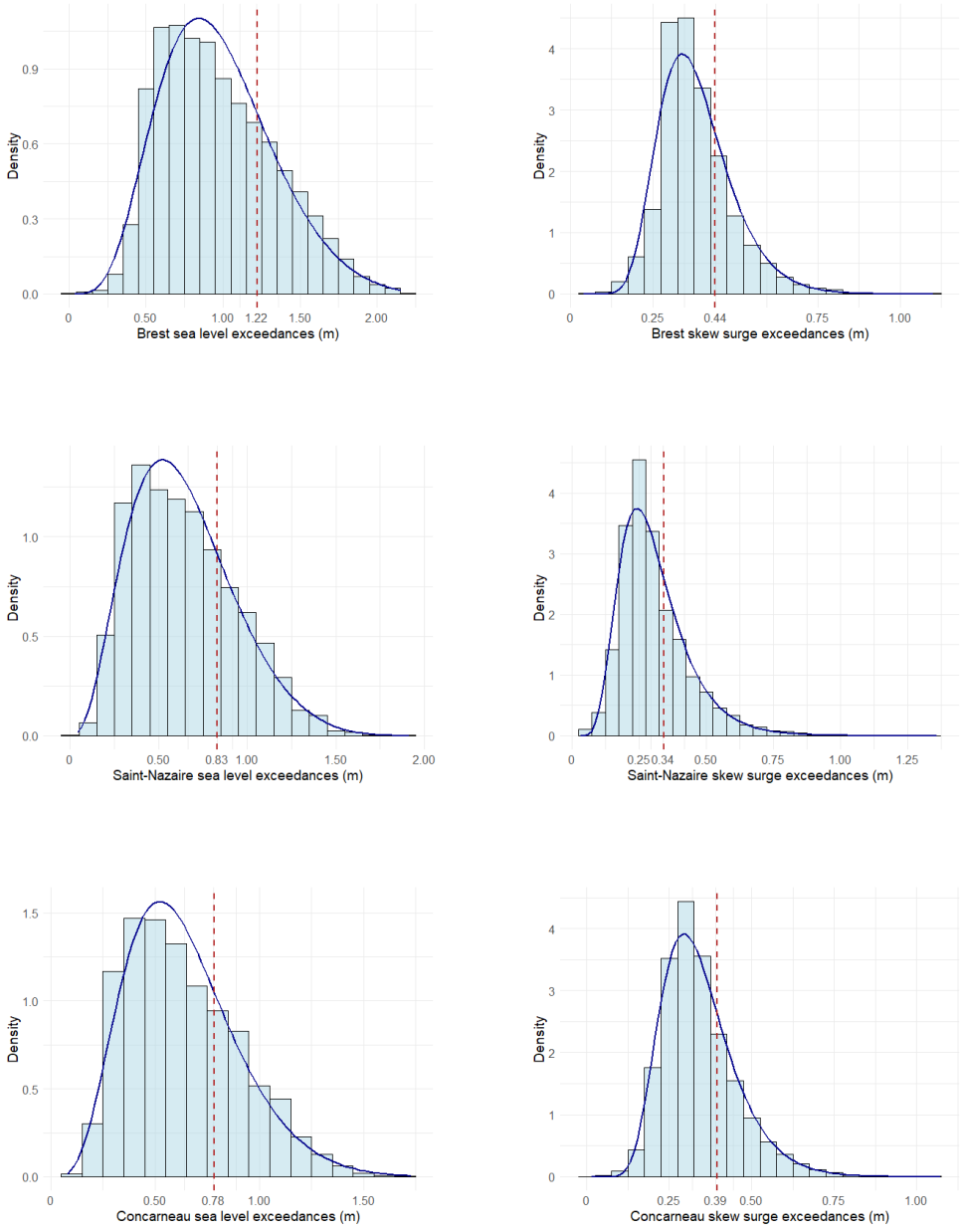


Figure 9.7: Histograms of sea level exceedances (left column) and skew surge exceedances (right column) of the training set at the three stations. The blue curves represent the fitted EGP densities, with parameters specified in Table 9.4. The dotted vertical red lines represent the first convexity point of each fitted density, which correspond to the chosen marginal threshold.

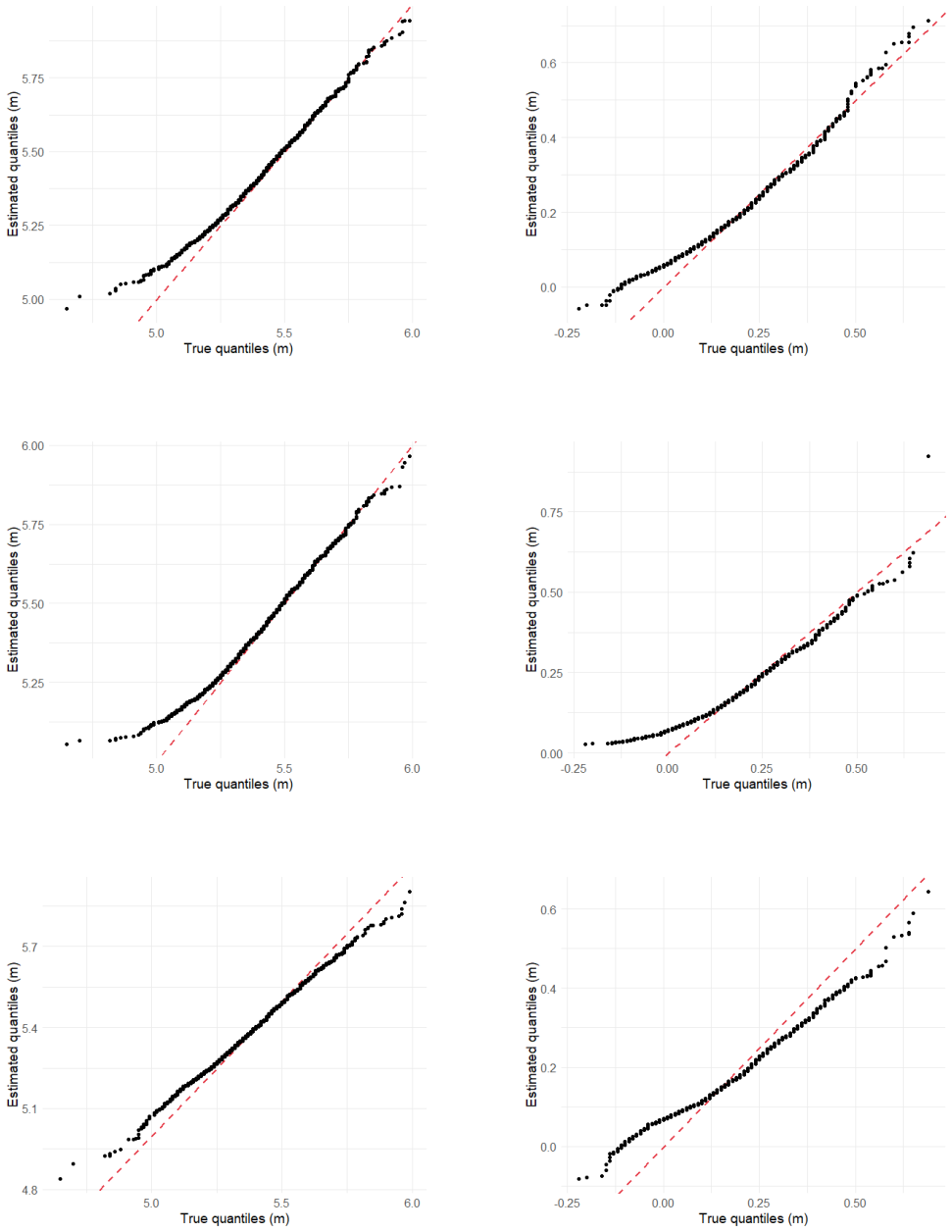


Figure 9.8: QQ-plots comparing observed sea level quantile (left column, x-axis) and skew surge quantile (right column, x-axis) to estimated quantile (y-axis) from the predictions given by the algorithms of Sections 9.2.3 and 9.2.2. The plots show results from the ROXANE procedure with RF regression (top row), ROXANE procedure with OLS regression (middle row), and MGP procedure (bottom row). The red line represents the identity line $x = y$.

9.A.2 Le Crouesty study

Table 9.6 presents the EGP parameters and the selected thresholds provided by Algorithm 9.1. The fitted EGP densities and the selected thresholds are illustrated in Figure 9.9. Figure 9.10 displays the QQ-plots comparing the true quantiles against the quantiles estimated with the OLS ROXANE, RF ROXANE and the MGP procedures. The RMSE and MAE associated with the three procedure are summarized in Table 9.7.

Parameters/Stations		Brest	Saint-Nazaire	Le Crouesty
SL:	$\hat{\sigma}$	0.54	0.41	0.39
	$\hat{\xi}$	-0.21	-0.21	-0.21
	$\hat{\kappa}$	5.66	3.06	2.88
	t	$q_B^{0.88} \approx 7.12$	$q_N^{0.87} \approx 5.94$	$q_T^{0.87} \approx 5.48$
SS:	$\hat{\sigma}$	0.12	0.11	0.09
	$\hat{\xi}$	-0.073	-0.01	0.01
	$\hat{\kappa}$	24.56	6.78	9.76
	t	$q_B^{0.88} \approx 0.20$	$q_N^{0.85} \approx 0.15$	$q_T^{0.85} \approx 0.12$

Table 9.6: Point estimates of EGP parameters for sea levels and skew surges at the three stations. The chosen thresholds, determined using the procedure described in Algorithm 9.1, are shown in the fourth row of each sub-table.

Training models/Errors	RMSE(80%)	MAE(80%)	RMSE(90%)	MAE(90%)
SL: ROX RF	0.066	0.051	0.061	0.048
ROX OLS	**0.059**	**0.044**	**0.056**	**0.043**
MGP	0.065	0.048	0.067	0.048
SS: ROX RF	0.065	0.050	0.062	0.046
ROX OLS	**0.058**	**0.044**	**0.055**	**0.040**
MGP	0.060	0.046	0.060	0.045

Table 9.7: RMSE and MAE of predicted sea levels and skew surges at Le Crouesty station from the ROXANE procedure with RF regression (ROX RF), ROXANE procedure with OLS regression (ROX OLS) and MGP procedure (MGP). Errors are computed for the entire test set (first two columns) and for the subset of the test set comprising the most extreme observations with respect to the Le Crouesty value (last two columns).

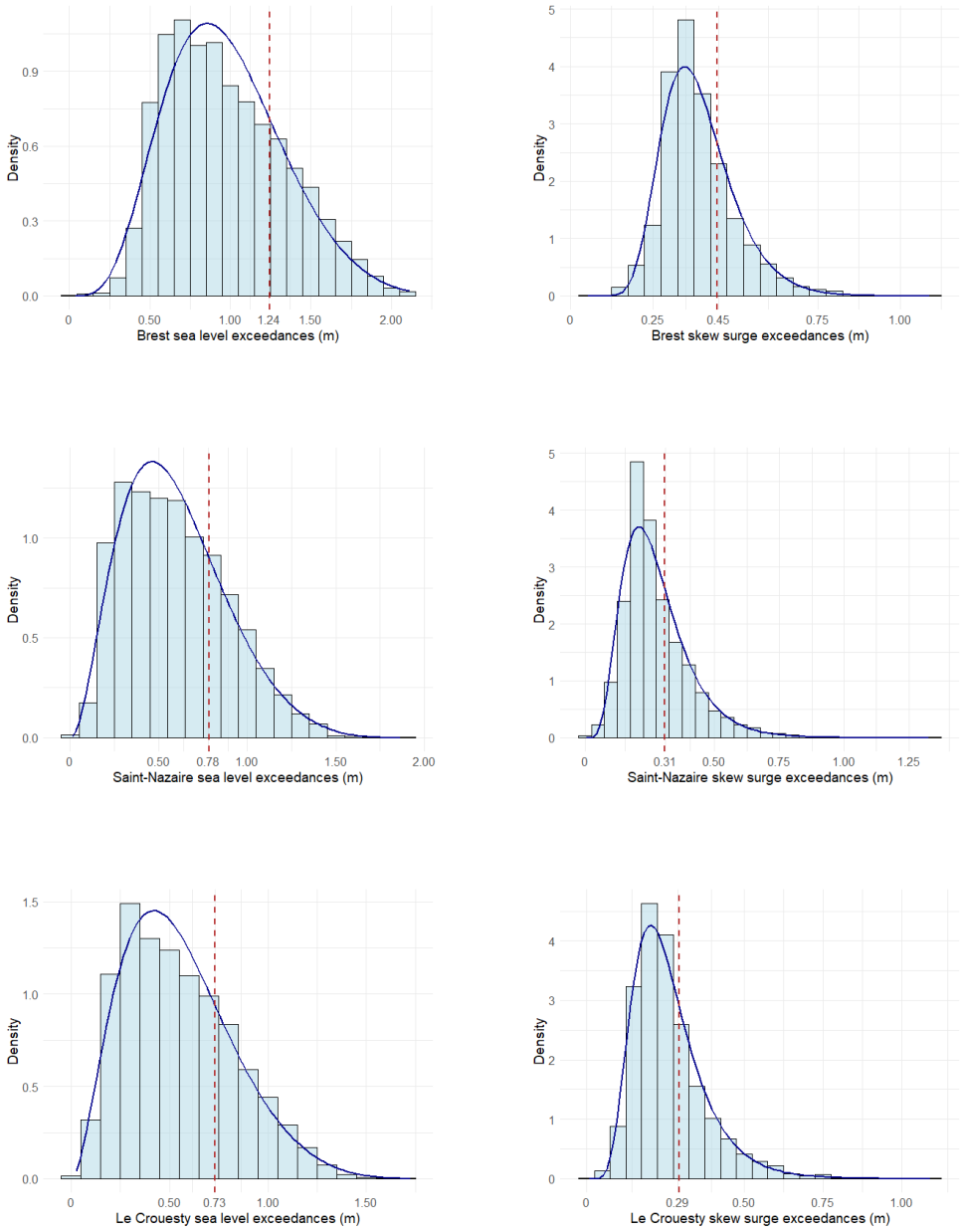


Figure 9.9: Histograms of sea level exceedances (left column) and skew surge exceedances (right column) of the training set at the three stations. The blue curves represent the fitted EGP densities, with parameters specified in Table 9.6. The dotted vertical red lines represent the first convexity point of each fitted density, which correspond to the chosen marginal threshold.

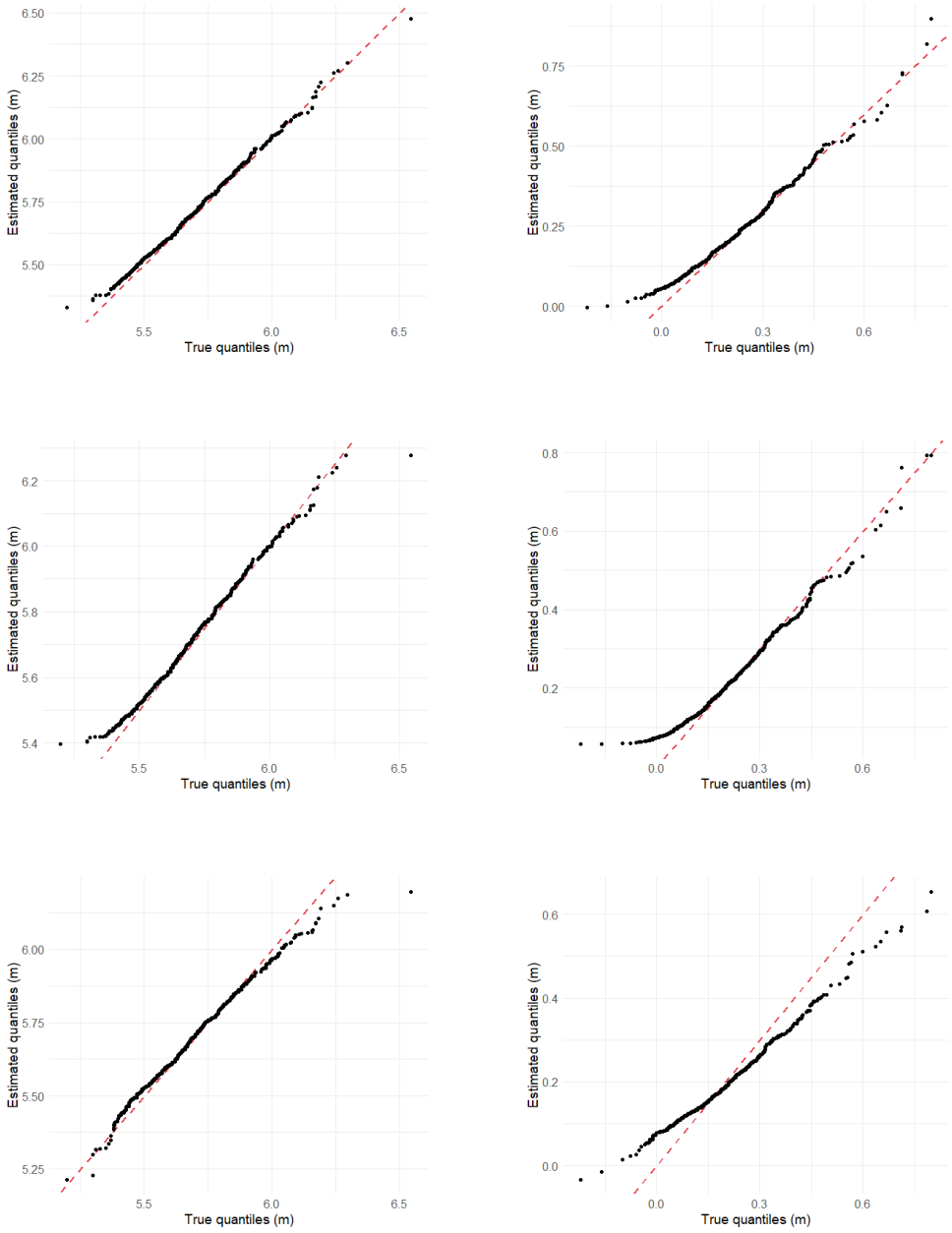


Figure 9.10: QQ-plots comparing observed sea level quantiles (left column, x-axis) and skew surge quantiles (right column, x-axis) to estimated quantile (y-axis) from the predictions given by the algorithms of Sections 9.2.3 and 9.2.2. The plots show results from the ROXANE procedure with RF regression (top row), ROXANE procedure with OLS regression (middle row), and MGP procedure (bottom row). The red line represents the identity line $x = y$.

Conclusions and Perspectives

This dissertation focuses on the development of probabilistic frameworks and practical tools for statistical learning tasks involving extremes. It revolves around two main aspects: dimension reduction of extremes and regression on extremes.

The concept of Regular Variation (RV) in Hilbert spaces has been studied extensively. Several characterisations have been developed, including some that rely only on finite-dimensional convergences, while manipulating possible infinite-dimensional objects. A wide range of examples are provided to illustrate these requirements. Dimension reduction techniques for a regularly varying random element in a Hilbert space are then investigated. The convergence of the eigenstructure of its covariance operator to the eigenstructure of a limit covariance operator is proved. In particular, the distance between the PCA decomposition of the limit random element and an empirical finite-range PCA decomposition is controlled by concentration bounds. These results ensure the theoretical validity for reducing the dimension of functional extremes. The empirical validity is demonstrated by various experiments on simulated and real data.

One particularly compelling prospect is the development of an anomaly detection method tailored to the functional setting. This method would hinge on our dimension reduction results, with the anomaly score depending on the reconstruction error, mirroring methodologies prevalent in finite-dimensional contexts Goix et al. (2017). In addition, the flexible nature of our supervised approach, which quantifies the "normalcy" of new observations in relation to a "normal" profile learned from our PCA procedure, extends its applicability to diverse data analysis tasks, including classification and clustering.

In numerous applications, such as audio and image compression, dimension reduction is frequently achieved through the decomposition of data using wavelet bases rather than PCA bases. This preference stems from the fact that, unlike PCA, signal approximations utilizing wavelets are non-linear, providing greater flexibility in representing the signal with fewer components (see Mallat (1999)). A promising direction for future research is to explore the properties of wavelet decomposition for regularly varying random functions, aiming to achieve more efficient representations of extreme values at a lower cost.

A second project investigates regression tasks in extreme regions. In this context, we have developed a suitable framework based on the novel assumption of regular variation w.r.t. a component. Its validity is illustrated by several examples of regression problems that fall within the scope of this working hypothesis. Properties of regression functions in this framework are then investigated. Under the assumption of RV w.r.t. a component assumption, we prove the existence of an optimal regression function in extreme regions which is angular. Empirical guarantees on the finite-range counterpart

of this regression function are proved by means of concentration inequalities. Based on the developed results, a practical algorithmic approach, namely the ROXANE algorithm, is constructed and its soundness is illustrated by numerical experiments.

As discussed throughout this part, several extensions of this work merit deeper investigation to further enhance the present study. First, we assume that the data are regularly varying with the same RV index, a condition that may only hold after marginal standardization (2.9). In practice, however, this transformation is unknown, and its empirical counterpart (2.10) must be used, introducing bias. In the context of binary classification, Clémençon et al. (2023) provide statistical guarantees concerning the error related to this bias. Extending such results to the regression setting would significantly complement this study, substantially improving the reliability of the ROXANE algorithm. Additionally, adapting the ROXANE algorithm to high-dimensional settings could broaden its applicability. This could be achieved by considering penalized versions of the risks, which would also help mitigate potential overfitting of the model. Finally, as highlighted in Remark 8.9, our concentration results are derived using general inequalities. However, there are inequalities specifically suited for regression problems that could potentially offer better convergence rates. Currently, the convergence rate is of the order $O(1/\sqrt{k})$, but based on results in Györfi et al. (2002), a more detailed study could yield an improved convergence rate of the order $O(\log(k)/k)$.

The last part of this manuscript is an applied study of extreme sea levels from tide-gauges on the French Atlantic coast. The aim of the study is to capture the dependence between extreme sea levels and skew surges at different sites, in order to allow the prediction of values at a site, given extreme values at nearby sites. First, a regression function is constructed using the ROXANE routine. The results of this predictive function are compared with a parametric approach which consists of fitting a Multivariate Generalized Pareto density to the data and then obtaining a prediction by sampling according to the conditional density. Both methods are compared and discussed in detail.

Several extensions can be derived from this work. The main output of our prediction procedure is the ability to reconstruct past extreme events at sites where no measurements have been made. By investigating the bias introduced by considering point estimates rather than true values, these estimates can be used to improve the precision and reduce the uncertainty of inferences, such as the original inference of return periods, made at sites with limited records. Our methods suffer from poor performance in non-extreme regions, which is to be expected, as mentioned above. The reason for this is that the models cannot distinguish observations with two extreme values from observations with only one extreme value. We believe that a more detailed study, including wind-related variables, could solve this problem.

Chapter 10

Introduction en français

Contents

10.1 Motivations

Le 1er février 1953, une tempête dévastatrice a frappé l'Europe du Nord, touchant les Pays-Bas et le Royaume-Uni. La tempête a submergé la plupart des défenses côtières, provoquant une inondation sans précédent qui a coûté la vie à plus de 2000 personnes, dont plus de 1800 juste aux Pays-Bas. Suite à ce tragique événement, le gouvernement néerlandais s'est posé une question cruciale : quelle hauteur devraient atteindre les nouvelles digues afin de prévenir de futures catastrophes de cette ampleur à moindre coût? La réponse repose sur la détermination des niveaux de mer maximaux pouvant être atteints au cours des cent ou mille prochaines années. Les méthodes statistiques traditionnelles ne suffisent pas à répondre à ce problème, car elles nécessitent de faire des inférences sur une période plus longue que les données d'observation disponibles.

La théorie des valeurs extrêmes fournit les outils statistiques nécessaires pour analyser ce type d'événements rares. Cette théorie se concentre sur la compréhension des événements de faible probabilité qui se situent en dehors du cœur d'une distribution, mais qui jouent une importance cruciale dans de nombreux domaines. Ces événements, bien qu'étant hors du centre de masse de la distribution, peuvent être essentiels dans divers domaines pratiques, qu'il s'agisse de la gestion des risques en finance ou en

Figure 10.1: Inondation de 1953 en mer du Nord, 1953 (photo de *Watersnoodmuseum*).

assurance, de la modélisation des événements extrêmes en climatologie (comme les fortes précipitations ou les vagues de chaleur), ou encore de la prévision des niveaux extrêmes de pollution de l'air ou des surcharges du trafic sur un réseau en sciences de la santé ou des télécommunications.

Dans cette thèse, nous proposons une étude à l'intersection de la théorie des valeurs extrêmes et de l'apprentissage statistique, une branche des statistiques dédiée à la prédiction et à la modélisation des structures dans les données. Notre attention se concentre sur deux domaines principaux de l'apprentissage statistique : l'analyse des données fonctionnelles et la régression. L'analyse des données fonctionnelles concerne les données sous forme de fonctions, dépendant de variables continues comme le temps ou l'espace. Avec les avancées technologiques des capteurs, qui fournissent des mesures massives et de plus en plus fines, il est devenu essentiel de modéliser les extrêmes fonctionnels, tels que les surcharges énergétiques ou les fortes précipitations au cours d'une période. La régression, l'une des tâches fondamentales de l'apprentissage statistique, consiste à apprendre des fonctions de prédiction à partir d'exemples labellisés pour faire des prédictions sur de nouvelles données non labellisés. Bien que ces fonctions prédictives ciblent généralement leurs performances sur le cœur des données, il est crucial dans de nombreuses applications de développer des modèles qui traitent spécifiquement aux exemples situés en dehors du cœur de la distribution, en particulier ceux de nature extrême.

10.2 État de l'art

La théorie des valeurs extrêmes (Extreme Value Theory, EVT) et l'apprentissage statistique sont deux branches des statistiques qui ont été activement étudiées pendant de nombreuses décennies. L'EVT se concentre sur la modélisation des événements rares, tandis que l'apprentissage statistique englobe des méthodes permettant d'apprendre des comportements et des caractéristiques à partir des données. Récemment, un intérêt croissant s'est manifesté pour l'application des outils d'apprentissage statistique à l'étude des extrêmes, en particulier dans des contextes d'apprentissage non supervisé.

Parmi les exemples figurent la réduction de dimension à l'aide de techniques de clustering dans plusieurs sous-espaces Goix et al. (2016, 2017); Chiapino et al. (2019); Simpson et al. (2020); Meyer and Wintenberger (2021, 2023), ainsi que l'analyse en composantes principales (ACP) Cooley and Thibaud (2019); Drees and Sabourin (2021). Centrale dans Chapitre 6, la réduction de dimension pour les extrêmes est présentée en détails dans Section 10.2.2. En outre, des études notables ont été réalisées dans les méthodes de clustering Janßen and Wan (2020); Vignotto et al. (2021), les modèles graphiques Engelke and Hitz (2020), et avec des applications comme la détection d'anomalies Chiapino et al. (2020); Vignotto and Engelke (2020) (voir Section 10.2.2 pour des références supplémentaires).

Dans un cadre supervisé, la littérature se concentre principalement sur la prédiction des valeurs extrêmes de la variable cible Y Aghbalou et al. (2024a) ou sur la régression quantile extrême via des méthodes comme le gradient boosting Velthoen et al. (2023) ou les forêts aléatoires Gnecco et al. (2024).

À notre connaissance, le seul travail qui traite de la prédiction d'une variable cible Y basée sur les valeurs extrêmes de la variable d'entrée $\mathbf{X}$[1] est celui de Jalalzai et al. (2018). Cette étude développe un cadre probabiliste pour la classification binaire avec des covariables extrêmes basé sur l'étude du risque empirique, en supposant que les distributions conditionnelles de $\mathbf{X}$ sachant $Y = \pm 1$ sont à variation régulière (Regular Variation, RV) (voir Chapitre 2 pour plus de détails). Les auteurs construisent ensuite une fonction de régression adaptée au problème d'estimation du risque empirique $\min_g L_t(g) = \mathbb{P}(Y \neq g(\mathbf{X}) \,|\, |\mathbf{X}| \geq t)$ pour une certaine norme $|\cdot|$. Partie III de cette thèse vise à étendre ces résultats au problème de régression, en établissant des conditions suffisantes et raisonnables pour traiter de la régression avec une sortie continue et une fonction de perte appropriée. Plus précisément, nous cherchons à étendre les garanties statistiques non asymptotiques fournies pour les classificateurs extrêmes aux fonctions de régression extrêmes.

Les développements récents dans les inégalités de concentration pour les extrêmes sont à souligner. À notre connaissance, les premières de ce genre sont dues à Boucheron and Thomas (2012) (voir aussi Boucheron and Thomas (2015)), qui prouvent des bornes de concentration pour les statistiques d'ordre extrême. Une autre approche pionnière, qui a influencé de nombreuses études ultérieures, est celle de Goix et al. (2015), qui présente des inégalités de concentration générales pour des événements de faible probabilité et les applique à des contextes de classification. Ces résultats forment la base des travaux non asymptotiques de Jalalzai et al. (2018). En outre, les auteurs de Clémençon et al. (2023) fournissent des bornes statistiques concernant l'utilisation de la standardisation empirique des marginales (voir Équation (2.10) et Remarque 7.3 pour plus de détails) au lieu de la vraie standardisation des marginales (inconnue) dans la procédure de classification. Les inégalités de concentration sont également utilisées pour des problèmes de validation croisée extrême Aghbalou et al. (2023) (également basés sur Goix et al. (2015)) et pour la classification déséquilibrée Aghbalou et al. (2024b), où la classe minoritaire correspond aux données extrêmes. Des inégalités de concentration générales, issues de la théorie de Vapnik-Chervonenkis (VC), pour les extrêmes ont également été développées dans Lhaut et al. (2022) et Lhaut and Segers

[1] Pour plus de clarté, tout au long de cette thèse, les quantités multivariées sont mises en gras lorsque nécessaire, par exemple $\mathbf{x} \in \mathbb{R}^d$, afin de distinguer les observations d'échantillons des coordonnées vectorielles. Les quantités univariées ou définies dans un espace de Hilbert sont notées de manière traditionnelle, telles que $x \in \mathbb{R}$ ou $h \in \mathbb{H}$, aucune confusion n'étant susceptible de survenir dans ces cas.

(2021). Dans Chapitre 6, les inégalités de concentration sont utilisées pour contrôler l'erreur de reconstruction liée à la décomposition en composantes principales (ACP) d'un élément aléatoire extrême dans un espace de Hilbert, ainsi que pour encadrer la déviation maximale entre un risque de régression extrême et son équivalent empirique dans Chapitre 8. Plus de détails sur les inégalités de concentration peuvent être trouvés dans Chapitre 4.

Dans le reste de cette section, nous approfondissons deux axes de recherche particulièrement actifs à l'intersection de l'apprentissage statistique et de l'EVT. Section 10.2.1 discute des approches fonctionnelles pour l'EVT, en couvrant spécifiquement la théorie générale dans les espaces métriques généraux et dans l'espace des fonctions continues sur $[0,1]$. Dans Section 10.2.2, des techniques de réduction de dimension, telles que le clustering ou l'ACP, pour les extrêmes sont présentées, avec un focus particulier sur la détection d'anomalies. Enfin, une dernière section examine les recherches existantes dans un domaine clé d'application de l'EVT : la modélisation des niveaux de mer extrêmes et l'estimation cruciale des périodes de retour.

10.2.1 Extrêmes fonctionnels

L'omniprésence des capteurs fournissant des mesures de plus en plus précises et massives de quantités dépendant du temps ou de l'espace a mis en évidence l'importance de comprendre les données continues, connues sous le nom de données fonctionnelles. L'analyse des données fonctionnelles (Functional Data Analysis, FDA) est une branche des statistiques qui étudie les données de dimension infinie et suscite l'intérêt de la recherche depuis de nombreuses années. Les livres Hsing and Eubank (2015), Horváth and Kokoszka (2012) et Ramsay and Silverman (2005) offrent une vue d'ensemble complète de ce domaine, allant des bases théoriques aux diverses applications de la FDA. La disponibilité croissante de données de nature fonctionnelle ouvre de nouvelles voies de recherche, notamment l'étude des extrêmes fonctionnels. Cette thématique est un domaine bien établi et actif en statistique spatiale, comme le souligne la récente synthèse de Huser and Wadsworth (2022).

La plupart des études existantes sur les extrêmes fonctionnels se concentrent sur le cas continu, dans la lignée des travaux fondateurs sur les processus max-stables (De Haan (1984); De Haan and Ferreira (2006)) : les objets aléatoires étudiés sont des fonctions aléatoires dans l'espace $\mathcal{C}[0,1]$, *i.e.*, l'espace des fonctions continues sur $[0,1]$ muni de la norme du supremum. Dans le cadre des Dépassement au dessus d'un Seuil (Peaks-over-Threshold, PoT), l'intérêt se porte sur la distribution limite des observations normalisées, conditionnellement au fait que leur norme dépasse un seuil, lorsque ce seuil tend vers l'infini. L'extrémalité d'une observation est mesurée par sa norme du supremum. Le processus limite résultant dans ce contexte est un processus de Pareto généralisé (voir par exemple Ferreira and de Haan (2014)). Contrairement à la dimension finie, la définition des extrêmes dans les espaces de dimension infinie nécessite de choisir une norme spécifique en raison de la non-équivalence des normes. Ce choix revêt une importance pratique significative ; par exemple, pour évaluer les risques d'inondation, il peut être plus pertinent d'analyser les précipitations totales quotidiennes plutôt que les précipitations maximales journalières sur une courte période. Cet important choix de la norme motive les recherches de Dombry and Ribatet (2015), qui proposent des définitions alternatives des événements extrêmes via une fonction de coût homogène, conduisant à la naissance des processus r-Pareto. Des détails supplémentaires et des définitions précises sur les extrêmes dans $\mathcal{C}[0,1]$ sont

fournis dans Section 2.2.2.

Quelques exceptions au cas continu existent. Par exemple, l'espace fonctionnel de Skorokhod $\mathbb{D}[0,1]$ muni de la topologie J_1 a été étudié dans plusieurs travaux (voir Davis and Mikosch (2008); Hult and Lindskog (2005) et les références citées), et les fonctions semi-continues supérieures munies de la topologie de Fell sont examinées dans Resnick and Roy (1991); Molchanov and Strokorb (2016); Sabourin and Segers (2017); Samorodnitsky and Wang (2019).

Une hypothèse classique en EVT adaptée au cadre PoT consiste à supposer que la RV de la variable aléatoire observée X, c'est-à-dire que la distribution normalisée $t^{-1}X$, conditionnellement au dépassement de sa norme au-delà d'un seuil $\|X\| \geq t$, converge vers une certaine variable aléatoire limite X_∞ lorsque le seuil tend vers l'infini, *i.e.*, $\mathcal{L}(t^{-1}X \mid \|X\| \geq t) \to \mathcal{L}(X_\infty)$ lorsque $t \to +\infty$ (voir les livres Resnick (1987, 2007) pour une présentation exhaustive de la variation régulière dans le cas multivarié). Hult and Lindskog (2006b) étendent la notion de RV, initialement définie dans un espace euclidien, aux mesures sur des espaces métriques complets et séparables. Dans ce contexte, les auteurs caractérisent la RV d'un élément aléatoire X par deux conditions : la RV de sa norme $\|X\|$ et la convergence en distribution de son angle $\Theta = \|X\|^{-1}X$ étant donné que $\|X\|$ dépasse un seuil $\|X\| \geq t$ vers un élément angulaire limite Θ_∞ lorsque le seuil tend vers l'infini, $\mathcal{L}(\Theta \mid \|X\| \geq t) \to \mathcal{L}(\Theta_\infty)$ lorsque $t \to +\infty$ (voir par exemple Segers et al. (2017); Davis and Mikosch (2008)).

Alors que la théorie RV a été largement étudiée dans $\mathcal{C}[0,1]$ et repose sur des bases théoriques solides dans les espaces métriques généraux, elle a reçu beaucoup moins d'attention dans $L^2[0,1]$, l'espace des fonctions réelles de carré intégrable sur $[0,1]$, et plus généralement dans les espaces de Hilbert séparables. Nous proposons dans Chapitre 5 de formaliser ce concept grâce aux résultats Hult and Lindskog (2006b).

Un des principaux intérêts de travailler dans un espace de Hilbert séparable réside dans la décomposition en composantes principales d'un élément aléatoire (voir Section 3.3 de Chapitre 3 pour plus de détails). L'analyse des valeurs extrêmes (Extreme Value Analysis, EVA) de l'ACP fonctionnelle avec des fonctions aléatoires à valeurs dans $L^2[0,1]$ a déjà été étudiée dans la littérature, mais sous des perspectives assez différentes, laissant certaines questions sans réponse. Dans Kokoszka and Xiong (2018), les auteurs supposent la RV des scores d'une décomposition en composantes principales (*i.e.*, les coordonnées aléatoires des observations projetées sur une famille orthogonale de $L^2[0,1]$) et examinent le comportement extrême de leurs équivalents empiriques. Dans Kokoszka et al. (2019) et Kokoszka and Kulik (2023), la RV est supposée, et divers résultats de convergence concernant les opérateurs de covariance empiriques de la fonction aléatoire X (et non de sa composante angulaire Θ) sont établis, sous la condition que l'indice de RV appartienne à un intervalle restreint, respectivement $2 < \alpha < 4$ et $0 < \alpha < 2$. Dans Kim and Kokoszka (2022), la dépendance extrémale entre les scores de la PCA fonctionnelle de X est étudiée. Ils démontrent à cette occasion que la RV dans $L^2[0,1]$ implique la RV multivariée des projections de dimension finie de X. Cependant, la réciproque de cette affirmation conditionnelle n'est pas examinée. Kim and Kokoszka (2024) généralisent la notion de coefficient de corrélation pour les extrêmes fonctionnels.

Les travaux mentionnés impliquent la PCA des extrêmes des fonctions aléatoires à valeurs dans $L^2[0,1]$, d'une manière ou d'une autre, mais il y a eu peu d'étude de la décomposition en composantes principales d'un élément à variation régulière. Dans

Chapitre 6, sous l'hypothèse de RV de X, nous proposons d'étudier la convergence de l'ACP associée à Θ vers l'ACP de Θ_∞. Ici, la valeur de l'indice de RV est sans importance, car l'ACP que nous considérons est celle de la composante angulaire Θ des fonctions aléatoires.

10.2.2 Réduction de dimension pour extrêmes

Les améliorations des dispositifs d'acquisition de données ont entraîné une augmentation de la disponibilité de mesures massives, ce qui motive le développement des statistiques pour les données fonctionnelles tout en présentant quelques défis. D'une part, la disponibilité accrue des données permet des études plus précises. D'autre part, l'analyse de données de grande dimension pose des difficultés dues à l'identification des parties informatives et aux lourdes exigences computationnelles dans le traitement de ces données dans des tâches complexes d'apprentissage automatique. Cette ambivalence en statistique des grandes dimensions est souvent désignée sous le nom de "malédiction de la dimensionnalité" (Giraud (2021)). Dans des domaines tels que les neurosciences et le traitement d'images, où les dimensions des données peuvent exploser, il devient crucial de réduire la dimension pour ne conserver que l'essence des données.

Ces dernières années, les problèmes d'extrêmes en grande dimension ont suscité un intérêt croissant. Du fait que l'EVA se concentre sur une partie restreinte des données, la taille effective de l'ensemble de données utilisé pour l'inférence peut être relativement limitée, soulignant l'importance des techniques de réduction de dimension adaptées aux contextes extrêmes. Une ligne de recherche active concerne la réduction de dimension non supervisée, pour laquelle diverses méthodes ont été proposées ces dernières années, certaines garnies de garanties statistiques non asymptotiques reposant sur des inégalités de concentration appropriées. Parmi ces stratégies, on peut citer l'identification d'un support parcimonieux pour la distribution limite des observations extrêmes renormalisés (Goix et al. (2017); Simpson et al. (2020); Meyer and Wintenberger (2021); Drees and Sabourin (2021); Cooley and Thibaud (2019); Medina et al. (2021)), la modélisation par la théorie des graphes et l'inférence causale basées sur la notion d'indépendance conditionnelle dans les queues (Hitz and Evans (2016); Segers (2020); Gnecco et al. (2021)), ou encore le clustering (Chautru (2015); Janßen and Wan (2020); Chiapino et al. (2020)), voir également l'article de synthèse Engelke and Ivanovs (2021). Dans ces travaux, la dimension de l'espace de l'échantillon, bien que potentiellement élevée, est finie, et la réduction de dimension constitue une étape clé, voire l'objectif principal, de l'analyse.

L'EVA caractérise le comportement des données extrêmes, qui se situent loin du centre de masse de la distribution. Cela rend les outils d'EVA naturellement adaptés au développement de procédures de détection d'anomalies, car les observations anormales se trouvent également en dehors du centre de masse de la distribution. La réduction de dimension pour les extrêmes vise à découvrir des régions qui capturent l'essence des grandes données. Des algorithmes tels que DAMEX (Goix et al. (2016, 2017)) et CLEF (Chiapino et al. (2020); Chiapino and Sabourin (2016)) identifient des sous-espaces où les composantes du vecteur observé peuvent être grandes ensemble. Une anomalie est ainsi détectée lorsque des points de données ne se trouvent pas dans ces sous-espaces malgré une grande norme. Une autre approche caractérise les observations anormales comme celles se trouvant en dehors des MV-sets extrêmes (qui peuvent être recherchés parmi les espaces résultats des algorithmes CLEF ou DAMEX), qui sont de

petits volumes mais de grandes masses (Thomas et al. (2017)).

Une technique classique de réduction de dimension en traitement du signal consiste à décomposer les données sur une base choisie en fonction du problème, puis à ne conserver que les composantes les plus importantes. Parmi les bases usuelles figurent les bases de Fourier (Exemple 3.4) et les bases d'ondelettes. Le lecteur est invité à consulter le livre complet et facile à lire Mallat (1999) pour plus de détails sur le traitement du signal. Les propriétés bénéfiques de ces bases sont nombreuses et variées, mais choisir une base précise adaptée à un problème peut parfois s'apparenter à chercher une aiguille dans une botte de foin. Ainsi, pour les tâches où la structure des données est linéaire, ou nécessite une réduction de dimension efficace sans perte d'informations importantes, l'ACP peut être particulièrement avantageuse. L'ACP détermine automatiquement un ensemble de composantes orthogonales capturant un maximum de la variance des données, fournissant une représentation simplifiée souvent bien alignée avec la structure sous-jacente des données (voir Mallat (1999) pour une comparaison entre l'ACP, les bases de Fourier et les bases d'ondelettes, ainsi que Hsing and Eubank (2015) ou Section 3.3 pour des notions de base sur l'ACP).

Plusieurs travaux ont appliqué l'ACP à l'EVA au fil des années. Dans les contextes de dimension infinie, des études telles que Kokoszka and Xiong (2018); Kokoszka et al. (2019); Kokoszka and Kulik (2023); Kim and Kokoszka (2022) ont exploré l'ACP pour les extrêmes fonctionnels, mais aucune ne propose une méthode applicative de l'ACP spécifiquement aux données extrêmes. Ces travaux ont déjà été mentionnés et présentés dans la section précédente. À notre connaissance, seuls deux travaux impliquent une ACP pour les extrêmes en dimension finie. Dans Cooley and Thibaud (2019), les auteurs proposent une ACP d'une matrice composée de coefficients de dépendance en queue de distribution par paires d'un vecteur aléatoire à valeurs positives et à variation régulière, résultant d'une transformation d'un vecteur aléatoire à variation régulière à valeurs dans tout l'espace ambiant. Dans Drees and Sabourin (2021), les auteurs étudient les relations entre l'ACP de l'angle aléatoire Θ, la composante angulaire d'un vecteur aléatoire régulièrement variable à valeurs dans $\mathbb{R}^d$, et l'ACP de sa limite extrême Θ_∞, puisque la variation régulière de X implique $\mathcal{L}(\Theta \mid \|X\| \geq t) \to \mathcal{L}(\Theta_\infty)$ lorsque $t \to +\infty$. Un argument clé de leur preuve est que Θ appartient à la sphère unité de $\mathbb{R}^d$, qui est un ensemble compact. En vertu du lemme de Riesz, leurs techniques de preuve ne peuvent pas être étendues aux espaces de dimension infinie, bien que les objets mathématiques impliqués dans cet article soient définis *mutatis mutandis* dans un espace de Hilbert séparable général. L'objectif de Chapitre 6 est d'étendre leurs résultats aux espaces non finis-dimensionnels en contournant l'argument de compacité en prouvant que la structure propre de Θ converge vers la structure propre de Θ_∞ sous l'hypothèse de variation régulière de l'élément aléatoire X dans un espace de Hilbert séparable.

10.2.3 Théorie des valeurs extrêmes pour l'étude des niveaux de mer

Les niveaux de mer peuvent être décomposés en une composante déterministe liée aux marées et une composante stochastique, correspondant aux surcotes. Les surcotes sont définies comme les différences instantanées entre les marées astronomiques prédites et les niveaux de mer observés. Les surcotes importantes sont causées par des pressions atmosphériques basses et des vents forts (en intensité ou en direction). Lorsque ces conditions météorologiques coïncident avec les niveaux élevés des marées de vive-eau, elles peuvent entraîner des inondations dévastatrices. Un exemple notable est l'inondation

de la mer du Nord de 1953, connue aux Pays-Bas sous le nom de *Watersnoodramp*, qui a causé plus de 2000 morts en Europe du Nord (McRobie et al. (2005)). Étudier ces événements pour en déduire leur intensité et leur fréquence constitue donc un défi crucial pour la surveillance des risques côtiers, afin de prévenir des pertes humaines et matérielles importantes (Genovese and Przyluski (2013); Chadenas et al. (2014); Karamouz et al. (2019)). Cette tâche est d'autant plus importante compte tenu du réchauffement climatique, qui augmente à la fois la fréquence et l'amplitude de ces événements extrêmes (voir Seneviratne et al. (2021)).

L'étude des niveaux de mer extrêmes est un domaine de recherche actif depuis des décennies. Un concept central dans ce domaine est l'inférence des niveaux de retour, qui correspondent aux niveaux maximaux attendus sur une période donnée (l'"inverse" de la période de retour, tel que détaillé dans Coles et al. (2001)). Deux études pionnières dans ce domaine sont Lennon et al. (1963) et Suthons (1963), qui utilisent les méthodes des maxima annuels pour les estimer. Comme s'appuyer uniquement sur les maxima annuels limite la quantité de données utilisables, de nouvelles méthodes pour les extrêmes ont été développées. Smith (1986) et Tawn (1988) introduisent l'utilisation des r-maxima annuels, tandis que Davison and Smith (1990) sont les premiers à considérer les dépassements d'un seuil comme des extrêmes. Ces études utilisent des méthodes directes, qui consistent à analyser directement les niveaux marins sans prendre en compte leur structure en composantes déterministe et stochastique.

Les méthodes indirectes consistent à analyser séparément les composantes de marée et de surcote. Ces méthodes sont souvent préférées aux approches directes car elles nécessitent moins de données pour mener efficacement une étude des valeurs extrêmes. Les méthodes de convolution, par exemple, permettent de considérer les niveaux de mer extrêmes en combinant les surcotes extrêmes avec les niveaux marins extrêmes. Comme souligné dans Dixon and Tawn (1999), les méthodes directes peuvent introduire des erreurs supplémentaires d'estimation. Des travaux précurseurs dans ce domaine incluent Pugh and Vassie (1978) et Pugh and Vassie (1980), qui ont introduit la méthode des probabilités conjointes pour combiner les surcotes aux niveaux de mer par convolution. Cependant, ces études supposaient que les surcotes horaires étaient indépendantes, une hypothèse jugée irréaliste par Tawn et al. (1989). Pour résoudre ce problème, Tawn (1992) ont proposé une méthode révisée des probabilités conjointes, intégrant l'indice extrémal (Leadbetter (1982)) pour tenir compte de la dépendance temporelle. Pour une comparaison complète entre les méthodes directes et indirectes, voir Haigh et al. (2010).

Modéliser la dépendance marée-surcote dans les méthodes indirectes peut s'avérer difficile (Idier et al. (2012)). Par conséquent, les surcotes de pleine mer, définies comme la différence entre les niveaux de mer maximaux observés pendant une marée et les niveaux de mer astronomiques maximaux pendant cette même marée, sont souvent utilisées à la place. Cette approche présente l'avantage que les hautes marées n'ont généralement pas d'impact sur les surcotes de pleine mer (voir Williams et al. (2016)). Dans cette optique, Batstone et al. (2013) ont proposé la méthode des probabilités conjointes des surcotes de pleine mer pour contourner la modélisation de l'interaction marée-surcote. Notons que l'indépendance entre la marée haute et la surcote de pleine mer a été empiriquement prouvée pour la plupart des stations côtières françaises, à l'exception de la station de Saint-Malo (voir Kergadallan et al. (2014); Kergadallan (2022)).

Toutes les méthodes mentionnées ci-dessus sont appliquées individuellement à chaque station de mesure, appelée *marégraphe*, ignorant la dépendance spatiale entre les stations. Cela peut représenter une limitation importante, car la survenue d'un événement extrême en un lieu augmente la probabilité d'un autre événement extrême à une station voisine. La communauté des valeurs extrêmes s'intéresse depuis longtemps à la modélisation de la structure de dépendance multivariée. La littérature dans ce domaine se concentre généralement sur des modèles pour des données asymptotiquement dépendantes ou indépendantes. La dépendance asymptotique est évaluée à l'aide de la mesure de dépendance (Coles et al. (1999)). De manière générale, les extrêmes dans des régimes asymptotiquement indépendants ont tendance à se produire séparément, tandis que les extrêmes dans des régimes asymptotiquement dépendants ont tendance à se produire simultanément. Pour les données asymptotiquement dépendantes, le travail fondateur est le célèbre modèle conditionnel de Heffernan and Tawn (2004), qui caractérise la distribution d'un vecteur aléatoire étant donné qu'une de ses composantes est extrême. Ce travail a été affiné au fil des ans, donnant lieu à de nombreux modèles conditionnels, tels que ceux proposés par Keef et al. (2013), Tawn et al. (2018) et Shooter et al. (2021).

Bien que certains des modèles mentionnés ci-dessus s'appliquent également aux données asymptotiquement indépendantes, d'autres modèles sont mieux adaptés pour capturer des connexions fortes entre les composantes, comme le modèle hiérarchique max-stable de Reich and Shaby (2012), le processus généralisé de Pareto de Ferreira and de Haan (2014), et la distribution multivariée généralisée de Pareto de Rootzén and Tajvidi (2006) (voir aussi Kiriliouk et al. (2019); Rootzén et al. (2018)). Cette liste de modèles de valeurs extrêmes multivariées n'est pas exhaustive, compte tenu du nombre important de modèles existants. Des références supplémentaires incluent Davison et al. (2012); Huser and Wadsworth (2022) pour les avancées dans les extrêmes spatiaux, Engelke and Ivanovs (2021) pour les structures parcimonieuses, Hao et al. (2018) pour les extrêmes composés, et de Carvalho and Ramos (2012) pour les données bivariées asymptotiquement indépendantes.

10.3 Résumé des contributions

Cette section vise à résumer les principaux résultats de la thèse, en laissant la motivation et la mise en contexte à Section 10.2 ainsi qu'aux sections introductives de chaque chapitre.

Chapitre 5 se concentre sur la caractérisation de la variation régulière dans un espace de Hilbert. Chapitre 6 s'appuie sur le formalisme introduit en Chapitre 5 pour établir la consistance et fournir des garanties statistiques pour l'ACP d'éléments à variation régulière dans un espace de Hilbert. Les travaux des chapitres 5 et 6 ont été publiés dans le journal *Stochastic Processes and their Applications* (voir Clémençon et al. (2024)). Les principaux résultats de ces recherches sont résumés dans les sections 10.3.1 et 10.3.2.

Chapitre 7 propose un nouveau cadre de variation régulière, appelé variation régulière par rapport à une composante, essentiel à la formalisation d'un cadre de régression pour les extrêmes. Dans Chapitre 8, nous développons des résultats pour la régression en contexte extrême, prouvant l'optimalité d'une fonction de régression dépendant uniquement de l'angle des covariables, la consistance de cet estimateur, ainsi que des garanties statistiques sur l'erreur associée à cet estimateur. Les travaux des chapitres 7 et 8 ont fait l'objet d'une pré-publication Huet et al. (2023), actuellement en cours

d'examen dans une revue évaluée par les pairs. Un résumé des principaux résultats est fourni aux sections 10.3.3 et 10.3.4.

Partie IV propose deux approches pour la prédiction des niveaux de mer extrêmes : la première repose sur le cadre de régression développé dans Chapitre 8, et la seconde sur un modèle de Pareto multivariée généralisé. Les travaux de Chapitre 9 font l'objet d'une soumission en cours dans une revue scientifique et sont résumés en Section 10.3.5.

10.3.1 Variation régulière dans un espace de Hilbert

Chapitre 5 a pour objectif principal de développer un cadre probabiliste général pour les extrêmes d'éléments à variation régulière dans un espace de Hilbert séparable $\mathbb{H}$, tel que l'espace $L^2[0,1]$, qui est l'espace de Hilbert des fonctions à valeurs réelles, de carré intégrables sur $[0,1]$. Ce cadre peut être immédiatement généralisé à d'autres domaines compacts, tels que des domaines spatiaux.

Dans ce travail, nous nous plaçons dans le contexte général de la variation régulière (RV) défini par la convergence $\mathbb{M}_0$ introduite dans Hult and Lindskog (2006b), et nous concentrons notre analyse sur des fonctions aléatoires à valeurs dans l'espace de Hilbert $L^2[0,1]$. Cet espace a reçu beaucoup moins d'attention dans la théorie des valeurs extrêmes (EVT) par rapport aux espaces de fonctions continues, semi-continues ou *càdlàg*. L'un des principaux avantages du cadre proposé, en plus de permettre des fonctions discontinues, est de préparer le terrain pour une réduction de dimension des observations par l'intermédiaire de l'ACP fonctionnelle appliquée à la composante *angulaire* Θ (voir Chapitre 6).

Plusieurs questions se posent dans ce contexte. Tout d'abord, dans le cas d'observations fonctionnelles, le choix de la norme (et donc de l'espace fonctionnel) n'est pas anodin, car toutes les normes ne sont pas équivalentes. Par exemple, il n'y a aucune raison pour que la variation régulière dans un espace fonctionnel (comme $\mathcal{C}[0,1]$) soit équivalente à celle dans un espace plus grand tel que $L^2[0,1]$. Par ailleurs, un problème récurrent dans le contexte de la convergence faible des processus stochastiques est de vérifier les conditions de tension, en plus de la convergence faible des projections finies, afin d'assurer la convergence faible du processus dans son ensemble. Les variables aléatoires à valeurs dans un espace de Hilbert ne font pas exception à cette règle (voir par exemple Chapitre 1.8 dans van der Vaart and Wellner (1996)). Une question naturelle à poser est alors : "Quelles conditions concrètes sur les composantes angulaire et radiale $(\Theta, |X|)$, dans un cadre PoT, peuvent être vérifiées dans des exemples génératifs spécifiques ou même sur des données réelles, et sont suffisantes pour assurer la tension et donc la RV globale ?"

Pour répondre à ces questions, nous proposons une description complète de la notion de RV dans un espace de Hilbert séparable, s'inscrivant dans le cadre de Hult and Lindskog (2006b). Plus précisément, nous proposons des caractérisations de la RV impliquant des conditions sur les projections finies et les moments de la variable angulaire Θ via le résultat important suivant (Théorème 5.8 dans Chapitre 5).

Theorem. *Soit X un élément aléatoire dans $\mathbb{H}$ et soit Θ_t un élément aléatoire dans $\mathbb{H}$ distribué sur la sphère $\mathbb{S}$ selon la loi conditionnelle $P_{\Theta,t} := \mathcal{L}(X/\|X\| \mid \|X\| \geq t)$. Soit $P_{\Theta,\infty}$ une mesure de probabilité sur $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$ et soit Θ_∞ un élément aléatoire distribué selon $P_{\Theta,\infty}$. Les énoncés suivants sont équivalents*

1. *X est à variation régulière avec indice α et limite angulaire limite $P_{\Theta,\infty}$, telle que $P_{\Theta,t} \xrightarrow{w} P_{\Theta,\infty}$.*
2. *$\|X\|$ est à variation régulière dans $\mathbb{R}$ avec indice α, et*

$$\forall h \in \mathbb{H}, \langle \Theta_t, h \rangle \xrightarrow{w} \langle \Theta_\infty, h \rangle \quad \text{as } t \to +\infty.$$

3. *$\|X\|$ est à variation régulière dans $\mathbb{R}$ avec indice α, et*

$$\forall N \geq 1, \pi_N(\Theta_t) \xrightarrow{w} \pi_N(\Theta_\infty) \quad \text{as } t \to +\infty,$$

avec $\pi_N : \mathbb{H} \to \mathbb{R}^N$ la projection sur les N premiers éléments de la base $(e_i)_{i\geq 1}$.

Pour valider ces résultats, nous fournissons plusieurs exemples d'éléments aléatoires à variation régulière dans $\mathbb{H}$, tels que les sommes aléatoires $\sum_{i=1}^{D} Z_i A_i$, où les Z_i sont des variables aléatoires réelles à variation régulière, les A_i sont des éléments aléatoires dans $\mathbb{H}$ et D est une constante ou une variable aléatoire d'espérance finie (propositions 5.1 et 5.2). Nous soulignons la nécessité des conditions de tension pour obtenir une RV globale, en construisant un élément aléatoire dans $\mathbb{H}$ ayant des projections finies et une norme à variation régulière, mais qui n'est pas à variation régulière dans $\mathbb{H}$ (Proposition 5.4).

Dans la section finale, nous discutons des relations entre la RV dans $\mathcal{C}[0,1]$ et dans $L^2[0,1]$. Nous démontrons que la RV dans $\mathcal{C}[0,1]$ implique la RV dans $L^2[0,1]$ et que les variables aléatoires limites dans ces deux cadres peuvent être reliées par une formule explicite (résultats de Dombry and Ribatet (2015), Proposition 5.9). L'inverse n'est cependant pas vrai : nous construisons un exemple d'une fonction aléatoire à variation régulière dans $L^2[0,1]$ qui n'est pas à variation régulière dans $\mathcal{C}[0,1]$ (Proposition 5.10).

10.3.2 ACP pour extrêmes fonctionnels

Dans Chapitre 5, une caractéristique majeure du cadre proposé est la possibilité de projeter les observations sur un espace fonctionnel de dimension finie via une modification de l'ACP fonctionnelle standard, adaptée pour traiter des observations à queues lourdes, pour lesquelles les moments d'ordre supérieur (ou même d'ordre premier) peuvent ne pas exister. Cette technique de réduction de dimension étend le travail de Drees and Sabourin (2021), qui a appliqué l'ACP dans en dimension finie et a dérivé des garanties non-asymptotique pour les sous-espaces propres de l'opérateur de covariance empirique pour Θ. Cependant, les techniques utilisées par Drees and Sabourin (2021) ne peuvent pas être directement appliquées ici, car elles reposent sur la compacité de la sphère unité dans $\mathbb{R}^d$, tandis que la sphère unité dans un espace de Hilbert de dimension infinie n'est pas compacte.

L'extension naturelle de la matrice de covariance des angles extrêmes $C_{t,\mathbb{R}^d} = \mathbb{E}[\Theta\Theta^\top \mid \|X\| > t]$ dans Drees and Sabourin (2021), lorsque $X \in \mathbb{R}^d$, est l'opérateur de covariance $C_t = \mathbb{E}[\Theta \otimes \Theta \mid \|X\| > t]$ pour $X \in \mathbb{H}$, comme expliqué dans les sections 3.1 et 3.3 du Chapitre 3. Sous l'hypothèse de RV pour X, où la distribution angulaire de Θ converge, *i.e.*, $P_{\Theta,t} := \mathcal{L}(\Theta \mid \|X\| > t) \to \mathcal{L}(\Theta_\infty) = P_{\Theta,\infty}$, une question naturelle se pose : la structure propre de C_t converge-t-elle quand $t \to +\infty$ vers celle de $C_\infty = \mathbb{E}[\Theta_\infty \otimes \Theta_\infty]$, où $\Theta_\infty \sim P_{\Theta,\infty}$? De plus, les résultats de concentration pour les sous-espaces propres empiriques dans Drees and Sabourin (2021) peuvent-ils être étendus au cadre des espaces de Hilbert de dimension infinie ?

Theorem. *La convergence suivante en norme Hilbert-Schmidt est vérifiée,*

$$\|C_t - C_\infty\|_{HS(\mathbb{H})} \to 0,$$

quand $t \to +\infty$.

En utilisant le Théorème 3 de Zwald and Blanchard (2005) (Théorème 3.19) et le théorème de Weyl (Théorème 3.11), nous prouvons que les valeurs propres et les sous-espaces propres de C_t convergent vers ceux de C_∞ lorsque $t \to +\infty$ (Corollaire 6.3).

Ensuite, nous étudions la convergence de l'opérateur de covariance empirique associé à la distribution de $P_{\Theta,t}$. Supposons que nous observons $n \geq 1$ réalisations indépendantes $X_1, ..., X_n$ de la fonction aléatoire X. Nous cherchons à estimer l'opérateur de covariance sous-asymptotique associé à un seuil radial $t_{n,k}$, qui est un quantile de la variable radiale $\|X\|$ au niveau $1 - k/n$, défini par:

$$\widehat{C}_k := \frac{1}{k} \sum_{i=1}^{n} \Theta_i \otimes \Theta_i \mathbb{1}\{\|X_i\| \geq \hat{t}_{n,k}\},$$

où $\hat{t}_{n,k}$ est la $k$ème plus grande norme parmi les $\|X_i\|$'s, *i.e.*, la version empirique de $t_{n,k}$. Nous fournissons des garanties statistiques sous forme d'inégalités de concentration concernant la norme de Hilbert-Schmidt de l'erreur d'estimation. Les termes principaux des bornes impliquent le nombre $k \leq n$ des statistiques d'ordre extrême utilisées pour calculer l'estimateur. Plus précisément, nous présentons le résultat suivant (Théorème 6.8 du Chapitre 6)

Theorem. *Soit* $\delta \in (0, 1)$. *Avec probabilité plus grande que* $1 - \delta$, *on a*

$$\|\widehat{C}_k - C_{t_{n,k}}\|_{HS(\mathbb{H})} \leq C(\delta)/\sqrt{k} + +o(1/\sqrt{k}),$$

avec $C(\delta)$ *une constante dépendent seulement de* δ.

Ces bornes, combinées avec la variation régulière de la fonction aléatoire observée X, assurent la consistance de la procédure d'estimation empirique, comme indiqué dans Corollaire 6.10 du Chapitre 6.

Enfin, nous présentons des résultats expérimentaux sur des données réelles et simulées. Plus précisément, nous analysons un ensemble de données sur la demande électrique et des données simulées, comme détaillé dans Chapitre 5. Ces expériences montrent la pertinence du cadre de réduction de dimension proposé, en comparant ses performances avec l'ACP standard appliquée à l'échantillon complet (non limité aux observations extrêmes). Les résultats mettent en évidence l'utilité du cadre proposé dans les applications pratiques, en particulier lorsqu'il s'agit de données à queues lourdes.

10.3.3 Un cadre de variation régulière pour la régression dans les extrêmes

Dans Chapitre 7, nous introduisons un cadre probabiliste, en particulier la variation régulière par rapport à une composante, pour la régression dans les extrêmes extrêmes. Nous proposons également une approche algorithmique dédiée, qui est analysée plus en détail dans Chapitre 8.

Pour motiver l'analyse qui suit, l'algorithme Regression On eXtreme ANglEs (ROXANE) est introduit au début de Partie III. Cette méthode traite du problème de régression pour la paire entrée/sortie $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ dans les régions extrêmes, spécifiquement lorsque $\|\mathbf{X}\| \gg 1$. L'objectif principal de l'algorithme est de minimiser un risque quadratique extrême empirique en utilisant uniquement l'angle de la variable d'entrée $\mathbf{X}/\|\mathbf{X}\|$. Cela est réalisé sans perte d'information sous une hypothèse spécifique de variation régulière (RV), comme justifié dans Chapitre 8. Pour simplifier, nous supposons que la sortie est bornée, *i.e.*, il existe un $M > 0$ tel que $Y \in I := [-M, M]$.

Notre hypothèse centrale, qui justifie l'algorithme ROXANE, est la variation régulière par rapport à la première composante de $(\mathbf{X}, Y)$. Cette hypothèse de RV modifiée, où l'extrémalité du vecteur aléatoire est définie uniquement par rapport à la variable d'entrée, est détaillée dans la définition suivante (Hypothèse 7.2 dans Chapitre 7). Soit $E := \mathbb{R}^d \backslash \{\mathbf{0}\}$.

Definition (Variation régulière par rapport à la première composante)**.** *Un vecteur aléatoire $(\mathbf{X}, Y) \in \mathbb{O} := E \times I$ est à variation régulière par rapport à la première composante avec indice $\alpha > 0$, si il existe une fonction à variation régulière b avec indice α et une mesure Borélienne non nulle μ sur $\mathbb{O}$, sur tout Borel de $\mathbb{O}$ à distance positive de $\mathbb{C} = \{\mathbf{0}\} \times I$, tels que*

$$\lim_{t \to +\infty} b(t) \mathbb{P}\Big(t^{-1}\mathbf{X} \in A, Y \in C\Big) = \mu(A \times C),$$

pour tout $A \in \mathcal{B}(E)$ à distance positive de zéro et $C \in \mathcal{B}(I)$ tels que $\mu(\partial(A \times C)) = 0$.

Cette hypothèse est un cas particulier de la théorie développée dans Lindskog et al. (2014) pour les mesures à variation régulière sur des espaces métriques séparables avec un ensemble fermé $\mathbb{C}$ retiré ; dans notre contexte, $\mathbb{C} = \{\mathbf{0}\} \times I$. Nous clarifions cette connexion dans Section 7.5 avec des énoncés équivalents de RV par rapport à la première composante. Les implications similaires à celles de la RV classique sont démontrées après Hypothèse 7.2, telles que l'homogénéité d'ordre $-\alpha$ de la mesure limite μ par rapport à la première composante, ce qui mène à une décomposition de la mesure :

$$\mu(\{\mathbf{X} \in E : \|\mathbf{x}\| \geq r, \boldsymbol{\theta}(\mathbf{x}) \in B\} \times C) = r^{-\alpha} \Phi(B \times C),$$

avec $\boldsymbol{\theta}(\mathbf{x}) := \mathbf{x}/\|\mathbf{x}\|$ et pour tout $C \in \mathcal{B}(I), B \in \mathcal{B}(\mathbb{S})$, $r > 0$. Cela engendre l'existence d'une paire de variables aléatoires limite $(\mathbf{X}_\infty, Y_\infty)$

$$\mathcal{L}(t^{-1}\mathbf{X}, Y \mid \|\mathbf{X}\| \geq t) \to \mathcal{L}(\mathbf{X}_\infty, Y_\infty),$$

lorsque $t \to +\infty$. Nous supposons également la convergence de la fonction de régression de Bayes $f^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ vers la fonction limite de régression de Bayes $\mathbb{E}[Y_\infty \mid \mathbf{X}_\infty]$ qui satisfait (Hypothèse 7.5)

$$\mathbb{E}\Big[|f^*(\mathbf{X}) - f^*_{P_\infty}(\mathbf{X})| \mid \|\mathbf{X}\| \geq t\Big] \to 0. \tag{10.1}$$

Trois conditions impliquant cette hypothèse, telles que la convergence uniforme des densités proposées dans De Haan and Resnick (1987), sont fournies pour soutenir la validité d'Équation (10.1) (Proposition 7.6 dans Chapitre 7).

Enfin, nous proposons quatre scénarios pratiques qui satisfont toutes les hypothèses de Section 7.4, y compris un exemple adapté à la prédiction d'une composante extrême manquante dans un vecteur aléatoire à variation régulière (Proposition 7.10 dans Chapitre 7).

10.3.4 Régression dans les extrêmes

Chapitre 8 a pour objectif de développer un cadre de régression dans des régions où les covariables extrêmes, en s'appuyant sur les hypothèses discutées en Chapitre 7, telles que la RV par rapport à la première composante du couple d'entrée/sortie $(\mathbf{X}, Y)$. L'objectif fondamental est de justifier l'algorithme ROXANE, *i.e.*, de prouver qu'une fonction de régression peut être construite de manière optimale dans les régions extrêmes, en utilisant uniquement l'angle de la variable d'entrée.

La régression est un problème prédictif essentiel en apprentissage statistique, couvrant une grande variété d'applications. Dans le cadre classique, le problème d'apprentissage prédictif consiste à construire, à partir d'un ensemble de données d'apprentissage $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), ; \ldots, ; (\mathbf{X}_n, Y_n)\}$, composé de $n \geq 1$ copies indépendantes de deux variables aléatoires $(\mathbf{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$, une fonction $f : \mathcal{X} \to \mathbb{R}$ permettant de produire une "bonne" prédiction $f(\mathbf{X})$ de Y, en minimisant le risque quadratique

$$R_P(f) = \mathbb{E}\Big[(Y - f(\mathbf{X}))^2\Big] \tag{10.2}$$

aussi proche que possible de la fonction de régression de Bayes $f^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$, qui minimise (10.2).

Une stratégie naturelle consiste à résoudre le problème de minimisation du risque empirique (Empirical Risk Minimization, ERM) $\min_{f \in \mathcal{F}} R_{\hat{P}_n}(f)$, où $\mathcal{F}$ est une classe de fonctions suffisamment riche pour inclure une bonne approximation de f^* et $\hat{P}_n$ est une version empirique de P basée sur $\mathcal{D}_n$.

Ce chapitre s'intéresse à la régression dans les régions extrêmes, lorsque la variable d'entrée est extrême. Les covariables sont considérées comme extrêmes lorsque leur norme $\|\mathbf{X}\|$ dépasse un seuil (asymptotiquement) grand $t > 0$ (voir Chapitre 7). Le choix de la norme dépend généralement du contexte applicatif.

Le seuil t dépend des observations, car "grand" doit être compris relativement à la majorité des données observées. Par conséquent, les observations extrêmes sont rares et sous-représentées dans l'ensemble d'apprentissage, ce qui signifie que les erreurs de prédiction dans les régions extrêmes ont généralement un impact négligeable sur l'erreur globale de régression de $\hat{f}$. En effet, la loi des probabilités totales donne :

$$R_P(f) = \mathbb{P}(\|\mathbf{X}\| \geq t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big] + \mathbb{P}(\|\mathbf{X}\| < t)\mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| < t\Big].$$

Cette décomposition met en évidence un terme d'erreur conditionnelle relatif aux dépassements de $\|\mathbf{X}\|$ au-delà de t, que nous appelons le *risque quadratique conditionnel* (ou simplement risque conditionnel) :

$$R_t(f) := \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big].$$

Le but de l'analyse qui suit est de construire une fonction prédictive $\hat{f}$ qui minimise (approximativement) $R_t(f)$ pour tout $t > t_0$, où t_0 est un seuil élevé. Il est important de noter qu'un estimateur du minimiseur de R_t peut ne pas être adapté pour minimiser $R_{t'}$ lorsque $t' > t$. Pour garantir des performances robustes d'extrapolation, nous visons à obtenir une fonction prédictive $\hat{f}$ qui minimise le *risque quadratique conditionnel asymptotique* défini par :

$$R_\infty(f) := \limsup_{t \to +\infty} R_t(f) = \limsup_{t \to +\infty} \mathbb{E}\Big[(Y - f(\mathbf{X}))^2 \mid \|\mathbf{X}\| \geq t\Big].$$

Ainsi, l'objectif est d'établir des liens entre les risques R_t et R_∞ ainsi que leurs minimiseurs respectifs, en s'appuyant sur les propriétés avantageuses de l'hypothèse de RV. Le théorème suivant (Théorème 8.2 dans Chapitre 8) fournit les premières motivations pour l'algorithme ROXANE. Soit $\boldsymbol{\theta}(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$.

Theorem. *Sous les hypothèses 7.1, 7.2 et 7.5, les deux assertions suivantes sont vraies :*

1. *Quand $t \to +\infty$, la valeur minimale de R_t converge vers celle de R_∞, i.e.,* $\inf_f R_t(f) \to \inf_f R_\infty(f)$.

2. *L'infimum de R_∞ sur toutes les fonctions mesurables est égal à son infimum sur toutes les fonctions mesurables angulaires, i.e.,* $\inf_f R_\infty(f) = \inf_h R_\infty(h \circ \boldsymbol{\theta})$.

L'utilisation de la RV par rapport à la première composante est cruciale pour prouver ce résultat. Ce concept clé permet de relier les deux risques, R_t et R_∞, et montre la nature angulaire d'un minimiseur de R_∞. Par conséquent, il est raisonnable de restreindre la recherche d'un minimiseur dans les régions extrêmes aux fonctions angulaires, comme le suggère l'algorithme ROXANE.

Notre stratégie consiste à résoudre le problème d'optimisation ERM associé à $\min_{h\in\mathcal{H}} R_{t_{n,k}}(h\circ\boldsymbol{\theta})$, où $t_{n,k}$ est un quantile de la variable radiale $|\mathbf{X}|$ au niveau $1-k/n$. Pour ce faire, nous considérons sa version empirique :

$$\hat{R}_k(f) = \frac{1}{k}\sum_{i=1}^{n}(Y_i - f(\mathbf{X}_i))^2 \mathbb{1}\{\|\mathbf{X}_i\| \geq \hat{t}_{n,k}\},$$

où $\hat{t}_{n,k}$ est la k-ième plus grande norme parmi les $|\mathbf{X}i|$'s, servant de contrepartie empirique du quantile tn,k. Comme pour l'ERM classique, nous analysons la minimisation du risque empirique sur une classe de fonctions à complexité contrôlée. Soit $\mathcal{H}$ une classe de fonctions continues, réelles, angulaires et uniformément bornées par M: $f \in \mathcal{C}(\mathbb{S}, I)$. Pour valider pleinement la stratégie empirique de l'algorithme ROXANE, nous fournissons une borne non asymptotique sur la déviation maximale entre $R_{t_{n,k}}$ et $\hat{R}_k$ sur $\mathcal{H}$, comme indiqué dans le théorème suivant (Théorème 8.4 dans Chapitre 8).

Theorem. *Supposons que les hypothèses 7.1 et 8.3 soient satisfaites. Soit $\delta \in (0,1)$. Avec une probabilité supérieure à $1-\delta$:*

$$\sup_{h\in\mathcal{H}}\left|\hat{R}_k(h\circ\boldsymbol{\theta}) - R_{t_{n,k}}(h\circ\boldsymbol{\theta})\right| \leq C(\mathcal{H}, M, \delta)/\sqrt{k} + o(1/\sqrt{k}),$$

où $C(\mathcal{H}, M, \delta)$ est une constante dépendant de $\mathcal{H}, M$ et δ.

En outre, avec une hypothèse supplémentaire sur la classe $\mathcal{H}$, qui est satisfaite en particulier par les fonctions issues de régression avec pénalisation Ridge ou Lasso contraintes (voir Remarque 8.6 dans Chapitre 8), nous montrons que le biais $\sup_{h\in\mathcal{H}}|R_{t_{n,k}}(h\circ\boldsymbol{\theta}) - R_\infty(h\circ\boldsymbol{\theta})|$ converge vers zéro lorsque $n \to +\infty$ (Proposition 8.5 dans Chapitre 8). Cela conduit à un contrôle maximal de l'excès de risque R_∞ d'une fonction de régression produite par l'algorithme ROXANE (Corollaire 8.8 dans Chapitre 8).

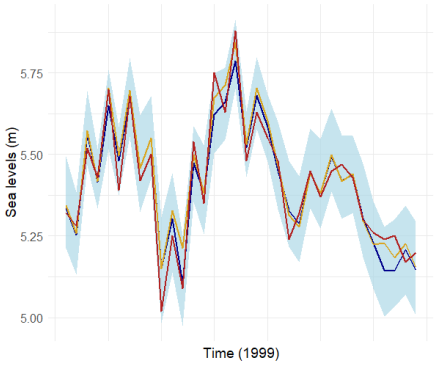


Figure 10.2: Prédictions des niveaux de mer à la station de Port-Tudy pour l'année 1999. La courbe rouge représente les valeurs réelles du jeu de test ; la courbe orange représente les valeurs prédites par la procédure ROXANE avec l'algorithme OLS ; la courbe bleue représente les valeurs prédites par la procédure MGP avec des intervalles de prédictions bootstrap de niveau 95% (bleu clair).

10.3.5 Modélisation et prédiction de niveaux de mer extrêmes

Chapitre 9 se concentre sur la prédiction des niveaux de mer extrêmes et des surcotes aux marégraphes situés le long de la côte atlantique française. Nous proposons d'apprendre la structure de dépendance spatiale des observations extrêmes entre différentes stations sur leur plage temporelle commune. Ce modèle appris est ensuite utilisé pour reconstruire les niveaux de mer et les surcotes aux stations avec des enregistrements historiques limités, à partir des observations extrêmes de stations voisines disposant d'enregistrements temporels plus étendus. En particulier, nous visons à prédire les valeurs à la station de Port-Tudy à partir des valeurs extrêmes mesurées aux stations de Brest et Saint-Nazaire (voir Figure 9.1). Une observation est déclarée extrême si au moins une de ses composantes est extrême, car un seul niveau de la mer élevé ou une surcote importante peut provoquer une inondation à la station, indépendamment des conditions aux autres stations. Nous décrivons deux procédures différentes pour apprendre la structure de dépendance extrême.

Dans la première méthode, nous ajustons une distribution extrême appropriée aux données. Compte tenu de la dépendance asymptotique clairement observée dans les données (voir Figures 9.3 et 9.4), nous modélisons les observations à l'aide d'une distribution *Multivariate Generalized Pareto* (MGP) H (Rootzén and Tajvidi (2006); Rootzén et al. (2018)), définie comme suit :

$$H(\mathbf{x}) = \frac{\log G(\mathbf{x} \wedge \mathbf{0}) - \log G(\mathbf{x})}{\log G(\mathbf{0})},$$

où G est une distribution multivariée de valeurs extrêmes (Définition 2.10). En particulier, nous suivons la procédure de modélisation paramétrique proposée par Kiriliouk et al. (2019), en tirant parti de la décomposition de la distribution MGP (2.14). Les prédictions finales sont obtenues en moyennant des simulations générées à partir de la densité conditionnelle apprise, conditionnellement aux deux valeurs d'entrée.

Dans la seconde approche, nous utilisons la procédure de régression proposée dans Partie III (Huet et al. (2023)), conçue pour des problèmes de prédiction de valeurs extrêmes où l'extrémalité est mesurée par rapport aux covariables — ici, les valeurs des stations disposant de longues séries temporelles. Nous apprenons une fonction de

prédiction à l'aide de l'algorithme ROXANE (Algorithme 7.1) sur la période commune des données, afin de prédire les valeurs aux stations de sortie à partir des valeurs extrêmes aux stations d'entrée.

Les deux procédures nécessitent une mise à l'échelle des observations marginales sur une échelle commune : des échelles exponentielles unitaires pour la procédure MGP et des distributions de Pareto unitaires pour la procédure ROXANE. Conformément à Legrand et al. (2023), nous utilisons une distribution *Extended Generalized Pareto* (EGP) comme modèle marginal pour répondre aux différentes exigences décrites dans l'introduction de Chapitre 9. Plus précisément, nous considérons le modèle EGPD3 de Papastathopoulos and Tawn (2013) dont la fonction de répartition est donnée par :

$$F_{\sigma,\xi,\kappa}(x) = \left(1 - \left(1 + \frac{\xi x}{\sigma}\right)^{-1/\xi}\right)^{\kappa},$$

avec $\sigma > 0$, $\xi \in \mathbb{R}$, $\kappa \in \mathbb{R}$ et $x \in [0, +\infty[$ si $\xi \geq 0$ et $x \in [0, -\sigma/\xi]$ sinon.

Les procédures de prédiction multivariées sont synthétisées dans deux algorithmes, Algorithme 9.2 et Algorithme 9.3. En complément des étapes de prétraitement marginal, l'Algorithme 9.1 introduit une nouvelle méthode pour sélectionner des seuils adaptés dans les études de valeurs extrêmes, basée sur les propriétés de la distribution EGP.

Les méthodes proposées sont appliquées aux données des niveaux de mer, et leurs performances sont évaluées par rapport à la racine carrée de l'erreur quadratique moyenne (RMSE) et d'erreur absolue moyenne (MAE) sur un jeu de données test constitué des premières observations extrêmes. Les deux procédures de prédiction multivariée donnent des résultats concluants et significatifs pour les praticiens, chacune présentant des avantages distincts : l'une fournit de meilleures estimations ponctuelles, tandis que l'autre offre un modèle génératif robuste. En particulier, Figure 10.2 donne une évaluation visuelle de la qualité des prédictions des deux méthodes (Table 9.3) et Figure 9.5 présente des QQ-plots pour une validation supplémentaire. Enfin, des études similaires menées pour les stations de Concarneau et Le Crouesty sont présentées en Annexe 9.A.

10.4 Plan de la thèse

Le manuscrit de thèse est organisé comme suit.

Chapitre 1 fournit un résumé de l'état de l'art ainsi que des contributions de cette thèse.

Partie I introduit les prérequis nécessaires à la compréhension et à la démonstration des résultats de la thèse.

Chapitre 2 traite des notions fondamentales de la théorie des valeurs extrêmes, depuis les extrêmes univariés jusqu'aux extrêmes dans des espaces de dimension infinie, en passant par les extrêmes multivariés.

Chapitre 3 aborde les concepts de l'analyse des données fonctionnelles, y compris les opérateurs et la théorie des probabilités dans les espaces de Hilbert.

Chapitre 4 présente les bases de l'apprentissage statistique, en mettant particulièrement l'accent sur ses applications aux extrêmes.

Partie II concerne les extrêmes hilbertiens.

Chapitre 5 développe la théorie de la variation régulière dans les espaces de Hilbert séparables.

Chapitre 6 utilise les principes de Chapitre 5 pour établir des résultats de consistance et de concentration pour l'analyse en composantes principales d'extrêmes fonctionnels.

Partie III étudie la tâche de régression dans les régions extrêmes.

Chapitre 7 propose un nouveau cadre de variation régulière permettant de traiter les extrêmes par rapport à certaines composantes.

Chapitre 8 exploite le cadre de Chapitre 7 pour développer un nouveau cadre adapté à la régression dans les régions extrêmes.

Partie IV est une application à la reconstruction des niveaux de mer extrêmes.

Chapitre 9 applique la procédure de régression pour les extrêmes de Chapitre 8 ainsi qu'une procédure de modélisation des extrêmes aux données de niveaux de mer extrêmes provenant des marégraphes situés le long de la côte atlantique française.

Le manuscrit se termine par une discussion sur les conclusions générales et les perspectives des résultats développés dans cette thèse, suivie d'une section d'annexes comprenant des démonstrations techniques et d'une introduction en français.

Le contenu de cette thèse repose sur les travaux suivants :

Partie II : Clémençon, S., Huet, N., et Sabourin, A., (2024) Regular Variation in Hilbert Spaces and Principal Component Analysis for Functional Extremes, *Stochastic Processes and their Applications*, 174, 104375 ;

Partie III : Huet, N., Clémençon, S., et Sabourin, A., (2024) On Regression in Extreme Regions, arXiv:2303.03084 (soumis).

Bibliography

A. Aghbalou, P. Bertail, F. Portier, and A. Sabourin. Cross Validation for Extreme Value Analysis. *arXiv:2202.00488*, 2023. pages 15, 61, 124, 174

A. Aghbalou, F. Portier, A. Sabourin, and C. Zhou. Tail Inverse Regression: dimension reduction for prediction of extremes. *Bernoulli*, 30(1):503–533, 2024a. pages 14, 88, 174

A. Aghbalou, A. Sabourin, and F. Portier. Sharp Error Bounds for Imbalanced Classification: How many Examples in the Minority Class? In *AISTATS*, pages 838–846. PMLR, 2024b. pages 15, 174

M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Appl. Math.*, 47(3):207–217, 1993. page 67

S. Arlot. Fondamentaux de l'apprentissage statistique. *Apprentissage statistique et données massives*, 2018. pages 62, 63, 66

A. A. Balkema and L. De Haan. Residual life time at great age. *Ann. Probab.*, 2(5): 792–804, 1974. page 35

B. Basrak, R. A. Davis, and T. Mikosch. A characterization of multivariate regular variation. *Ann. Appl. Probab.*, 12(3):908–920, 2002a. page 77

B. Basrak, R. A. Davis, and T. Mikosch. Regular variation of GARCH processes. *Stochastic Process. Appl.*, 99(1):95–115, 2002b. pages 73, 74

C. Batstone, M. Lawless, J. Tawn, K. Horsburgh, D. Blackman, A. McMillan, D. Worth, S. Laeger, and T. Hunt. A UK best-practice approach for extreme sea-level analysis along complex topographic coastlines. *Ocean Eng.*, 71:28–39, 2013. pages 20, 179

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, 2006. pages 91, 101

P. Bertail, S. Clémençon, Y. Guyonvarch, and N. Noiry. Learning from Biased Data: A Semi-Parametric Approach. In *ICML*, pages 803–812. PMLR, 2021. page 102

P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, 2013. page 45

N. H. Bingham, C. M. Goldie, J. L. Teugels, and J. Teugels. *Regular Variation*. Cambridge University Press, 1989. pages 36, 42

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66(2-3):259–294, 2007. pages 84, 89

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electron. Commun. Probab.*, 17:1–12, 2012. pages 15, 174

S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. 2015. pages 15, 174

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: A Survey of Some Recent Advances. *ESAIM: P&S*, 9:323–375, 2005. page 66

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. pages 61, 63, 64, 100

L. Breiman. Random Forests. *Mach. Learn.*, 45:5–32, 2001. page 128

L. Breiman. *Classification and Regression Trees*. Routledge, 2017. page 128

C. Brownlees, E. Joly, and G. Lugosi. Empirical Risk Minimization for Heavy-Tailed Losses. *Ann. Stat.*, 43(6):2507–2536, 2015. page 107

J. Cadima and I. Jolliffe. On Relationships Between Uncentred And Column-Centred Principal Component Analysis. *Pak. J. Stat.*, 25(4):473–503, 2009. pages 57, 89

F. Caeiro and M. I. Gomes. Threshold Selection in Extreme Value Analysis. *Extreme Value Modeling and Risk Analysis: Methods and Applications*, 1:69–86, 2016. page 86

J.-J. Cai, J. H. Einmahl, and L. De Haan. Estimation of extreme risk regions under multivariate regular variation. *Ann. Stat.*, 39(3):1803–1826, 2011. pages 108, 109

C. Chadenas, A. Creach, and D. Mercier. The impact of storm Xynthia in 2010 on coastal flood prevention policy in France. *J. Coast. Conserv.*, 18:529–538, 2014. pages 19, 179

E. Chautru. Dimension reduction in multivariate extreme value analysis. *Electron. J. Stat.*, 9(1):383–418, 2015. pages 18, 177

V. Chavez-Demoulin, P. Embrechts, and S. Sardy. Extreme-quantile tracking for financial time series. *J. Econom.*, 181(1):44–52, 2014. page 107

M. Chiapino and A. Sabourin. Feature Clustering for Extreme Events Analysis, with Application to Extreme Stream-Flow Data. In *NFMCP*, pages 132–147. Springer, 2016. pages 18, 177

M. Chiapino, A. Sabourin, and J. Segers. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22:193–222, 2019. pages 14, 125, 174

M. Chiapino, S. Clémençon, V. Feuillard, and A. Sabourin. A multivariate extreme value theory approach to anomaly clustering and visualization. *Comput. Stat.*, 35(2): 607–628, 2020. pages 14, 18, 61, 174, 177

S. Clémençon, P. Bertail, and G. Papa. Learning from Survey Training Samples: Rate Bounds for Horvitz-Thompson Risk Minimizers. In *ACML*, pages 142–157. PMLR, 2016. page 102

S. Clémençon, H. Jalalzai, S. Lhaut, A. Sabourin, and J. Segers. Concentration bounds for the empirical angular measure with statistical learning applications. *Bernoulli*, 29(4):2797–2827, 2023. pages 15, 61, 105, 170, 174

S. Clémençon, N. Huet, and A. Sabourin. Regular variation in Hilbert spaces and principal component analysis for functional extremes. *Stochastic Process. Appl.*, 174: 104375, 2024. pages 20, 180

S. Coles, J. Heffernan, and J. Tawn. Dependence Measures for Extreme Value Analyses. *Extremes*, 2:339–365, 1999. pages 20, 180

S. Coles, J. Bawa, L. Trenner, and P. Dorazio. *An Introduction to Statistical Modeling of Extreme Values*, volume 208. Springer, 2001. pages 19, 91, 142, 179

S. G. Coles and J. A. Tawn. Statistical Methods for Multivariate Extremes: An Application to Structural Design. *J. R. Stat. Soc., Series C: Appl.*, 43(1):1–31, 1994. page 86

D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019. pages 14, 18, 125, 174, 177, 178

D. J. Daley, D. Vere-Jones, et al. *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*. Springer, 2003. page 42

A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19:2557–2589, 2013. page 107

A. Daouia, S. A. Padoan, G. Stupfler, et al. Optimal weighted pooling for inference about the tail index and extreme quantiles. *Bernoulli*, 2023. page 107

J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.*, 12(1):136–154, 1982. pages 59, 60

R. A. Davis and T. Mikosch. Extreme value theory for space-time processes with heavy-tailed distributions. *Stochastic Process. Appl.*, 118(4):560–584, 2008. pages 16, 71, 176

A. C. Davison and R. L. Smith. Models for Exceedances Over High Thresholds. *J. R. Stat. Soc., Series B: Stat. Methodol.*, 52(3):393–425, 1990. pages 19, 179

A. C. Davison, S. A. Padoan, and M. Ribatet. Statistical Modeling of Spatial Extremes. *Stat. Sci.*, 27(2):161 – 186, 2012. pages 20, 180

M. de Carvalho and A. Ramos. Bivariate Extreme Statistics, II. *REVSTAT Stat. J.*, 10(1): 83–107, 2012. pages 20, 180

L. De Haan. A Spectral Representation for Max-stable Processes. *Ann. Probab.*, 12(4): 1194–1204, 1984. pages 16, 175

L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer New York, 2006. pages 16, 33, 45, 101, 175

L. De Haan and S. I. Resnick. On regular variation of probability densities. *Stochastic Process. Appl.*, 25:83–93, 1987. pages 24, 108, 109, 119, 184

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013. page 61

M. J. Dixon and J. A. Tawn. The effect of non-stationarity on extreme sea-level estimation. *J. R. Stat. Soc., Series C: Appl. Stat.*, 48(2):135–151, 1999. pages 19, 179

C. Dombry and M. Ribatet. Functional regular variations, Pareto processes and peaks over threshold. *Stat. Interface*, 8(1):9–17, 2015. pages 16, 22, 45, 46, 71, 79, 175, 182

H. Drees and A. Sabourin. Principal component analysis for multivariate extremes. *Electron. J. Stat.*, 15:908–943, 2021. pages 14, 18, 22, 61, 71, 72, 83, 125, 174, 177, 178, 182

J. H. Einmahl and D. M. Mason. Strong Limit Theorems for Weighted Quantile Processes. *Ann. Probab.*, pages 1623–1643, 1988. page 88

J. H. Einmahl, L. de Haan, and V. I. Piterbarg. Nonparametric estimation of the spectral measure of an extreme value distribution. *Ann. Stat.*, 29:1401–1423, 2001. page 105

J. H. J. Einmahl and J. Segers. Maximum Empirical Likelihood Estimation of the spectral Measure of an Extreme-Value Distribution. *Ann. Stat.*, 37:2953–2989, 2009. page 105

J. El Methni, L. Gardes, S. Girard, and A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions. *J. Stat. Plan. Infer.*, 142(10):2735–2747, 2012. page 107

S. Engelke and A. S. Hitz. Graphical Models for Extremes. *J. R. Stat. Soc., Series B Stat. Methodol.*, 82(4):871–932, 2020. pages 14, 174

S. Engelke and J. Ivanovs. Sparse Structures for Multivariate Extremes. *Annu. Rev. Stat. Appl.*, 8:241–270, 2021. pages 18, 20, 61, 177, 180

G. Faÿ, B. González-Arévalo, T. Mikosch, and G. Samorodnitsky. Modeling teletraffic arrivals by a Poisson cluster process. *Queueing Syst.*, 54:121–140, 2006. page 74

W. Feller. *An Introduction to Probability Theory and its Applications, Volume 2*, volume 81. John Wiley & Sons, 1991. page 37

A. Ferreira and L. de Haan. The generalized Pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717 – 1737, 2014. pages 16, 20, 45, 175, 180

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Math. Proc. Cambridge Philos. Soc.*, volume 24, pages 180–190. Cambridge University Press, 1928. page 34

E. Genovese and V. Przyluski. Storm surge disaster risk management: the Xynthia case study in France. *J. Risk Res.*, 16(7):825–841, 2013. pages 19, 179

J. Gertheiss, D. Rügamer, B. X. W. Liew, and S. Greven. Functional data analysis: An introduction and recent developments. *arXiv preprint:2312.05523*, 2023. pages 48, 70

E. Giné and A. Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. *Ann. de l'IHP Probab. et Stat.*, 37(4):503–522, 2001. pages 66, 135, 136

C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2021. pages 17, 177

N. Gnecco, N. Meinshausen, J. Peters, and S. Engelke. Causal discovery in heavy-tailed models. *Ann. Stat.*, 49(3):1755–1778, 2021. pages 18, 177

N. Gnecco, E. M. Terefe, and S. Engelke. Extremal Random Forests. *J. Am. Stat. Assoc.*, pages 1–14, 2024. pages 15, 174

B. Gnedenko. Sur la distribution limite du terme maximum d'une série aleatoire. *Ann. Math.*, pages 423–453, 1943. page 34

I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Classes of Linear Operators*, volume 63. Birkhäuser, 2013. page 52

N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study. *COLT*, pages 843–860, 2015. pages 15, 135, 174

N. Goix, A. Sabourin, and S. Clémençon. Sparse Representation of Multivariate Extremes with Applications to Anomaly Ranking. *AISTATS*, pages 75–83, 2016. pages 14, 18, 125, 174, 177

N. Goix, A. Sabourin, and S. Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. *J. Multivariate Anal.*, 161:12–31, 2017. pages 14, 18, 61, 125, 169, 174, 177

U. Grömping. Variable importance in regression models. *Wiley Interdiscip. Rev., Comput. Stat.*, 7(2):137–152, 2015. page 128

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002. pages 61, 100, 104, 126, 170

I. D. Haigh, R. Nicholls, and N. Wells. A comparison of the main methods for estimating probabilities of extreme still water levels. *Coast. Eng.*, 57(9):838–849, 2010. pages 19, 145, 179

Z. Hao, V. P. Singh, and F. Hao. Compound Extremes in Hydroclimatology: A Review. *Water*, 10(6):718, 2018. pages 20, 180

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 2. Springer, 2009. page 61

J. E. Heffernan and J. A. Tawn. A conditional approach for multivariate extreme values (with discussion). *J. R. Stat. Soc., Series B: Stat. Methodol.*, 66(3):497–546, 2004. pages 20, 148, 180

A. Hitz and R. Evans. One-Component Regular Variation and Graphical Modeling of Extremes. *J. Appl. Probab.*, 53(3):733–746, 2016. pages 18, 105, 177

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*, volume 200. Springer Science & Business Media, 2012. pages 16, 48, 57, 59, 175

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, 2015. pages 16, 18, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 78, 175, 178

N. Huet, S. Clémençon, and A. Sabourin. On Regression in Extreme Regions. *arXiv preprint:2303.03084*, 2023. pages 21, 27, 143, 148, 180, 187

H. Hult and F. Lindskog. Extremal behavior of regularly varying stochastic processes. *Stochastic Process. Appl.*, 115:249–274, 2005. pages 16, 176

H. Hult and F. Lindskog. On Kesten's counterexample to the Cramér-Wold device for regular variation. *Bernoulli*, pages 133–142, 2006a. page 77

H. Hult and F. Lindskog. Regular Variation for Measures on Metric Spaces. *Publ. Inst. Math.*, 80(94):121–140, 2006b. pages 16, 17, 21, 33, 39, 42, 43, 45, 46, 71, 72, 75, 77, 105, 176, 181

R. Huser and J. L. Wadsworth. Advances in statistical modeling of spatial extremes. *Wiley Interdiscip. Rev. Comput. Stat.*, 14(1):e1537, 2022. pages 16, 20, 70, 143, 175, 180

D. Idier, F. Dumas, and H. Muller. Tide-surge interaction in the English Channel. *Nat. Hazard Earth Sys.*, 12(12):3709–3718, 2012. pages 19, 142, 179

H. Jalalzai, S. Clémençon, and A. Sabourin. On Binary Classification in Extreme Regions. pages 3092–3100, 2018. pages 15, 61, 102, 105, 174

G. James, D. Witten, T. Hastie, R. Tibshirani, et al. *An Introduction to Statistical Learning*, volume 112. Springer, 2013. page 61

A. Janßen and P. Wan. k-means clustering of extremes. *Electron. J. Stat.*, 14(1):1211–1233, 2020. pages 14, 18, 174, 177

J. Karamata. Sur un mode de croissance régulière. Théorèmes fondamentaux. *B. Soc. Math. Fr.*, 61:55–62, 1933. page 37

M. Karamouz, M. Taheri, P. Khalili, and X. Chen. Building Infrastructure Resilience in Coastal Flood Risk Management. *J. Water Res. Plan. Man.*, 145(4):04019004, 2019. pages 19, 179

C. Keef, I. Papastathopoulos, and J. A. Tawn. Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model. *J. Multivariate Anal.*, 115:396–404, 2013. pages 20, 180

X. Kergadallan. Estimation des valeurs extrêmes de niveau d'eau : Littoral métropolitain. Technical report, CEREMA, 2022. URL https://doc.cerema.fr/Default/doc/SYRACUSE/593562/estimation-des-valeurs-extremes-de-niveau-d-eau-littoral-metropolitain?_lg=fr-FR. pages 20, 179

X. Kergadallan, P. Bernardara, M. Benoit, and C. Daubord. Improving the estimation of extreme sea levels by a characterization of the dependence of skew surges on high tidal levels. *Coast. Eng. Proc.*, 1:48, 2014. pages 20, 179

M. Kim and P. Kokoszka. Extremal dependence measure for functional data. *J. Multivariate Anal.*, 189:104887, 2022. pages 17, 18, 75, 176, 178

M. Kim and P. Kokoszka. Extremal correlation coefficient for functional data. *arXiv preprint:2405.17318*, 2024. pages 17, 176

A. Kiriliouk, H. Rootzén, J. Segers, and J. L. Wadsworth. Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions. *Technometrics*, 61(1):123–135, 2019. pages 20, 27, 40, 42, 143, 148, 149, 150, 154, 180, 187

P. Kokoszka and R. Kulik. Principal component analysis of infinite variance functional data. *J. Multivariate Anal.*, 193:105123, 2023. pages 17, 18, 176, 178

P. Kokoszka and Q. Xiong. Extremes of projections of functional time series on data–driven basis systems. *Extremes*, 21:177–204, 2018. pages 17, 18, 176, 178

P. Kokoszka, S. Stoev, and Q. Xiong. Principal components analysis of regularly varying functions. *Bernoulli*, 25(4B):3864–3882, 2019. pages 17, 18, 84, 176, 178

J. Kuelbs and V. Mandrekar. Domains of Attraction of Stable Measures on a Hilbert Space. *Studia Math.*, 50:149–162, 1974. pages 44, 45, 75

S. Kundu, S. Majumdar, and K. Mukherjee. Central Limit Theorems revisited. *Stat. Probab. Lett.*, 47(3):265–275, 2000. page 88

N. le Carrer. egpd4gamlss, 2022. URL https://github.com/noemielc/egpd4gamlss. page 152

M. R. Leadbetter. *Extremes and local dependence in stationary sequences*. Københavns Universitet, Inst. Math. Stat., 1982. pages 19, 179

G. Lecué and S. Mendelson. Learning subgaussian classes : Upper and minimax bounds. *arXiv:1305.4825*, 2013. page 100

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 1991. pages 54, 84

J. Legrand, P. Ailliot, P. Naveau, and N. Raillard. Joint stochastic simulation of extreme coastal and offshore significant wave heights. *Ann. Appl. Stat.*, 17(4):3363 – 3383, 2023. pages 27, 145, 148, 188

G. W. Lennon, E. Gumbel, N. Barricelli, and A. Jenkinson. A frequency investigation of abnormally high tidal levels at certain west coast ports. *P. I. Civil Eng.*, 25(4): 451–484, 1963. pages 19, 142, 179

S. Lhaut and J. Segers. Inégalités de concentration pour évènements rares. Master's thesis, Master's thesis, Faculté des sciences, Université catholique de Louvain . . . , 2021. pages 15, 174

S. Lhaut, A. Sabourin, and J. Segers. Uniform concentration bounds for frequencies of rare events. *Stat. Probab. Lett.*, 189:109610, 2022. pages 15, 67, 135, 174

Y. Li, Y. Qiu, and Y. Xu. From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. *J. Multivariate Anal.*, 188:104806, 2022. pages 48, 70

F. Lindskog, S. I. Resnick, and J. Roy. Regularly varying measures on metric spaces: Hidden regular variation and hidden jumps. *Probab. Surv.*, pages 270–314, 2014. pages 3, 24, 39, 103, 105, 110, 111, 112, 184

M. Loève. *Probability Theory II*. Springer New York, 1978. page 59

G. Lugosi and S. Mendelson. Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc.*, 22(3):925–965, 2019. pages 107, 126

S. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 1999. pages 18, 169, 178

P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, 2007. page 100

C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998. pages 64, 86, 87, 96, 134

A. McRobie, T. Spencer, and H. Gerritsen. The big flood: North Sea storm surge. *Philos. T. R. Soc. A, Math. Phys. Eng. Sci.*, 363(1831):1263–1270, 2005. pages 19, 142, 179

M. A. Medina, R. A. Davis, and G. Samorodnitsky. Spectral learning of multivariate extremes. *arXiv preprint:2111.07799*, 2021. pages 18, 177

M. M. Meerschaert. *Multivariate domains of attraction and regular variation*. PhD thesis, University of Michigan, 1984. pages 42, 43

T. Meinguet and J. Segers. Regularly varying time series in banach spaces. *arXiv preprint:1001.3262*, 2010. page 46

S. Mendelson. On aggregation for heavy-tailed classes. *Probab. Theory Rel.*, 168(3-4): 641–674, 2017. page 107

N. Meyer and O. Wintenberger. Sparse regular variation. *Adv. Appl. Probab.*, 53(4): 1115–1148, 2021. pages 14, 18, 125, 174, 177

N. Meyer and O. Wintenberger. Multivariate sparse clustering for extremes. *J. Am. Stat. Assoc.*, 0:1–23, 2023. pages 14, 127, 129, 174

T. Mikosch. Regular Variation, Subexponentiality and Their Applications in Probability Theory. 1999. pages 38, 79

J. Mikusiński. *The Bochner Integral*. Birkhäuser Basel, 1978. page 53

I. Molchanov and K. Strokorb. Max-stable random sup-measures with comonotonic tail dependence. *Stochastic Process. Appl.*, 126(9):2835–2859, 2016. pages 16, 176

P. Naveau, R. Huser, P. Ribereau, and A. Hannart. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52 (4):2753–2769, 2016. pages 42, 147, 148

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE T. on Knowl. Data Eng.*, 22 (10):1345–1359, Oct. 2010. page 102

I. Papastathopoulos and J. A. Tawn. Extended generalised pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143(1):131–143, 2013. pages 27, 42, 145, 188

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011. page 127

N. Pfister and P. Bühlmann. Extrapolation-Aware Nonparametric Statistical Inference, 2024. page 102

Y. V. Prokhorov. Convergence of Random Processes and Limit Theorems in Probability Theory. *Theory Probab. Appl.*, 1(2):157–214, 1956. page 55

D. Pugh and J. Vassie. Applications of the joint probability method for extreme sea level computations. *P. I. Civil Eng.*, 69(4):959–975, 1980. pages 19, 142, 179

D. Pugh and P. Woodworth. *Sea-Level Science: Understanding Tides, Surges, Tsunamis and Mean Sea-Level Changes*. Cambridge University Press, 2014. page 158

D. T. Pugh and J. Vassie. Extreme sea levels from tide and surge probability. In *Coast. Eng.*, pages 911–930. 1978. pages 19, 142, 179

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. *J. R. Stat. Soc., Series C: Appl. Stat.*, 54:507–554, 2005. page 152

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer-Verlag New York, 2005. pages 16, 48, 57, 175

B. J. Reich and B. A. Shaby. A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.*, 6(4):1430, 2012. pages 20, 180

S. Resnick and L. de Haan. Second-order regular variation and rates of convergence in extreme-value theory. *Ann. Probab.*, 24(1):97 – 124, 1996. page 126

S. I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag New York, 1987. pages 16, 33, 34, 38, 40, 42, 71, 101, 120, 176

S. I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media, 2007. pages 16, 33, 37, 38, 39, 44, 71, 176

S. I. Resnick and R. Roy. Random USC Functions, Max-Stable Processes and Continuous Choice. *Ann. Appl. Probab.*, 1(2):267–292, 1991. pages 16, 176

H. Rootzén and N. Tajvidi. Multivariate generalized Pareto distributions. *Bernoulli*, 12 (5):917–930, 2006. pages 20, 27, 40, 143, 180, 187

H. Rootzén, J. Segers, and J. L. Wadsworth. Multivariate peaks over thresholds models. *Extremes*, 21(1):115–145, 2018. pages 20, 27, 40, 41, 42, 180, 187

A. Sabourin and J. Segers. Marginal standardization of upper semicontinuous processes. with application to max-stable processes. *J. Appl. Probab.*, 54(3):773–796, 2017. pages 16, 176

G. Samorodnitsky and Y. Wang. Extremal theory for long range dependent infinitely divisible processes. *Ann. Probab.*, 47(4):2529–2562, 2019. pages 16, 176

C. Scarrott and A. MacDonald. A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *Revstat Stat. J.*, 10(1):33–60, 2012. page 86

J. Segers. One-versus multi-component regular variation and extremes of Markov trees. *Adv. Appl. Probab.*, 52(3):855–878, 2020. pages 18, 177

J. Segers, Y. Zhao, and T. Meinguet. Polar decomposition of regularly varying time series in star-shaped metric spaces. *Extremes*, 20:539–566, 2017. pages 16, 44, 71, 176

S. I. Seneviratne, X. Zhang, M. Adnan, W. Badi, C. Dereczynski, A. D. Luca, S. Ghosh, I. Iskandar, J. Kossin, S. Lewis, et al. Weather and climate extreme events in a changing climate. 2021. pages 19, 158, 179

SHOM. Refmar. URL https://data.shom.fr/donnees/refmar. page 145

R. Shooter, J. Tawn, E. Ross, and P. Jonathan. Basin-wide spatial conditional extremes for severe ocean storms. *Extremes*, 24:241–265, 2021. pages 20, 180

E. S. Simpson, J. L. Wadsworth, and J. A. Tawn. Determining the Dependence Structure of Multivariate Extremes. *Biometrika*, 107(3):513–532, 2020. pages 14, 18, 174, 177

R. L. Smith. Extreme value theory based on the r largest annual events. *J. Hydrol.*, 86 (1-2):27–43, 1986. pages 19, 179

C. Stărică. Multivariate extremes for models with constant conditional correlations. *J. Empir. Financ.*, 6(5):515–553, 1999. page 86

A. Stephenson. Simulating multivariate extreme value distributions of logistic type. *Extremes*, 6(1):49–59, 2003. page 127

C. Suthons. Frequency of occurence of abnormally high sea levels on the east and south coasts of England. *P. I. Civil Eng.*, 25(4):433–450, 1963. pages 19, 142, 179

J. Tawn, J. Vassie, and E. Gumbel. Extreme sea levels; the joint probabilities method revisited and revised. *P. I. Civil Eng.*, 87(3):429–442, 1989. pages 19, 179

J. Tawn, R. Shooter, R. Towe, and R. Lamb. Modelling spatial extreme events with environmental applications. *Spat. Stat.*, 28:39–58, 2018. pages 20, 180

J. A. Tawn. An extreme-value theory model for dependent observations. *J. Hydrol.*, 101 (1-4):227–250, 1988. pages 19, 179

J. A. Tawn. Estimating probabilities of extreme sea-levels. *J. R. Stat. Soc., Series C: Appl. Stat.*, 41(1):77–93, 1992. pages 19, 142, 179

A. Thomas, S. Clemencon, A. Gramfort, and A. Sabourin. Anomaly Detection in Extreme Regions via Empirical MV-sets on the Sphere. In *AISTATS*, volume 54, pages 1011–1019. PMLR, 2017. pages 18, 178

K. Tsukuda. A change detection procedure for an ergodic diffusion process. *Ann. Inst. Stat. Math.*, 69(4):833–864, 2017. pages 56, 78

N. Vakhania, V. Tarieladze, and S. Chobanyan. *Probability Distributions on Banach Spaces*, volume 14. Springer Science & Business Media, 2012. pages 53, 54, 55

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer-Verlag New York, 1996. pages 21, 54, 55, 56, 72, 73, 80, 124, 136, 137, 181

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 1999. page 65

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity: festschrift for Alexey Chervonenkis*, pages 11–30. Springer, 2015. page 66

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *Extremes*, 0(0):1–29, 2023. pages 15, 174

E. Vignotto and S. Engelke. Extreme value theory for anomaly detection–the GPD classifier. *Extremes*, 23(4):501–520, 2020. pages 14, 174

E. Vignotto, S. Engelke, and J. Zscheischler. Clustering bivariate dependencies of compound precipitation and wind extremes over great britain and ireland. *Weather Clim. Extrem.*, 32:100318, 2021. pages 14, 174

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional Data Analysis. *Annu. Rev. Stat. Appl.*, 3:257–295, 2016. page 70

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a Few Examples: A Survey on Few-shot Learning. 53(3), 2020. page 100

P. Wei, Z. Lu, and J. Song. Variable importance analysis: A comprehensive review. *Reliab. Eng. Syst. Saf.*, 142:399–432, 2015. page 128

J. Williams, K. J. Horsburgh, J. A. Williams, and R. N. Proctor. Tide and skew surge independence: New insights for flood risk. *Geophys. Res. Lett.*, 43(12):6410–6417, 2016. pages 20, 142, 179

L. Zwald and G. Blanchard. On the Convergence of Eigenspaces in Kernel Principal Component Analysis. In *NIPS*, 2005. pages 23, 60, 85, 183

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Apprentissage Statistique des Extrêmes Multivariés et Fonctionels

Mots clés : Théorie des Valeurs Extrêmes, Apprentissage Statistique, Analyse des Données Fonctionnelles

Résumé : Dans un monde où le réchauffement climatique provoque de plus en plus de phénomènes météorologiques extrêmes d'ampleurs croissantes, cette thèse explore la modélisation des événements extrêmes à travers des méthodes statistiques enrichies par l'apprentissage statistique. Elle se divise en deux grandes parties.
Dans un premier temps, les extrêmes fonctionnels sont étudiés, c'est-à-dire les extrêmes de données dépendant explicitement d'une variable continue comme le temps. Nous travaillons dans un espace de Hilbert séparable, avec un focus sur l'espace $L^2[0,1]$. Des résultats sur la variation régulière, hypothèse fondamentale en théorie des valeurs extrêmes, sont développés, et des caractérisations ainsi que des exemples non triviaux sont présentés. De plus, une méthode de réduction de dimension adaptée aux données fonctionnelles extrêmes est proposée, avec des garanties probabilistes et statistiques. Dans un second temps, nous développons un cadre probabiliste pour la régression dans des régions où la variable d'entrée est extrême, contrairement aux approches classiques qui se concentrent sur les régions où la variable de sortie est extrême. Des résultats sur les risques et les fonctions de régression dans les régions extrêmes, ainsi qu'un algorithme adapté, sont établis. Ce dernier est comparé à des méthodes classiques et appliqué à la prédiction des extrêmes maritimes en Bretagne, où nous cherchons à compléter les données extrêmes passées pour réduire les incertitudes liées à certaines estimations.

Title : Statistical Learning of Multivariate and Functional Extremes

Keywords : Extreme Value Theory, Statistical Learning, Functional Data Analysis

Abstract : In a world where climate change is causing more and more extreme weather events of increasing magnitude, this thesis explores the modeling of extreme events through statistical methods enhanced by statistical learning. It is divided into two main parts. First, functional extremes are studied, that is, the extremes of data explicitly dependent on a continuous variable such as time. We work in a separable Hilbert space, with a focus on the space $L^2[0,1]$. Results on regular variation, a fundamental hypothesis in extreme value theory, are developed, along with characterizations and non-trivial examples. Additionally, a dimensionality reduction method tailored to functional extreme data is proposed, with probabilistic and statistical guarantees. In the second part, we develop a probabilistic framework for regression in regions where the input variable is extreme, in contrast to classic approaches that focus on regions where the output variable is extreme. Results on risks and regression functions in extreme regions, as well as an adapted algorithm, are established. This algorithm is compared to classical methods and applied to the prediction of extreme sea levels in Brittany, where the goal is to reconstruct past extreme data to reduce uncertainties associated with certain estimates.

Institut Polytechnique de Paris
91120 Palaiseau, France